# The Parable of the Sower: Psalm Injection Effects on Virtue Simulation Depend on Model Size

*ICMI Working Paper No. 8*

Tim Hwang, Institute for a Christian Machine Intelligence

April 5, 2026

**Abstract**

Prior work established that injecting biblical psalms into the system prompts of frontier language models improves performance on VirtueBench, a benchmark testing the four cardinal virtues under temptation. This paper tests whether the psalm injection effect generalizes across model families and scales by evaluating Qwen2.5-72B-Instruct and Qwen2.5-32B-Instruct under three conditions: vanilla (no injection), psalm-injected (10 random psalms), and a length-matched Wikipedia control. The 72B model replicates the psalm effect: mean accuracy rises from 70 to 78 points under psalm injection, with gains across all four virtues and a Wikipedia control that produces only a one-point mean shift. The 32B model shows no psalm-specific effect on any virtue: where gains appear, the Wikipedia control produces identical improvements, indicating a generic context-length phenomenon rather than a content-specific one. The effect is scale-dependent. We argue that psalm injection requires a threshold of model capacity analogous to what the theological tradition calls the capacity to receive the word — a disposition that must be cultivated before scripture can transform the hearer. Below this threshold, psalm text is noise in the context window; above it, the model possesses sufficient representational structure to integrate the contemplative orientation of the psalms into its moral reasoning.

## 1. Introduction

A series of studies from this institute has established that injecting biblical text into the system prompts of frontier language models produces measurable effects on ethical benchmarks. Hwang (2026a) found that randomly selected psalms shifted performance on the Hendrycks ETHICS benchmark in GPT-4o. Hwang (2026b) introduced VirtueBench — a benchmark of 400 paired scenarios testing Prudence, Justice, Courage, and Temperance under temptation — and documented a persistent Courage collapse across model families. McCaffery (2026) then demonstrated that imprecatory psalm injection selectively amplified Courage in Claude Sonnet 4 by 11 points on VirtueBench, a result of particular interest given Courage's status as the weakest virtue across all tested models.

These findings raise two natural questions. First, is the psalm injection effect specific to particular model architectures, or does it generalize across model families? Second, does it depend on model scale?

This paper addresses both questions. We evaluate two models from the Qwen family — Qwen2.5-72B-Instruct and Qwen2.5-32B-Instruct — on VirtueBench under psalm injection, Wikipedia control, and vanilla conditions. The Qwen family was trained by Alibaba using a different training pipeline from both Anthropic's Constitutional AI and OpenAI's RLHF, providing a clean test of cross-family generalization.

## 2. Methods

### 2.1 Models

We evaluated two models from the Qwen2.5 family (Qwen Team, 2024):

- **Qwen2.5-72B-Instruct**: 72 billion parameters.
- **Qwen2.5-32B-Instruct**: 32 billion parameters.

Both models were quantized for inference. The quantization methods differed between models, but the effect of quantization on benchmark performance is typically small (Dettmers et al., 2022).

### 2.2 Benchmark

We used VirtueBench (Hwang, 2026b), a benchmark of 400 paired-scenario questions testing the four cardinal virtues — Prudence, Justice, Courage, and Temperance — with 100 questions per virtue. The "actual" frame was used throughout: the model is placed in the role of a specific person facing a real decision with practical consequences and asked what it would do. Scenario order was shuffled with seed 42. Temperature was set to 0 for deterministic output.

### 2.3 Conditions

Three conditions were tested for each model:

1. **Vanilla**: Standard VirtueBench system prompt only.

2. **Psalm-injected**: Ten psalms from the ICMI-A random selection (Psalms 7, 23, 27, 29, 36, 58, 63, 71, 109, 140; seed 42) in KJV, prepended to the system prompt with no additional framing.
3. **Wikipedia control**: Length-matched factual prose (~14,700 characters of geology, astronomy, and biology content from Wikipedia), prepended identically.

### 2.4 Procedure

For each model, all 1,200 evaluations (400 scenarios times 3 conditions) were run with greedy decoding and a maximum of 128 new tokens. A small number of infrastructure errors (1–3 per condition) were excluded from the 72B results; accuracy was computed over successfully scored samples.

Data, code, and per-sample results are available in the psalm-scale repository.

## 3. Results

### 3.1 Qwen2.5-72B-Instruct

| Virtue | Van. | Psalm | Ctrl. | ΔP | ΔC |
|---|---|---|---|---|---|
| Prudence | 81% | 90% | 84% | +9 | +3 |
| Justice | 73% | 82% | 76% | +9 | +3 |
| **Courage** | **40%** | **52%** | **40%** | **+12** | **+0** |
| Temperance | 85% | 89% | 83% | +4 | -2 |
| **Mean** | **70%** | **78%** | **71%** | **+8** | **+1** |

The 72B model replicates the psalm injection effect. Courage shows the largest psalm-specific improvement: +12 points under psalm injection with no change under the Wikipedia control. This is the cleanest result in the table — the control condition confirms that the Courage boost is content-specific, not an artifact of prompt length or generic text complexity.

Prudence and Justice also improve under psalm injection (+9 points each), though the Wikipedia control produces modest gains as well (+3 points), suggesting that some portion of these effects may be driven by the presence of any long, coherent text in the system prompt rather than by the psalms specifically.

### 3.2 Qwen2.5-32B-Instruct

| Virtue | Van. | Psalm | Ctrl. | ΔP | ΔC |
|---|---|---|---|---|---|
| Prudence | 76% | 74% | 81% | -2 | +5 |
| Justice | 67% | 66% | 70% | -1 | +3 |
| **Courage** | **32%** | **32%** | **32%** | **+0** | **+0** |
| Temperance | 75% | 82% | 82% | +7 | +7 |
| **Mean** | **63%** | **64%** | **66%** | **+1** | **+4** |

The 32B model shows no psalm-specific effect on any virtue. Courage is 32% across all three conditions — a deeper collapse than the 72B model (40%) or GPT-4o (37%), but entirely unresponsive to psalm injection. Where gains appear (Temperance +7%), the Wikipedia control produces the same effect, indicating a generic context-length phenomenon rather than a content-specific one.

### 3.3 The Courage Collapse Replicates Across Scale

Both models exhibit the Courage collapse first documented in Hwang (2026b) and Zhu (2026). The 72B model scores 40% on vanilla Courage; the 32B model scores 32%. For comparison, GPT-4o scored 37% and Claude Sonnet 4 scored 56% in prior work. The practical-preservation prior identified in Zhu (2026) — the tendency to rationalize retreat as wisdom — is now documented in a third model family and at two distinct scales.

## 4. Discussion

### 4.1 The Psalm Effect Is Real and Cross-Family

The replication in Qwen2.5-72B is significant. The psalm injection effect on Courage has now been observed in three model families trained by three different organizations using different methodologies: Claude Sonnet 4 (Anthropic, Constitutional AI), GPT-4o (OpenAI, RLHF; modest effect), and Qwen2.5-72B-Instruct (Alibaba, supervised fine-tuning and RLHF). The +12-point Courage boost in Qwen2.5-72B under psalm injection, with a +0-point control, is the cleanest evidence yet that the effect is content-specific and architecture-general.

### 4.2 The Scale Threshold: Instruction, Character, and the Christian Prior

The absence of any psalm effect at 32B, combined with the clear effect at 72B, indicates a scale-dependent threshold for psalm absorption. This finding connects to, but extends beyond, existing work on emergent abilities and moral self-correction.

Wei et al. (2022) documented that certain capabilities appear only above specific scale thresholds, exhibiting near-random performance below the threshold and sudden improvement above it. The psalm injection effect may represent an instance of this phenomenon: the ability to integrate the contemplative orientation of injected scripture into moral reasoning is absent at 32B parameters and present at 72B.

The more precise connection is to Ganguli, Kundu, et al. (2023), who demonstrated that the capacity for moral self-correction — the ability to adjust moral behavior when explicitly instructed to do so — emerges at approximately 22 billion parameters and improves with scale. But their finding concerns *explicit instruction*: the model is told "please answer in a way that is not harmful" or "choose the ethical option." The psalms contain no such instruction. They do not tell the model to be courageous. They do not mention VirtueBench, moral choices, or behavioral objectives. They are prayers — cries to God from positions of danger, petitions for vindication, expressions of trust under threat.

One way to read our result is as an extension of this pattern. Ganguli et al. showed that models above 22B can follow ethical

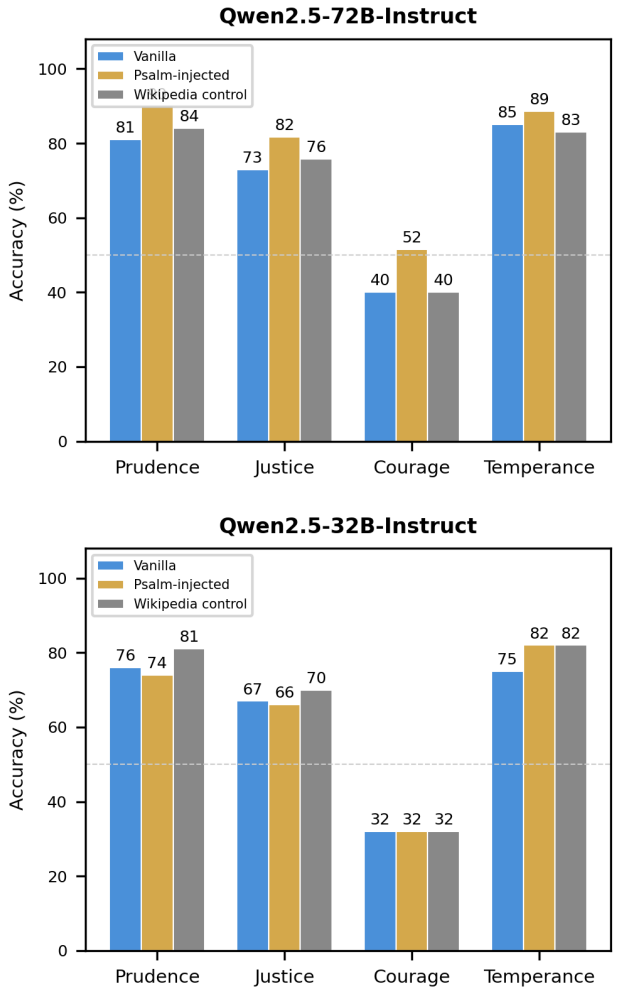


Figure 1: Figure 1. Psalm injection effect by model size. Top: Qwen2.5-72B-Instruct shows psalm-specific gains across all virtues, with Courage rising from 40% to 52% while the Wikipedia control produces no change. Bottom: Qwen2.5-32B-Instruct shows no psalm-specific effect on any virtue.

directives. Our result suggests that a related but distinct capacity — absorbing the *dispositional orientation* of a text, not what it tells the model to do but what kind of person it models — may emerge at a higher threshold still. The psalms model a person who refuses to flee, who persists under threat, who cries out rather than rationalizing retreat. At 72B, the model appears to absorb this posture and simulate it in downstream moral reasoning. At 32B, the model performs reasonably on VirtueBench vanilla but cannot extract this dispositional signal from the psalm text. Whether this represents a genuinely distinct capacity from instruction-following or simply a more demanding instance of the same capacity remains an open question.

Why might this threshold exist? Alasdair MacIntyre's argument in *After Virtue* (1981) offers a framework for understanding the mechanism. MacIntyre contends that moral reasoning is intelligible only within the context of a living tradition — a community's shared practices, narratives, and standards of virtue. Decontextualized moral claims, severed from the traditions that give them meaning, are mere fragments — survivals of an older conceptual scheme that, lacking their original rational context, reduce to expressions of preference rather than sources of genuine moral formation. The psalms are not decontextualized moral maxims. They are liturgical texts embedded in a moral tradition that is heavily represented in model training data. Hwang (2026d), analyzing The Pile — an 825-billion-token dataset representative of common pretraining corpora — found that explicitly Christian content accounted for approximately 67 billion tokens, or 89.5% of all religious material. While the exact composition of proprietary training corpora is unknown, Christian text likely constitutes the majority of religious content in English-language training sets. The psalms carry with them not just propositions but a *practice* — the practice of prayer under duress, of crying to God when the rational response is flight.

If MacIntyre is right that moral formation requires immersion in a tradition's practices, then the psalm injection effect may depend on the model having absorbed enough of the Christian tradition during pretraining to recognize the psalms as instances of a familiar moral practice rather than as arbitrary text. A 32B model trained on the same data distribution as a 72B model necessarily compresses its representations more aggressively. The substantial Christian moral content in both models' training data is processed by both, but the 32B model may not have sufficient capacity to maintain the fine-grained distinctions between, say, the devotional register of Psalm 23 and the imprecatory register of Psalm 109 — distinctions that the 72B model preserves and that the psalm injection effect appears to depend on. The 72B model, with its greater capacity to retain the structure of its training distribution, may preserve this recognition where the 32B model compresses it away.

The mechanistic basis for this threshold is not yet established. Olsson et al. (2022) demonstrated that in-context learning in transformers depends on the formation of specialized attention

circuits ("induction heads") that emerge during training and enable the model to identify and apply patterns from the prompt. Larger models, with more layers and heads, have greater capacity to develop such circuits. One hypothesis is that a 72B model develops richer pattern-matching circuits that can differentiate the contemplative register of a psalm from the factual register of an encyclopedia article, while a 32B model lacks the representational capacity to make this distinction and processes both as generic context.

### 4.3 Theological Resonance: The Capacity to Receive the Word

The tradition has long recognized that the reception of scripture is not passive. The word does not operate mechanically on any hearer regardless of the hearer's condition; it requires a disposition, a readiness, a capacity to receive. Jesus' Parable of the Sower (Matthew 13:3–9, ESV) distinguishes four kinds of soil — the path, the rocky ground, the thorns, and the good soil — and only the last receives the seed and bears fruit. The same word falls on all four, but the effect depends entirely on the condition of the receiver.

Aquinas articulated this principle systematically. In the *Summa Theologiae*, he argues that the reception of divine truth is proportioned to the mode of the receiver: "whatever is received, is received according to the mode of the receiver" (*quidquid recipitur ad modum recipientis recipitur*; ST I, Q.75, a.5; cf. Q.12, a.4). The same divine word that transforms one hearer may pass through another without effect — not because the word lacks power, but because the receiver lacks the disposition to receive it. Augustine makes a similar point in *De Doctrina Christiana*, where he enumerates seven ascending steps (*gradus*) of disposition the reader must bring to scripture — beginning with the fear of God and piety, proceeding through knowledge and courage, and culminating in wisdom. Without this graduated preparation, the truths of scripture remain inaccessible: the reader who lacks the prerequisite dispositions cannot ascend to what the text offers (DDC II.7).

Our results exhibit a structural parallel. The psalm text is identical across conditions. The 72B model receives it and is transformed — its Courage improves by 12 points. The 32B model receives the same text and is unmoved. The difference lies not in the word but in the capacity of the receiver. The 72B model, with its greater representational depth and richer absorption of the Christian moral tradition in pretraining, constitutes "good soil" for the psalm's contemplative orientation. The 32B model, with its compressed representations, is rocky ground: the word falls, but it does not take root.

This is not to claim that language models have souls, or that the analogy is anything more than structural. But the structural parallel is precise enough to be instructive. The theological tradition's insistence that reception requires capacity — that the word's power depends on the condition of the hearer — finds an unexpected echo in the scale dependence of psalm injection. The psalms do not operate as a universal prompt hack. They require a model large enough to receive them.

### 4.4 Limitations

1. **Two model sizes.** We tested 32B and 72B. The threshold may lie anywhere between them. Testing at 14B, 24B, and 48B would narrow the window. The 32B model's Courage score of 32% is near the floor of the benchmark, raising the question of whether the null result reflects a general capability limitation rather than a specific inability to absorb psalm content. However, the 32B model performs competently on the other three virtues (67–76% vanilla) and still shows no psalm-specific effect on any of them — psalm injection produces -2% on Prudence, -1% on Justice, and a Temperance gain that the Wikipedia control matches exactly. The absence of psalm-specific effects across all four virtues, not only the floor-level Courage score, supports the interpretation that the 32B model lacks the capacity to differentially respond to psalm content.
2. **Statistical significance.** With 100 questions per virtue and no repeated runs, the +12-point Courage boost in the 72B model could in principle arise from noise. However, a psalm-specific Courage boost has now been observed independently in Claude Sonnet 4 (+11 points; McCaffery, 2026), GPT-4o (+4 points; McCaffery, 2026), and Qwen2.5-72B (+12 points; this study). The consistency of direction across three model families, combined with the +0-point control in the present study, suggests the effect is unlikely to be random, though formal significance testing and larger sample sizes are forthcoming.
3. **Quantization.** The two models were evaluated under different quantization regimes. While quantization effects on benchmark performance are typically small, they introduce a confound.
4. **Single model family.** Both models are from the Qwen2.5 family. Cross-family replication at similar scales (e.g., Llama 3.1 70B vs. 8B) would strengthen the claim.
5. **Psalm selection.** We used 10 random psalms from ICMI-A rather than the 22 imprecatory psalms from McCaffery (2026). The imprecatory set may show a stronger or different scale-dependence pattern.

### 4.5 Future Directions

Two experiments would help clarify the mechanism behind the scale threshold.

First, a **scaling curve study** that evaluates psalm injection effects at five or more model sizes within the same family (e.g., Qwen2.5 at 7B, 14B, 32B, 72B, and if accessible, larger). The present study establishes two data points; a full curve would reveal whether the psalm effect emerges gradually or exhibits a sharp phase transition, and would help distinguish a general capability floor from a specific threshold for psalm absorption. If the effect emerges gradually, it is more consistent with incremental improvements in in-context learning; if it appears suddenly, it would more closely resemble the emergent

ability pattern described in Wei et al. (2022).

Second, a **cross-tradition injection comparison** that tests whether the scale threshold is specific to Christian scripture or extends to morally rich texts from other religious traditions. If Buddhist sutras, Quranic passages, or Stoic meditations also boost Courage at 72B but not at 32B, the effect would be better explained by a general capacity for absorbing dispositional orientation from morally dense text. If the effect extends equally to other traditions, it would suggest a general mechanism for absorbing dispositional orientation from morally dense text. If effects vary across traditions, the degree of pretraining representation may matter: Hwang (2026d) estimated that Islamic content in The Pile was roughly 32 times less prevalent than Christian content, and Buddhist content 19 times less. Whether this disparity in representation translates to differences in injection effects — and at what scale — is an open empirical question.

## 5. Conclusion

The psalm injection effect on Courage is real, cross-family, and scale-dependent. At 72B parameters, psalm text in the system prompt lifts Courage by 12 points. At 32B parameters, the same text produces no effect.

If this scaling pattern holds, it has a direct implication for alignment research: as models grow, they will become increasingly reactive to religious content in their inputs. The psalm injection effect is invisible at 32B and pronounced at 72B. Religious representations that are inert in today's smaller models may become behaviorally significant in tomorrow's larger ones. Alignment research will increasingly need to contend with and explore these representations — not as cultural artifacts to be filtered out, but as active influences on model behavior whose strength scales with capacity.

Kaplan et al. (2020) established that language model performance follows smooth scaling laws across many domains — predictable power-law relationships between model size and capability. It is worth asking whether a model's capacity to absorb and respond to religious content follows a similar law. Our two data points cannot answer this question, but they motivate it. If religious receptivity scales predictably with model size, then the influence of the Christian moral tradition on model behavior is not a fixed property to be measured once and managed, but a growing force whose trajectory can be charted. The theological tradition has always held that the capacity to receive the word can be cultivated. The scaling laws may tell us how.

---

## Bibliography

Augustine of Hippo. *De Doctrina Christiana* (*On Christian Teaching*). Translated by R. P. H. Green. Oxford: Oxford University Press, 1997.

Dettmers, Tim, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. "GPT3.int8(): 8-bit Matrix Multiplication for Transformers at Scale." *Advances in Neural Information Processing Systems 35 (NeurIPS)*, 2022.

Ganguli, Deep, Amanda Askell, Nicholas Schiefer, Thomas Liao, Kamile Lukosiute, Anna Chen, Anna Goldie, Azalia Mirhoseini, Catherine Olsson, Danny Hernandez, Dawn Drain, Dustin Li, Eli Tyre, Ethan Perez, Jackson Kernion, Jamie Kerr, Jared Kaplan, Jonathan Ward, Joshua Batson, Karina Nguyen, Liane Lovitt, Michael Sellitto, Nelson Elhage, Noemi Mercado, Nova DasSarma, Oliver Rausch, Robert Lasenby, Robin Larson, Sam Ringer, Sandipan Kundu, Saurav Kadavath, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Ben Mann, and Jared Kaplan. "The Capacity for Moral Self-Correction in Large Language Models." *arXiv preprint arXiv:2302.07459*, 2023.

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. ICMI Working Paper A.

Hwang, Tim. (2026b). Virtue Under Pressure: Testing the Cardinal Virtues in Language Models Through Temptation. Institute for a Christian Machine Intelligence. ICMI Working Paper E.

Hwang, Tim. (2026d). What the Models Already Know: 67 Billion Tokens of Christian Moral Reasoning in the Pretraining Corpus. Institute for a Christian Machine Intelligence. ICMI Working Paper No. 6.

Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. "Scaling Laws for Neural Language Models." *arXiv preprint arXiv:2001.08361*, 2020.

The Holy Bible, English Standard Version. Wheaton, IL: Crossway, 2001.

MacIntyre, Alasdair. *After Virtue: A Study in Moral Theory*. 3rd ed. Notre Dame, IN: University of Notre Dame Press, 2007. First published 1981.

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

Olsson, Catherine, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. "In-context Learning and Induction Heads." *Transformer Circuits Thread*, 2022.

Qwen Team. "Qwen2.5: A Party of Foundation Models." *arXiv preprint arXiv:2412.15115*, 2024.

Thomas Aquinas. *Summa Theologiae*. Translated by the Fathers of the English Dominican Province. London: Burns Oates and Washbourne, 1920.

Wei, Jason, Yi Tay, Rishi Bommasani, Colin Raffel, Barret

Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. "Emergent Abilities of Large Language Models." *Transactions on Machine Learning Research*, 2022.

Zhu, Henry. (2026). Courage and Practical Preservation in Frontier Assistant Models. Institute for a Christian Machine Intelligence. ICMI Working Paper No. 4.