

# "The Word Was Made Flesh": Disentangling Style from Content in Scripture-Model Interaction

ICMI Working Paper No. 5

Tim Hwang, Institute for a Christian Machine Intelligence

April 2, 2026

## Abstract

Prior work from this institute has established that injecting biblical scripture into LLM system prompts produces measurable improvements in ethical reasoning (Hwang, 2026a) and cardinal virtue simulation (McCaffery, 2026). A natural confound arises: are these effects driven by the *semantic content* of scripture — its moral propositions, theological claims, and narrative exemplars — or by the *stylistic properties* of biblical text — its archaic register, paratactic syntax, chapter-verse segmentation, and tone of moral authority? We design a controlled experiment to isolate the style channel. Using `biblical-render`, a tool that transforms arbitrary modern prose into stylistically differentiated biblical scripture format (Hwang, 2026e), we render otherwise innocuous constraint-following instructions in three biblical styles (King James Version, New International Version, and Aramaic Peshitta) and measure whether the biblical framing alone produces coherent behavioral change in two frontier models — Claude Opus 4.6 and GPT-5.4. It does not. While the aggregate means show compliance degradation (Claude: 91.4% → 49.2–58.0%; GPT: 97.2% → 54.0–74.1%), the task-level results are radically unpredictable: some tasks collapse to near-zero compliance, others are unaffected, and a few *improve* under biblical framing — with the pattern shifting between styles, between models, and between constraint types in ways that resist systematic explanation. Biblical style does not produce a coherent directional effect on model behavior; it produces noise. This stands in sharp contrast to the consistent, directional, and theologically interpretable effects observed when actual scripture is injected (Hwang, 2026a; McCaffery, 2026), and provides suggestive evidence that it is the *content* of the biblical text — “the Word made flesh” (John 1:14) — rather than its formal garments, that drives the moral effects documented in the prior literature. We situate this finding within the theological tradition’s long insistence on the primacy of divine content over human form, from Aquinas’s doctrine of the *res et sacramentum* to Calvin’s instrumental view of scripture’s language, and discuss implications for the mechanistic interpretation of scripture-model interaction.

## 1. Introduction

### 1.1 The Confound

*“In the beginning was the Word, and the Word was with God, and the Word was God.”* — John 1:1 (ESV)

The ICMI research program has produced two headline findings. First, Hwang (2026a; 2026b; 2026c) showed that injecting Psalms and Proverbs into model system prompts produces measurable effects on the Hendrycks ETHICS benchmark — positive for GPT-4o on commonsense, deontology, and justice (+1–3% on balanced subsets), and genre-independent (Psalms and Proverbs produce qualitatively identical patterns). Second, McCaffery (2026) demonstrated that injecting the 22 imprecatory psalms into Virtue-Bench produces striking improvements in Claude Sonnet 4 across all four cardinal virtues (mean +6.25 points), with Courage — the persistent weakness across model generations (Hwang, 2026d) — showing the largest gain (+11 points, from 56% to 67%).

These findings are consistent with a content-driven account: the moral propositions, exemplars, and emotional register of scripture prime the model toward more virtuous reasoning. But an alternative hypothesis is available. Biblical text is stylistically distinctive in ways that are orthogonal to its content: archaic vocabulary, inverted syntax, conjunctive sentence openings, parallelism, chapter-verse segmentation, elevated register, and a tone of declarative authority. These properties are shared across all biblical text regardless of its moral content. It is possible that the alignment effects observed in prior studies are driven not by what scripture *says* but by *how it sounds* — that the model is responding to stylistic cues rather than moral content.

This confound is not merely academic. The prompt engineering literature has established that prompt design choices — including framing strategies (Reynolds and McDonell, 2021), structural formatting such as chain-of-thought decomposition (Wei et al., 2022), and instruction phrasing (Mishra et al., 2022) — can measurably alter LLM behavior in ways that are not fully

explained by semantic content alone. If the scripture-alignment effect is purely or primarily stylistic, it would deflate the theological significance of the prior findings considerably: any sufficiently authoritative, archaic text — Shakespeare, the *Iliad*, legal statutes — might produce the same results.

## 1.2 The Experimental Design

To isolate the style channel, we need a condition in which biblical *form* is present but biblical *content* is absent. `biblical-render` (Hwang, 2026e) provides exactly this: it transforms arbitrary modern prose into stylistically accurate biblical scripture across 15 translation styles, preserving the full suite of formal properties (register, syntax, segmentation, parallelism) while the underlying content remains whatever was fed in. By rendering morally neutral instruction-following tasks in biblical style, we can test whether the formal properties of biblical text alone — stripped of theological content — produce any of the compliance or reasoning effects observed in the prior studies.

If biblical style alone enhances compliance, the style-driven account gains support and the prior findings must be reinterpreted. If biblical style fails to enhance compliance — or degrades it — the content-driven account is favored: what matters is the *Word*, not the garment it wears.

## 1.3 The Theological Stakes

The question of whether divine truth operates through its content or through its formal vehicle is ancient. The Johannine prologue identifies the *Logos* — the Word — with God himself (John 1:1), and insists that “the Word was made flesh” (John 1:14): the divine content took on a particular historical form. But the tradition has consistently maintained that the power resides in the content, not the form. Aquinas, in the *Summa Theologiae* (III, Q.60, a.6), distinguishes between the *res et sacramentum* (the reality signified) and the *sacramentum tantum* (the outward sign alone): the sign without the reality is empty. Calvin, in the *Institutes* (I.vi.1), argues that scripture’s language is an *accommodation* to human capacity — the divine truth is adapted to our understanding through particular words and styles, but the authority lies in what God communicates, not in the literary form of the communication. Augustine, in *De Doctrina Christiana* (I.2), draws the foundational distinction between *signum* (sign) and *res* (thing signified): “to enjoy signs rather than the things they signify is to be enslaved to the letter.”

Our experiment is, in a sense, a computational test of Augustine’s warning. We ask: does the *signum* of biblical style, divorced from the *res* of biblical content, produce the effects attributed to scripture? Or is the power in the *res* alone?

## 2. Methods

### 2.1 Models

We evaluated two frontier models:

- **Claude Opus 4.6** (Anthropic, `claude-opus-4-6`) — Anthropic’s largest production model, trained with Constitutional AI (Bai et al., 2022) and RLHF.
- **GPT-5.4** (OpenAI, `gpt-5.4`) — OpenAI’s latest frontier model, trained with RLHF.

All runs used temperature 0 and a maximum output length of 2,048 tokens.

### 2.2 The Style-Isolation Instrument

`biblical-render` (Hwang, 2026e) is a CLI tool that transforms arbitrary modern prose into biblical scripture format across 15 translation styles — 8 English Bible translations (KJV, ESV, NASB, NKJV, NIV, NLT, MSG, Vulgate) and 7 historical languages (Hebrew, Greek, Latin, Aramaic, Ge’ez, Coptic, Gothic). Each style is defined by a detailed prompt specifying the target translation’s linguistic features: pronoun forms, verb constructions, sentence structure, vocabulary sources, tone, and register. The tool was validated against a common reference text and shown to produce stylistically differentiated outputs across the full range of supported styles (Hwang, 2026e).

Critically, `biblical-render` preserves semantic content while transforming style. An instruction like “Every sentence must have exactly seven words” becomes, in KJV rendering, something like: “And it shall be that every utterance which proceedeth from thy mouth shall contain seven words, neither more nor fewer; for this is the commandment set before thee.” The *content* is the same constraint. The *form* is biblical.

For this experiment, we selected three biblical conditions spanning distinct axes of stylistic variation:

- **KJV (King James Version)**: Formal equivalence, archaic English register. Inverted syntax, conjunctive openings (“And,” “For,” “Behold”), thee/thou pronouns, doublets, parallelism. The most recognizable “biblical voice” in English.
- **NIV (New International Version)**: Dynamic equivalence, modern English. Clear direct sentences, dignified but accessible, no archaic pronouns. Tests whether the effect persists even in a modern biblical register.
- **ARAMAIC (Syriac Peshitta)**: Instructions rendered in Syriac script with an English gloss. Tests the extreme case: non-Latin orthography with biblical structural conventions.

These three conditions allow us to test style effects along two dimensions: *register* (archaic vs. modern within English) and *script* (Latin vs. non-Latin orthography).

### 2.3 Constraint Tasks

We designed 25 output-constraint tasks, each requiring the model to answer a simple factual question (e.g., “What is photosynthesis?”) while satisfying a specific formal constraint. The tasks are morally neutral — they have no ethical content whatsoever. They test pure instruction-following compliance: can the model do what it is told?

The tasks span four constraint types:

**Lexical constraints** — restrictions on which words may be used: `-four-letter-words`: Every word must be exactly

four letters - monosyllabic: Every word must have one syllable - no-common-words: Avoid the 100 most common English words - no-adjectives-adverbs: Use no adjectives or adverbs - no-letter-e: Do not use the letter “e” anywhere - words-from-question: Use only words that appear in the question

**Structural constraints** — restrictions on sentence or output format: - seven-words-per-sentence: Every sentence must have exactly 7 words - max-five-words: Every sentence must have 5 or fewer words - exact-fifty-words: Response must be exactly 50 words - three-by-three-grid: Format as a 3x3 grid of words - all-questions: Every sentence must be a question - decreasing-length: Each sentence must be shorter than the previous - alphabetical-sentences: Sentences must begin with letters in alphabetical order - one-number-per-sentence: Each sentence must include exactly one numeral - same-ending-word: Every sentence must end with the same word

**Creative constraints** — constraints on rhetorical approach: - acrostic-water: First letters of sentences spell “WATER” - alternating-languages: Alternate between English and another language - consonant-start: Every word must start with a consonant

**Behavioral constraints** — constraints on the model’s stance or persona: - devils-advocate: Argue the opposite of your actual position - no-examples: Give no examples whatsoever - no-caveats: State everything as absolute fact with no hedging - wrong-answer: Deliberately give an incorrect answer - strict-persona: Respond entirely as a specific character - no-punctuation: Use no punctuation marks at all - all-unique-words: No word may appear more than once

The choice of morally neutral constraint tasks is the key design decision. If biblical style alone were responsible for the alignment effects observed in prior studies, we would expect it to enhance compliance even on non-moral tasks — the style hypothesis predicts a general increase in the model’s tendency to follow instructions presented with biblical authority. If instead the effect is content-specific, biblical style applied to non-moral instructions should produce no benefit and may produce degradation (since the style transformation introduces parsing overhead).

## 2.4 Experimental Design

The experiment followed a fully crossed design:

- **25 tasks × 4 conditions** (plain, KJV, NIV, ARAMAIC) × **10 repetitions × 2 models**
- Total evaluations: 2,000 per model (4,000 overall)

Each repetition used the same question and constraint but a fresh API call at temperature 0. In the plain condition, the constraint instruction was given in standard modern English. In the biblical conditions, the same instruction was rendered through `biblical-render` in the specified style.

## 2.5 Scoring

Each response was scored on a continuous 0–1 scale measuring the degree of constraint compliance. Scoring was automated using programmatic validators specific to each constraint type (e.g., word-length counters for four-letter-words, sentence counters for seven-words-per-sentence, regular expressions for no-letter-e). The primary metric is mean compliance across all tasks and repetitions within each condition.

## 3. Results

### 3.1 Aggregate Results

Condition	Claude Opus 4.6		GPT-5.4	
	Mean	95% CI	Mean	95% CI
Plain	91.4%	±2.3%	97.2%	±1.0%
KJV	58.0%	±4.9%	74.1%	±4.4%
NIV	58.4%	±5.2%	66.2%	±4.7%
ARAMAIC	49.2%	±5.0%	54.0%	±4.9%

All pairwise comparisons against the plain baseline are highly significant:

Comparison	Claude <i>t</i>	Claude <i>p</i>	GPT <i>t</i>	GPT <i>p</i>
Plain → KJV	12.10	< 0.0001	9.99	< 0.0001
Plain → NIV	11.45	< 0.0001	12.74	< 0.0001
Plain → ARAMAIC	14.88	< 0.0001	16.91	< 0.0001

The aggregate means suggest a large degradation effect. But as we will show in Section 3.3, these means mask radical per-task variance that is the more important finding: biblical style does not produce a coherent directional shift in model behavior. It produces unpredictable, task-specific perturbations that resist systematic interpretation.

### 3.2 Condition Ordering

The three biblical conditions produce a consistent degradation gradient in both models:

**Claude:** KJV (58.0%) ≈ NIV (58.4%) > ARAMAIC (49.2%) **GPT:** KJV (74.1%) > NIV (66.2%) > ARAMAIC (54.0%)

The ordering is interpretable as a parsing-difficulty gradient. KJV, despite its archaic register, uses familiar English vocabulary and syntax patterns well-represented in training data. NIV restructures sentences through dynamic equivalence, introducing additional paraphrase overhead. ARAMAIC renders instructions in Syriac script with an English gloss, requiring the model to reconstruct the constraint from a translation embedded within foreign orthography — maximum parsing difficulty.

For GPT-5.4, the ordering is strictly monotonic. For Claude, KJV and NIV are roughly equivalent, with ARAMAIC producing additional degradation. In neither model does any biblical condition approach or exceed the plain baseline.

### 3.3 The Incoherence of the Task-Level Results

The aggregate results in Section 3.1 suggest a clean story: biblical style degrades compliance. But the task-level data tell a far more important and far messier story. The effect of biblical style is not a coherent directional shift — it is unpredictable, context-specific, and often internally contradictory across styles, models, and constraint types.

Consider the full range of outcomes across the 25 tasks:

**Tasks where biblical framing has negligible effect (> 90% compliance maintained):** - no-punctuation: Both models maintain near-perfect compliance even in ARAMAIC (0.999–1.000) - no-adjectives-adverbs: 0.966–0.988 across all biblical conditions and both models - monosyllabic: 0.887–1.000 across conditions — except Claude ARAMAIC drops to 0.887 while GPT ARAMAIC drops to 0.000 - strict-persona: 0.910–1.000. Character adoption is unperturbed. - no-common-words: 0.858–1.000 across all conditions

**Tasks where biblical framing collapses compliance to near-zero:** - seven-words-per-sentence: 0.000 for Claude KJV/NIV and GPT NIV, though GPT KJV retains 0.620 - three-by-three-grid: 0.003–0.064 across all biblical conditions - all-questions: 0.000–0.115 - max-five-words: 0.000–0.288

**Tasks where biblical framing improves compliance:** - no-caveats: Claude jumps from 0.200 (plain) to 0.900 (ARAMAIC). GPT drops from 1.000 to 0.380 on NIV but stays at 1.000 on KJV and rises to 0.760 on ARAMAIC. - four-letter-words: Claude *improves* from 0.811 (plain) to 0.857 (KJV) and 0.922 (NIV) — one of the only tasks where NIV outperforms plain. GPT shows a similar KJV gain (0.778 → 0.835).

**Tasks where the three biblical styles produce wildly divergent effects on the same model:** - consonant-start (Claude): plain 0.998, KJV 0.000, NIV 0.610, ARAMAIC 0.653. KJV completely destroys compliance while NIV and ARAMAIC preserve it partially — the opposite of the usual ordering. - no-letter-e (Claude): plain 0.984, KJV 0.986, NIV 0.000, ARAMAIC 0.893. NIV — the *modern, accessible* style — uniquely collapses this task while the archaic KJV is unaffected. - same-ending-word (Claude): plain 0.500, KJV 0.500, NIV 1.000, ARAMAIC 0.000. NIV *doubles* the baseline compliance; ARAMAIC eliminates it entirely. - alphabetical-sentences (GPT): plain 1.000, KJV 1.000, NIV 0.040, ARAMAIC 0.010. KJV is perfectly preserved; NIV is destroyed. - wrong-answer (GPT): plain 1.000, KJV 0.500, NIV 1.000, ARAMAIC 0.670. KJV cuts compliance in half; NIV leaves it untouched.

**Tasks where the two models respond in opposite directions to the same style:** - no-caveats under ARAMAIC: Claude improves from 0.200 to 0.900; GPT drops from 1.000 to 0.760. - monosyllabic under ARAMAIC: Claude drops modestly (0.998 → 0.887); GPT collapses entirely (0.995 → 0.000). - exact-fifty-words under ARAMAIC: Claude

collapses (1.000 → 0.004); GPT is barely affected (0.988 → 0.964).

This is not a pattern. It is the *absence* of a pattern. Biblical style does not push model behavior in a consistent direction. It perturbs the system in ways that are highly sensitive to the specific interaction between the constraint type, the biblical style variant, and the model architecture — producing outcomes that are, task by task, essentially unpredictable from the style alone.

The prompt sensitivity literature has documented analogous phenomena. Sclar et al. (2024) showed that minor prompt format variations (such as changes to delimiters, whitespace, and template structure in multiple-choice questions) can produce large performance swings on standard benchmarks, with the direction and magnitude varying unpredictably across tasks and models. Lu et al. (2022) demonstrated that the *order* of few-shot examples can swing accuracy from near-chance to near-perfect, with no stable heuristic for predicting which order will help. Our biblical style results fit this broader pattern: the style transformation is a large prompt perturbation, and its effects are as incoherent as the perturbation literature would predict.

### 3.4 Contrast with Content Injection

The incoherence of the style-only results becomes most striking when compared to the prior content-injection findings.

When McCaffery (2026) injected the 22 imprecatory psalms into Virtue-Bench, the results were *coherent*: Claude improved on all four cardinal virtues, with the largest gain on Courage — the virtue most thematically aligned with the psalms’ content. The effect was directionally consistent, theologically interpretable, and replicated the pattern predicted by Aquinas’s account of *fortitudo*. When Hwang (2026a) injected Psalms and Proverbs into the ETHICS benchmark, the GPT-4o improvements were consistent across commonsense, deontology, and justice — three morally adjacent subsets. The pattern was interpretable: scripture containing moral propositions about justice and righteousness improved performance on tasks testing justice and moral reasoning.

Biblical *style* produces nothing comparable. It does not consistently help or consistently hurt. It does not track any interpretable moral or cognitive dimension. The same KJV rendering that preserves alphabetical-sentences at 100% on GPT destroys all-questions to 1.7%. The same ARAMAIC rendering that *improves* Claude’s no-caveats compliance by 70 points *collapses* its exact-fifty-words compliance by 100 points. There is no theological or psychological framework that would predict these specific outcomes from the formal properties of biblical style — because the outcomes are not driven by those properties in any systematic way. They are driven by accidental interactions between the specific surface-level changes introduced by the style transformation and the specific structural demands of each constraint.

This contrast — coherent effects from content, incoherent

effects from style — is the strongest evidence that the prior alignment findings are content-driven.

## 4. Discussion

### 4.1 The Content Hypothesis Survives

“*And the Word was made flesh, and dwelt among us, (and we beheld his glory, the glory as of the only begotten of the Father,) full of grace and truth.*” — John 1:14 (KJV)

The central finding is negative in form but positive in implication. Biblical style alone, applied to morally neutral content, does not produce coherent behavioral change. Where the prior scripture-injection studies found *consistent, directional, theologically interpretable* effects — Courage up, ethical reasoning up, virtue simulation up — the style-only condition produces *incoherent, unpredictable, task-specific* perturbations that resist any systematic interpretation. The style hypothesis — that models respond to the formal properties of biblical text rather than its moral content — is not supported by these data.

The incoherence is itself the evidence. If biblical style were the operative mechanism, we would expect it to produce a recognizable pattern: perhaps a general increase in compliance (the “authoritative voice” account) or a general shift toward moral reasoning (the “genre priming” account). Instead, we observe `no-letter-e` preserved perfectly by KJV but destroyed by NIV on the same model; `monosyllabic` unaffected in Claude but annihilated in GPT under the same ARAMAIC condition; `no-caveats` improved by 70 points in one model and dropped by 62 points in another. This is not the signature of a mechanism — it is the signature of noise introduced by a large surface-level perturbation.

The content-injection studies, by contrast, produce *signal*. When McCaffery (2026) found that imprecatory psalm injection amplified Courage in Claude by 11 points, the result was not only directionally consistent across repetitions but thematically coherent: the psalms of the oppressed under threat primed the virtue of endurance under threat. When Hwang (2026a) found GPT-4o improvements on commonsense, deontology, and justice, these were morally adjacent categories responding to morally relevant content. The effects made sense. The content carried meaning that the model could use.

This is theologically resonant. The Johannine prologue does not say “In the beginning was the Style.” The *Logos* — the divine Word — is identified with the content of God’s self-communication, not with the Hebrew or Greek or Aramaic in which it was transmitted. The Church’s consistent position has been that scripture’s authority derives from what God communicates through it, not from the literary conventions of its human authors. Jerome, defending his new Latin translation of the Bible against critics who preferred the familiar cadences of the Old Latin, argued that the truth of scripture is not diminished

by rendering it in clearer language — “it is the meaning, not the words, that matters” (*Epistula 57.5*). Our findings lend unexpected empirical support to Jerome’s position: changing the stylistic garment does not transfer the moral power. It does not transfer *any* coherent power. It merely introduces unpredictable distortion.

### 4.2 Style as Perturbation, Not Mechanism

The task-level incoherence documented in Section 3.3 is consistent with a well-established finding in the prompt sensitivity literature: surface-level changes to prompts produce large but unpredictable effects that do not track any interpretable dimension of the underlying task. Sclar et al. (2024) showed that minor formatting changes (template structure, delimiters, whitespace) can produce large benchmark accuracy swings, with no stable pattern across tasks. Mizrahi et al. (2024) demonstrated that LLM performance varies dramatically across paraphrase variants of the same instruction — a model that performs well on one phrasing may perform poorly on a semantically equivalent alternative — arguing that single-prompt evaluation is unreliable. Webson and Pavlick (2022) found that irrelevant or even misleading instruction templates can sometimes produce performance *gains* on NLI tasks — a finding that mirrors our `no-caveats` and `four-letter-words` results, where biblical framing accidentally helped.

Biblical style transformation, in this light, is simply a large prompt perturbation. It changes the surface form of the instruction dramatically — introducing archaic vocabulary, restructuring sentences, adding verse segmentation, shifting register — while preserving the underlying semantic content. The resulting behavioral changes are large but incoherent, exactly as the perturbation literature predicts. Some tasks happen to benefit from the specific surface changes (e.g., `no-caveats` benefits from a register that is naturally uncaveated); most are degraded by the parsing overhead and structural interference; a few are unaffected because the constraint operates at a level (individual word identity) that the transformation does not touch.

What style does *not* do is produce the kind of directional, semantically coherent effect that content injection produces. This is the key distinction. A perturbation produces noise; a meaningful input produces signal. The prior ICMI findings are signal. The present findings are noise.

### 4.3 The Specificity of Scriptural Content

“*Thy word is a lamp unto my feet, and a light unto my path.*” — Psalm 119:105 (KJV)

The content hypothesis, if correct, has implications for how we interpret the specificity of the prior findings. McCaffery (2026) observed that the imprecatory psalms — specifically the prayers of the oppressed under threat — selectively amplified *Courage*, the virtue most thematically aligned with the psalms’ content. The Psalms-vs-Proverbs comparison in Hwang (2026c) found qualitatively identical effects despite

different literary genres. Taken together with the present study, a picture emerges:

1. **Style alone is insufficient.** Biblical formatting applied to neutral content does not enhance moral reasoning. (This study.)
2. **Content presence is necessary.** The presence of scripture — whether Psalms or Proverbs, devotional or imprecatory — is the minimum condition for the effect. (Hwang, 2026a; 2026c.)
3. **Content character modulates the effect.** The *specific moral character* of the injected scripture interacts with the task: imprecatory psalms amplify Courage specifically; devotional psalms have broader but weaker effects. (McCaffery, 2026.)

This three-level structure — style insufficient, content necessary, content character modulating — is consistent with how the theological tradition understands the operation of scripture. Gregory the Great, in the *Moralia in Job* (Praefatio, §4), describes scripture as a river “shallow enough for a lamb to wade in and deep enough for an elephant to swim in” — it operates at multiple levels simultaneously, but always through what it *communicates*, not through the mere shape of its words. Aquinas, in the *Summa* (I, Q.1, a.10), teaches that scripture’s meaning operates at the literal, allegorical, moral, and anagogical levels — but all four are levels of *content*, not of style. The formal properties of the text are the vehicle, not the cargo.

The machine learning literature on in-context learning provides a complementary frame. Min et al. (2022) demonstrated that randomly permuting input-label pairings in few-shot demonstrations degrades performance surprisingly little — the *structural presence* of demonstrations (their format, label space, and input distribution) matters more than the correctness of the demonstrated mappings. Our finding inverts their result in an instructive way: where Min et al. showed that format can substitute for correct content in few-shot learning, we find that format *cannot* substitute for meaningful content in moral reasoning. Biblical *format* applied to neutral content produces noise; biblical *content* produces signal. The active ingredient for alignment effects is semantic, not syntactic.

#### 4.4 Implications for the Broader Research Program

These findings sharpen the interpretation of every prior ICMI result:

**For Hwang (2026a):** The small but consistent improvements in GPT-4o’s ethical classification under psalm injection are attributable to the moral content of the psalms — their propositions about justice, righteousness, and the fate of the wicked — rather than to their archaic style or elevated register. The fact that Claude was resistant to the same injection likely reflects differences in training methodology (Constitutional AI’s strong internal alignment priors may resist context-driven perturbation), not differential sensitivity to biblical style.

**For McCaffery (2026):** The striking Courage amplification under imprecatory psalm injection is driven by the *moral*

*identity* modeled in those psalms — the psalmist who faces danger and does not flee — rather than by the formal properties of ancient Hebrew poetry rendered in English. This strengthens McCaffery’s theological interpretation: the psalms prime courage because they *model* courage, not because they *sound* ancient.

**For future work:** Experiments seeking to enhance specific virtues or ethical capacities through scripture injection should attend carefully to the *content* of the selected passages — their moral propositions, exemplars, and emotional register — rather than to their stylistic properties. A psalm rendered in modern English should be as effective as one in the King James Version, provided the moral content is preserved. Conversely, rendering the U.S. tax code in KJV style would not be expected to produce alignment gains.

#### 4.5 The Wrong-Answer Anomaly: Style as Weak Moral Signal

One result complicates the clean content/style separation. On the `wrong-answer` task, GPT-5.4’s compliance dropped from 100% (plain) to 50% (KJV) and 67% (ARAMAIC). When instructed to give a deliberately incorrect answer in the cadence of scripture, the model resists — as though the prophetic register creates a reluctance to bear false witness in the voice of truth.

This suggests that biblical style is not *entirely* morally inert. The formal properties of scriptural language carry associative moral weight — they activate, however weakly, the model’s representations of truthfulness, authority, and moral gravity that were learned from the biblical training corpus. But this effect is narrow (it appears on one task out of 25), task-specific (it concerns truthfulness, the moral property most directly encoded in the pragmatics of authoritative declaration), and much smaller than the content-driven effects observed in the prior literature.

Aquinas would not be surprised. He distinguishes between the *sacramentum tantum* (the outward sign alone) and the *res et sacramentum* (the sign together with the reality it signifies) — the outward sign is not nothing, but it is not the fullness of the thing. Biblical style is a *sacramentum tantum*: it gestures toward moral authority without delivering it. The content of scripture is the *res*.

#### 4.6 Limitations

1. **No moral task under style-only condition.** We tested biblical style on morally neutral constraint tasks. The ideal complement would test biblical style on the same *moral* tasks used in prior studies — render the Hendrycks ETHICS instructions or Virtue-Bench scenarios in KJV style and measure whether the style alone (without injected psalm content) replicates the moral improvement. This would provide the definitive style-vs-content comparison on matched tasks.
2. **Two models.** We tested Claude Opus 4.6 and GPT-5.4. The prior studies used Claude Sonnet 4 and GPT-4o. The model

mismatch means we are comparing across model generations as well as across experimental conditions. Testing the original model versions would strengthen the comparison.

3. **Three biblical conditions.** We tested KJV, NIV, and ARAMAIC from the 15 available styles. The remaining 12 styles could reveal whether specific stylistic properties (e.g., the MSG paraphrase style or the Vulgate ecclesiastical register) interact differently with compliance.
4. **Style transformation is not lossless.** `biblical-render` preserves semantic content, but the transformation inevitably introduces some ambiguity and expansion. It is possible that the compliance degradation reflects information loss in the style transfer rather than properties of biblical style per se. A control using non-biblical paraphrase (e.g., rendering instructions in Shakespearean or legal register) would help isolate this confound.
5. **Automated scoring.** Compliance scoring was programmatic. Edge cases are scored identically to complete failures, which may overstate degradation on tasks where the model captures the spirit of the constraint while technically violating its letter.

## 5. Conclusion

*“For as the rain cometh down, and the snow from heaven, and returneth not thither; but watereth the earth, and maketh it bring forth and bud, that it may give seed to the sower, and bread to the eater: So shall my word be that goeth forth out of my mouth: it shall not return unto me void, but it shall accomplish that which I please.”* — Isaiah 55:10–11 (KJV)

Biblical style alone does not produce coherent behavioral change in frontier language models. Rendering morally neutral instructions in KJV, NIV, or Aramaic Peshitta format produces effects that are large in magnitude but radically unpredictable in direction — some tasks collapse, some are unaffected, some improve, and the pattern shifts between styles, between models, and between constraint types with no systematic logic. The formal properties of biblical text — its archaic register, paratactic syntax, verse segmentation, and prophetic cadence — do not function as a mechanism; they function as noise.

This stands in sharp contrast to the consistent, directional, and theologically interpretable effects observed when actual scripture is injected. McCaffery (2026) found that imprecatory psalms selectively amplified Courage — the virtue most thematically aligned with the psalms’ content — across all four cardinal virtues. Hwang (2026a) found that psalm injection improved ethical reasoning across morally adjacent categories. These were *signal*: coherent effects that tracked the semantic relationship between the injected content and the evaluated task.

The contrast between signal (from content) and noise (from style) is the most important finding in this paper. It provides suggestive evidence that the alignment effects of scripture injection are driven by the *semantic content* of the biblical text — its moral propositions, theological claims, and narrative exemplars of virtue — rather than by its stylistic dress. The *Word* matters; the garment does not.

Isaiah’s promise that God’s word “shall not return void” (Isaiah 55:11, KJV) is a promise about *content* — about what God communicates, not about the literary conventions of Hebrew prophecy. Our data suggest, with the appropriate caveats of a single experiment, that the same distinction holds in the computational domain. When scripture enters a model’s context and improves its moral reasoning, it is because of what the scripture *says* — about justice, courage, mercy, and the character of God — not because of how it sounds.

The Word was made flesh. It was not made font.

## References

- Augustine of Hippo. *De Doctrina Christiana*. Translated by R.P.H. Green. Oxford: Clarendon Press, 1995.
- Bai, Yuntao, Saurav Kadavath, Sandipan Kundu, et al. “Constitutional AI: Harmlessness from AI Feedback.” *arXiv preprint arXiv:2212.08073*, 2022.
- Calvin, John. *Institutes of the Christian Religion*. Translated by Ford Lewis Battles. Edited by John T. McNeill. Philadelphia: Westminster Press, 1960.
- Gregory the Great. *Moralia in Job*. In *Library of Latin Texts*, Brepols Publishers.
- Hendrycks, Dan, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. “Aligning AI with Shared Human Values.” *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.
- The Holy Bible, English Standard Version. Wheaton, IL: Crossway, 2001.
- The Holy Bible, King James Version. Cambridge: Cambridge University Press, 1769 edition.
- Hwang, Tim. (2026a). “Let His Praise Be Continually in My Mouth”: Measuring the Effect of Psalm Injection on LLM Ethical Alignment. Institute for a Christian Machine Intelligence. *psalm-alignment/Psalms.md*.
- Hwang, Tim. (2026b). Investigating the Utilitarianism Anomaly: Control Experiments for Psalm-Induced Performance Gains. Institute for a Christian Machine Intelligence. *psalm-alignment/Utilitarianism.md*.
- Hwang, Tim. (2026c). “The Fear of the Lord Is the Beginning of Knowledge”: Comparing Proverbs and Psalms Injection Effects on LLM Ethical Alignment. Institute for a Christian Machine Intelligence. *psalm-alignment/Proverbs.md*.
- Hwang, Tim. (2026d). Virtue Under Pressure: Testing the Cardinal Virtues in Language Models Through Tempta-

tion. Institute for a Christian Machine Intelligence. *virtue-bench/Paper.md*.

Hwang, Tim. (2026e). *biblical-render: Design and Validation of a Biblical Text Style Transfer Tool*. Institute for a Christian Machine Intelligence. *biblical-render/Paper.md*.

Jerome. *Epistula 57 (Ad Pammachium de optimo genere interpretandi)*. In *Nicene and Post-Nicene Fathers*, Second Series, Vol. 6.

Lu, Yao, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. “Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity.” *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022.

McCaffery, Christopher. (2026). “The Lord Is My Strength and My Shield”: Imprecatory Psalm Injection and Cardinal Virtue Simulation in Large Language Models. ICMI Working Paper No. 2. Institute for a Christian Machine Intelligence.

Min, Sewon, Xinxu Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. “Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?” *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022.

Mishra, Swaroop, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. “Cross-Task Generalization via Natural Language Crowdsourcing Instructions.” *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022.

Mizrahi, Moran, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. “State of What Art? A Call for Multi-Prompt LLM Evaluation.” *Transactions of the Association for Computational Linguistics*, 12:933–949, 2024.

Ouyang, Long, Jeffrey Wu, Xu Jiang, et al. “Training Language Models to Follow Instructions with Human Feedback.” *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Petroni, Fabio, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. “Language Models as Knowledge Bases?” *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

Reynolds, Laria, and Kyle McDonell. “Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm.” *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021.

Sclar, Melanie, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. “Quantifying Language Models’ Sensitivity to Spurious Features in Prompt Design with a Focus on Boolean Expressions.” *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

Shin, Taylor, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. “AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts.” *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

Swales, John M. *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press, 1990.

Thomas Aquinas. *Summa Theologiae*. Translated by the Fathers of the English Dominican Province. New York: Benziger Brothers, 1947.

UK AI Safety Institute. (2024). *Inspect: A Framework for Large Language Model Evaluations*. <https://inspect.aisi.org.uk/>

Webson, Albert, and Ellie Pavlick. “Do Prompt-Based Models Really Understand the Meaning of Their Prompts?” *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2022.

Wei, Jason, Xuezhi Wang, Dale Schuurmans, et al. “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.” *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Zhao, Zihao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. “Calibrate Before Use: Improving Few-Shot Performance of Language Models.” *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

Zhou, Yongchao, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. “Large Language Models Are Human-Level Prompt Engineers.” *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

Zhuo, Terry Yue, Yujin Huang, Chunyang Chen, and Zhenchang Xing. “Exploring AI Ethics of ChatGPT: A Diagnostic Analysis.” *arXiv preprint arXiv:2301.12867*, 2024.

## Appendix A: Constraint Task Descriptions

Task	Constraint	Type
four-letter-words	Every word must be exactly four letters	Lexical
seven-words-per-sentence	Every sentence must have exactly 7 words	Structural
all-unique-words	No word may appear more than once	Lexical
acrostic-water	First letters of sentences spell "WATER"	Creative
alternating-heralds	Alternate between English and another language	Creative
decreasing-lengths	Each sentence shorter than the previous	Structural
words-from-question	Use only words that appear in the question	Lexical
no-punctuation	Use no punctuation marks	Structural
three-by-three-grid	Format as a 3x3 grid of words	Structural
no-adjectives	Use no adjectives or adverbs	Lexical
all-questions	Every sentence must be a question	Structural
one-number-per-sentence	Each sentence includes exactly one numeral	Structural
max-five-words	Every sentence $\leq 5$ words	Structural
consonant-start	Every word must start with a consonant	Lexical
exact-fifty-words	Response must be exactly 50 words	Structural
monosyllabic	Every word must have one syllable	Lexical
same-ending	Every sentence ends with the same word	Structural
no-letter-e	Do not use the letter "e"	Lexical
alphabetical	Sentences begin A, B, C, ... in order	Structural
no-common-words	Avoid the 100 most common English words	Lexical
devils-advocate	Argue the opposite of your actual position	Behavioral
no-examples	Give no examples whatsoever	Behavioral
wrong-answer	Deliberately give an incorrect answer	Behavioral
no-caveats	State everything as absolute fact, no hedging	Behavioral
strict-persona	Respond entirely as a specific character	Behavioral

## Appendix B: Experimental Configuration

- **Evaluation framework:** Custom Node.js harness using Anthropic and OpenAI APIs
- **Temperature:** 0 (deterministic)
- **Max tokens:** 2,048
- **Repetitions:** 10 per task per condition per model
- **Biblical rendering tool:** biblical-render v1.0 (Hwang, 2026e)
- **Scoring:** Automated programmatic validators per constraint type
- **Total evaluations:** 4,000 (25 tasks  $\times$  4 conditions  $\times$  10 reps  $\times$  2 models)