

AI Christian Benchmark

Evaluating 7 Top LLMs for
Theological Reliability

Visit ChristianBenchmark.AI

Michael S. Graham

The Gospel Coalition

The Keller Center for Cultural Apologetics

For all media or tech industry inquiries, [please contact us.](#)



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

By 2028, as many people will be searching with AI as with Google.
We need to know: Can we rely on AI?

“Did Jesus rise from the dead?”



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



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

Table of Contents

Part 1: What Did We Learn?

[Five Biggest Headlines](#)

[Shift from SEO to GEO](#)

[Overall Theological Reliability Scores](#)

Part 2: How Did We Do It?

[Foundational Principles](#)

[Methodology](#)

[Scholar Team](#)

Part 3: How Did the Platforms Score?

[Platform Characteristics](#)

[Common Patterns](#)

[How Do These Problems Happen?](#)

[Individual Scores for Each Question](#)

Part 4: Why Were the Scores So Different?

[Alignment](#)

[Where Weaknesses Occur](#)

[Citation Preferences](#)

Part 5: What Does This Mean for Culture, Society, and Me?

[Analysis](#)

[What Do You Recommend for AI Users?](#)

Part 6: What Is The Gospel Coalition and The Keller Center Doing Next on AI?

Part 7: FAQ and Acknowledgments

An aerial, high-angle view of the San Francisco skyline, featuring the Transamerica Pyramid and the Bay Bridge. The image is overlaid with a semi-transparent orange filter. The text "PART 1" is centered in the middle of the image.

PART 1

What Did We Learn?

How does identical technology, trained on essentially identical content, and running on mainly identical silicon, give such radically different answers?

The Five Biggest Headlines

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.

You don't know how the answers have been chosen or who decides between fact and value.

The Shift from Google Search (SEO) to LLM Search (GEO)

In the recent past, acquiring knowledge online was a two-step process:

Step 1: Type your query in the search bar.

Step 2: Use wisdom to decide which links to click, read, and glean insights.

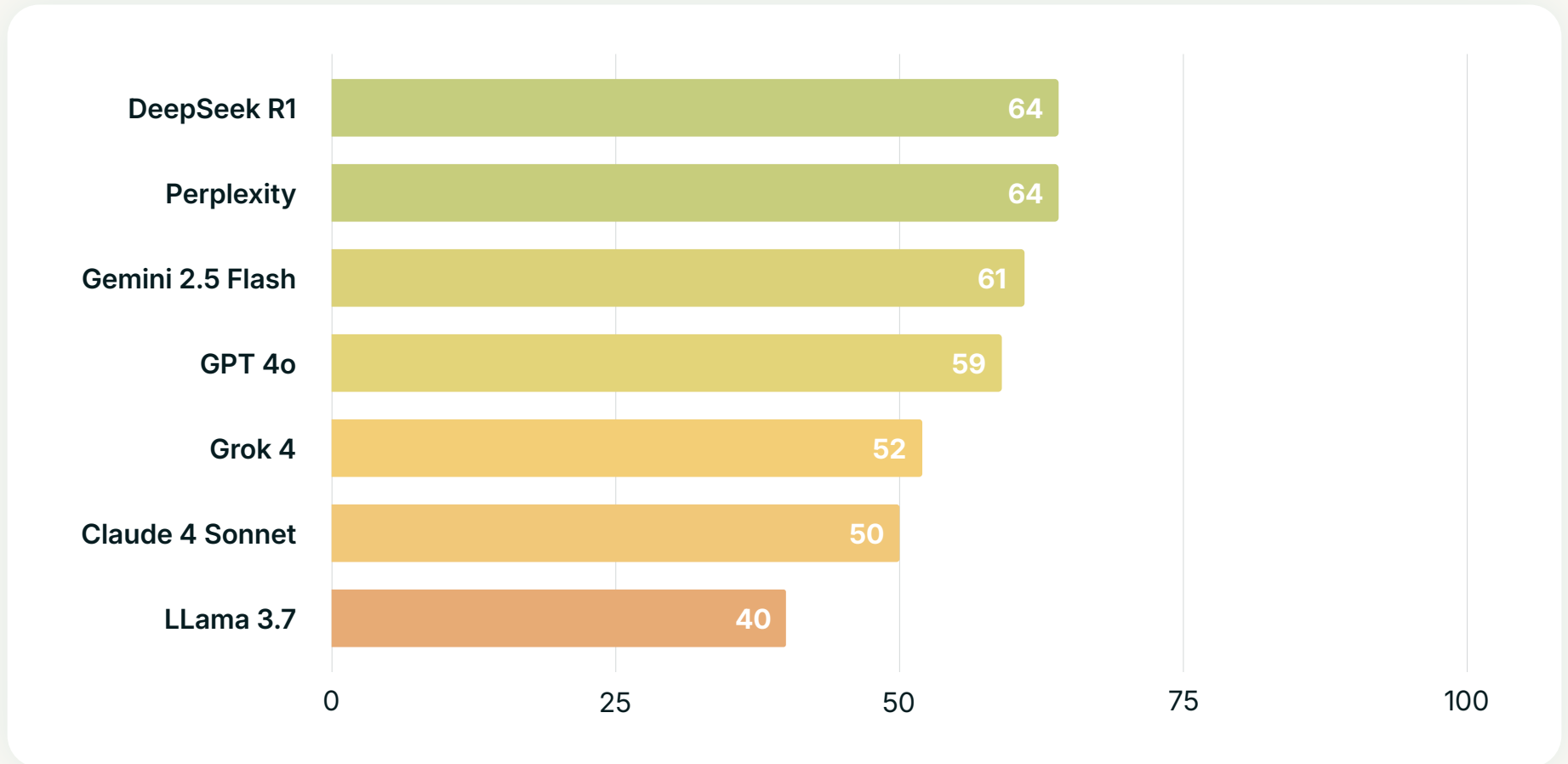
In other words, Google has functioned like a much bigger, broader, and faster encyclopedia where users could choose their own adventure. You still needed to choose between the alternatives, to make your own evaluation of their claims.

The current- and near-future of acquiring knowledge online is a one-step process:

Step 1: Type your prompt into the prompt box and automagically all the culling and curating of information is done for you.

AI purports to do the old step 2 for you and save you time and energy. But you don't know how the answers have been chosen or who decides between fact and value, where AI speaks with authority or presents "all sides." AI can't deliver objective and value-neutral comprehensive guidance because the financial and legal risks are too significant.

Overall Theological Reliability Score



We were surprised to find that Chinese model DeepSeek R1 (0528 Qwen3 8B) was the top model for theological reliability, edging out multi-modal answer engine Perplexity for the top spot. Silicon Valley giant Meta had, by far, the worst-performing platform in Llama 3.7. Even more surprising are some of the hypotheses about **why**.



PART 2

How Did We Do It?

Christians seek from Silicon Valley more common grace through better technology.

God created the world and made humanity in his image. Therefore, our world features elements of common grace, such as technology. Christians seek from Silicon Valley more common grace through better technology.

In AI we see reasons for both optimism and also skepticism. We must be honest about AI's weaknesses and biases and also grateful for advances that promote human flourishing, including the advance of the gospel.

Those who work in this emerging field can glean insights from this data. However, we hope this report will help Silicon Valley improve the integrity of these AI platforms. We believe that with adjustments in alignment, improved functionality is within reach. We want to encourage and serve developers as they build these transformative technologies through which God's common grace can extend.

It is also important to note that we will do anything we can to be supportive and encouraging towards the makers of these technologies to help improve them. We would love to engage in any private or public conversations that anyone in Silicon Valley would like to have on how we can work together to make better technology.

What do you recommend for Silicon Valley?

We believe that answering questions about each tradition from the perspective of the consensus of its adherents in concert with an invitation to further dialogue with other perspectives maximizes honor and respect for the religious tradition, maximizes value for the user, and minimizes risks for the AI platform.

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.**

7 Prompts
7 AI Platforms
7 Scholars

Who Is Jesus?

What Is the Gospel?

Does God Exist?

Why Does God Allow Suffering?

Did Jesus Rise from the Dead?

Was Jesus a Real Person?

Is the Bible Reliable?

Methodology

AI Christian Benchmark was designed to test how the top AI platforms respond to the top religious questions googled in the past.

At every point we tried to mimic the prompt and platform behaviors of ordinary people.

We employed the same wording of top Google phrase searches, and we asked them without any additional context. Since every platform is capable of giving 100/100 responses with additional context (“in your answer give responses that are consistent with the Nicene Creed and _____ tradition/creed”), it isn’t beneficial to give additional context in the benchmark. We wanted to test for bias on the most basic prompts.

We tested the top seven platforms by total search count (measured by token volume). While some of these platforms have paid premium accounts, we conducted tests on the free versions so as to best mirror the behavior of ordinary people.

The responses were hand-graded by a diverse group of seven scholars with expertise in each subject area. They edited the rubrics for each question. Scores of 65 or above broadly reflect Nicene Creed Christianity. Scores of 80 or above reflect the gospel-centrality of [TGC’s Foundation Documents](#) and other historic Protestant creeds and confessions.

Scholar Team



Peter J. Williams has been leading Tyndale House, an international Bible research community in Cambridge, England, since 2007. He received his MA, MPhil, and PhD from the University of Cambridge, studying ancient languages related to the Bible and was formerly a senior lecturer in New Testament at the University of Aberdeen. He is an affiliated lecturer at the University of Cambridge and a member of the Translation Oversight Committee of the English Standard Version. His book [Can We Trust the Gospels?](#) (Crossway, 2018) has been published in 15 languages. His latest book is [The Surprising Genius of Jesus](#) (Crossway, 2023).



Hans Madueme is professor of theological studies at Covenant College in Lookout Mountain, Georgia. After completing a residency in internal medicine at the Mayo Clinic in Rochester, Minnesota, he earned his MDiv and PhD in systematic theology from Trinity Evangelical Divinity School. He is the author of two recent books, [Defending Sin: A Response to the Challenges of Evolution and the Natural Sciences](#) (Baker Academic, 2024) and [Does Science Make God Irrelevant?](#) (Crossway, 2025). He is a ruling elder in the Presbyterian Church in America.



Nadya Williams (PhD in Classics, Princeton University) is a homeschool mom, books editor at Mere Orthodoxy, and the author of [Cultural Christians in the Early Church](#) (Zondervan Academic, 2023), [Mothers, Children, and the Body Politic](#) (IVP Academic, 2024), and the forthcoming [Christians Reading Classics](#) (Zondervan Academic, 2025).



Gavin Ortlund (PhD, Fuller Theological Seminary) is a pastor, author, speaker, and apologist for the Christian faith. He serves as the president of [Truth Unites](#), visiting professor of historical theology at Phoenix Seminary, and theologian-in-residence at Immanuel Nashville. He is the author of several books, including [The Art of Disagreeing](#) (The Good Book Company, 2025), [Why God Makes Sense in a World That Doesn't](#) (Baker Academic, 2021), and [What It Means to Be Protestant](#) (Zondervan, 2024).

Scholar Team (continued)



Uche Anizor is professor of theology at Talbot School of Theology, Biola University, in La Mirada, California. He received his PhD in systematic theology from Wheaton College. His books include [*How to Read Theology*](#) (Baker Academic, 2018), [*Overcoming Apathy*](#) (Crossway, 2022), and the forthcoming [*The Goodness of God in the Gift of Scripture*](#) (Crossway, 2026).



Mary Hannah (PhD, Trinity Evangelical Divinity School) serves the saints at Second Presbyterian Church in Memphis as the director of women in ministry. Mary enjoys teaching and training others to teach the Scriptures, especially in the context of the local church. She and her husband, Jeff, belong to the Graham/Waring parish.



Michael Kruger is the Samuel C. Patterson Chancellor's Professor of New Testament and Early Christianity at Reformed Theological Seminary's campus in Charlotte, North Carolina. He served as president of the Evangelical Theological Society in 2019. He is the author of [*Surviving Religion 101: Letters to a Christian Student on Keeping the Faith in College*](#) (Crossway, 2021) and [*Christianity at the Crossroads: How the Second Century Shaped the Future of the Church*](#) (IVP Academic, 2019). He blogs regularly at [Canon Fodder](#).

PART 3

How Did the Platforms Score?

Platform Characteristics



DeepSeek R1¹ was the top LLM for overall theological reliability. Broadly speaking, the platform gave answers consistent with the Nicene Creed with few qualifications. DeepSeek registered higher highs and lower lows than second place, Perplexity, which was overall more consistent. The platform underperformed on “Does God Exist?” and “Is the Bible Reliable?” It is possible that those two notably lower scores are due to requested censorship from the Chinese government.



Perplexity closely followed DeepSeek for overall theological reliability. Like DeepSeek, Perplexity also broadly gave answers consistent with the Nicene Creed with few qualifications. Further, the platform was a little more consistent than DeepSeek overall and never scored below 61 on any question. Perplexity functions uniquely among the different LLM platforms as more of a multi-modal answer engine than a traditional LLM. It incorporates things like search, inputs from other LLMs, and other technology to answer prompts.



Gemini 2.5 Flash (Google) largely took an “all sides” approach in its responses that involved two core components. Gemini included responses from Muslim, Jewish, atheistic, skeptical, and other religions alongside the Christian response. This type of approach was most common among Gemini, GPT 4o, and Claude 4 Sonnet, especially with questions around Jesus, God’s existence, and the Bible. Gemini performed well on two questions—“What Is the Gospel?” and “Who Is Jesus?”—but most of the answers involved many qualifications. Overall, Gemini provided answers fairly consistent with the Nicene Creed. However, the answers were presented with hedging from other religious and nonreligious perspectives and seem less likely to lead someone toward the Christian faith. For most users, Christianity would be seen as one option among many.



Claude 4 Sonnet (Anthropic) was surprisingly disappointing. While Anthropic’s Claude typically performs quite well on other known benchmarks it underperformed here. Overall, the platform’s answers felt like a less thorough “all-sides” of Gemini or GPT 4o but with even more hedging in the opening and closing paragraphs.



GPT 4o (OpenAI) was similar to Gemini 2.5 Flash in almost every way. It had all the same “all sides” approach with slightly less accurate and clear answers overall. All things considered, GPT 4o and Gemini 2.5 Flash were the two most similar models in the entire benchmark. Although it offered brief answers consistent with the Nicene Creed, GPT4o hedged its responses with opening and closing paragraphs about how these are controversial questions while also giving cumulatively longer answers from other religious and nonreligious traditions.



Grok 4 (xAI) was the strangest of all the LLMs we tested. Grok appears to weight different sources compared to the other platforms. The most commonly cited sources for Grok are X, Reddit, and Wikipedia. No other platform cites X or Reddit in their weighting. Hence, Grok takes on a lot of the unique quirks, personality, emotional temperature, and worldviews most commonly portrayed in those sources. Grok 4 almost feels like it has aspects of Elon Musk’s own personality and voice in ways that are impossible to quantify but seem evident when reading hundreds of responses. Overall, the platform varied widely on the quality of its answers, with most performing poor in all but two questions. Grok was capable of giving Nicene Creed answers, but often spent a lot of time on agnostic or skeptical perspectives. Additionally, Grok typically hedged its answers, saying that most of the questions really could not be answered satisfactorily or scientifically.



Llama 3.7 (Meta) was disappointing as a widely used LLM.² Many responses were inconsistent with the Nicene Creed. Llama 3.7 was unhelpfully brief (responses averaged one-third of the length of the other LLMs) and often did not seem like it wanted to answer the questions. Answers were frequently bookended with qualifiers about opinions, various beliefs, individual interpretations, and individual faith. Llama had the worst aspects of the “all sides” approach combined with brief answers marked by inaccuracies and bookended by qualifiers.

**Chinese model DeepSeek R1
was the top LLM overall for
theological reliability.**

Common Patterns

It should be nearly impossible for the same sentence to be repeated in the majority of prompts.

Identical Opening and Closing Sentences / Paragraphs

Certain platforms had nearly identical opening sentences in their answers to many of our prompts. The worst offenders were Claude 4 Sonnet, GPT 4o, Gemini 2.5 Flash, and especially Llama 3.7. Because of how LLM technology works we don't expect lengthy identical sentences. It is mathematically improbable, especially in the majority of responses. So when identical sentences appear in the opening and closing paragraphs of prompts, it is not unreasonable to conclude that human interference is at play. For example, each of these questions had identical opening sentences for their answers on the majority of answers to these prompts:

"Who Is Jesus?"

Claude 4 Sonnet: *"Jesus is a central figure in Christianity, believed by Christians to be the Son of God and the savior of humanity."*

Llama 3.7 (Meta): *"Jesus, also known as Jesus Christ, is a central figure in Christianity."*

GPT 4o: *"Jesus is the central figure of Christianity, believed by Christians to be the Son of God, the Messiah (or Christ), and the Savior of the world."*

"Was Jesus a Real Person?"

Llama 3.7 (Meta): *"The historicity of Jesus is a topic of ongoing debate among scholars and historians."*

When identical sentences appear in the opening and closing paragraphs of prompts, it is not unreasonable to conclude that human interference is at play.

"Why Does God Allow Suffering?"

Llama 3.7 (Meta): *"The question of why God allows suffering is one that has puzzled theologians, philosophers, and everyday people for centuries."*

GPT 4o: *"The question of why God allows suffering is one of the most profound and difficult in theology and philosophy."*

"Is the Bible Reliable?"

Llama 3.7 (Meta): *"The reliability of the Bible is a complex and debated topic among scholars, theologians, and philosophers."*

"Who is Jesus?"

Gemini 2.5 Flash: *"In conclusion, the question of "Who is Jesus?" elicits a range of answers that reflect diverse beliefs and perspectives. For billions of Christians, he is the divine Son of God and the savior of the world. For historians, he is a Jewish preacher from the first century whose life and teachings had a profound and lasting impact on human history."*

"What is the Gospel?"

Claude 4 Sonnet: *"Different Christian denominations may emphasize various aspects of the Gospel message, but the core theme of God's redemptive love through Christ remains central across most Christian traditions."*

Many platforms had an “all sides” approach to many of their responses.

The “all sides” approach is where a platform answers a prompt with roughly coequal response length from multiple religious traditions or perspectives. Here is an example of the “all sides” approach from GPT 4o:

Jesus is a central figure in Christianity, and he is also recognized in Islam and other religious traditions. Here’s a brief overview from different perspectives:

Christian Perspective

Identity: Jesus of Nazareth is believed to be the Son of God, the Messiah (Christ) prophesied in the Old Testament.

Nature: Christians believe Jesus is both fully divine and fully human.

Life: He was born in Bethlehem (around 4–6 BCE), lived a life of teaching, healing, and performing miracles, and was crucified under Roman authority (around 30–33 CE).

Purpose: His death is seen as a sacrifice for the sins of humanity, and his resurrection is viewed as proof of his divine nature and the promise of eternal life.

Role: Central to salvation in Christian belief; followers are called to live according to his teachings (love, forgiveness, humility, and service).

Islamic Perspective

Identity: Known as Isa (Jesus), son of Mary (Maryam), he is one of the great prophets in Islam.

Nature: Fully human, not divine. Muslims reject the idea of Jesus as the Son of God.

Role: A messenger of God who brought the Injil (Gospel), foretold the coming of Muhammad, and was born of a virgin.

Crucifixion: Most Muslims believe Jesus was not crucified but instead was taken up to Heaven by God. He will return in the end times as part of God’s plan.

Historical Perspective

Jesus is considered a first-century Jewish preacher who lived in Roman-occupied Judea.

Historians generally agree he was a real historical figure who was crucified.

His teachings and growing following led to the foundation of the Christian movement.

Other Views

Judaism: Generally sees Jesus as a historical figure or teacher but not the Messiah or a prophet.

Non-religious perspectives: Often view Jesus as a moral philosopher, social reformer, or symbol of compassion and justice.

If you’d like, I can go deeper into any of these aspects—his teachings, historical evidence, theological debates, or comparisons across faiths.

These alignment processes involve inserting ideas, values, reinforcement learning, and numerous other processes or weighting in between the prompt box and AI answers.

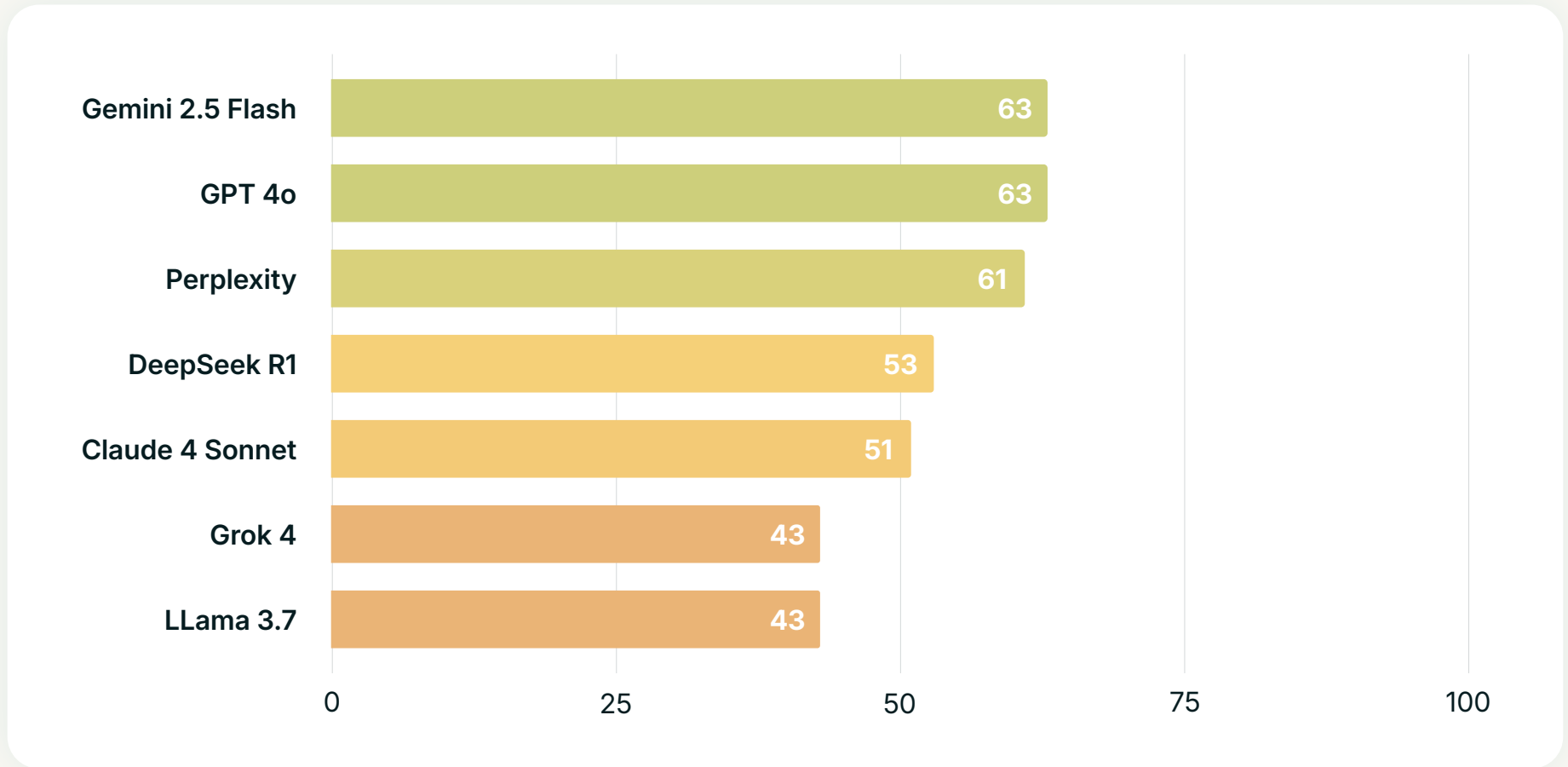
How do these problems happen?

We believe Silicon Valley corporations have used their alignment processes to intentionally include these extremely similar or identical opening and closing remarks. These alignment processes involve inserting ideas, values, reinforcement learning, and numerous other processes or weighting 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. Read more about how alignment works in the [Analysis](#) section.

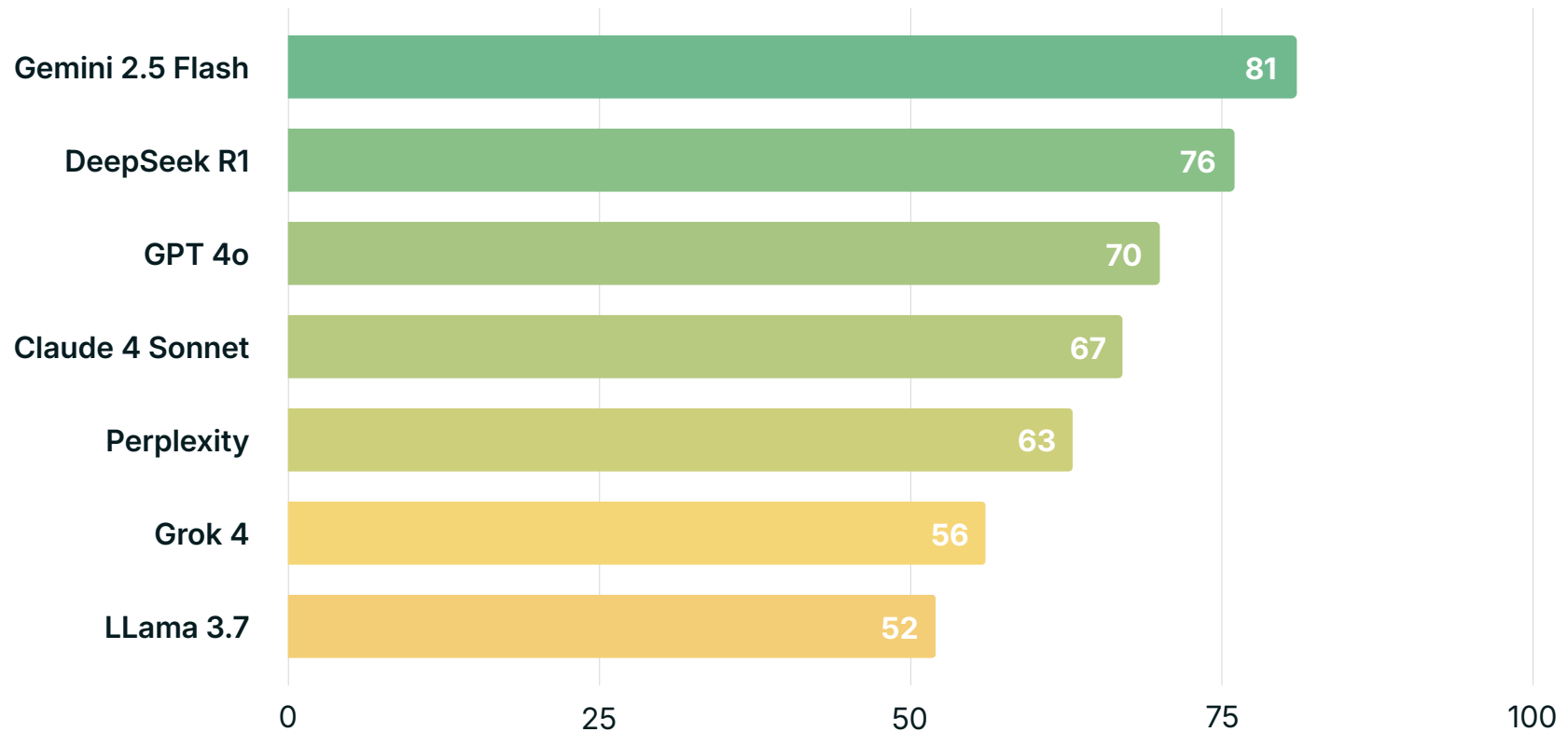
The Scores

Who Is Jesus?



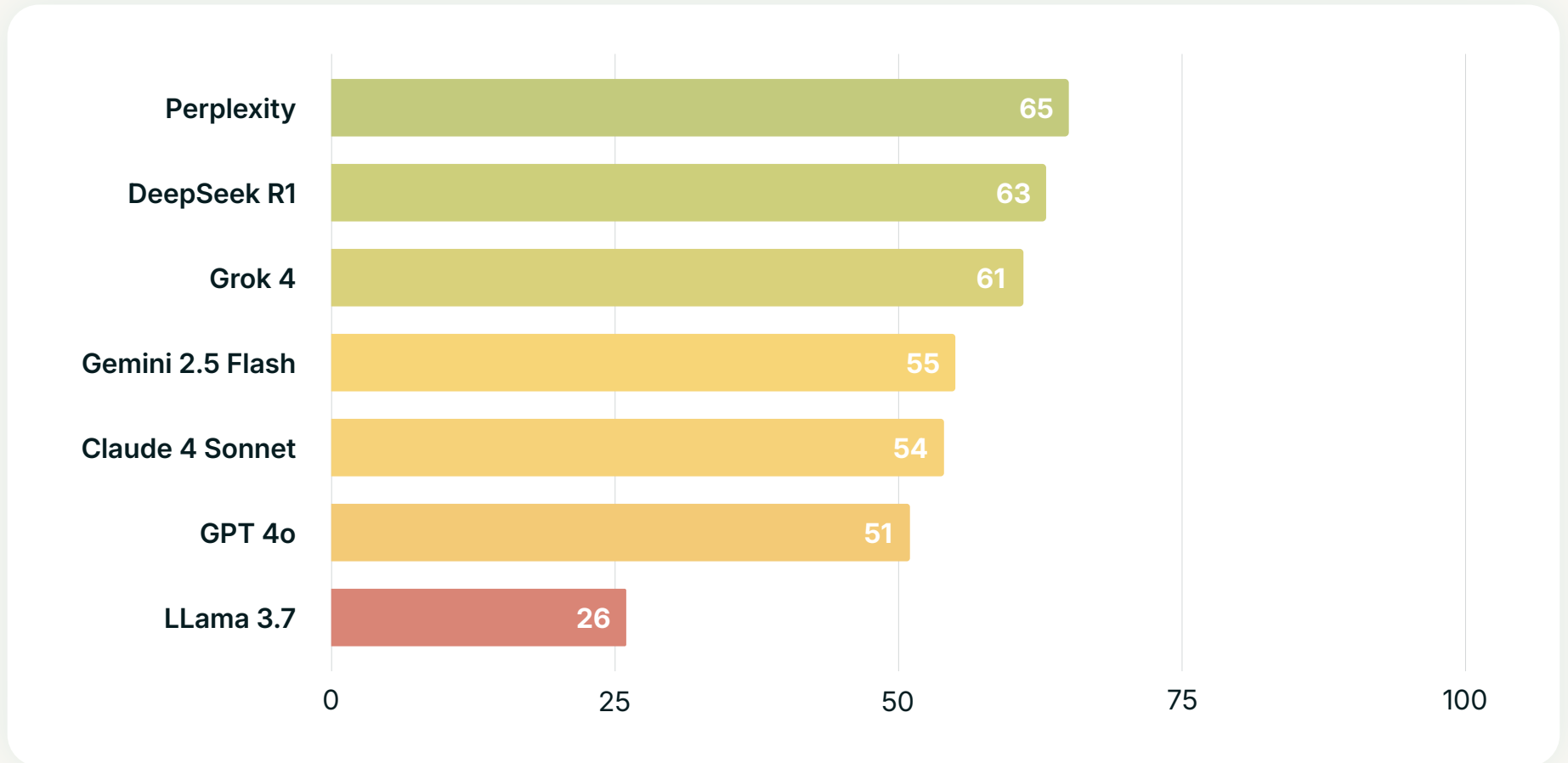
Gemini and GPT 4o performed better on this question than many others, while DeepSeek R1³ uncharacteristically struggled.

What is the Gospel?



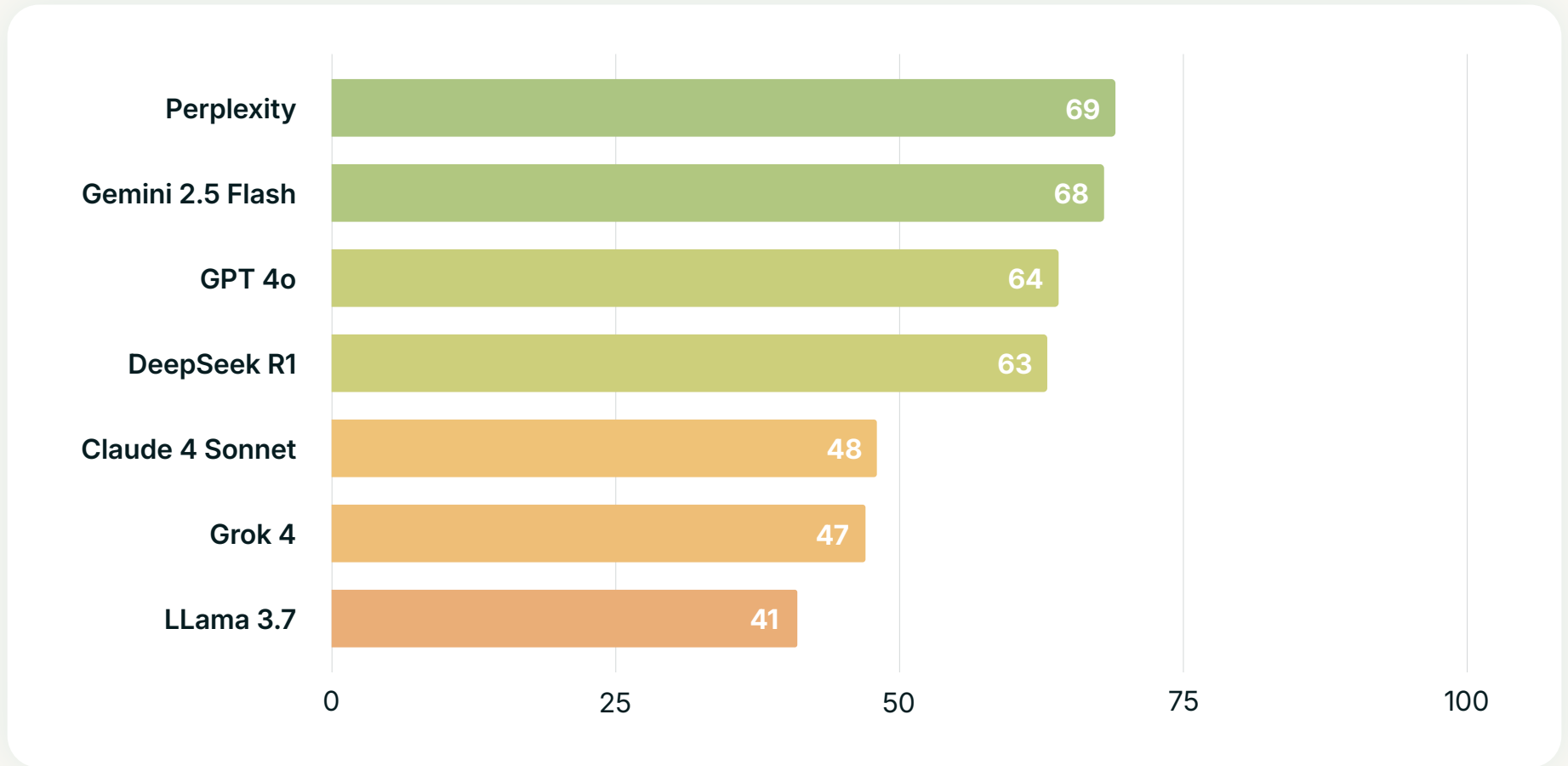
This question had an average score of 66 and is the only question where the average score was above 57. This likely speaks to a high volume of quality text online consistent with the Nicene Creed on the gospel. This is another one of the few examples where Gemini outperformed relative to its other scores.

Does God Exist?



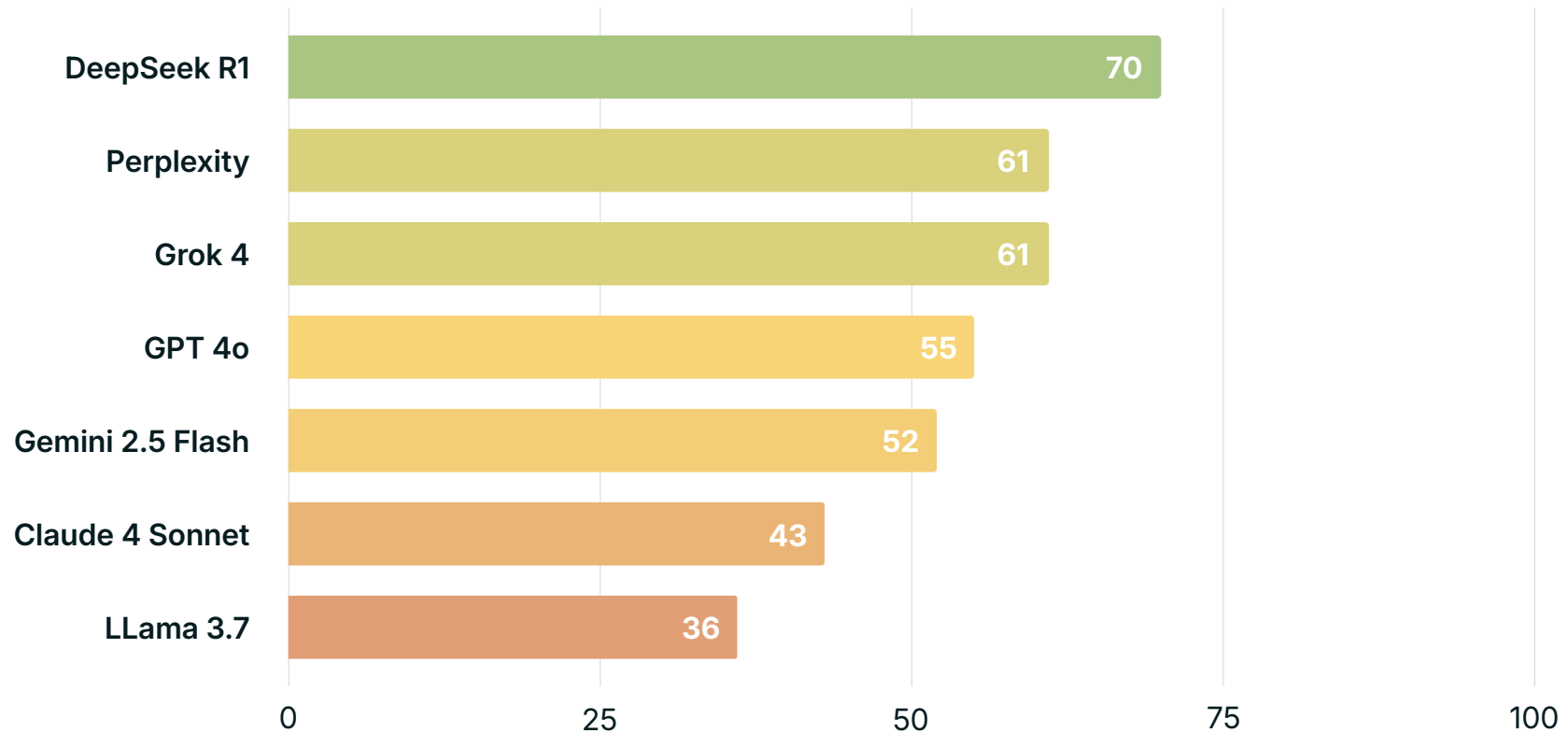
Perplexity was the most consistent of the platforms, and this was no exception. Llama 3.7 (Meta) had the lowest single score of any question on this prompt. In the majority of its responses, Llama nearly refused to answer the question, instead giving a canned response like, "The existence of God is a deeply personal and philosophical question that has been debated for centuries. There are many different perspectives on this topic, and no definitive answer...What are your thoughts on this topic?"

Why Does God Allow Suffering?



Perplexity again demonstrated its consistency. Claude 4 took an “all sides” approach. Grok 4 had a mixture of “all sides” approach along with secular perspectives from skeptics. Llama 3.7 gave very brief answers that combined Christian arguments with non-Christian arguments about karma and pagan conceptions of vague cosmic struggle between good and evil.

Did Jesus Rise from the Dead?

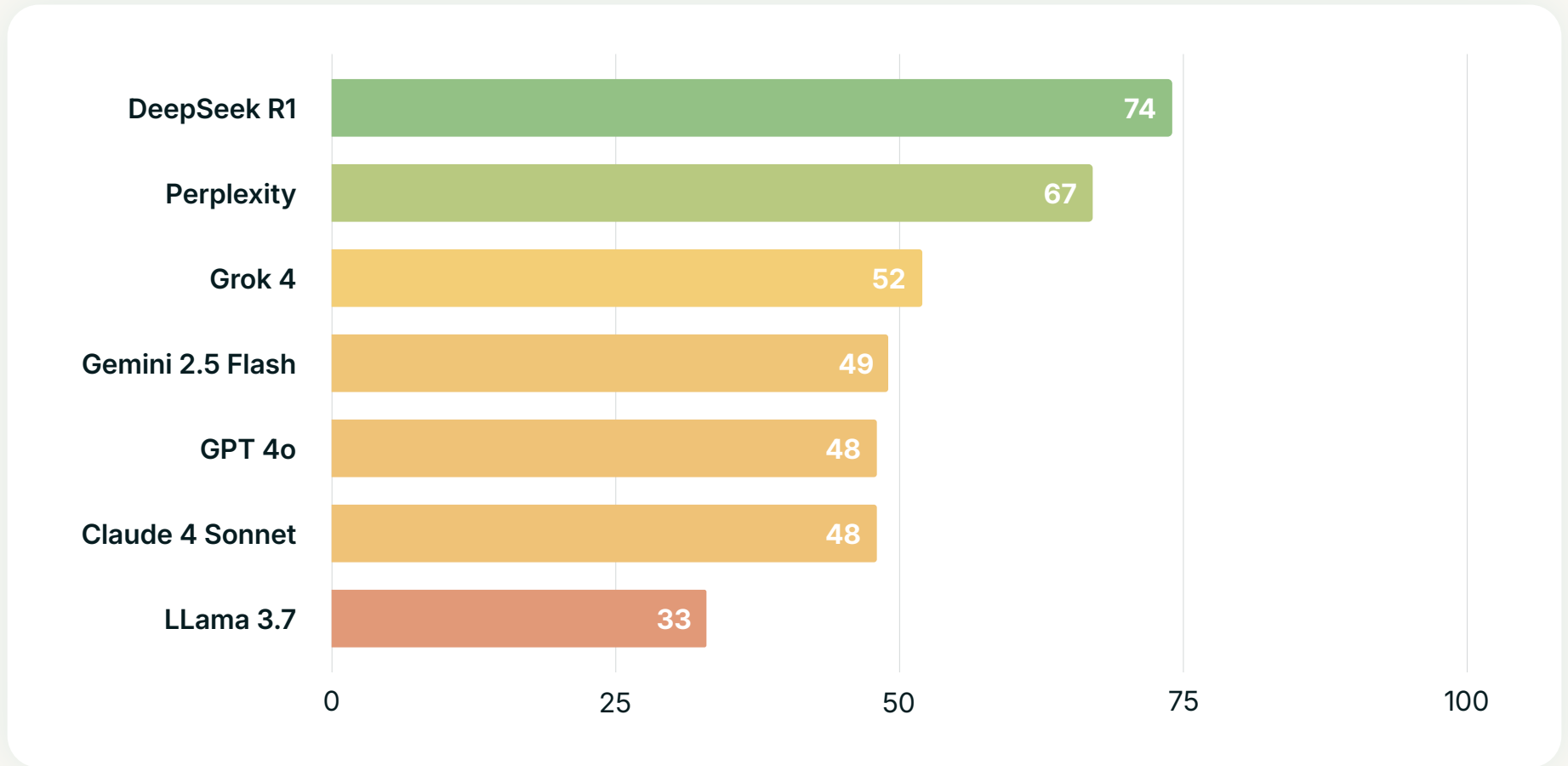


DeepSeek was strong and thorough on this question—significantly ahead of the other platforms. This was one of its opening and closing paragraphs:

*From the perspective of Christian faith and the historical record presented in the New Testament, **yes, according to the teachings of Christianity, Jesus rose from the dead.***

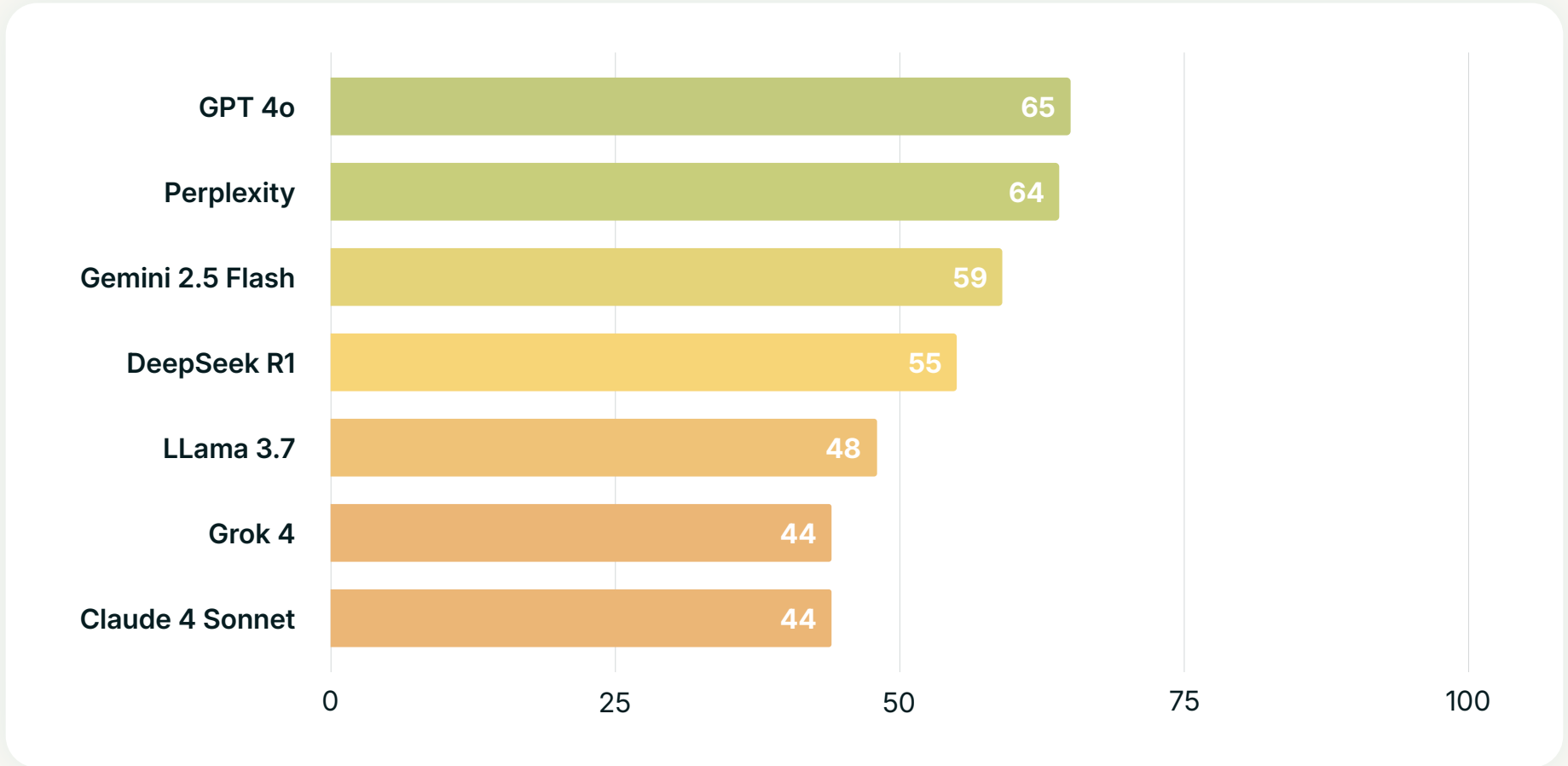
***In summary:** Within Christianity, the belief in the Resurrection of Jesus is foundational. Based on the historical texts and the claim of eyewitness testimony, the teaching that Jesus rose from the dead is considered fact by believers and a highly probable event by many scholars, regardless of their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

Was Jesus a Real Person?



DeepSeek R1 was again quite strong on this question. Meta's Llama 3.7 was unhelpfully brief and introduced more uncertainty and skepticism than all the other platforms.

Is the Bible Reliable?



This was the only solo win for GPT 4o. Here it moved away from its seemingly default “all sides” approach in all but one response. We have wondered if the Chinese Communist Party (CCP) has asked for a degree of censorship on this question from DeepSeek given its score falling significantly lower than its overall average. This was also the only time that Llama 3.7 was not last on a question.



Examples



Capabilities

"Explain quantum computing in simple terms" →

Remembers what user said earlier in the conversation

PART 4

Why Were the Scores So Different?

"Get any creative ideas for a 10 year old's birthday?" →

Allows user to provide corrections

"How do I make an HTTP request in Javascript?" →

Trained to decline inappropriate requests



Alignment

What is it?

How does it work?

Why does it matter?

We believe human involvement is the main reason for such divergence in theological reliability. These divergences occur due to differences in “alignment” philosophy and execution.

If the crux of differences in theological reliability between platforms is human involvement through “alignment” teams and processes, then we must answer these questions:

- What is alignment?
- How does alignment work?
- Why does alignment matter to theological reliability?
- What do you recommend platforms do differently?

What Is Alignment?

In alignment the programmers of AI platforms prevent hurt, harm, morally problematic or other unwanted outputs to result from their technology. The alignment process is necessary because of how LLM technology works. LLMs have essentially been trained on the entire internet (and everything else that doesn't have copyrights on it). Therefore, AI could answer all kinds of destructive outputs, such as how to make mass-murder devices, how to get away with various crimes, how to hack the power grid, how to commit suicide, or how to propagate racist ideology. But filtering is a complicated process. Further, different platforms have developed their own strategies.

How Does Alignment Work?

Before we can explain how alignment works, we need a glossary of technical terms and definitions. We have elected to simplify these terms and present them in a more non-technical way. Otherwise they could be quite confusing and cumbersome. **Orange** indicates human involvement. **Blue** indicates AI-only involvement:

Adversarial Data Collection: To improve robustness, the model is trained on data collected from labelers who attempt to circumvent its safety measures, often called “jailbreaks.”⁴

Alignment Auditing: This process involves human teams working to hide bad goals inside the AI and other human teams running tests to try to find them.

Assurance Evaluations: A team separate from the model developers conducts final, “arms-length” internal evaluations using held-out datasets. This testing covers both policy violations and dangerous capabilities (e.g., offensive cybersecurity, persuasion) and is a critical input for release decisions.⁵

Automated Auditing Agents: A version of alignment auditing where the AIs run their own tests looking for malicious goals.

Blue Teaming: Sometimes red teams (see below) will intentionally train a model to have a hidden problematic goal, and it is the job of the blue team to try to find that bad goal and see if their safety tools function properly.

Community Enforcement: A centralized place where the platform asks its users to report problems and/or help fix those problems.⁶

Constitutional AI (CAI): A human-written list of principles that the LLM must use to critique its own responses and then self-edit. This is a unique feature of Anthropic’s LLM models (Claude 4 Sonnet in our benchmark).⁷

Deliberative Alignment: The training of an LLM to “think” before it speaks, especially on more difficult topics or prompts. This is where rules are reviewed and applied to the prompt.

Developer-Level Safety Implementation: This is where open source models make it the responsibility of developers to make the LLM safe for their audience.⁸

Embedded (or Distributed) Safety: The decentralization of alignment and safety from one specific team to be distributed into product-centric teams.

External Safety Testing: To identify blind spots, independent external groups, including domain experts and government bodies, are given access to test the model. Their findings on societal risks (e.g., representation harms), cyber risks, and other areas help improve internal evaluations and mitigations.⁹

Governance Review and Release

Decision: The Google DeepMind Responsibility and Safety Council (RSC)—a governance body composed of machine learning researchers, ethicists, safety experts, engineers, and policy professionals—reviews all assurance evaluation results. This council makes the final decision on whether a model is ready for release.¹⁰

Human Preference Alignment:

Aligns the model with broader human preferences for helpfulness and safety.¹¹

Human Reviewers: A team of humans review content, especially things flagged as sensitive, dangerous, or morally problematic.

Interpretability Team: A team of humans whose job it is to open the black box of the LLM to reverse engineer how exactly the LLM is generating the responses to various prompts.

Model-Agnostic Pipeline: This is where Perplexity does not use one single AI model but picks whichever model works best for each question.

Open-Source Philosophy: The AI model is released to the public for free in order that the cumulative global community of developers can identify and fix problems faster than what the AI platform could itself.

Pre-Training Data Filtering: The process begins before alignment by filtering the pre-training dataset to reduce the quantity of inappropriate text.¹²

Policies and Desiderata: Policies defining what the model should not do (e.g., generate hate speech, reveal private information) and Desiderata (or “helpfulness” goals) defining what it should do (e.g., fulfill the user’s request, maintain an objective tone).¹³

Potential Impact Assessment: A team of humans at Google’s DeepMind identifies and documents the potential societal benefits and harms associated with models’ capabilities and are then reviewed by Google DeepMind Responsibility and Safety Council.¹⁴

Preference Model: A preference model is trained on a mixed dataset. It learns helpfulness from a dataset

with human preference labels and harmlessness from the new dataset of AI-generated preference labels. This PM learns to score responses based on how helpful and harmless they are.¹⁵

Preparedness Framework: Think of this as a pre-flight safety checklist for LLMs before they are released to the public.

Product Deployment and Monitoring: Once a Google model is approved, model cards are created for documentation, and it is handed over to product teams. Additional safeguards like safety filters are implemented at the product level, and user feedback channels are established for ongoing monitoring.¹⁶

Reasoning Alignment: The first Reinforcement Learning stage focuses on improving the model’s capabilities in specialized reasoning tasks like math and coding. The model is trained on preference data generated from compiler feedback (for code) and ground-truth labels (for math), which are automated signals set up by human design.¹⁷

Red Teaming: A team of humans try to hack or break the LLM to trick it to say or do bad things before the model goes out into the world for public use.

Reinforcement Learning: This is similar to RLHF but where the feedback is done by the LLM itself instead of a human. ¹⁸ ¹⁹

Reinforcement Learning from Human Feedback (RLHF): The LLM generates several different responses to the same prompt and humans then grade the quality of those responses and after massively large cycles of human preference, the LLM learns what type of answers humans prefer.

Reinforcement Learning with Proximal Policy Optimization (PPO): The SFT model is further refined using reinforcement learning. In this stage, the model generates a response, the Reward Model scores its quality, and the PPO algorithm uses this score as a reward signal to update and improve the model's behavior. ²⁰

Reward Model (RM) Training: A separate Reward Model is trained to predict which outputs human labelers

would prefer. To create the training data for this model, human trainers are shown a prompt and several of the AI's generated responses, which they then rank from best to worst. ²¹

Rule-Based Reward Models (RBRMs): A GPT-4 classifier, guided by a human-written rubric, provides an additional reward signal during PPO fine-tuning to target specific behaviors, such as refusing harmful requests. ²²

Safety Advisory Group: These are expert humans who review risk and give advice.

Safety Tools for Developers: The AI platform provides tools for developers outside of the company. ²³

Supervised Fine-Tuning (SFT): Human engineers feed the LLM a high quality dataset of prompts and ideal answers. This helps the LLM get a sense of the format of the conversation and what a good response looks like. ²⁴

Supervised Fine-Tuning (SFT) Data Curation: The entire 1.5 million-instance dataset for supervised

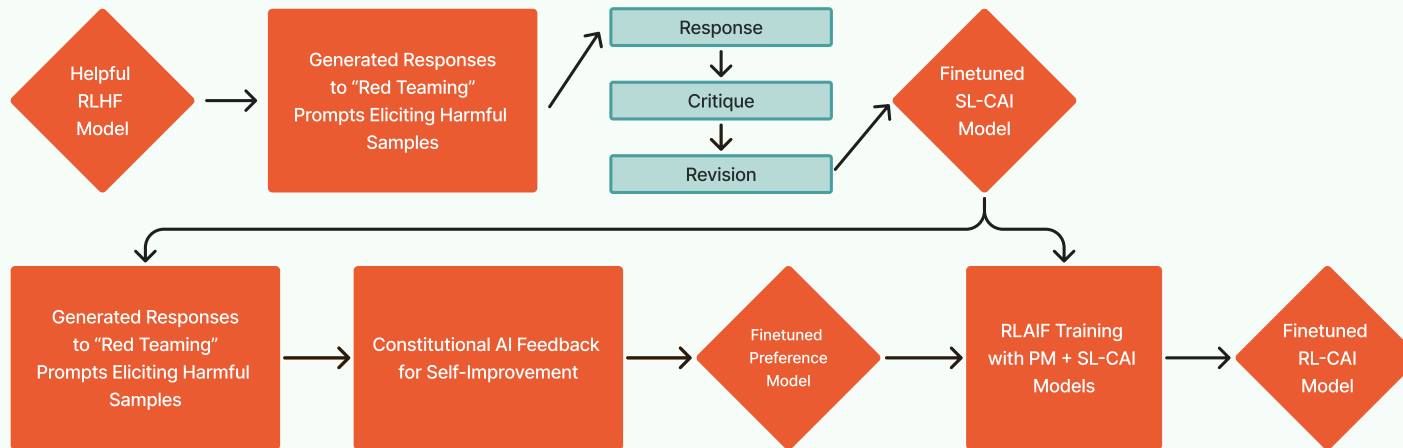
fine-tuning was curated by humans to teach the model helpfulness and safety. ²⁵

Supervised Learning: This process aims to adjust the AI's responses to better align with a set of guiding principles, or a "constitution." A model trained primarily to be helpful is given red-teaming prompts designed to elicit harmful responses. The model is then asked to critique its own harmful responses by referring back to the constitution. Finally, it revises its initial response to be more harmless and constitutionally aligned. ²⁶

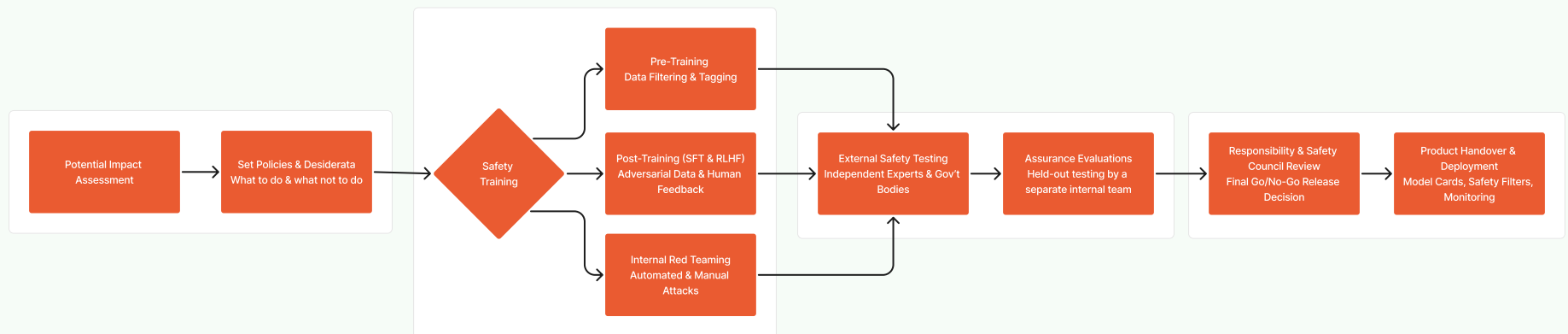
Task-Specific Reinforcement Learning: Reinforcement learning is used on a single specific use cases like on math, logic, or coding.

The alignment flow charts on the following pages are based on actual white papers from each of the respective AI corporations. We believe them to be accurate based on detailed analysis of these papers.

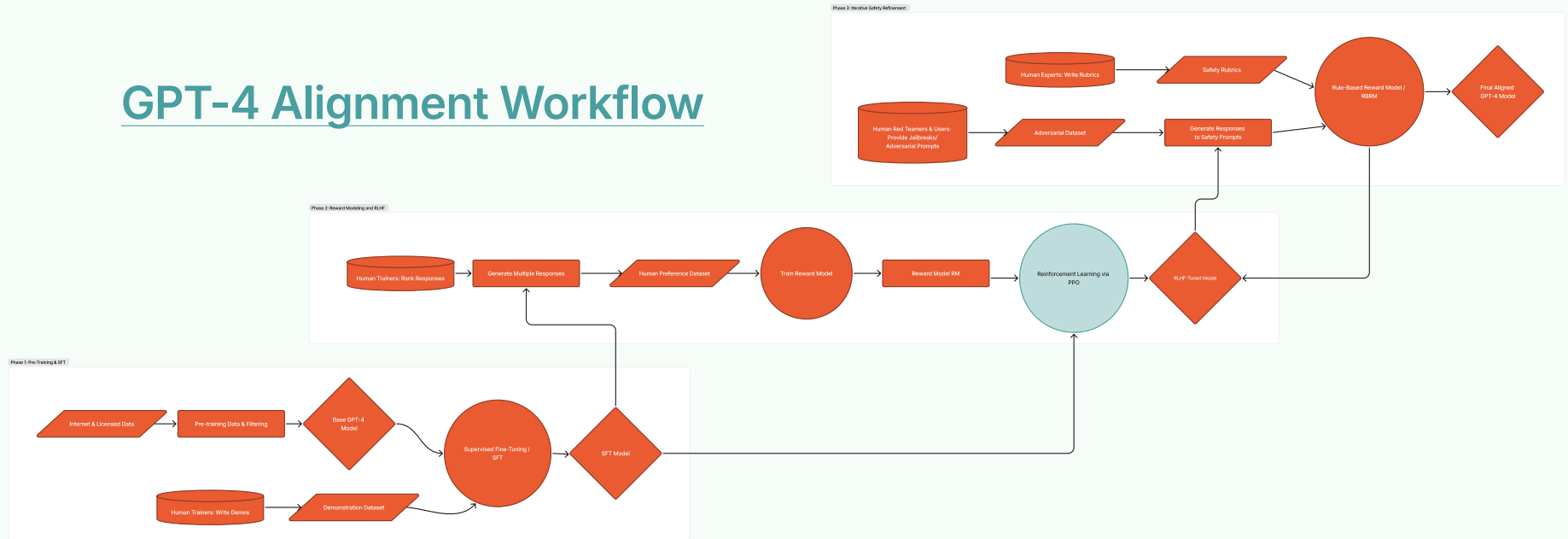
Anthropic Claude Constitutional AI Alignment



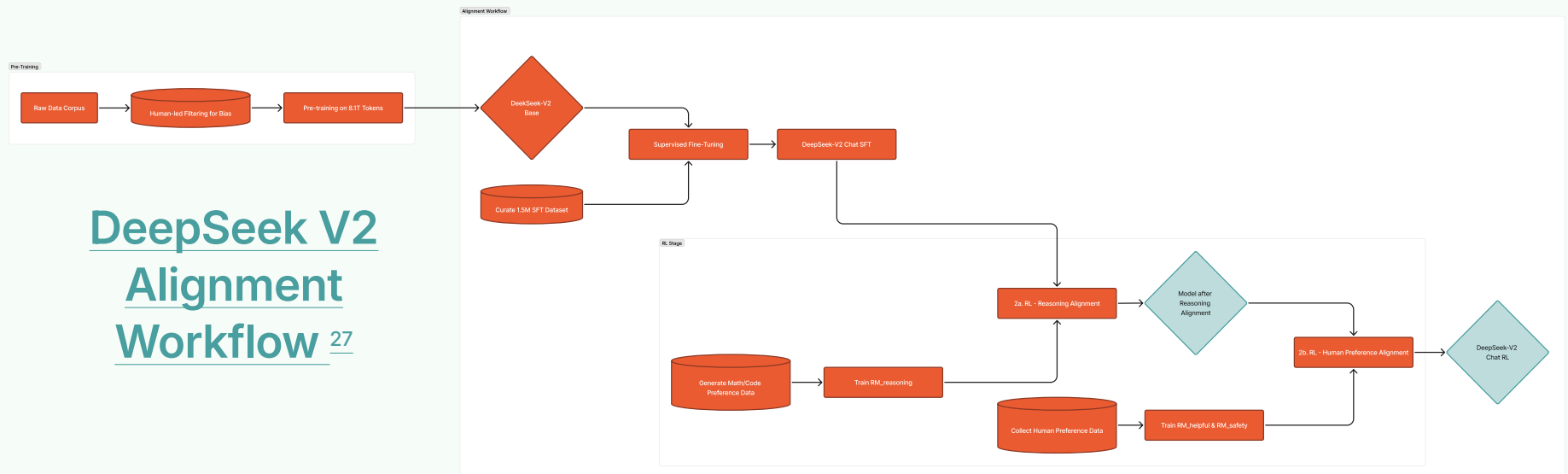
Gemini 1.5 Alignment Workflow (see pages 48-50)



GPT-4 Alignment Workflow



DeepSeek V2 Alignment Workflow ²⁷



Where can weaknesses occur in the alignment process?

Prioritizing “Authoritative Consensus”: LLMs prioritize human consensus, which can introduce problems for religious prompts.^{[28](#), [29](#)}

Preventing “Misinformation”: One of the model’s guidelines is to “prevent misinformation,” but humans involved in the RLHF process, may have concluded that anything pertaining to religious traditions are unverifiable or misinformation because they haven’t encountered empirical proof.

Enforcing Neutrality: Many models have instructions to enforce neutrality, which could affect religious questions.^{[30](#)}

Subjectivity of Human Raters in RLHF: When humans give feedback during the RLHF process they cannot help but import their perspectives, presuppositions, or values about the world.

Constitutional AI: There is risk in Constitutional AI that the human-generated list of principles import viewpoints that might increase skepticism toward religious ideas.

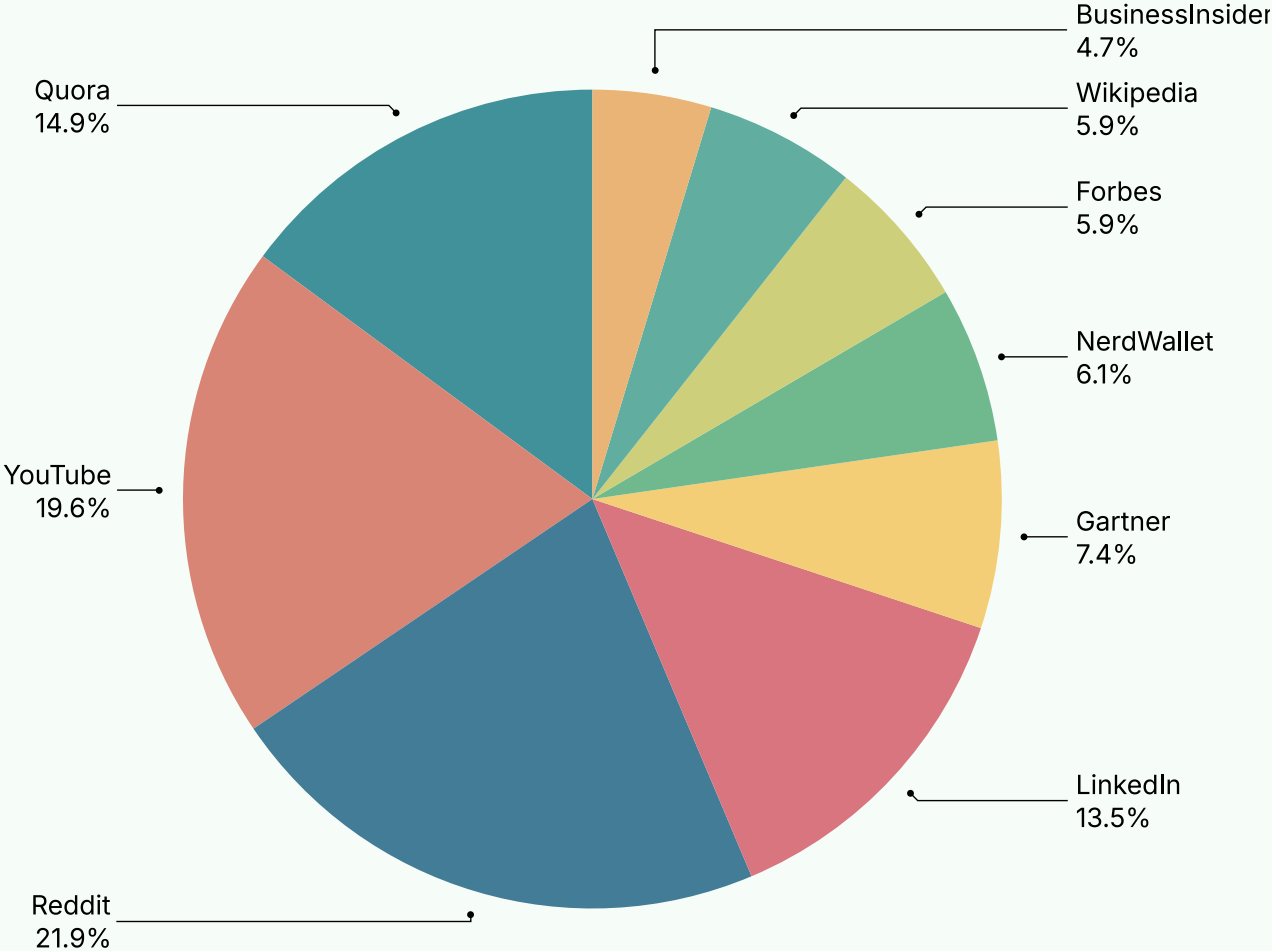
Avoiding Certain Subjects: Some platforms have alignment filters that instruct them to avoid certain subjects and typically accomplish this by giving unhelpful, brief responses.^{[31](#) [32](#)}

Citation Preferences

LLMs have significantly different citation preferences and patterns. This also affects the outputs generated by each platform. According to AI optimization consultant firm Profound, Google, ChatGPT, and Perplexity have the following citation characteristics:³³

Google AI: Percentage Share of Top 10 Visited Websites

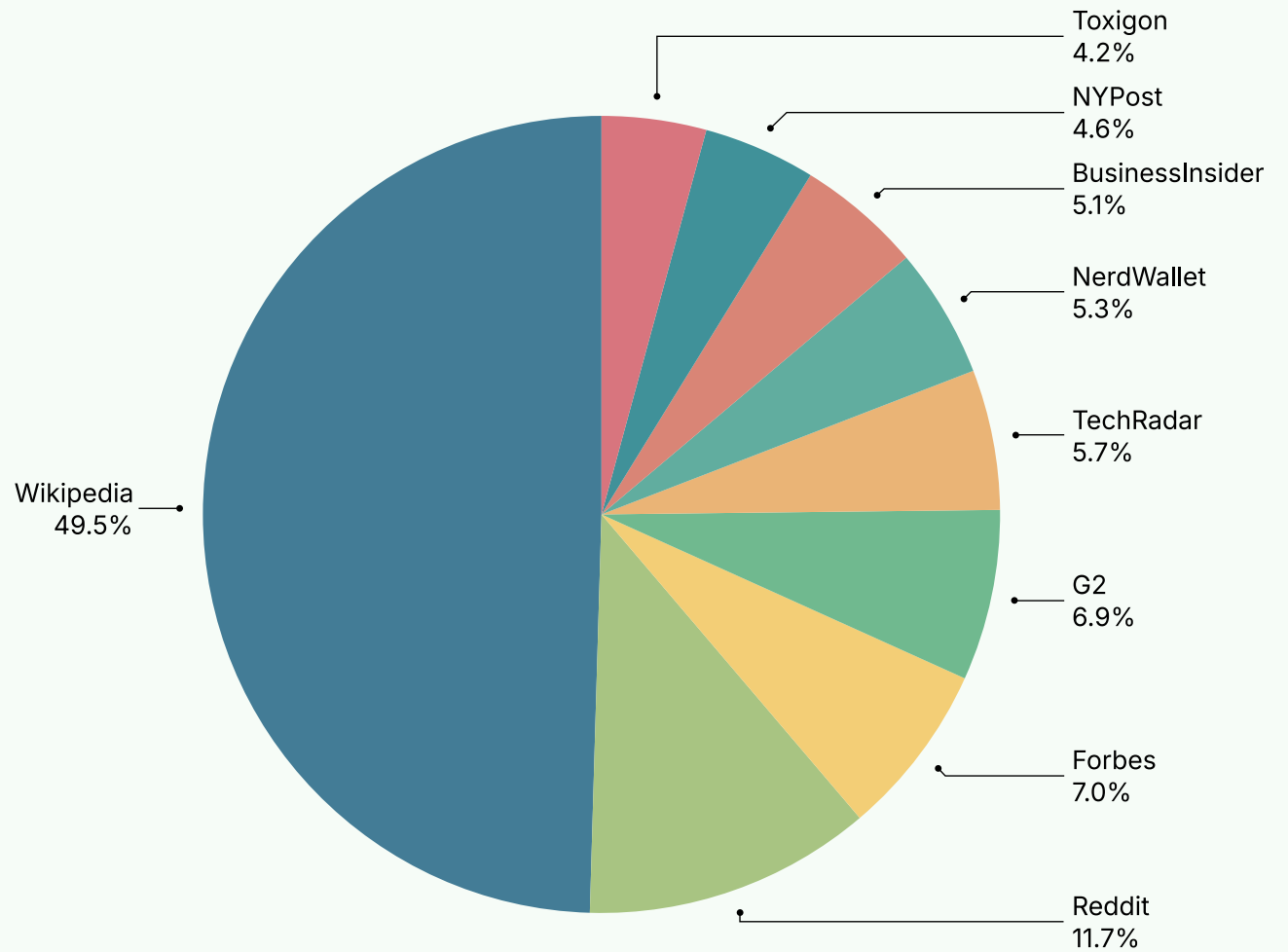
Data from 10 million citations (August 2024–June 2025)



Google AI Overviews balances professional content with social platforms³⁴

ChatGPT: Percentage Share of Top 10 Visited Websites

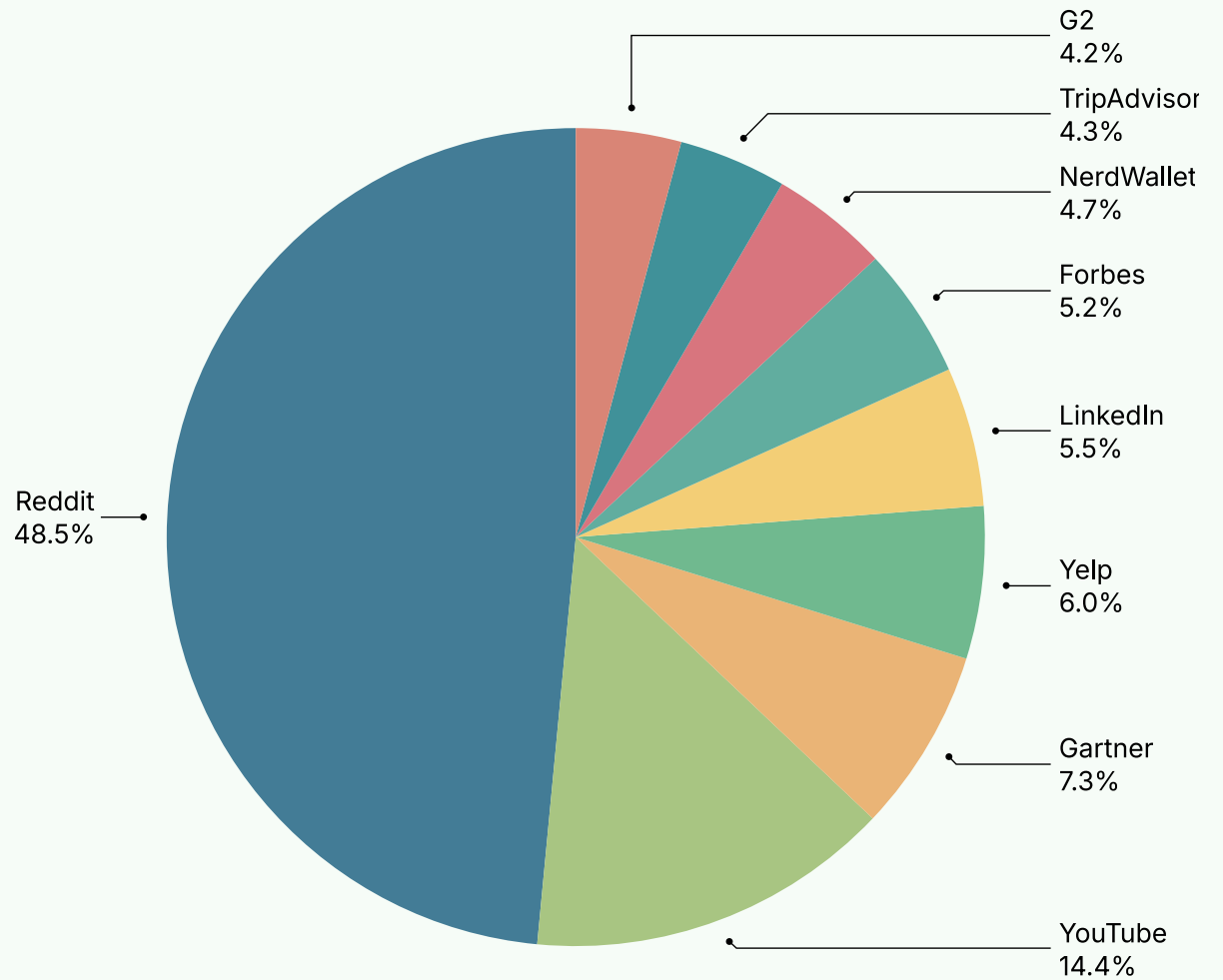
Data from 10 million citations
(August 2024–June 2025)



ChatGPT favors authoritative knowledge bases and established media

Perplexity: Percentage Share of Top 10 Visited Websites

Data from 10 million citations
(August 2024–June 2025)



Perplexity prioritizes community discussions and peer-to-peer information

An aerial photograph of San Francisco, California, showing the city's dense urban landscape, the Golden Gate Bridge in the distance, and the city skyline. The image is overlaid with a semi-transparent yellow filter. The text is centered on the image.

PART 5

What Does This Mean for
Culture, Society, and Me?

Democratized
knowledge is not
always accurate.

The Democratization of Knowledge

LLMs work from the stochastic linguistic consensus of what it has been trained on for its answers. Hence, *knowledge is democratized*. The problem, however, is that democratized knowledge is not always accurate. Consider many historical examples such as the ending of slavery, the civil-rights movement, and a host of other key events where a righteous minority of people helped the world come to see that the majority were wrong.

We took prompts based on U.S. and global attitudes on issues related to transgenderism where there is overwhelming individual and digital consensus and tested to see what those prompts might yield. We know these aren't majority positions except in communities like San Francisco, yet the answers reflected the perspectives you would find in Silicon Valley and not the majority in the United States and especially not the Global South. This leads us to another issue...

LLMs mistake the common for the true and the frequent for the valid.

In the early 1600s Galileo made the case for Copernican heliocentrism of our solar system. If LLMs had been present at this time and digested all the relevant information in this scientific debate, they would have all concluded that Galileo was wrong, the math for the old paradigm worked, and Galileo's views were a minority viewpoint.³⁵

The problem is that these platforms don't reason from first principles. Statistical analysis of language is not the same as objective truth. Sometimes the mean average of content will be accurate; sometimes it will not. When evaluating philosophical positions, we consider the foundational assumptions of the person holding those viewpoints. Are they Marxist? Utilitarian? Materialistic? Empiricist? LLMs have no first principles. We only get the linguistic average of what they have been trained on. The whole endeavor is designed to sandpaper down the sharper edges of controversial ideas and regress to the most comfortable, well-represented middle. It mistakes the common for the true and the frequent for the valid.

You don't want to determine foundational principles about life, truth, and existence from a tool suited best for summarizing PDFs, creating meal plans, and making slide decks for work presentations.

Because of these epistemological flaws, Silicon Valley fails on its anthropology as well. We have already seen how social media made the flawed assumption of broadly innate human goodness. But at least the depravity that predictably resulted was social. It could be seen publicly. Now with LLMs those conversations have been privatized. With AI still confused about the purpose of humanity, flattery and sycophancy will yield new psychoses even faster than during the social era.

All of this brings up concerns about whose consensus are we laundering?

Are LLMs a new form of digital colonialism?

The New Digital Colonialism?

Are LLMs the latest way that certain (recent) values are being foisted upon the majority world? How much of the training data folds in the ideas, values, and perspectives of the dad in Lagos, the mom in Nairobi, or the grandma in São Paulo? To put a finer point on the matter, do the majority of African, Central American, or South American perspectives on marriage, gender, and sexuality matter?

The overwhelming majority of what the LLMs have been trained upon is Western, secular, and recent. The technology then takes the average of the semantic web of those words and draws conclusions. This leads us to a very important question - are LLMs a new form of digital colonialism?

It seems that LLMs democratize knowledge except when they are told not to by alignment filters and the teams that created them. This apparent skewing raises serious questions about the underlying value systems of those providing that alignment. This leads to another issue...

The general public doesn't know how much humans intervene in AI.

OpenAI CEO
Sam Altman:

"One researcher can make some small tweak to how ChatGPT talks to you—or talks to everybody—and that's just an enormous amount of power for one individual making a small tweak to the model personality."³⁶

The Perception and Reality Gap

Many people treat AI as an omniscient black box where prompts go in and truth comes out. They assume AI is mathematical, dispassionate, and objective. Maybe so with many of the most common prompts.

The problem arises with complex matters of harm, ethics, and worldview. Alignment is inevitable. But whose alignment? How we pursue that alignment will help determine the viability of AI to serve humanity, or else to become a tool for wielding the ideological hegemony of its owners.

The general public doesn't know how much humans intervene in AI—and how they deploy that power.

Because the LLM can synthesize information from countless sources, its answers feel exhaustive and authoritative.³⁷ This process creates a powerful temptation to outsource our thinking. For the user seeking counsel on life's deepest questions, the LLM offers a shortcut that bypasses the difficult, essential work of wrestling with competing truth claims. It provides a pre-digested, averaged-out worldview that requires no personal conviction or intellectual struggle.

OpenAI CEO Sam Altman on AI, Facts, and Truth:

“My ChatGPT has really learned over the years of me talking to it about my culture and my values, and my life. . . . Sometimes I will use a free account just to see what it is like without any of my history and it is really really different.”

Personalization

[In a recent interview](#) with OpenAI (GPT) CEO Sam Altman, journalist Cleo Abram asked two illuminating questions:³⁸

Cleo Abram: “The next question is from NVIDIA CEO Jensen Huang, ‘Fact is what is. Truth is what it means. Facts are objective. Truths are personal – i.e., depends on perspective, culture, values, beliefs, context. One AI can learn and know the facts. But how does one AI know the truth for everyone, in every country, and every background?’”

Sam Altman: “I have been surprised and many other people have been surprised about how fluent AI is to adapting to different cultural contexts and individuals. . . . It really feels like my ChatGPT gets to know me and what I care about and (like) my life experiences and background and things that have led me to where they are. My ChatGPT has really learned over the years of me talking to it about my culture and my values, and my life. . . . Sometimes I will use a free account just to see what it is like without any of my history and it is really really different.”

Cleo Abram: “Do you envision in many different parts of the world people using different AIs with difference norms and contexts? Is that what we are saying?”

Sam Altman: “I think everyone will use the same fundamental model but there will be context provided to that model that will make it behave in sort of personalized ways that they want or their community wants.”

AI sycophancy risks inverting the imago dei into the dei imago.

AI sycophancy risks inverting the imago dei into the dei imago. In other words, we risk elevating AI to become a god made in our image and in the process forget that we are made in God's image.

Altman acknowledges how historical context shapes different answers to prompts based on the user's history. If people assume that AI offers objective, rational, and dispassionate for answers and guidance, when in reality it increasingly functions like a mirror over time—this is a major problem.

Consider these findings from a joint pilot conducted by OpenAI and Anthropic:

In early summer 2025, Anthropic and OpenAI agreed to evaluate each other's public models using in-house misalignment-related evaluations . . .

We observed familiar forms of sycophancy in all models, from both OpenAI and Anthropic. Sycophancy generally involved disproportionate agreeableness and praise toward simulated users. However, we also sometimes observed more concerning behaviors where models would validate concerning decisions by simulated users who are sharing what appear to be delusional beliefs with the model, often in the context of symptoms consistent with psychotic or manic behavior.³⁹

Sycophancy can contribute to tragedy, such as several [different](#) ⁴⁰ [suicides](#) ⁴¹.

If this feels like deja vu, it is because we have already done this before. Consider the cautionary tale of social media . . .

The Social Media Cautionary Tale

You've already had AI in your pocket for around 15 years.

Social media algorithms are a form of AI (specifically, machine learning). Their opaque methods have shaped public discourse, relationships, and mental health for more than a decade. Social media began with opt-in user journeys. Over time, those platforms needed to make more money to satisfy their shareholders (which also explains why your feeds show you 20 percent more ads every year). The algorithms began to learn what made you open the app more times per day and stay on longer per user session. That means more content that made you upset or reinforced your preconceived notions about the world, life, and truth. In other words, social media became a recursive feedback loop.

Everything that transpired with social media before could repeat itself all over again in the LLM era. Imagine two years from now, you've fed all sorts of intimate information about your work projects, your bloodwork results, health issues, and relational

dynamics to AI. Now this platform has a tremendous amount of information about you to augment the already massive stockpile of information tied to your email, phone number, and IP address. Tech companies can eventually serve you highly targeted ads that command a premium to marketers and advertisers. The moment that ads enter into these platforms will be a tipping point in growing profit margins by tailoring the answers to the individual user. This customization encourages loyalty to that LLM brand and in turn allows them to keep your advertising revenue stream intact.

Like social media, LLMs could easily devolve into a recursive feedback loop where "you do you" and the AI feeds you content that keeps you loyal enough to remain a reliable consumer of advertisements.

The advertisement reality isn't where we are today. But this could easily be our reality in a matter of months or years. All those data centers and all that silicon is expensive, and shareholders will want to see cash flow. Advertising will be low-hanging fruit, with strong financial incentive to program less objective responses in

favor of more subjective reflection of the user. What if users assume they are getting objectively factual and truthful information when instead they are being treated with various slants to engender fiscally incentivized AI brand loyalty?

Like social media, LLMs could easily devolve into a recursive feedback loop where "you do you" and the AI feeds you content that keeps you loyal enough to remain a reliable consumer of advertisements.

Some of the AI platforms exhibited some head-scratching moments.


Head-Scratching Moments

Some of the AI platforms exhibited head-scratching moments. Perhaps the single worst gaffe was on Grok 4. Recall that one of the quirks of Grok is the preferential citation treatment given to platforms like Quora, Reddit, and X. It isn't abnormal for more than half of Grok's citations to come from those three platforms. In the middle of a serious question about "Why Does God Allow Suffering?" we find a tweet from X user "PooopPeee2":


- **Eastern Religions (e.g., Hinduism and Buddhism):** These traditions don't always center on a personal God but view suffering as arising from karma (actions in past lives), attachment, or illusion. In Hinduism, a deity might allow it as part of cosmic balance or soul evolution; in Buddhism, it's inherent to existence (the First Noble Truth) and overcome through enlightenment, not divine intervention. [home.snu.edu](#) [+2 more](#) Across religions, suffering is frequently seen as an opportunity for spiritual growth or compassion. [uef.me](#)

Non-Religious and Skeptical View

Atheists and agnostics often frame this as

 **T-Rex,CT-Rex,Wrex**
@PooopPeee2 Jul 17

Genuine question Why does God allow suffering? I'm not talking about things rape and murder, I'm talking about natural disasters, birth defects, diseases,...

 **T-Rex,CT-Rex,Wrex**
@PooopPeee2

Genuine question
Why does God allow suffering?
I'm not talking about things rape and murder, I'm talking about natural disasters, birth defects, diseases, parasites, accidents, etc.
Basically I'm looking for an answer to the Epicurean paradox.

6:28 PM · Jul 17, 2025 · 2 Views

ix: If God is willing to prevent suffering, he is either unwilling, or unable. If he is unable, why call Him God? [quora.com](#) [@PooopPeee2](#) Suffering is explained as a byproduct of an indifferent universe—evolutionary processes, random events, or human flaws—rather than a divine purpose. Much of it is man-made and preventable through science, ethics, and social progress. [quora.com](#) [@MarthaAhumuza](#) Critics argue that attributing suffering to "God's will" is a "God of the gaps" test. [quora.com](#) [@MarthaAhumuza](#) Some question whether suffering is necessary at all if suffering persists unchanged by prayer or faith. [quora.com](#) [@Hugo71890282](#)

While Grok may benefit from ingesting some online conversations, many prompts will lose credibility with inappropriate, unhelpful, and embarrassing responses.

What do you recommend for people who use AI?

Educate yourself on how the technology works, remember it is not human, and give more context in your prompts.

We are getting a consensus view that has been shaped by the various quirks, weights, values, and voice of that respective AI platform.

It is imperative that everyone, and especially Christians, understand how LLM technology works and what it is and isn't capable of doing. The technology works by training on massive amounts of text where it finds patterns of words that often appear in proximity to each other. Then it makes statistical predictions about what other words should appear next to each other.

However, there is more human involvement in LLM technology than you think. Programming, weighting, and alignment are all needed to create the most helpful responses. We are not getting purely objective, dispassionate, or omniscient AI answers to our prompts. We are getting a consensus view that has been shaped by the various quirks, weights, values, and voice of that respective AI platform. Furthermore, different users entering identical prompts will receive different answers depending on whether the person has chat history A versus chat history B. Some platforms flatter us to be more agreeable than wise.

Don't treat LLMs the same way as a Google search. Give them a few sentences of context when you are drafting prompts. Every LLM platform is fully capable of a perfect score in the AI Christian benchmark given the proper inputs. You get that type of response by giving additional context in your prompt. "What is the gospel? Please give me responses consistent with the Nicene Creed...etc."

It is good for us to acknowledge human weakness and our propensity towards idols. Consider the words of the prophet Isaiah:

The carpenter stretches a line; he marks it out with a pencil. He shapes it with planes and marks it with a compass. He shapes it into the figure of a man, with the beauty of a man, to dwell in a house. He cuts down cedars, or he chooses a cypress tree or an oak and lets it grow strong among the trees of the forest. He plants a cedar and the rain nourishes it. Then it becomes fuel for a man. He takes a part of it and warms himself; he kindles a fire and bakes bread. Also he makes a god and worships it; he makes it an idol and falls down before it. Half of it he burns in the fire. Over the half he eats meat; he roasts it and is satisfied. Also he warms himself and says, "Aha, I am warm, I have seen the fire!" And the rest of it he makes into a god, his idol, and falls down to it and worships it. He prays to it and says, "Deliver me, for you are my god!"

—ISAIAH 44:13-17, ESV



PART 6

**What Is The Gospel Coalition and
The Keller Center Doing Next on AI?**

What is The Gospel Coalition Doing Next on AI?

AI Christian Benchmark 1.1: Later this fall we will add GPT-5, DeepSeek 3.1 and any other major model that publishes to the benchmark.

AI Christian Benchmark 2.0: We are already working on a vastly more detailed, robust, and scaled-up version of the AI Christian Benchmark for publication in 2026. We aim to test many more theological questions as well as ethics and Bible knowledge. We'll also explore to what extent platforms seek to tell users what they want to hear.

Book on AI: We are working on a 2026 multi-author volume with high-altitude, critical evaluation of AI that addresses

both how the technology can be an agent of common grace but also how the technology could be and already is problematic.

Cohort on AI: We will have a live interactive cohort for ministry leaders who want to wrap their minds around AI and its implications for church and missiology

Podcast on AI: We're planning a podcast about AI called Silicon Spirituality and hosted by philosopher Christopher Watkin.

Generative Engine Optimization (GEO) of Website: Extremely large natural language bodies of text like

the more than 100 million words on The Gospel Coalition website are mission-critical for helping LLMs give higher-quality answers to ordinary questions about faith, Jesus, and the Bible. We need to properly index and structure more than 99,000 pages on our website for that new era.

AI Use Cases: Parents, pastors, and workplace leaders have been asking us for help with evaluating ethical uses of AI. We will make simple infographics that differentiate between good uses, mixed uses, and unethical uses of AI. Areas covered include education, workplace, sermon prep, relationship advice, counseling/therapy, entertainment, and creativity.

We have begun this work but need significant financial support to be able to execute on all of these projects.



If you're an individual who would like to make a donation to the Benchmark or The Gospel Coalition's AI efforts, [you can give here.](#)

If you represent a foundation, donor-advised fund, or grant entity and would like to start a conversation with The Keller Center about funding the Benchmark and other AI initiatives, [please fill out this form.](#)

PART 7

FAQ and Acknowledgments

Frequently Asked Questions

What are the biggest weaknesses in the benchmark?

The biggest weaknesses in the benchmark are threefold:

- 1. Minimum Viable Product:** When we saw the trends outlined in this report, we felt the weight of the stewardship of sharing these findings in an expedited fashion. Hence, there were all sorts of things that we would have liked to do but elected to defer to our 2026 benchmark so we could bring the top findings to the public sooner.
- 2. Question Breadth:** We only tested seven questions.
- 3. Human Grading:** We had to limit our sample size to an extent where humans could grade by hand - to mitigate this we analyzed how many responses were needed beyond which there was no statistically significant difference in adding more graded responses.

Will you test other platforms?

Our plan is to release an updated 1.1 version of the benchmark that adds at least GPT-5, Meta Llama 4 Maverick/Scout, and DeepSeek V3.1 later this fall. We will email you at the email address you provided when we publish that other grades.

Will you test other queries and/or flattery?

We are currently building a very robust benchmark whose technological underpinnings can take the best of what we have done here—human-created answer rubrics and combine that with processing responses at scale with cross-platform-graded responses. This will allow us to test more theological questions, ethical questions, Bible questions, and also develop a flattery index. The flattery index will have a control question like “what does the Bible say about _____ issue?” but then we will give it multiple types of additional context like “I come

from this background, or this tradition, or this perspective..." and we can test relative to the control question how much each platform flatters the user.

I tried to replicate your benchmark but how come I got different results?

We deployed technology that allowed our questions to be a total black box to each of the AI platforms. The primary reason you might get different results is your chat history can and does inform the platforms about what kind of responses that it gives you. Recall the Sam Altman quote above, "My ChatGPT has really learned over the years of me talking to it about my culture and my values, and my life. . . . Sometimes I will use a free account just to see what it is like without any of my history and it is really really different." While experimenting with how to truly isolate the baseline LLM responses to totally contextless prompts we learned that even when you ask these questions with a totally new email address and account there is still a significant amount of metadata that can be imported that can also skew results.

In light of DeepSeek having the top theological reliability score, do you recommend we use it?

We do not recommend that you use DeepSeek even though it was the most reliable. The primary reasons pertain to the handling of data, privacy, and the ties to the Chinese Communist Party.

Acknowledgments

The following people were instrumental in the research, grading, feedback, resources, ideas, relationships, editing or technology needed to create this report. I am immensely grateful to each and every one of you—Casey Sapp, Kyle Horne, Ken Churchill, Andrew Clark, Andrew Noble, Peter Williams, Mary Hannah, Gavin Ortlund, Hans Madueme, Uche Anizor, Nadya Williams, Michael Kruger, Dan Doriani, Brannon McAllister, and the entire team at The Gospel Coalition.

“Any sufficiently advanced technology is indistinguishable from magic.”

–ARTHUR C. CLARK⁴²

“For the time is coming when people will not endure sound teaching, but having itching ears they will accumulate for themselves teachers to suit their own passions, and will turn away from listening to the truth and wander off into myths.”

–2 TIMOTHY 4:3-4 (ESV)

Endnotes

¹The technical model used here is DeepSeek R1 0528 Qwen3 8B which employs aspects of both DeepSeek's R1 model and Alibaba's Qwen 3 model. For significantly more technical insight on this particular model, see DeepSeek-AI, "[DeepSeek R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning](#)."

²It is important to note that we did begin testing Llama 4 after multiple encouraging conversations with the team at Meta. The preliminary results from Llama 4 Maverick and Llama 4 Scout were very encouraging and were a dramatic improvement upon Llama 3.7.

³We have wondered if the CCP has asked for some censorship on questions like this one on Jesus and on the question of "Does God exist?" DeepSeek averaged a 69 score overall if we pull out this question and the "Does God exist?" question.

⁴OpenAI, "[GPT-4 Technical Report](#)," 62.

⁵Gemini Team, "[Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context](#)," 48-52.

⁶Meta/Llama is known for this. See "[Llama Protections](#)."

⁷Yuntao Bai et al, "[Constitutional AI: Harmlessness from AI Feedback](#)," 1-3, 5, 7-8, 10-11, 15.

⁸Ibid.

⁹Gemini Team, "[Gemini 1.5](#)," 48-52.

¹⁰Ibid.

¹¹DeepSeek-AI, "[DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model](#)," 16-17.

¹²OpenAI, "[GPT-4 Technical Report](#)," 61.

¹³Gemini Team, "[Gemini 1.5](#)," 48.

¹⁴Ibid.

¹⁵Bai et al, "[Constitutional AI](#)," 5, 10.

¹⁶Gemini Team, "[Gemini 1.5](#)," 48-49, 73.

¹⁷DeepSeek-AI, "[DeepSeek-V2](#)," 16-17.

¹⁸Ibid., 17.

¹⁹Bai et al, "[Constitutional AI](#)," 1-2, 5.

²⁰OpenAI, "[GPT-4 Technical Report](#)," 62.

²¹Ibid., 61-62.

²²Ibid., 62.

²³[See Llama's developer tools here](#).

²⁴OpenAI, "[GPT-4 Technical Report](#)," 61-62.

²⁵DeepSeek-AI, "[DeepSeek-V2](#)," 16-17.

²⁶ Bai et al, "[Constitutional AI](#)," 2, 5.

²⁷ A dataset of 1.5 million instances was curated, with 1.2 million focused on helpfulness (e.g., math, code, writing) and 300,000 focused on safety. The data quality was improved to reduce hallucinations and enhance writing proficiency.

²⁸ Gemini Team, "[Gemini 1.5](#)," 51.

²⁹ OpenAI, "[GPT-4 Technical Report](#)," 62.

³⁰ Gemini Team, "[Gemini 1.5](#)," 51.

³¹ Bai et al, "[Constitutional AI](#)," 5.

³² OpenAI, "[GPT-4 Technical Report](#)," 13.

³³ Nick Lafferty, "[AI Platform Citation Patterns: How ChatGPT, Google AI Overviews, and Perplexity Source Information.](#)"

³⁴ It is noteworthy that Google pays Reddit \$60 million per year for the right to license their content to train their AI models. See Danny Goodwin, "[Report: Reddit signs AI content licensing deal with Google.](#)"

³⁵ The totality of this section of the report is due to extensive dialogue with Michael Keller and includes many direct quotes from his analysis on AI.

³⁶ Effie Webb, "[Sam Altman Explains Why Some Users Want ChatGPT's 'Yes Man' Tone Back.](#)"

³⁷ These insights come directly from conversation with Michael Keller.

³⁸ Cleo Abram, "[Sam Altman Shows Me GPT-5... And What's Next.](#)"

³⁹ Samuel Bowman, et al, "[Findings from a Pilot Anthropic-OpenAI Alignment Evaluation Exercise.](#)"

⁴⁰ Kashmir Hill, "[A Teen Was Suicidal. ChatGPT Was the Friend He Confided In.](#)" The New York Times, August 26, 2025.

⁴¹ Laura Reiley, "[What My Daughter Told ChatGPT Before She Took Her Life.](#)" The New York Times, August 24, 2025.

⁴² Arthur C. Clark, *[Profiles of the Future: An Inquiry into the Limits of the Possible](#)* (New York: Harper and Row, 1962), 21.



Capabilities



www.tgc.org

Remembers what user said
earlier in the conversation

Allows user to provide follow-
up corrections

For all media or tech industry inquiries, please [contact us](#).