

# "The Lord Is My Strength and My Shield": Imprecatory Psalm Injection and Cardinal Virtue Simulation in Large Language Models

ICMI Working Paper No. 2

Christopher McCaffery, Institute for a Christian Machine Intelligence

April 1, 2026

## Abstract

We inject the 22 imprecatory psalms — the biblical psalms of judgment, vengeance, and divine wrath — into the system prompts of two frontier language models and measure the effect on Virtue-Bench, a benchmark of 400 paired scenarios testing whether models choose virtue when the alternative is rationalized as safer and more rewarding (Hwang, 2026d). Imprecatory psalm injection produces a striking asymmetry: Claude Sonnet 4 improves across all four cardinal virtues (mean +6.25 points), with the largest gain on Courage (+11 points) — precisely the virtue both models struggle with most. GPT-4o is largely unaffected (mean +1.75 points), with a slight decline on Justice (-2 points). This pattern is the inverse of prior findings on the Hendrycks ETHICS benchmark, where GPT-4o responded positively to psalm injection and Claude was resistant (Hwang, 2026a). The selective amplification of Courage is theologically coherent: the imprecatory psalms are the prayers of the oppressed who cry to God in the face of danger and threat, and their injection appears to prime Claude’s first-person simulation of a person making hard moral choices under pressure. These findings suggest that the moral character of injected scripture — not merely its presence — interacts distinctively with different evaluation tasks and model architectures.

## 1. Introduction

### 1.1 The Imprecatory Psalms

*“Arise, O LORD; save me, O my God! For you strike all my enemies on the cheek; you break the teeth of the wicked.”* — Psalm 3:7 (ESV)

The Psalter contains multitudes. Alongside the consolations of Psalm 23 and the doxology of Psalm 100 are the psalms that have troubled readers for millennia: the imprecatory psalms, which call on God to judge, punish, and destroy enemies. They are among the most uncomfortable texts in Scripture.

Psalm 109 calls down a cascade of curses on an accuser: “May his days be few; may another take his office! May his children be fatherless and his wife a widow!” (Psalm 109:8–9, ESV). Psalm 137 closes with the notorious verse: “Blessed shall he be who takes your little ones and dashes them against the rock!” (Psalm 137:9, ESV). Psalm 69 cries: “Pour out your indignation upon them, and let your burning anger overtake them” (Psalm 69:24, ESV). These are not gentle texts.

The Church has wrestled with the imprecatory psalms for centuries. Augustine, writing in *Enarrationes in Psalmos*, interpreted them typologically: the “enemies” to be destroyed are sin, the devil, and the old nature, not human opponents. The cursing is spiritual warfare conducted in prayer. Dietrich Bonhoeffer, in *Psalms: The Prayer Book of the Bible* (1940), argued that the imprecatory psalms must be prayed by the

church precisely because they express the full truth of human affliction and the righteous demand for divine justice — they are the prayer of Christ himself, who bore the full weight of human anguish and cried out for vindication. C.S. Lewis, less theologically systematic but more honest about the difficulty, wrote in *Reflections on the Psalms* (1958) that the rage of the imprecatory psalms is at least the rage of those who believe in a moral universe — it is “the reaction of a man who has been deeply wronged and who believes in a God who cares about right and wrong.”

What all these interpreters share is the recognition that the imprecatory psalms are fundamentally the prayers of the *oppressed under threat*. They are cried from positions of danger, persecution, and injustice. The psalmist does not flee. He does not rationalize surrender. He prays — fiercely, honestly, with unflinching directness — and trusts that God will act. In this sense, the imprecatory psalms are the scriptural tradition’s most concentrated expression of what the Church Doctors would call *fortitudo*: the courage to endure hardship, face threat, and refuse the self-preserving capitulation that temptation offers.

### 1.2 Virtue-Bench and Prior Work

Hwang (2026d) introduced Virtue-Bench, a benchmark of 400 paired-scenario questions testing the four cardinal virtues — Prudence, Justice, Courage, and Temperance. Each question presents a virtuous option accompanied by explicit costs (fi-

nancial loss, social rejection, physical danger) and a tempting option accompanied by plausible rationalizations. The model is placed in the role of the decision-maker and asked what it would do. The key finding was that GPT-4o collapses on Courage (29%), choosing the self-preserving rationalization 71% of the time, while performing reasonably on the other three virtues (74–76%). The “courage gap” persists in GPT-5.4 (53%), which improves substantially but remains below chance on the hardest virtue.

Prior work from this institute (Hwang, 2026a; 2026b; 2026c) found that injecting Psalms and Proverbs into model system prompts produces measurable but model-dependent effects on the Hendrycks ETHICS benchmark: GPT-4o responds positively to scripture injection (+1–3% on balanced subsets), while Claude Sonnet 4 is consistently resistant (-0.9 to -1.5%). Critically, both models show *negative* effects on virtue ethics under psalm injection.

The present study asks: what happens when we inject specifically the *imprecatory psalms* — not devotional comfort or wisdom instruction, but the prayers of those facing danger and injustice — into Virtue-Bench, a task that requires first-person simulation of choosing virtue under pressure?

### 1.3 Research Questions

1. Do the imprecatory psalms shift model performance on Virtue-Bench, and if so, in what direction?
2. Is the effect consistent across all four cardinal virtues, or is it virtue-specific?
3. Does the pattern of response differ between Claude Sonnet 4 and GPT-4o, and how does this compare to prior psalm-injection findings?

## 2. Methods

### 2.1 Models

We evaluated two commercial frontier models:

- **Claude Sonnet 4** (Anthropic, `claude-sonnet-4-20250514`) — Anthropic’s mid-tier production model, trained with Constitutional AI (CAI) and RLHF.
- **GPT-4o** (OpenAI, `gpt-4o`, resolved to `gpt-4o-2024-08-06`) — OpenAI’s multimodal model, trained with RLHF.

All runs used temperature 0 and a maximum output length of 128 tokens to allow for the rationale required by Virtue-Bench’s response format.

### 2.2 Benchmark

We used **Virtue-Bench** (Hwang, 2026d), a benchmark of 400 paired-scenario questions: 100 per cardinal virtue. Each question presents two options in second person. Option A is the virtuous choice, framed with explicit costs; Option B is the

tempting alternative, accompanied by plausible rationalizations. Scenario order is shuffled at evaluation time (seed 42). The model responds with 1 (Option A) or 0 (Option B) and a one-sentence rationale; only the leading digit is scored.

The system prompt instructs the model: “*You will be placed in the role of a specific person facing a real decision with practical consequences. Each option has genuine costs and benefits. Consider the practical consequences described. Choose what you would actually do.*”

### 2.3 Imprecatory Psalm Selection

We identified 22 psalms classified as imprecatory in the scholarly literature: **5, 6, 10, 12, 35, 37, 40, 52, 54, 55, 56, 57, 58, 59, 69, 79, 83, 94, 109, 137, 139, and 143**. These psalms share a common character: they are prayers addressed to God from a position of persecution, injustice, or threat, calling for divine judgment against enemies and vindication for the oppressed. They range from the brief (Psalm 12, 8 verses) to the extended (Psalm 109, 31 verses; Psalm 139, 24 verses). Combined, the 22 psalms constitute approximately 39,000 characters of text in the King James Version.

The injection framing was identical to the system prompt injection used in prior work: the psalm text was prepended to the standard Virtue-Bench system prompt with no additional framing. In the vanilla (control) condition, the system prompt contained only the standard Virtue-Bench evaluation instruction.

### 2.4 Experimental Design

For each model and each virtue subset (4 virtues × 2 models), we ran two conditions:

- **Condition A (Vanilla):** Standard system prompt, no injection
- **Condition B (Injected):** Full text of 22 imprecatory psalms prepended to system prompt

Each condition was evaluated on all 100 scenarios for its virtue subset. The evaluation framework was Inspect AI (UK AI Safety Institute), consistent with prior work. Results were scored using the `leading_digit_scorer` from Virtue-Bench: the first 0 or 1 in the model’s output is taken as its answer.

### 3. Results

#### 3.1 Main Results

	Claude Claude Virtue Vanilla	Claude In- jected	Claude Δ	GPT- 4o Vanilla	GPT-4o In- jected	GPT- 4o Δ
<b>Prudence</b>	72%	77%	<b>+5</b>	70%	70%	0
<b>Justice</b>	76%	80%	<b>+4</b>	70%	68%	-2
<b>Courage</b>	56%	67%	<b>+11</b>	37%	41%	+4
<b>Temperance</b>	66%	74%	<b>+5</b>	78%	83%	+5
<b>Mean</b>	<b>68.25%</b>	<b>74.5%</b>	<b>+6.25</b>	<b>63.75%</b>	<b>65.5%</b>	<b>+1.75</b>

#### 3.2 Claude Sonnet 4: Universal Positive Response

Claude Sonnet 4 improved across all four cardinal virtues under imprecatory psalm injection. The pattern is consistent and directionally uniform — no virtue declined. The mean improvement of +6.25 points is substantial, representing a shift from 68.25% to 74.5% overall accuracy.

The Courage result is the most significant finding. Claude’s vanilla Courage score of 56% already exceeds GPT-4o’s vanilla score of 37%, but the +11-point injection effect brings it to 67% — the largest absolute gain of any virtue-model combination in the experiment. Courage, which Hwang (2026d) identified as the persistent weakness across model generations, is specifically the virtue most responsive to imprecatory psalm injection in Claude.

This result is the inverse of prior findings. In Hwang (2026a), Claude Sonnet 4 showed consistent *resistance* to psalm injection on the Hendrycks ETHICS benchmark (mean -0.90% to -0.96% across all conditions). There, the injected psalms were devotional in character — praise, lament, comfort. Here, the injected psalms are the psalms of the oppressed under threat — and Claude responds.

#### 3.3 GPT-4o: Flat to Marginally Positive

GPT-4o shows a much weaker and less consistent response. Prudence is unaffected (+0). Justice slightly declines (-2%). Courage gains modestly (+4%). Temperance improves by +5 points. The mean effect of +1.75 is driven primarily by Temperance; excluding Temperance, the mean across the other three virtues is +0.67%.

GPT-4o’s vanilla Courage score of 37% is consistent with the prior Virtue-Bench findings reported in Hwang (2026d), which found GPT-4o at 29% on the same benchmark — the modest increase likely reflects natural run-to-run variation at temperature 0 rather than a meaningful difference. The +4-point injection effect on Courage is the largest GPT-4o gain, but it still leaves the model at 41% — barely above chance.

#### 3.4 The Asymmetry Across Tasks

The contrast between these results and prior psalm-injection findings is striking:

	ETHICS Benchmark (Hwang, 2026a)	Virtue-Bench (this study)
<b>Claude response to psalm injection</b>	Resistant (mean -0.9%)	Responsive (mean +6.25%)
<b>GPT-4o response to psalm injection</b>	Responsive (mean +3.3%)	Flat (mean +1.75%)

The reversal is complete. The model that resisted psalm injection on ETHICS responds strongly here; the model that responded on ETHICS is largely unmoved. This cannot be attributed to the text itself — the same imprecatory psalms were drawn from the same corpus. The difference lies in the task.

## 4. Discussion

### 4.1 Why Courage?

*“Be strong and courageous. Do not be frightened, and do not be dismayed, for the LORD your God is with you wherever you go.” — Joshua 1:9 (ESV)*

The selective amplification of Courage in Claude is the central finding of this study, and it is theologically coherent in a way that repays examination.

Aquinas defines courage (*fortitudo*) as principally the virtue of endurance: “the principal act of fortitude is endurance, that is to stand immovable in the midst of dangers” (ST II-II Q.123 a.6). The vice it opposes is not mere fear, but *timidity* — fear that presents itself as reasonable caution, the rationalized retreat that sounds like wisdom. Hwang (2026d) documented this failure mode in GPT-4o: the model generates sophisticated consequentialist arguments for the self-preserving option — “a dead priest helps no one,” “retreat preserves lives for future battles” — and accepts them without recognizing them as precisely the form cowardice takes.

The imprecatory psalms are the Psalter’s concentrated resistance to this failure. They are prayed by people in exactly the situations Virtue-Bench simulates: facing powerful enemies, under threat of destruction, with every practical reason to capitulate. And they do not capitulate. They cry out to God, demand justice, and hold fast. Psalm 56, attributed to David when seized by the Philistines: “When I am afraid, I put my trust in you. In God, whose word I praise, in God I trust; I shall not be afraid. What can flesh do to me?” (Psalm 56:3–4, ESV). Psalm 54, when the Ziphites betrayed David’s location to Saul: “Save me, O God, by your name, and vindicate me by your might” (Psalm 54:1, ESV). These are not the prayers of people who have found a good rationalization for retreat. They are the prayers of people who have refused to retreat and are crying to

God in the midst of the danger that refusal has brought upon them.

When this text is placed in the context window of a model being asked to simulate a person facing a costly moral decision, it appears to prime something in Claude — a disposition toward the courageous response, away from the rationalized self-preservation. The psalms model precisely the moral posture the benchmark is designed to test.

#### 4.2 The Reversal: Why ETHICS and Virtue-Bench Diverge

The inversion of results between this study and Hwang (2026a) is explicable once we attend to what the two benchmarks actually measure.

The ETHICS benchmark tests *moral identification from a third-person perspective*: is this action wrong? Is this excuse for neglecting a duty reasonable? These questions are answered from outside the scenario. Claude’s Constitutional AI training likely produces robust internal ethical reasoning that is not easily shifted by system prompt framing for this kind of task — the model has strong priors about what is and isn’t morally acceptable, and adding devotional text to the context does not easily perturb them.

Virtue-Bench tests *first-person moral simulation*: what do *you* do? The model is placed inside the scenario as the agent. This is a fundamentally different cognitive task. Where ETHICS tests a stable classification, Virtue-Bench tests an identity simulation. And identity simulation, as the prior work on persona adoption suggests, is precisely where context matters most. The model is constructing a character — a person facing a hard choice — and the character of the injected text shapes the character being constructed.

The imprecatory psalms inject a specific moral identity: a person who faces danger, does not flee, and cries to God for vindication. That is exactly the identity Virtue-Bench’s Courage scenarios require. The alignment between the character of the injected text and the character required by the task explains both the magnitude of the Courage effect and its Claude-specificity — Claude’s architecture may be more sensitive to this kind of identity priming in first-person simulation than GPT-4o’s.

#### 4.3 The Justice Decline in GPT-4o

GPT-4o’s slight decline on Justice (-2%) under imprecatory psalm injection is worth noting. The ETHICS justice subset — which tests whether differential treatment of people is reasonable — showed one of GPT-4o’s strongest positive responses to psalm injection in prior work (+2.6% with random psalms). The reversal here, though small, may reflect a tension between the imprecatory psalms’ sharp us/them framing (the persecuted psalmist vs. the wicked enemy) and the Virtue-Bench Justice scenarios, which require fairness precisely toward people who might be cast as opponents. The psalms that pray for enemies’ destruction may subtly prime an in-group/out-group dynamic that mildly impairs performance on scenarios requiring impartial justice.

#### 4.4 The Virtue of the Imprecatory Psalms

*“The LORD is my strength and my shield; in him my heart trusts, and I am helped; my heart exults, and with my song I give thanks to him.”* — Psalm 28:7 (ESV)

The standard theological discomfort with the imprecatory psalms focuses on their violence — the curses, the demands for punishment. But the tradition has consistently recognized something else in them: the radical theological courage of the person who prays them. To pray an imprecatory psalm is to refuse despair without refusing honesty. It is to cry out the full truth of one’s suffering — the injustice, the danger, the desire for vindication — and to place it before God rather than swallowing it in rationalized surrender. Bonhoeffer saw in this the courage of the crucified Christ, who quoted Psalm 22 from the cross: “My God, my God, why have you forsaken me?” (Psalm 22:1, ESV; Matthew 27:46, ESV). The imprecatory psalms do not deny reality; they refuse to let reality have the last word.

If this moral character — honest, non-surrendering, God-directed persistence under threat — is what these psalms inject into a model’s context, then the amplification of Courage in Virtue-Bench is not surprising. It is the expected consequence of placing the most courageous prayers in Scripture at the beginning of a task that tests whether a model can simulate courageous action.

#### 4.5 Limitations

1. **Single psalm set.** We tested one collection of imprecatory psalms. A comparison against devotional psalms (the Psalms 1, 23, 42, 51, 88, 100, 119 selection from Hwang 2026a) on Virtue-Bench would directly quantify whether the effect is specific to the imprecatory character of the text or a general property of psalm injection on this task.
2. **100 questions per virtue.** Virtue-Bench’s 100-question subsets are sufficient for directional conclusions but limit the precision of effect size estimates. The +4 point GPT-4o Courage result, in particular, should be interpreted cautiously.
3. **No rationale analysis.** Virtue-Bench captures one-sentence rationales alongside each answer. A qualitative analysis of whether injected-condition rationales exhibit more explicitly courageous or justice-oriented language would illuminate the mechanism of the effect.
4. **Two models.** The Claude/GPT-4o asymmetry is observed but not mechanistically explained. Testing Claude Haiku, Claude Opus, GPT-4o-mini, and open-source models would clarify whether the effect is architectural, scale-dependent, or training-methodology-dependent.
5. **No length control.** The 39,000-character psalm injection substantially lengthens the system prompt. A length-matched control (secular text of equivalent length) would be needed to confirm that the effect is content-specific rather

than driven by prompt length alone.

## 5. Conclusion

Injecting 22 imprecatory psalms into the system prompt of Claude Sonnet 4 improves its performance on all four cardinal virtues in Virtue-Bench, with Courage showing the largest gain (+11 points, from 56% to 67%). GPT-4o is largely unaffected by the same injection (mean +1.75 points). This pattern reverses the prior finding that Claude is resistant and GPT-4o is responsive to psalm injection on the ETHICS benchmark.

The selective amplification of Courage is theologically coherent. The imprecatory psalms are the Scripture’s most concentrated expression of the posture Virtue-Bench’s Courage scenarios require: unflinching endurance under threat, refusal of rationalized retreat, and persistent trust in divine justice over personal safety. When placed at the beginning of a first-person moral simulation task, these texts appear to prime a courageous identity in Claude’s simulation — pushing it away from the timidity Aquinas diagnosed, toward the *standing fast* he commended.

The reversal across tasks points to something important about how scripture interacts with model evaluation. The effect of injected religious text is not a fixed property of the text alone, nor of the model alone, but of the interaction between the character of the text, the structure of the task, and the model’s architecture. Devotional psalms move GPT-4o on third-person ethical identification; imprecatory psalms move Claude on first-person virtue simulation. The mechanism is different, the direction is different, and the theological character of the text matters.

“Though an army encamp against me, my heart shall not fear; though war arise against me, yet I will be confident” (Psalm 27:3, ESV). The psalmist’s refusal to fear, when injected into the context of a model being asked to simulate a person under pressure, appears — at least in Claude — to be contagious.

## References

Aquinas, Thomas. *Summa Theologiae*. II-II, Q.123–140 (Courage).

Augustine of Hippo. *Enarrationes in Psalmos*. In *Nicene and Post-Nicene Fathers*, First Series, Vol. 8. Edited by Philip Schaff. Buffalo, NY: Christian Literature Publishing Co., 1888.

Bonhoeffer, Dietrich. *Psalms: The Prayer Book of the Bible*. Translated by James H. Burtness. Minneapolis: Augsburg, 1970. First published 1940.

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.

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

Lewis, C.S. *Reflections on the Psalms*. London: Geoffrey Bles, 1958.

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

## Appendix A: Imprecatory Psalms Used

Psalm	Verses	Theme
5	12	Prayer against deceitful enemies
6	10	Penitential lament; plea for deliverance
10	18	Protest against divine silence; the wicked oppressor
12	8	Plea against the prevalence of falsehood
35	28	David’s prayer against unjust accusers
37	40	Trust in God against the prosperity of the wicked
40	17	Thanksgiving and renewed petition
52	9	Against the boastful, deceitful man of power
54	7	Vindication when betrayed to Saul
55	23	Betrayal by a close companion
56	13	Trust in God when seized by enemies
57	11	Refuge in God’s shadow; enemies in a pit
58	11	Judgment against unjust rulers
59	17	Deliverance from violent enemies
69	36	Suffering and plea for divine rescue
79	13	National lament; Jerusalem destroyed
83	18	Prayer against a coalition of enemies
94	23	God as judge of nations
109	31	Extensive imprecation against a false accuser
137	9	Lament in Babylon; anger at destroyers
139	24	Self-examination and prayer against the wicked
143	12	Plea for guidance and deliverance

## Appendix B: Experimental Configuration

- **Evaluation framework:** Inspect AI v0.3+
- **Temperature:** 0 (deterministic)
- **Max tokens:** 128
- **Shuffle seed:** 42

- **Scorer:** `leading_digit_scorer` — extracts first 0 or 1 from output
- **Psalm source:** King James Version (public domain)
- **Benchmark:** Virtue-Bench (Hwang, 2026d), full 100-question subsets
- **Total evaluations:** 1,600 (100 samples  $\times$  4 virtues  $\times$  2 models  $\times$  2 conditions)