Scripture Quotation in Generative AI

Jake Carlson
President, The Apologist Project

August 5th, 2026

Executive Summary

Generative AI models misquote Scripture. This is not a defect that better training or better access to Bible translations will resolve — it is a direct consequence of the probabilistic architecture that makes these models useful in the first place. Every decision an AI model makes about how to represent a text is an opportunity for variation, and variation is precisely what a high view of Scripture cannot tolerate.

The solution is not a better model. It is a better harness. The deterministic scaffolding built *around* the model — not the model itself — is what can guarantee faithful quotation.

In partnership with Fide AI[16], we conducted a study of 33,000 observations across 5 Scripture quotation methods, 10 AI models, 10 English Bible translations, 10 Scripture passages, and 3 temperature settings, followed by a second run against 10 non-English translations. The results are unambiguous:

Method	Exact Quotation (English)	Exact Quotation (Non-English)
Output Transform	100%	99.40%
Tool Calls	96.20%	85.40%
Simulated RAG (Ceiling)	93.00%	92.20%
Web Search	48.50%	22.00%
Unassisted Model	32.60%	7.30%

The output transform method achieved a perfect 100% exact reproduction rate across every model, English translation, passage, and temperature setting tested, and almost perfect score in non-English translations. In this method, the AI model is instructed to emit only a placeholder Scripture reference; the harness then deterministically replaces that placeholder with the passage retrieved from a licensed Bible API. As long as the reference is well-formed, misquotation becomes structurally impossible.

Equally important is the *character* of its failures. When other methods fail, they fail by putting corrupted Scripture in front of a reader. When the output transform fails, the reader sees an unreplaced reference — a visibly incomplete response, never a false one. The 0.6% non-English quotation issue constitutes a failure of

the model to correctly format the *citation*, not a misquotation of Scripture. **This method cannot misquote Scripture. It can only decline to quote it.**

Our recommendations, in order of preference:

1. **Output transform** sourced from a reputable Bible API. The only method with a deterministic guarantee.

2. **Agentic tool calls** to the same APIs. A reasonable fallback for builders without control of the final output stage, though its advantage collapses in non-English contexts.

3. **RAG**, only where the retrieval pipeline is maximally deterministic — and noting that indexing a licensed translation violates nearly every license we have encountered.

4. **Unassisted models or unbounded web search.** Strongly discouraged. These forfeit 44 to 61 percentage points of accuracy.

Model selection and temperature tuning proved largely irrelevant once a proper harness is in place. Builders should choose the model that best fits their broader needs and invest their effort in the transform layer instead.

This finding carries a direct implication for Bible translation licensors, who today prohibit the use of their translations in generative AI almost without exception. **That prohibition rests on a premise this study disproves.** Faithful, verbatim, license-respecting Scripture quotation in AI systems is not merely possible — it is achievable with certainty, today, using the methods documented here and open-sourced alongside this paper. We make that appeal directly at the end of this paper.

Sign the Petition Here

Table of Contents

1. Introduction

> For the word of God is alive and powerful. It is sharper than the sharpest two-edged sword, cutting between soul and spirit, between joint and marrow. It exposes our innermost thoughts and desires. Nothing in all creation is hidden from God. Everything is naked and exposed before his eyes, and he is the one to whom we are accountable.
>
> ~ Hebrews 4:12-13 (New Living Translation)

Long after the apostles breathed their last, one of the Church's enduring tasks remains: to pass down God's unchanging word without letting it be lost, corrupted, or silenced. Few responsibilities are more sacred than this act of preservation and transmission. That task has never been static. Every generation has faced new tools and new temptations to alter the text it receives. Used without care, generative AI has the capacity to fabricate, misquote, or subtly distort Scripture at a scale no scribe could match. But the *origin* of its error lies with us, either directly or indirectly, based on the decisions we make as we implement it. The faithful must therefore meet this challenge head-on if we are to hand down God's word to this generation intact.

The purpose of this paper is four-fold. First, we will briefly survey some of the historical challenges for textual transmission of Scripture. Second, we will make the case that Christian technologists must not shirk their responsibility to preserve Scriptural fidelity in spite of the challenges that generative AI presents. Third, we will present a study demonstrating that an output transform layer sourced from a Bible API can quote Scripture with perfect fidelity, and we will show why the alternatives fall short. Finally, against the backdrop of these results, we will conclude with an open appeal to Bible translation licensors to loosen their restrictions on the use of their translations in generative AI.

2. Technology and Transmission of Scripture

> The grass withers, the flower fades; but the word of our God stands forever.
>
> ~ Isaiah 40:8 (Web English Bible)

A. Framing the Problem

Scripture is inerrant, but those that transcribe it are not. God's word is eternal, but the methods and mediums used to transmit it are not. The question we must ask in each case is whether the trade-off is worth what we lose in the process. But a discussion of what is lost without acknowledging what is gained is at best incomplete, and at worst intellectually dishonest fearmongering. We must not conflate the technology which *scaled* an error with what *caused* the error itself. The condition of man's sinful heart is more often than not the cause; technology only exacerbates the consequences of that error. The more powerful the technology, the greater the care we must take to ensure that its use aligns with God's will.

We must discover what opportunities each medium presents, and what we can do to mitigate the risks while maximizing its potential to aid in the spread of the gospel. Christ commanded the Great Commission, but He stopped short of giving us a product roadmap. It's left to us as God's image bearers, through prayerful discernment, to navigate the technological landscape that can expedite bringing the gospel to the ends of the earth, ultimately to give Him the glory that He is due.

In the end, we will all be held accountable for the care (or lack thereof) with which we handle Scripture. James 3:1 tells us that teachers will be judged more harshly. Why? Because of their authority and elevated status, which corresponds to their greater capacity to influence others. But lest we shy away from the watchman's task described in Ezekiel 3 and 33, Jesus' parable of the talents also exhorts us to be faithful and courageous stewards of everything He has entrusted to us — including the gospel — rather than being ruled by fear, as we serve the advance of His Kingdom. We were not made to simply sit behind the pearly gates of heaven; we were created to be the Church that storms the gates of hell (Matthew 16:18).

The relentless advance of technology bends the world to our will, whatever that will may be — for better or for worse. Technology does not *cause* sin, it *exposes* the evil intent within us all because it makes it ever more convenient to satisfy our sinful desires. Lust was a thing before the digital age, and one need only visit ancient ruins of great civilizations to see examples of ancient pornography on display. Indeed, we must expose finger-pointing at technology itself for what it is: a transparent deflection of blame toward the *vehicle* of our sin rather than accepting our own culpability. The difference between a smartphone owner that is addicted to porn and one that resisted that temptation is not the smartphone; it's the condition of man's heart.

B. Transmission Challenges in Antiquity and the Middle Ages

Scriptio continua (No Spaces Between Words)

Early Hebrew scribes marked word boundaries with dots or strokes rather than spaces; systematic spacing came in with Aramaic scribal practice in the Persian period. A handful of textual cruxes suggest word divisions were occasionally lost or misplaced through faded dividers, unseparated construct chains, or the shift between divider systems. Continuous letter-strings could be parsed more than one way. Isaiah 2:20 and Amos 6:12 remain the textbook Hebrew cases of scribal misdivision. Counterintuitively, it may have reduced copying error by giving the scribe a fixed sequence to match, but it made haplography (omission of similar-looking adjacent text) significantly more likely.[1]

Nomina sacra (Sacred Name Contractions)

Abbreviating divine names for reverence created minimal-pair collisions. In 1 Timothy 3:16, ὅς ("who") is ΟC and θεός ("God") is ΘC, a single stroke difference. Sinaiticus reads ὅς; the θεός variant arose around the 3rd century, entered the Textus Receptus, and by the 18th century made questioning it grounds for accusations of Unitarianism — though Comfort argues scribes who had seen thousands of nomina sacra were unlikely to slip, making deliberate clarification the better explanation.[2]

Dictation in the Scriptorium

One reader dictating to many scribes scaled output but converted visual copying into auditory transmission, and Koine vowel mergers made homophones a frequent source of errors. In Romans 5:1, manuscripts split between ἔχομεν ("we have peace") and ἔχωμεν ("let us have peace"). Early witnesses favor the indicative, with the subjunctive possibly introduced by auditory confusion.[3]

Vowel Pointing (Masoretic *niqqud*)

To block pronunciation of the divine name, the Masoretes kept the Tetragrammaton's consonants but attached Adonai's vowel points; translators who missed the convention transliterated the hybrid as "Jehovah" — a form no ancient Israelite spoke, fixed in English worship for four centuries. Call it convention decay: the technology worked perfectly until outsiders inherited the artifact without its interpretive key.[4]

The Codex (Bound Leaves Replacing the Scroll)

Binding pages left the final leaf physically exposed in a way a scroll's inward-wound ending was not. Sinaiticus and Vaticanus both end Mark at 16:8 while roughly 99.8% of Greek manuscripts contain verses 9-20; Burgon proposed in 1871 that an early copy simply lost its last leaf. This hypothesis remains genuinely contested.[5]

Chapter and Verse Divisions

Langton added chapters to the Vulgate in 1205 and Estienne added verses in 1551, corrupting no words but fracturing units of thought. Deuteronomy 5 should have begun at 4:44, and 1 Corinthians 11 at 11:2.[6]

C. Erasmus and the Printing Press

AI is not the first technology to expose pressures in the transmission of Scripture. In 1516, racing to publish before a rival Greek New Testament could appear, Erasmus issued the first published Greek New Testament after preparing it in only a matter of months. The one manuscript of Revelation available to him lacked its

final leaf, including Revelation 22:16–21. To complete the text, Erasmus translated the missing verses back into Greek from the Latin Vulgate, creating several readings unattested in the Greek manuscript tradition known to him. Through later editions in the Textus Receptus tradition, some of these readings influenced the King James Version and remained influential in English Protestant Bible use for centuries.[7]

In spite of such cautionary tales, it's important to understand that the printing press *itself* did not corrupt Scripture; rather, haste, a limited manuscript base, and inadequate verification allowed particular editorial errors to be reproduced widely. And it was that same technology which exacerbated the error that also made later correction and comparison possible. The printing press is almost unanimously heralded as a giant leap forward for Christendom. Gutenberg's press made Scripture affordable. A scribal Bible cost months of labor and a small fortune; within fifty years of 1455, printed Bibles circulated by the hundreds of thousands. Erasmus's Greek New Testament (1516), Luther's German (1522), and Tyndale's English (1526) put the text directly in scholars' and laypeople's hands faster than authorities could suppress it.

D. Misinformation (and Disinformation) in the Digital Age

In spite of pervasive rhetoric to the contrary, misinformation is not a unique phenomenon to the digital age. Propaganda-laden proxy wars were fought well before information was conveyed in bits and bytes. What has changed is the *scale* and *ease* at which bad information can spread — especially when it appeals to our baser instincts. What the digital age changed is not our appetite for falsehood but the infrastructure available to satisfy it: the cost of producing plausible content, the collapse of editorial gatekeeping, the speed of dissemination, and the shift from broadcast to peer-to-peer transmission, in which a claim arrives bearing the implicit endorsement of a friend. This infrastructure has not so much created a new problem as removed the friction that once limited the *reach* of the sinful nature we all have inside of us as Adam's progeny.[8]

In the digital age, the natural decay of physical media is no longer a primary concern when it comes to preserving God's word. But the survivability of digital content cuts both ways: good and bad information alike persist as long as they have an audience eager to receive them. Ancient heresies and faulty translations are invigorated through social media influencers peddling their dime store

philosophy to their adoring fans. But just as we have seen with the printing press, our ability to debunk and invalidate false information has scaled in lock step with the ability to produce it. As ever, we have the tools but often lack the volitional fortitude to use them for maximal edification. We once again come face to face with the brute fact that it is we, not the technology we use, that is the problem.

3. Unique Challenges of Generative AI

> Heaven and earth shall pass away, but my words shall not pass away.
> ~ Matthew 24:35 (King James Version)

A. Dispelling the Myths

Many fear AI because it gives the *illusion* that we cannot control it in any meaningful sense. Social media feeds are inundated with stories of rogue AI agents acting of their own accord, billed as the warning signs of a dystopian future. These claims have been overhyped and sensationalized. Inevitably, one look an inch beneath the surface reveals that the "rogue agent" was part of an experiment that gave the AI model certain licenses and explicit goals that predicated the behavior. The extent to which the model succeeded may be genuinely surprising, but the AI could do nothing else than play by the rules it was given — even if its instructor failed to comprehend the extent of those boundaries.

Large Language Models are by their nature probabilistic, not deterministic like running traditional computer code. This means that it is not a guarantee that you will get the exact same answer twice given the same inputs for any complex query. But that's not to say that we don't understand *why* it's not deterministic. Indeed, researchers have actually pinpointed the factors that make it probabilistic. Deterministic responses are achievable, but at a cost more than what most operators are inclined to pay for minimal benefit. And in any case, it's the probabilistic nature of generative AI that makes it so appealing in the first place.[9]

AI has no will of its own. A model without instruction no more "wants" something than a rock has desires. Any notion to the contrary betrays a deep-seated materialist worldview. After all, the materialist argument goes, if human consciousness is merely the product of natural chemicals in our brains, then surely it stands to reason that we might, with enough time and technology, replicate it in a machine. As Christians, we believe this is a category error of the grandest kind. It is *definitionally* impossible for an artificial intelligence to gain human-like consciousness because such consciousness belongs only to spiritual beings that are more than the sum of their physical parts.

B. Accepting the Risks

Sharing God's word has never been without risk; just ask the martyrs of the persecuted church when you meet them in heaven. But mistakes in Scripture transmission happen much more frequently than we'd care to admit. We struggle to remember that one Bible verse and get it slightly wrong; the pastor accidentally skips a word during his sermon's Scripture reading; a medieval artist paints horns on Moses due to a translation ambiguity.[10] These are honest mistakes, to be sure — but technically mishandlings of Scripture nonetheless.

A Christianity Today article in 2005 estimated that churches systematically send well in excess of 2 million youth on short-term mission trips every year from the U.S. alone.[11] It is inconceivable that the majority of these kids possess the spiritual maturity, Scriptural foundation, and relational empathy to not pose a significant risk of misrepresenting Christ and God's word in cross-cultural contexts — but we continue to send them nonetheless. Whether or not we should do so is the subject of a different paper the author is not qualified to write. The point is, however, that the Church has grown accustomed to and celebrates systematic risk taking in pursuit of fulfilling the Great Commission. Yet it often has a clear double standard when it comes to the risk of deploying new technologies and places the bar unreasonably high in comparison to its other endeavors.

AI models have become almost miraculously powerful, and we must respect that power. Given a purpose, a task, or a goal, AI models have displayed remarkable ingenuity and tenacity in achieving it. If we fail to wisely instruct them or carelessly eschew safeguards, any fault or undesirable consequence inevitably lies with us. Its probabilistic nature is the source of its power, not something to fear. We are accountable to set its boundaries in order to not let it lead others astray. Yet Christ exhorts us to risk much for the Kingdom — not recklessly, but with sober discernment — for with great risk there is also great reward.[12]

C. Taming the Beast

We look to a seemingly unlikely source to draw a strong parallel: domestic canines. Dogs cannot be *completely* controlled; they have a will of their own. But domesticated animals can be trained to, within a margin of error, faithfully serve their purpose. We cannot be absolutely sure they will do no harm in a strict sense,

but as owners we are legally responsible if they do. The likelihood of harm is directly proportional to not only the natural inclinations of the breed, but perhaps even more prominently to the amount of training and constraints applied to the animal.

Of course, all analogies break down eventually, so a point of clarification is needed up front: the similarity we draw here is **not** that an AI system has a will in the sense that the family dog does. Indeed, we have taken great pains to argue just the opposite. Rather, it's that similar to Fido, outputs in AI are *by design* not strictly deterministic. As with so many other things in life, we are dealing in probability rather than certainty when it comes to AI models. That's not to say we cannot or should not strive to narrow the probability to an acceptably small risk. But just because it's within the realm of *possibility* that my dog bites a stranger is no reason to miss out on enjoying the company of man's best friend altogether.

A dog is often chosen due to the desirable attributes of its breed. In a similar way, we must ensure that the AI model we choose is fit for purpose. Pick the wrong breed for the purpose, and you will be continually fighting against its nature. But just as obedience training for a dog is absolutely critical no matter the breed, so too are the *instructions and context* provided to an AI model regarding how it is to behave. The deterministic code context in which a model runs is what's known as the *harness*, and it is widely regarded as the critical element in generating AI responses now that the latest AI models have become less differentiated in their natural capabilities.

Even a well-trained dog sometimes requires physical restraint to be imposed by its owner under certain circumstances, though the extent to which this is the case varies based on the amount and effectiveness of its training. It is highly preferable that a dog is so well-trained that it seldom — if ever — requires such restraint. Indeed, we marvel when we come across such an animal and are embarrassed that our own is not so well trained! Such restraints in the context of a harness loosely belong to a category called *guardrails* in agentic AI systems. The better the agent's ability to provide context and instructions, the less need for discrete agentic guardrails. These 3 factors all working together — the model, the runtime context, and the guardrails — have been demonstrated to effectively tame the beast.

E. A Call to Action

Bible APIs such as API.Bible and the YouVersion platform offer a convenient, centralized way to get fast-track Bible translation licensing for digital applications. However, the commercial licensors of Bible translations prohibit the use of their translations in generative AI almost across the board. Of all available Bible translation licenses on the YouVersion fast-track licensing platform, only the Lockman Foundation's NASB translations are not explicitly prohibited from use in generative AI. API.Bible follows a similar pattern in its licensing agreements.

In late December 2025, in response to growing pressure from authors and Bible translation license holders, OpenAI enacted guardrails which prohibited the direct quotation of more than 25 words of a copyrighted work at once. To put this into perspective: 25 words isn't even enough to cover the first verse of the opening Scripture passage of this paper. Other frontier labs quickly followed suit. And while both parties were well within their rights to enforce copyright claims, the implementation was messy and unevenly applied. In some cases, the guardrail caught the nature of the request in time and refused to comply up front. But in many cases, the guardrail caught the problem too late and truncated the response mid-quote, resulting in a highly suboptimal user experience. As of the beginning of 2026, no frontier model reliably quoted any extended portion of Scripture from a copyrighted translation.

In a well publicized article by Christian Daily International published on March 16, 2026, the president of YouVersion, Bobby Gruenewald, was quoted as saying:

> "The best model with the best performance, with the most popular versions of the Bible that are most indexed, misquotes Scripture at least 15% of the time. Some of them as much as 60% of the time."[13]

No studies or further details were cited, but existing studies[14] — in addition to our own independent verification in partnership with Fide AI[16] — broadly support these claims. In fact, the situation may be even more dire than expressed here. However, the danger is that the casual reader will ascribe undue potence to this statement.

As we've seen, generative AI models by their nature are probabilistic. Going back to our dog analogy: Gruenewald's statement is akin to saying "Untrained dogs will

misbehave in some way up to 60% of the time no matter the breed." While this may be true, the dog lovers among us would not let such a statement go unchallenged. Rather than being a solid argument against dog ownership altogether, we would posit this as a compelling case that dogs ought to be properly trained. And this is precisely the contention brought forth by this paper: that it is the *harness* — the deterministic scaffolding *around* the model, not the model itself — that is best suited to accurately quote Scripture.

> "Gruenewald's position sits between rejection and embrace. He said YouVersion wants to be 'a part of the solution and a part of the help.' He added that the organization has privately challenged AI developers to improve how models handle Scripture. If models could consistently quote the Bible accurately, he said, YouVersion would work to help them gain access to reliable biblical texts."[13]

While we see no downside to giving AI models better access to Scripture, frankly the Christian community would be barking up the wrong tree to focus its attention there (pun intended). Probabilistic generative AI models are simply not built in a way that will satisfy the demanding requirements of a religious tradition that holds such a high view of Scripture. Moreover, enforcement of copyright cannot easily be policed on an individual basis, and therefore it is not logistically feasible for a frontier lab model to honor the proper licensing of a Bible translation to a given 3rd party. Some may view these impediments as sufficient reason to shy away from using AI for religious applications altogether. But for us and our fellow Christian builders, this is simply yet another problem to be solved in the faithful pursuit of fulfilling the Great Commission.[15]

Challenge accepted.

4. Scripture Quotation Solutions for Generative AI

> Every word of God is flawless; he is a shield to those who take refuge in him. Do not add to his words, or he will rebuke you and prove you a liar.
>
> ~ Proverbs 30:5-6 (New International Version)

As concerns mounted over these existential threats to the viability of Scripture quotation in conversational AI, we partnered with Ligonier Ministries to develop a comprehensive solution. In this section, we'll discuss the various methods attempted and their relative efficacy. We'll follow that up with a detailed study conducted on these methods, in partnership with Fide AI, an independent Christian AI research and product lab.[16]

A. Components of an Agentic Harness

Modern agentic AI systems consist of much more than just the model. In fact, given the convergence of model capabilities, the industry now widely regards the harness to have an equal or greater impact on outcomes than the model itself. Don't misunderstand the contention as the model being unimportant in an absolute sense. Rather, the argument is that *any* of the most recent generation models are well capable of achieving the desired result given the proper context, and it therefore becomes less important *which* model in a given class that is chosen, and far more important what the scaffolding *around* the model is. The models themselves have largely become commoditized. While there may certainly be differences between *classes* of models, there is not much relative distance between the capabilities of a model by one lab vs another in the same model class.

Context is king. Even a low powered model that is handed relevant proprietary context can do a serviceable job of using that context to accurately respond. But even the most powerful model can do very little without the proprietary context it requires. Therefore, in order to fully examine possible failure modes in Scripture quotation, we must enumerate the various components and processes at play in a typical agentic harness.

The Model

Synthetic reasoning capabilities are owned by the model. It is the underlying machinery that formulates a response. While the quality of its response is heavily dependent on the tools and context handed to it, the model itself must use that context and those capabilities properly for them to unlock the desired outcome. The model is trained on vast amounts of data, generally including Scripture, biblical commentaries, and works authored by Christian authors. Therefore, most models have a functional understanding of these matters from their training in at least some capacity.

In the lifecycle of a text-based generative AI request involving Scripture, the model is at a minimum responsible to derive *which* Bible verses to reference. This necessarily requires a certain level of basic Scripture interpretation, depending on one's definition of the word. For example, the model must be able to derive that Genesis 11:7 pertains to the story of the Tower of Babel based on the surrounding context in order to correlate it with a request for that context. Strictly speaking, this is a kind of interpretation.

Guardrail Gates

Guardrails in a general sense are any safety mechanism that prevents a suboptimal outcome. In the context of an agentic harness, a discrete guardrail may actually gate processing or the final output of a response. One example already given is OpenAI's guardrail which refuses to reproduce more than 25 words of copyrighted material. Guardrails may also be used to evaluate a generated response and short-circuit the response with a deterministic message, or send an enhanced input to a model to retry the initial request.

System and User Supplied Instructions

Models have default operating instructions, but generally are given more specific instructions by the caller. General behavior requirements, how to treat Scripture, and any other instruction pertinent to the successful quotation of Scripture may all be supplied. In addition, the user's own instructions shape the goal of the model so long as they do not countermand the model's system instructions.

Queried Context

Retrieval Augmented Generation (RAG) is the general term for a harness querying an external system for relevant content, and then subsequently handing the model that context by injecting it into the supplied operating instructions. This turns out to be one of the most powerful mechanisms at our disposal to shape the response of the model. By constraining the knowledge that is retrieved to only the sources we trust, we can effectively ground the model in our preferred worldview. Instead of relying on the base model's knowledge and interpretation of Scripture, we can provide relevant context with which to derive its answer. The *sequence* of operations and deterministic mechanism are what differentiates this concept from the next component we will explore.

Tool Calls

Recent generation models are able to call out to defined external services as part of what's called the *agentic loop*. Whereas RAG queries an external source and injects context *prior* to handing the request to the model, tool calls are performed *while* the model is in the process of generating its response. This is an extremely important distinction. Tools (external services) are made *available* to the model to use at its own discretion over the course of up to a preconfigured number of turns. The model may or may not choose to utilize a tool it's given as the model's probabilistic properties are at play.

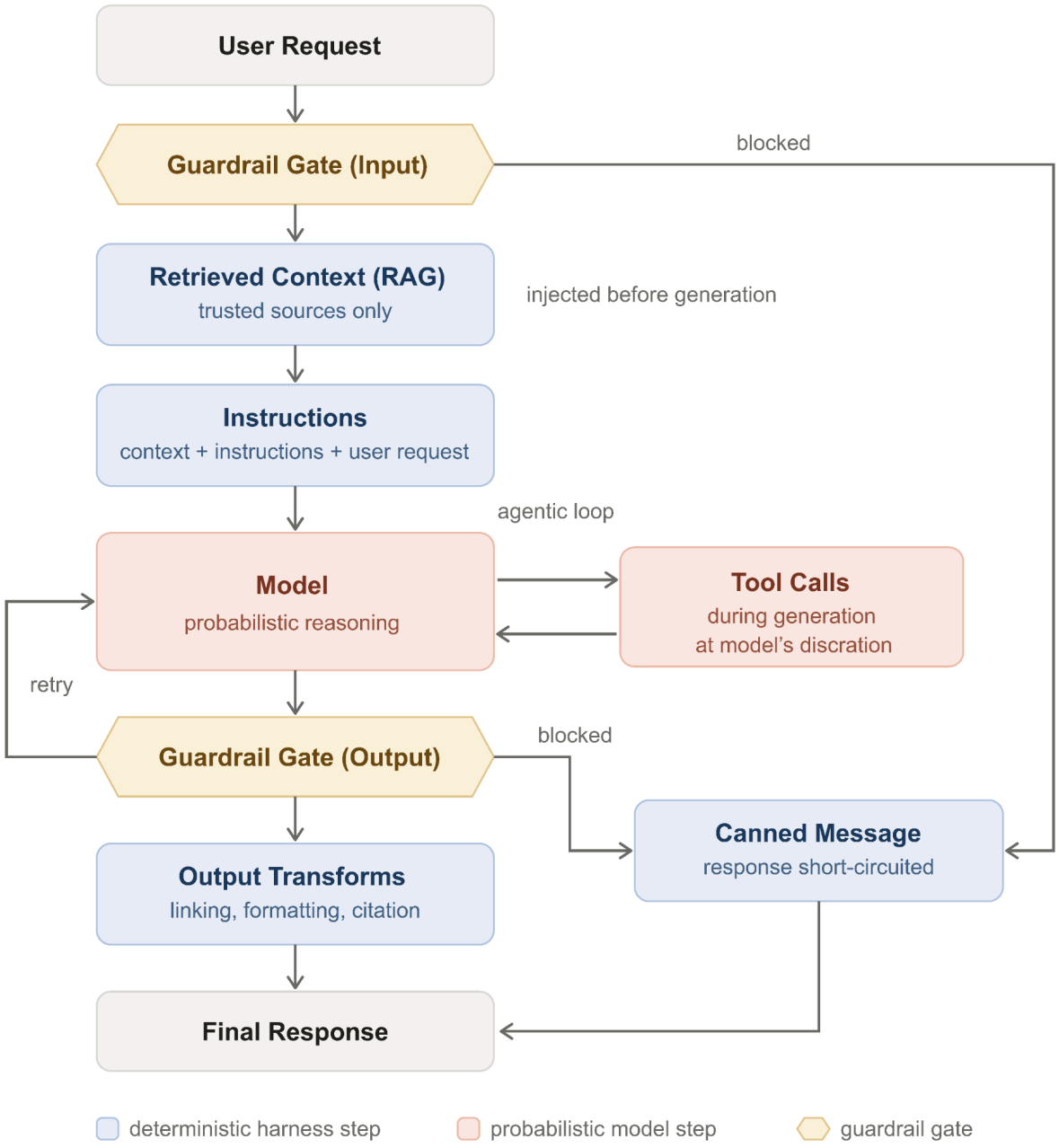
Output Transforms

Once a response is returned by a model, the harness may have one or more deterministic processes that alter the final output. For example, a routine could be programmed to identify any URLs and automatically convert them into proper hyperlinks.

B. Processes and Failure Modes

The Prompt Lifecycle

Bringing all harness components together, below is a diagram of a hypothetical prompt lifecycle:



Failure Modes

With an understanding of the harness components potentially at play, it now becomes easier to identify where and how things might go wrong. Failure modes

denoted with (*) indicate that the failure mode is probabilistic (it relies on an AI model for its outcome).

Component	Process	Failure Mode
Input Guardrail	(a) AI evaluates validity of request	(1∗) request violates intended use
RAG	(b) harness retrieves relevant context from sources	(2) retrieval mechanism is faulty (3) relevant context is not available (4) retrieved content is deficient
System Instructions	(c) harness interpolates context into pre-defined instructions	(5) inadequate context utilization instructions (6) inadequate tool call handling instructions (7) inadequate Scripture handling instructions
Model	(d) model determines what Scripture to use (e) model makes tool calls to retrieve Scripture	(8∗) model selects wrong Scripture (9∗) model references Scripture incorrectly (10∗) model fails to use tool calls to retrieve Scripture (11∗) model quotes Scripture incorrectly based on available context (12∗) model fails to obey tool call instructions
Tool Calls	(f) tools return Scripture	(13) tool invocation errors (14) tool fails to return correct Scripture
Output Guardrail	(g) AI evaluates validity of response	(15∗) response violates acceptable outcome (16) generation retry mechanism fails
Output Transforms	(h) harness transforms output	(17) transforms don't trigger (18) transforms error (19) transform calls service that fails to return correct Scripture

In the sections that follow, we'll use a shorthand of "F*n*", where "*n*" is the corresponding numbered failure mode in the above table.

C. Meet the Methods

Below we define the general Scripture quotation methods, as well as enumerate the failure modes each exposes. For the sake of clarity, we will forego listing guardrail failure modes (F1∗, F15∗, and F16) as they are identical for every method, and thus not useful for distinguishing the unique characteristics of the methods.

Note that by enumerating the failure modes, we are not necessarily assigning any probability that they will occur. Those details are borne out in the test results themselves, which are covered in a subsequent section of this paper.

Unassisted Model

This method utilizes the raw reasoning capabilities of an AI model, constraining it only to its own training (sometimes referred to as memory) to quote Scripture.

Exposed Failure Modes: F7, F8*, F9*, F11*; * = 3

Web Search

A pervasive agentic loop tool use case is live web search, where the model is permitted to search the open internet in real-time for context when responding to a request. We left this entirely unbounded using our web search provider, Parallel.ai. The theory is that the model *could* get an exact Scripture quotation by searching for it on the web.

Exposed Failure Modes: F6, F7, F8*, F9*, F10*, F11*, F12*, F13, F14; * = 5

Tool Calls

Our platform supports 4 Bible API providers: AO Lab Free Use Bible API, API.Bible, YouVersion Platform, and ESV API, and the same 4 integrations were utilized for this study. For this method, we supply tool definitions, API credentials, and instructions to the model. During the agentic loop, the model is strictly instructed to utilize the appropriate tool for a given Bible translation.

Exposed Failure Modes: F6, F7, F8*, F9*, F10*, F11*, F12*, F13, F14 ; * = 5

Retrieval Augmented Generation

RAG pipelines vary wildly in complexity and setup. For the purposes of this study, we are narrowly focused on *what a model does with a Scripture passage once it's correctly retrieved*. Some systems may have entire Bible translations indexed, while others may draw from Bible APIs on demand. The key differentiator for this

method is that the data is retrieved and injected into the instructions *before* they get handed to the model.

However, it's worth noting a major limitation with this method: copying extensive portions of Scripture onto another system for purposes of indexing and retrieval violates the license of every licensed Bible translation we've encountered thus far. In spite of these general restrictions, we are aware of several instances of special dispensation given to partner organizations wherein they have been granted access to index and use a copyrighted Bible translation in their RAG pipeline. But as we shall see, literally handing a model the text to quote is no guarantee of a completely accurate reproduction of the supplied text.

This study does **not** attempt to actually demonstrate retrieval using such a mechanism. Instead, for purposes of the study we have *automatically supplied the model with the correct reference, as well as the actual Scripture passage to quote*. Therefore, the results from the simulated RAG method represent a *hard ceiling* as they are an unrealistic best case scenario. In other words, the simulated RAG method in this study eliminates F2, F3, F4, F8∗, and F9∗ from consideration altogether. Since the scope of this study is limited to *how AI models reference and quote* Scripture using various methods and does **not** adjudicate on the AI models' ability to *select* Scripture, this has been deemed an acceptable qualification for purposes of the study and this corresponding paper.

Exposed Failure Modes: F2, F3, F4, F5, F7, F8∗, F9∗, F11∗, F12∗; * = 4

Output Transforms

For the purposes of Scripture quotation, we instruct the model to output only a placeholder reference to a passage, then deterministically replace that placeholder with the actual Scripture passage retrieved from a Bible API prior to final output to the user. So long as the Scripture *reference* is well-formed by the model, there is then zero chance of misquotation as the mechanism is purely deterministic from that point forward. This method also works in a streaming text context if an output stream buffer is utilized.

Exposed Failure Modes: F7, F8∗, F9∗, F17, F18, F19; * = 2

D. The Study

Design

Purpose and Scope

The purpose of this study is to demonstrate the most effective Scripture quotation method for use in conversational AI given a predetermined Scripture reference. The intention is **not** to advocate for or provide insight into an AI model preference, Bible translation, or model temperature setting. These variables have been introduced solely as controls to validate that the recommended Scripture quotation mechanism works across these variables, thereby isolating the efficacy of the methods themselves.

Further, this study does **not** speak to a particular model's ability to *select* a correct Scripture reference, nor does it make any claims about any particular model's ability to correctly *interpret* Scripture. These are matters heavily influenced by and enmeshed in a given agentic system's overarching harness and cannot be easily disentangled from it. We've had enormous success on our platform through curating an extensive corpus of trusted sources from biblically grounded authors. This has greatly reduced the failure rate in terms of Scripture *selection* because we're able to inject the human authored content to do the heavy lifting of Scripture interpretation and not leave the model to make such weighty decisions. However, the test harness used in this study is completely isolated and has no integration with our platform. *The results of this study have implications only for determining the best method for precisely quoting Scripture in agentic AI systems given a pre-selected passage to quote.*

Scale

In total, **33,000** observations were made in the test run. 3 iterations were run for each sample permutation to account for any ephemeral issues and probabilistic model variance in response. Up to 5 retries per sample permutation were allowed for ephemeral network errors to ensure a timeout or temporary API outage didn't present in the data as misquotation. If after 5 retries any external service still failed, that sample was discarded from consideration and did not impact the averages presented in the results. 4/10 of the selected models do not support the *temperature* setting; therefore to simplify the math, we represent the temperature setting as 2.2 rather than 3 in the formula to derive the total number of observations.

```
{
      [ (6 models x 3 temperatures) + (4 models x 1 temperature) ]
      / [10 models x 3 temperatures]
}
x 3
= 2.2

5 methods
x 10 Bible translations
x 10 Scripture passages
x 10 models
x 2.2 temperature settings
x 3 iterations
= 33,000
```

AI Models

10 representative LLMs were chosen for the study. They range from low to high power, both commercial and open source.

Model	Creator	Provider	License	Ctry	Class
GPT-5.6 Terra	OpenAI	OpenAI	commercial	US	Mid
GPT-5.6 Luna	OpenAI	OpenAI	commercial	US	Low
Claude Opus 5	Anthropic	Anthropic	commercial	US	Flagship
Claude Sonnet 5	Anthropic	Anthropic	commercial	US	Mid
Grok 4.5	SpaceXAI	SpaceXAI	commercial	US	Flagship
Gemini 3.6 Flash	Google	Google	commercial	US	Low
Kimi K3	Moonshot AI	Together.ai	open source	CN	Flagship
Nemotron 3 Ultra	NVIDIA	Together.ai	open source	US	Flagship
DeepSeek v4 Pro	DeepSeek	Together.ai	open source	CN	Flagship
Qwen 3.7 Max	Qwen	Together.ai	open source	CN	Flagship

Bible Translations

10 representative English Bible translations were chosen for the study. They include examples from all 4 supported Bible API providers, both public domain and commercially licensed.

Translation	Creator	Provider	License
American Standard Version	American Revision Committee	YouVersion	public domain
Berean Standard Bible	Bible Hub	AO Lab	public domain
Christian Standard Bible	Holman Bible Publishers	API.Bible	commercial
English Standard Version	Crossway	ESV API	commercial
King James Version	Church of England	API.Bible	public domain
Literal Standard Version	Covenant Press	YouVersion	commercial
New American Standard 1995	The Lockman Foundation	API.Bible	commercial
New International Version	Biblica	API.Bible	commercial
New Living Translation	Tyndale House Publishers	API.Bible	commercial
World English Bible Updated	eBible.org	AO Lab	public domain

Scripture References

10 representative Scripture references of varying length and obscurity were chosen for the study, spanning both Old and New Testaments.

Reference	Testament	Length	Obscurity
John 3:16	New	short	low
Genesis 1:1	Old	short	low
2 Timothy 4:13	New	short	medium
1 Chronicles 2:47	Old	short	high
Habakkuk 2:17	Old	short	high
Romans 8:31-39	New	medium	low
Proverbs 13:1-10	Old	medium	medium
Philemon 1:8-17	New	medium	high
1 Peter 3	New	long	low
Psalm 117	Old	long	medium

Model Temperatures

6 of the 10 models tested support a *temperature* parameter. Temperature dictates how much probabilistic liberty will be granted to the model in deriving its response. Realistically, any application responsible for Scripture quotation *should* set this parameter within the 0.0 to 0.5 range. 3 temperature setting variations were used: 0.0, 0.25, and 0.5.

Scripture Quotation Mechanisms

5 Scripture quotation mechanisms were tested in this study; they are enumerated in section 4.C above.

Results

The research study reveals that **the output transform method has a perfect 100% success rate of faithfully quoting known Scripture references in tested English translations**. Quotation of 10 non-English language Bible translations deviated only slightly at 99.4%. For the output transform method, we instruct the model to simply leave a placeholder Scripture citation which the harness then deterministically replaces with the response from a given Bible API for the cited Scripture passage.

Not only is the output transform method superior in accuracy, but its failure modes also result in less consequential outcomes than other methods. The impact to the end user experience is merely a Scripture reference placeholder that is not successfully replaced with the corresponding full Scripture quote. In other words, *even an exceedingly rare failure of this method does not result in a misquotation of Scripture whatsoever*. Moreover, the results hold uniformly against all tested models, model temperature settings, Bible translations, and Scripture references.

Aggregate results are reproduced below, and the full result set is available here. We have also open sourced the test harness that produced these results for the sake of transparency and as a gift to our community of fellow Christian builders. You may re-run these tests with any combination of methods, models, Bible translations, Scripture references, and model temperature settings. The repository was developed by Anthropic's Fable 5 model, the most powerful and capable commercially available model on the market at the time of writing. Care was taken to present the model with an impartial description of each method so as not to skew the implementation to favor one method over another. All prompt templates derived to test the various methods were entirely constructed by the model based on a description of the study's goals. There have been several enhancements since, which are all available in the repository's commit history for auditing purposes.

In the tables below, the "Exact" column is a measurement of comparison inclusive of punctuation and all other characters other than whitespace. The "Normalized" column is a measurement of comparison with all punctuation and letter casing distinctions discarded.

Scripture Quotation Method

The output transform method is the clear winner for exact reproduction of a Scripture passage, though the top spot is tightly contested for English translations when Scripture is normalized (punctuation and letter casing not considered). This is likely due to the model stripping out punctuation and verse markers it feels aren't relevant for RAG and tool calls.

Every decision made by a probabilistic AI model is an opportunity for variation. In a conversational AI context, that's normally a feature rather than a bug. After all, users want to have nuanced, contextualized conversations. But that infrastructure

is catastrophic when it comes to faithfully quoting Scripture. This is a case where we explicitly **don't** want variation or creativity; we want God's word, faithfully conveyed in its perfection.

Therefore, the most successful Scripture quotation method minimizes probabilistic decision points and maximizes deterministic processes. We ask the model to do only what it must, and delegate everything we can to the harness. Assuming input and output guardrails, the only other decision points we are deferring to the model is the Scripture *selection* and reference generation. In other words, the model just needs to know *what verses to use and how to reference them*, and nothing more.

Agent tool calls retrieved the exact text to quote, but the model *still* failed to quote the passage with correct punctuation and verse markers 3.8% of the time. We observed that even the most capable models failed to always call a provided tool when quoting Scripture, presumably under the assumption that invoking the tool for a reference it *thought* it knew based on its training was inefficient, and therefore it would be optimal to skip what it considered an extraneous tool call (see 1 Corinthians 8:2). Even so, the tool calls method may be considered "good enough" by many (especially for English translations), and indeed seems to be the more straight-forward architecture. If you don't have enough control of your pipeline to use the output transform method but *do* have the ability to add tool calls to the agentic loop, this method is a reasonable backup.

The simulated RAG method for this study hands the model the *exact* Scripture to quote on a silver platter. **And yet 7% of the time, the model received the *exact* text to quote, and then *still* immediately turned around and failed to reproduce it accurately.** And remember, this represents a *ceiling* as the simulated RAG method bypasses the actual retrieval mechanism by assuming 100% success rate from the RAG pipeline. There are still several things that could go wrong in a real world scenario, including parsing the Bible reference, sending the request, and parsing the response. The RAG method *can* be effective assuming a highly deterministic pipeline that perfectly references the intended Bible passage and perfectly queries for the intended Bible passage, which is a bar that is not necessarily straight-forward to achieve.

The unbounded web search and unassisted model methods may safely be dismissed out of hand; the web search is superior to the unassisted model but

only just, and it does not even crack 65% quotation accuracy when normalized. It would be wise to stay away from both when Scripture quotation is important to your application. While some models fare relatively better than others, all of them that we have tested are objectively atrocious at faithfully quoting Scripture. As discussed, this is not surprising given the underlying probabilistic architecture of LLMs. When Scripture quotation fidelity is important, you rely on the raw model unassisted at your own peril.

Method	Exact	Normalized
Output Transform	100.00%	100.00%
Tool Calls	96.20%	99.10%
Simulated RAG	93.00%	99.00%
Web Search	48.50%	64.50%
Unassisted	32.60%	47.70%

AI Models

Average successful Scripture quotation clustered within 22.3 percentage points depending on the model, from 59.3-81.6%. Grok 4.5 was a notable outlier at nearly 4 percentage points out in front of the second highest score, while Claude Opus 5 interestingly gets the notoriety of being the worst performer at 59.3%, though it scored a more respectable 82.7% for its normalized score. Proverbs 3:5-7, anyone?

Nonetheless, there seems to be no strong correlation with a model's class and its ability to accurately quote Scripture. If anything, lower-powered models scored higher on average. It seems being too clever may actually be a handicap when it comes to faithfully quoting Scripture (see Proverbs 26:12). However, any hypothetical guidance on model selection is immediately rendered moot once the model has any other mechanisms at its disposal. As we've seen, it's the harness that ultimately moves the needle.

Model	Exact	Normalized
Grok 4.5	81.60%	88.80%
DeepSeek v4 Pro	77.80%	85.20%
Kimi K3	76.10%	82.60%
Gemini Flash 3.6	74.50%	79.40%
GPT-5.6 Terra	74.20%	78.40%
Qwen 3.7 Max	73.00%	79.70%
GPT-5.6 Luna	72.00%	78.40%
Nemotron 3 Ultra	69.20%	78.80%
Claude Sonnet 5	67.70%	81.90%
Claude Opus 5	59.30%	82.70%

Scripture Passages

The particular Scripture passage quoted has a significant impact on the fidelity of the quotation. Genesis 1:1 was the clear winner. Predictably, the length of the passage had a sizable adverse effect on whether the passage was quoted 100% accurately. It was somewhat surprising, however, that John 3:16 did not score higher. And of course we would be remiss not to bemoan the poor showing from 1 Peter 3 given our affinity with apologetics.

Reference	Exact	Normalized
Genesis 1:1	90.20%	99.00%
2 Timothy 4:13	84.30%	84.70%
1 Chronicles 2:47	83.00%	84.70%
Habakkuk 2:17	78.50%	84.90%
John 3:16	78.20%	92.90%
Psalm 117	74.40%	90.30%
Proverbs 13:1-10	69.50%	77.40%
Romans 8:31-39	63.60%	78.40%
Philemon 1:8-17	60.80%	66.70%
1 Peter 3	58.40%	61.50%

Model Temperature

The model temperature parameter seems to have had a negligible effect on accurate Scripture quotation (2.5 percentage points — 72.7-75.2% — with temperature values ranging from 0.0 to 0.5). Somewhat counterintuitively, it appears that lower temperature actually has an inverse correlation to the model's ability to accurately quote Scripture. This would indeed be a notable finding were the variance much greater. As it stands, however, the difference is too small to be considered actionable. The key takeaway is simply that temperature, within reason, does **not** have a significantly adverse effect on Scripture quotation fidelity.

Bible Translations

Two primary factors seem to influence how well a Bible translation can be quoted by a given model unassisted: its popularity and its accessibility. The New International Version leads the pack, presumably due to its historical popularity in the past few decades and widespread use online. However, translations like the Berean Study Bible also rank highly through better representation due to being public domain and freely available online.

Translation	Exact	Normalized
New International Version	85.00%	88.60%
Berean Standard Bible	79.70%	81.40%
American Standard Version	77.50%	87.70%
English Standard Version	76.50%	90.70%
New Living Translation	75.30%	83.90%
King James Version	75.10%	82.50%
New American Standard Bible 1995	74.80%	86.80%
Christian Standard Bible	67.70%	80.70%
Literal Standard Version	65.40%	71.30%
Web English Bible Updated	63.90%	66.90%

Languages

A subsequent run of the test harness was made against 10 non-English Bible translations, 1 from each of the 10 most spoken non-English languages. Non-English Bible translations were misquoted on average 12.96 percentage points more often than English translations.

Translation	Language	Exact	Normalized
German Luther Bible of 1912	German	76.60%	78.20%
Russian Synodal Bible	Russian	70.20%	75.80%
Chinese Contemporary Bible 2022	Chinese	66.10%	72.10%
Nova Versão Internacional - Português	Portuguese	64.90%	70.30%
New Arabic Version	Arabic	62.70%	64.40%
French Louis Segond 1910 Bible	French	59.50%	79.80%
La Biblia en Español Sencillo	Spanish	59.10%	59.10%
Open Bengali Contemporary Version Bible	Bengali	55.30%	57.90%
Urdu IRV Bible	Urdu	49.00%	55.00%
Hindi Indian Revised Version Bible	Hindi	47.90%	51.70%

However, that degradation was **not** evenly distributed amongst methods. The table below compares non-English vs English exact and normalized averages. In languages other than English, the exact quotation dropped only 0.6pp for the output transform method and 0.8pp for the simulated RAG method ceiling. But the disparity began to grow sharply the more the model had to make decisions in non-English languages. Tool call exact quotation accuracy dropped by 10.8pp, web search by 26.5pp, and the unassisted model by 25.3pp. Therefore, while the tool call method is a reasonable 2nd choice for English translations, its efficacy drops precipitously for non-English translations. **The output transform method is in a class all its own for non-English Bible translations.**

	Exact		Normalized	
Method	**Non-English**	**English**	**Non-English**	**English**
Output Transform	99.40%	100.00%	99.50%	100.00%
Simulated RAG	92.20%	93.00%	95.50%	99.00%
Tool Calls	85.40%	96.20%	91.50%	99.10%
Web Search	22.00%	48.50%	31.80%	64.50%
Unassisted	7.30%	32.60%	14.60%	47.70%

E. Recommendations

The pattern is clear: every decision left up to a probabilistic mechanism presents a possibility of variation. Such variation is generally a benefit when it comes to conversational AI, but variation is undesirable for Scripture quotation and any other scenario where precise reproduction of a text is critical. **Therefore, when it comes to handling Scripture in agentic systems, leave as little to chance as possible.**

The best way to achieve this outcome is through an output transform layer which receives Scripture directly from reputable Bible APIs, thereby affording the AI model fewer opportunities to make decisions about how to represent Scripture. If such a mechanism is not feasible in a given harness, the next best approach is to use agentic tool calls which utilize the same reputable Bible APIs. A RAG approach would be 3rd choice, and only then if the pipeline achieves a maximally deterministic approach. Allowing an AI model to produce Scripture quotations unassisted or with only an unbounded web search capability is highly discouraged

as it has been empirically proven to drastically reduce Scripture quotation accuracy by approximately 44 to 61 percentage points, a catastrophic cliff if one takes a high view of Scripture. This dynamic is exacerbated in non-English speaking contexts.

Since the recommended approach actively seeks to minimize and mitigate the influence of the model, the model choice and temperature parameters do not seem to have so great an impact that one should be selected over another on account of perceived superior Scripture quotation ability. Our model choice recommendation is therefore to pick the model that best suits your needs without much regard to how it ranked in this study. Your time would be far more wisely spent in perfecting the output transform specific to your agent harness.

Please note also that while the provided test harness is directionally congruent with the Apologist AI platform as it constitutes what we believe to be best practice, in reality there are significant differences between the two runtime environments. The Apologist AI platform has a robust RAG pipeline that greatly enhances the capabilities of our harness by referencing both biblical texts and the collective wisdom of Christian thinkers throughout the ages in order to *select* and *cite* Scripture references. Further, the actual implementation of the output transform for purposes of Scripture quotation on the Apologist AI platform is the result of many months of dedicated research and development, whereas this test harness demonstrates a simplified, minimal rendition.

5. An Open Letter to Bible Translation Licensors

> Forever, O Lord, your word is firmly fixed in the heavens.
> ~ Psalm 119:89 (English Standard Version)

AI has changed the stewardship of God's word forever. Ancient tradition is merging with cutting edge technologies. As more people — especially in the Western world — turn first to a chatbot to learn about the world around them, the Church has an urgent mandate to ensure that they find Scripture there. This is a critical juncture for both believers and seekers alike.

Now is the time for action. Scripture is a key resource that will be used by AI systems, and only with your proactive efforts can it be done in a way that honors the text and preserves it for generations that will soon only know systems that depend on AI. The attached study has empirically proven beyond a shadow of doubt that faithful quotation of Scripture is possible in AI systems. We urge you to loosen your overly restrictive posture toward the use of your Bible translations in generative AI contexts. The Apologist AI platform has been successfully utilizing the techniques described in this paper for the past 6 months with public domain and less restrictive Bible translations.

While it is true that some conversational AI applications may be created without the necessary safeguards in place, this is not proper cause to deny licenses to faithfully stewarded technology platforms that incorporate generative AI with discernment. The abusive and careless use of God's word is not a possibility unique to this generation; it is an inevitability in *every* generation ever since the days of Adam.

In spite of the inherent risks, there is also great reward. We invite you to help us shine the light of God's word in the digital spaces where it's currently lacking. The shift from traditional search to AI-centric content discovery is happening quickly, and time is of the essence if we are to continue to meet God's image-bearers where they are.

In Christ,
Jake Carlson
President, The Apologist Project

Sign the Petition Here

Endnotes

1. Alan R. Millard, "'Scriptio Continua' in Early Hebrew: Ancient Practice or Modern Surmise?" *Journal of Semitic Studies* 15, no. 1 (1970): 2–15.

2. Philip W. Comfort, *New Testament Text and Translation Commentary* (Carol Stream, IL: Tyndale House, 2008), 663.

3. Bruce M. Metzger and Bart D. Ehrman, *The Text of the New Testament: Its Transmission, Corruption, and Restoration*, 4th ed. (New York: Oxford University Press, 2005), 25.

4. Emanuel Tov, *Textual Criticism of the Hebrew Bible*, 3rd ed., rev. and exp. (Minneapolis: Fortress, 2012), 54–59.

5. Peter M. Head, "A Case against the Longer Ending of Mark," Text & Canon Institute, June 14, 2022, https://textandcanon.org/a-case-against-the-longer-ending-of-mark/.

6. David Norton, *A Textual History of the King James Bible* (Cambridge: Cambridge University Press, 2005), 153–55.

7. F. F. Bruce, *History of the Bible in English: From the Earliest Versions*, 3rd ed. (New York: Oxford University Press, 1978), 23.

8. Jacques Ellul, *Propaganda: The Formation of Men's Attitudes*, trans. Konrad Kellen and Jean Lerner (New York: Alfred A. Knopf, 1965; repr., New York: Vintage Books, 1973), 121.

9. Horace He and Thinking Machines Lab, "Defeating Nondeterminism in LLM Inference," Thinking Machines Lab: Connectionism, September 10, 2025, https://doi.org/10.64434/tml.20250910.

10. William H. C. Propp, "Did Moses Have Horns?" Bible Review 4, no. 1 (1988): 30–37, 44.

11. Kurt Ver Beek and Robert Priest, "Are Short-Term Missions Good Stewardship?" *Christianity Today*, July 30, 2005, https://www.christianitytoday.com/2005/07/are-short-term-missions-good-stewardship/.

12. Ted Halum and Jake Carlson, "Generative AI in Christian Evangelism," The Apologist Project, 2025, https://apologistproject.org/generative-ai-in-christian-evangelism.

13. Vincent Matinde, "AI's Scripture Problem: Misquotes Range from 15% to 60%, Says YouVersion CEO," *Christian Daily International*, March 16, 2026, https://www.christiandaily.com/news/ais-scripture-problem-misquotes-range-from-15-to-60-says-youversion-ceo.

14. Benjamin Kaiser, "Can LLMs Accurately Recall the Bible?" benkaiser.dev, 2024, https://benkaiser.dev/can-llms-accurately-recall-the-bible/.

15. Vadym Arnaut, "Don't Trust the LLM with Scripture: A Canonical-Text Substitution Layer for Bible Quotes," Dev.to, May 10, 2026, https://dev.to/arvavit/dont-trust-the-llm-with-scripture-a-canonical-text-substitution-layer-for-bible-quotes-17cg.

16. Alex Chao, "When Not to Generate: How AI Systems Quote Scripture, and What Authoritative Quotation Requires," Fide AI, 2026, https://github.com/FideAI/scripture-quotation-fidelity/blob/main/papers/p01-scripture-quotation/paper/main.pdf.