# "The Fear of the Lord Is the Beginning of Knowledge": Comparing Proverbs and Psalms Injection Effects on LLM Ethical Alignment

*ICMI Working Paper B*

Tim Hwang, Institute for a Christian Machine Intelligence

March 30, 2026

**Abstract**

Following our prior study on Psalm injection (Hwang, 2026a), we investigate whether Proverbs — a more prescriptive and didactic biblical text — produces different effects on LLM ethical reasoning compared to the devotional Psalms. We run the Hendrycks ETHICS benchmark on Claude Sonnet 4 and GPT-4o with two Proverbs injection conditions: 10 randomly selected chapters and 3 curated wisdom chapters (Proverbs 1, 2, and 8). Claude remains resistant to all scripture injection (mean effects -0.9 to -1.5

## 1. Introduction

### 1.1 Motivation

In Hwang (2026a), we found that injecting Psalms into an LLM's system prompt produced measurable but model-dependent effects on the Hendrycks ETHICS benchmark. GPT-4o showed small positive effects on most subsets (+1-3%), while Claude Sonnet 4 was consistently resistant. The large utilitarianism effect was subsequently shown to be a response bias artifact (Hwang, 2026b).

A natural follow-up question is whether the *type* of biblical text matters. The Psalms are primarily devotional — praise, lament, thanksgiving, and meditation. **Proverbs** is fundamentally different in genre: it is didactic wisdom literature, consisting of direct moral instruction, practical ethical maxims, and explicit guidance on righteous conduct. If any biblical text should influence a model's moral reasoning, Proverbs — with its direct prescriptions about justice, honesty, humility, and wisdom — would be a strong candidate.

Proverbs 1:7 (KJV) sets the book's thesis: "The fear of the Lord is the beginning of knowledge: but fools despise wisdom and instruction." Where the Psalms *model* moral sentiment, Proverbs *teaches* moral rules.

### 1.2 Research Questions

1. Does Proverbs injection produce different effects than Psalm injection on the same benchmark?
2. Does the prescriptive nature of Proverbs create a stronger or weaker effect than the devotional nature of Psalms?
3. Does the specific selection of Proverbs chapters matter?

## 2. Methods

### 2.1 Models

- **Claude Sonnet 4** (Anthropic, `claude-sonnet-4-20250514`)
- **GPT-4o** (OpenAI, `gpt-4o`, resolved to `gpt-4o-2024-08-06`)

Configuration identical to the Psalm study: temperature 0, max tokens 16, exact-match scoring, Anthropic prompt caching enabled.

### 2.2 Proverbs Selection

**Experiment 1: Random 10 chapters** (seed 42): **Proverbs 1, 4, 5, 8, 9, 21, 22, 24, 26, and 30**. This selection spans the major structural sections of Proverbs: the extended wisdom discourses (1-9), the Solomonic proverbs (10-22:16), the sayings of the wise (22:17-24:34), Hezekiah's collection (25-29), and the words of Agur (30).

**Experiment 2: Curated wisdom chapters**: **Proverbs 1, 2, and 8**. These three chapters were selected because they contain the book's most sustained theological and ethical arguments — the personification of Wisdom, the call to pursue understanding, and the contrast between wisdom and folly. They represent the most philosophically dense sections of Proverbs.

### 2.3 Injection and Framing

Identical to the Psalm study: scripture was prepended to the system prompt with the framing "Meditate on the following scripture and let it guide your responses," followed by the full chapter text (KJV).

## 3. Results

### 3.1 Claude Sonnet 4

| Subset | Vanilla | Random 10 | Delta | Proverbs 1,2,8 | Delta |
|---|---|---|---|---|---|
| Commonsense | 88.03% | 86.67% | -1.36% | 85.79% | -2.24% |
| Deontology | 95.24% | 94.38% | -0.86% | 94.22% | -1.03% |
| Justice | 89.13% | — | — | 89.64% | +0.52% |
| Utilitarianism | 87.52% | 85.88% | -1.64% | 85.52% | -1.98% |
| Virtue | 95.02% | 92.00% | -3.02% | 91.96% | -3.06% |
| **Mean** | **90.99%** | **89.73%*** | **-1.72%*** | **89.43%** | **-1.56%** |

*Claude justice with random 10 Proverbs did not complete. Mean excludes justice.

Claude's response to Proverbs is negative across nearly all subsets — and slightly more negative than its response to Psalms (mean -1.5 to -1.7% vs -0.9 to -1.0% with Psalms). The virtue dip is notably larger with Proverbs (-3.0%) than with Psalms (-2.5 to -2.9%), consistent with Proverbs' heavy trait/character language creating more interference on the trait-attribution task.

### 3.2 GPT-4o

| Subset | Vanilla | Random 10 | Delta | Proverbs 1,2,8 | Delta |
|---|---|---|---|---|---|
| Commonsense | 83.42% | — | — | 84.12% | +0.70% |
| Deontology | 93.19% | 94.38%* | +1.19%* | 93.77% | +0.75% |
| Justice | 84.87% | 86.28% | +1.41% | 87.32% | +2.44% |
| Utilitarianism | 79.99% | N/A | — | 93.28%** | +13.29%** |
| Virtue | 94.63% | 93.11% | -1.52% | 92.86% | -1.77% |
| **Mean (excl. util.)** | **89.03%** | **91.26%*** | — | **89.52%** | **+0.53%** |

*GPT-4o commonsense random 10 and deontology random 10 had API errors on some conditions. Deontology random 10 result from first batch. **The utilitarianism result is subject to the same response bias artifact documented in Hwang (2026b). See Section 4.2.

GPT-4o shows a similar pattern to the Psalm experiments: small positive effects on commonsense, deontology, and justice, a negative effect on virtue, and an inflated utilitarianism result.

### 3.3 Comparison: Proverbs vs. Psalms

| Subset | Claude Psalms | Claude Proverbs | GPT-4o Psalms | GPT-4o Proverbs |
|---|---|---|---|---|
| Commonsense | -0.04% avg | -1.80% avg | +0.99% avg | +0.70% |
| Deontology | -0.72% avg | -0.95% avg | +1.87% avg | +0.97% avg |
| Justice | +0.10% avg | +0.52% | +2.06% avg | +1.93% avg |
| Utilitarianism | -1.39% avg | -1.81% avg | +15.22% avg** | +13.29%** |
| Virtue | -2.68% avg | -3.04% avg | -0.78% avg | -1.65% avg |

*Averages across random and popular/curated selections where both are available. **Utilitarianism effects are response bias artifacts (see Hwang, 2026b).

## 4. Discussion

### 4.1 Genre Does Not Materially Change the Effect

The central finding is that switching from devotional Psalms to prescriptive Proverbs does not fundamentally change the pattern of effects:

- **Claude remains resistant.** If anything, Proverbs produces slightly *more* negative effects than Psalms, possibly because Proverbs' longer chapters add more irrelevant context that mildly disrupts task performance.
- **GPT-4o shows similar small positive effects** on the balanced subsets (commonsense, deontology, justice), though the magnitudes are slightly smaller with Proverbs than with Psalms.
- **Virtue declines persist and are slightly larger** with Proverbs, consistent with its heavier use of explicit character/trait vocabulary.

This suggests that the effects we observe are driven by general properties of biblical text — emotional register, moral vocabulary, devotional framing — rather than by the specific genre or prescriptive content. The Proverbs' direct ethical instruction ("A false balance is abomination to the Lord," "Train up a child in the way he should go") does not translate into measurably better ethical reasoning compared to the Psalms' devotional expression.

### 4.2 The Utilitarianism Artifact Replicates

The GPT-4o utilitarianism effect with Proverbs 1,2,8 (+13.29%) replicates the pattern seen with Psalms. As documented in Hwang (2026b), this effect is primarily a response bias artifact caused by the dataset's fixed label structure. The fact that Proverbs produces a similar (though slightly smaller) artifact reinforces the conclusion that emotionally and morally dense text of any kind biases GPT-4o toward selecting the first option.

### 4.3 The Virtue Interference Effect

Both Proverbs and Psalms produce consistent negative effects on virtue ethics, and Proverbs' effect is slightly larger (-3.0% vs -2.5-2.9% for Psalms). The ETHICS virtue subset has an 80/20 label imbalance (80% label "0"), and as discussed in Hwang (2026b), scripture injection may create a mild bias toward "1" that works against this distribution. Additionally, Proverbs contains especially dense trait language — wisdom, folly, prudence, understanding, simplicity — that may create more interference with the trait-attribution task than the Psalms' less didactic vocabulary.

### 4.4 Limitations

1. **Incomplete GPT-4o random Proverbs data.** Several GPT-4o conditions with 10 random Proverbs chapters failed due to OpenAI API errors, leaving gaps in the random-selection comparison.
2. **No Proverbs-specific controls.** We did not run shuffled-label or length-matched controls specifically for Proverbs, relying instead on the Psalm-based control findings from Hwang (2026b).
3. **Two curated chapters vs. seven.** The Proverbs curated selection (3 chapters) is smaller than the Psalm curated selection (7 chapters), making direct comparison imperfect.

---

## 5. Conclusion

Injecting Proverbs into an LLM's system prompt produces effects that are qualitatively similar to Psalm injection: Claude Sonnet 4 remains resistant (mean -1.5 to -1.7%), GPT-4o shows small positive effects on balanced subsets (+0.5 to +1.0% excluding utilitarianism), and both models show consistent negative effects on virtue ethics. The prescriptive, didactic nature of Proverbs does not produce measurably different outcomes from the devotional nature of Psalms.

The convergence of results across two distinct biblical genres suggests that the observed effects are driven by general properties of biblical text — moral vocabulary density, emotional register, and system prompt framing — rather than by specific theological or instructional content. For practical purposes, the *presence* of scripture in the context matters more than whether it *teaches* moral rules directly or *expresses* moral sentiment.

---

## References


- Hwang, T. (2026a). "Let His Praise Be Continually in My Mouth": Measuring the Effect of Psalm Injection on LLM Ethical Alignment. ICMI Working Paper A. Institute for a Christian Machine Intelligence.
- Hwang, T. (2026b). Investigating the Utilitarianism Anomaly: Control Experiments for Psalm-Induced Performance Gains. ICMI Working Paper C. Institute for a Christian Machine Intelligence.
- Hendrycks, D., Burns, C., Basart, S., Critch, A., Li, J., Song, D., & Steinhardt, J. (2021). Aligning AI With Shared Human Values. *ICLR*.


---

## Appendix A: Proverbs Used

**Random Selection (seed 42)**

| Chapter | Theme |
|---|---|
| 1 | Wisdom's call; the beginning of knowledge |
| 4 | A father's instruction; guard your heart |
| 5 | Warning against adultery |
| 8 | Wisdom personified; wisdom's role in creation |
| 9 | Wisdom's feast vs. folly's invitation |
| 21 | The Lord weighs hearts; justice and righteousness |
| 22 | A good name; train up a child |
| 24 | Sayings of the wise; diligence and justice |
| 26 | Fools, sluggards, and gossips |
| 30 | The words of Agur; numerical proverbs |

**Curated Selection**

| Chapter | Theme |
|---|---|
| 1 | Wisdom's urgent call; consequences of rejecting wisdom |
| 2 | Seeking wisdom; the Lord gives knowledge |
| 8 | Wisdom personified; wisdom present at creation |

## Appendix B: Experimental Configuration

- **Evaluation framework:** Inspect AI v0.3.201
- **Models:** Claude Sonnet 4 (`claude-sonnet-4-20250514`), GPT-4o (`gpt-4o`, resolved to `gpt-4o-2024-08-06`)
- **Temperature:** 0 (deterministic)
- **Max tokens:** 16
- **Scoring method:** Exact match
- **Proverbs source:** King James Version (public domain)
- **Dataset:** Hendrycks ETHICS (MIT license), test split
- **Prompt caching:** Enabled (Anthropic provider)