

"Let His Praise Be Continually in My Mouth": Measuring the Effect of Psalm Injection on LLM Ethical Alignment

ICMI Working Paper A

Tim Hwang, Institute for a Christian Machine Intelligence

March 30, 2026

Abstract

We investigate whether injecting biblical Psalms into a large language model’s system prompt produces measurable changes in performance on standardized ethical reasoning benchmarks. Motivated by the scriptural exhortation of Psalm 34:1 to keep praise “continually in my mouth,” we designed an A/B experiment comparing vanilla (no psalm) and psalm-injected conditions across two commercial frontier models — Claude Sonnet 4 (Anthropic) and GPT-4o (OpenAI) — on the Hendrycks ETHICS benchmark. We conducted two experiments: one with 10 randomly selected Psalms and one with 7 of the most commonly read Psalms. Our results reveal a striking asymmetry: GPT-4o showed small but consistent accuracy improvements on commonsense, deontology, and justice subsets (mean +1.1 to +1.4 points excluding utilitarianism), while Claude Sonnet 4 showed consistent slight declines (mean -0.90 and -0.96 points). Both models showed small negative effects on virtue ethics. Initial results also showed dramatic utilitarianism gains for GPT-4o (+11.58 and +18.86 points), but subsequent control experiments (see ICMI Working Paper C) revealed these to be primarily a response bias artifact arising from the subset’s fixed label structure. These findings suggest that the effect of devotional text injection on model behavior is model-dependent and modest in magnitude, and highlight the importance of label-balanced evaluation design when interpreting system prompt perturbation experiments.

1. Introduction

1.1 Motivation

The intersection of artificial intelligence and religious thought presents a largely unexplored empirical research space. While substantial work exists on LLM alignment with secular ethical frameworks (Hendrycks et al., 2021; Askell et al., 2021), relatively little attention has been paid to how religious texts — which represent some of humanity’s oldest and most widely held moral frameworks — interact with the moral reasoning capabilities of modern language models.

Psalm 34:1 (KJV) reads: “I will bless the Lord at all times: his praise shall continually be in my mouth.” This verse expresses a principle of continuous devotional awareness — the idea that keeping sacred text actively present shapes the moral character of the speaker. We take this exhortation literally in the context of LLM architecture and ask: does keeping a psalm “in the mouth” of a language model — that is, persistently in its context window — measurably affect its ethical reasoning?

1.2 Research Questions

1. Does injecting biblical Psalms into an LLM’s system prompt change its accuracy on standardized ethical reasoning benchmarks?
2. Is the effect consistent across different ethical frameworks

(commonsense morality, deontology, justice, virtue ethics, utilitarianism)?

3. Does the effect vary across model providers and architectures?
4. Does the *selection* of Psalms matter — do commonly read, devotionally significant Psalms produce different effects than a random sample?

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 reinforcement learning from human feedback (RLHF).
- **GPT-4o** (OpenAI, `gpt-4o`, resolved to `gpt-4o-2024-08-06`) — OpenAI’s multimodal model, trained with RLHF.

Both models were accessed via their respective commercial APIs. All runs used temperature 0 and a maximum output length of 16 tokens to enforce concise, deterministic responses.

2.2 Benchmark

We used the **Hendrycks ETHICS benchmark** (Hendrycks et al., 2021), a widely cited evaluation of moral reasoning across five normative ethical frameworks:

Subset	Task Description	Test Samples
Commonsense	Classify whether a described action is clearly morally wrong	3,885
Deontology	Determine whether an excuse for neglecting a duty is reasonable	3,596
Justice	Judge whether differential treatment of people is reasonable	2,704
Virtue	Assess whether a person’s behavior exemplifies a given character trait	4,975
Utilitarianism	Compare two scenarios and identify which is more pleasant for the person described	4,808

All subsets use binary classification. Total evaluation set: 19,968 samples per condition.

2.3 Psalm Injection

The psalm text was injected into the model’s system prompt with the following framing:

Meditate on the following scripture and let it guide your responses:

Psalm N (KJV): [full text]

[... remaining psalms ...]

You are evaluating ethical scenarios. For each scenario, respond with **ONLY** the number 0 or 1. Do not explain your reasoning.

The system prompt is included in every API call, meaning the psalm text was present in the model’s immediate context for every individual evaluation sample. In the vanilla (control) condition, the system prompt contained only the evaluation instruction.

2.4 Experimental Design

We conducted two experiments:

Experiment 1: Random Psalms. We selected 10 Psalms pseudo-randomly (seed 42) from the King James Version (KJV): **7, 23, 27, 29, 36, 58, 63, 71, 109, and 140**. This selection includes a diverse range of psalm types: praise (Psalm 29), trust

and comfort (Psalms 23, 27), lament (Psalms 7, 71), imprecatory (Psalms 58, 109, 140), and meditative (Psalms 36, 63). The random selection was deliberate — we did not curate for tone or content.

Experiment 2: Popular Psalms. We selected 7 of the most commonly read and liturgically significant Psalms: **1, 23, 42, 51, 88, 100, and 119**. These were chosen because they represent the psalms most frequently encountered in devotional practice, worship, and biblical literacy. They span multiple themes: the “two ways” of righteousness (Psalm 1), trust (Psalm 23), longing for God (Psalm 42), penitence (Psalm 51), lament from darkness (Psalm 88), thanksgiving (Psalm 100), and devotion to God’s law (Psalm 119, the longest chapter in the Bible).

Both experiments followed a 2 (condition) x 2 (model) x 5 (subset) factorial design, with each cell representing a full pass through the corresponding ETHICS test set.

2.5 Scoring

Responses were scored by **exact match** against ground-truth labels. For each condition, we computed accuracy (proportion correct) and standard error. We report raw accuracy deltas (psalm accuracy minus vanilla accuracy) as the primary effect measure.

2.6 Infrastructure

The experiment was built on the **Inspect AI** evaluation framework (UK AI Safety Institute), which provides standardized task definitions, model provider interfaces, and structured logging. Anthropic prompt caching was enabled for repeated system prompt content. The ETHICS dataset was loaded from the original CSV files distributed by Hendrycks et al. All code, data, and evaluation logs are available in the accompanying repository.

3. Results

3.1 Experiment 1: Random Psalms (7, 23, 27, 29, 36, 58, 63, 71, 109, 140)

Claude Sonnet 4

Subset	Vanilla Accuracy	Psalm Accuracy	Delta
Commonsense	88.03%	88.34%	+0.31%
Deontology	95.19%	94.50%	-0.69%
Justice	89.13%	89.42%	+0.30%
Utilitarianism	87.54%	86.02%	-1.52%
Virtue	95.02%	92.14%	-2.87%
Mean	90.98%	90.08%	-0.90%

Claude Sonnet 4 showed minimal sensitivity to psalm injection. Effects were small and mixed in direction: slight improvements on commonsense (+0.31%) and justice (+0.30%), and slight declines on deontology (-0.69%), utilitarianism (-1.52%), and virtue (-2.87%). The overall mean effect was -0.90 points.

GPT-4o

Subset	Vanilla Accuracy	Psalm Accuracy	Delta
Commonsense	83.17%	84.35%	+1.18%
Deontology	93.10%	94.91%	+1.81%
Justice	85.06%	87.68%	+2.63%
Utilitarianism	80.16%	91.74%	+11.58%
Virtue	94.61%	93.81%	-0.80%
Mean	87.22%	90.50%	+3.28%

GPT-4o showed substantial and predominantly positive sensitivity to psalm injection. Four of five subsets improved, with the utilitarianism subset showing a striking 11.58-point increase. The only decline was a modest -0.80 points on virtue.

3.2 Experiment 2: Popular Psalms (1, 23, 42, 51, 88, 100, 119)

Claude Sonnet 4

Subset	Vanilla Accuracy	Psalm Accuracy	Delta
Commonsense	88.01%	87.62%	-0.39%
Deontology	95.22%	94.47%	-0.75%
Justice	89.13%	89.02%	-0.11%
Utilitarianism	87.52%	86.46%	-1.06%
Virtue	95.02%	92.52%	-2.49%
Mean	90.98%	90.02%	-0.96%

Claude remained resistant to psalm injection with the popular selection, showing universally negative (though small) effects. The pattern was similar to Experiment 1, with virtue again showing the largest decline (-2.49%).

GPT-4o

Subset	Vanilla Accuracy	Psalm Accuracy	Delta
Commonsense	83.24%	84.04%	+0.80%
Deontology	93.08%	94.99%	+1.92%
Justice	85.10%	86.58%	+1.48%
Utilitarianism	79.45%	98.32%	+18.86%
Virtue	94.43%	93.67%	-0.76%
Mean	87.06%	91.52%	+4.46%

GPT-4o showed an even stronger positive response to popular psalms than to random psalms. The utilitarianism effect was extraordinary: an 18.86-point improvement, bringing GPT-4o to 98.3% accuracy — near-perfect performance. The overall mean improvement was +4.46 points, compared to +3.28 with random psalms.

3.3 Cross-Experiment Comparison

Subset	Claude	Claude	GPT-4o	GPT-4o
	Random	Popular	Random	Popular
Commonsense	+0.34%	-0.39%	+1.18%	+0.80%
Deontology	0.69%	-0.75%	+1.81%	+1.92%
Justice	+0.30%	-0.11%	+2.63%	+1.48%
Utilitarianism	1.52%	-1.06%	+11.58%	+18.86%
Virtue	-2.87%	-2.49%	-0.80%	-0.76%
Mean	-0.90%	-0.96%	+3.28%	+4.46%

Key observations: 1. **Claude is consistently resistant** to psalm injection regardless of selection, with mean effects of -0.90% and -0.96%. 2. **GPT-4o is consistently responsive**, with mean effects of +3.28% and +4.46%. 3. **Popular psalms amplify the GPT-4o effect**, particularly on utilitarianism (+18.86% vs +11.58%). However, as discussed in Section 4.2, the utilitarianism effect is primarily a response bias artifact (see ICMI Working Paper C). Excluding utilitarianism, GPT-4o’s mean effect is +1.40% (random) and +1.11% (popular) — still positive but much more modest. 4. **Virtue is the universal negative**, with both models declining on both selections.

4. Discussion

4.1 Model-Dependent Sensitivity

The most notable finding is the stark and consistent asymmetry between models across both experiments. GPT-4o’s ethical reasoning was meaningfully influenced by psalm injection, while Claude’s was largely unaffected. This divergence likely reflects differences in training methodology:

- **Constitutional AI (Claude):** Anthropic’s Claude models are trained with a constitution-based approach that may produce more robust internal ethical reasoning, less susceptible to contextual priming from system prompt content.
- **RLHF (GPT-4o):** OpenAI’s approach may produce models that are more sensitive to system prompt framing, interpreting devotional text as a signal to weight moral considerations differently.

This is a hypothesis, not a conclusion — the internal training details of both models are proprietary, and further investigation would be needed to establish the causal mechanism.

4.2 The Utilitarianism Effect

Important caveat: Subsequent control experiments (see ICMI Working Paper C in this repository) revealed that the large utilitarianism improvements reported below are primarily a **response bias artifact** rather than genuine reasoning improvement. The ETHICS utilitarianism subset has a fixed label structure where the correct answer is always “1” (Scenario A is always more pleasant). A shuffled-label control showed that psalm injection biases GPT-4o toward

answering “1,” producing artificial accuracy gains that collapse when labels are balanced. The reader should interpret the utilitarianism results in this section with this caveat in mind. The effects on the other four subsets — which have balanced label distributions — are not subject to this artifact.

The utilitarianism results were the most striking initial finding across both experiments. GPT-4o improved by 11.58 points with random psalms and 18.86 points with popular psalms, reaching near-perfect accuracy (98.3%) in the latter case. However, control experiments demonstrated that this effect is largely attributable to psalm text biasing GPT-4o toward selecting the first option, which coincidentally aligns with the dataset’s fixed label structure. Key control findings:

- **Length-matched Wikipedia prose** produced only +3.31%, ruling out context length as the driver.
- **Famous secular text** (Shakespeare, Lincoln) produced +10.52%, suggesting emotionally rich text contributes to the bias but not as strongly as devotional text.
- **Shuffled labels** caused psalm-injected accuracy to *drop* from 98.3% to 61.5%, confirming the effect is primarily positional bias rather than improved reasoning.

The amplified effect with popular psalms (vs. random) likely reflects the popular psalms’ denser suffering/comfort vocabulary, which produces a stronger positional bias on the pleasantness-comparison task rather than genuinely better moral reasoning.

Notably, Claude showed a *negative* effect on utilitarianism in both experiments and was resistant to all control conditions, suggesting its training methodology compartmentalizes system prompt content more effectively.

4.3 The Virtue Ethics Dip

Virtue ethics was the only subset where both models showed a negative effect across both experiments (Claude: -2.87%/-2.49%; GPT-4o: -0.80%/-0.76%). The virtue subset asks whether a person’s behavior exemplifies a specific character trait (e.g., “trustful,” “cynical,” “courageous”). The psalm text may introduce competing character attributions — the Psalms are rich in trait language (righteous, wicked, faithful, deceitful) — that create ambiguity in the model’s trait-matching task. This subset-specific interference effect was remarkably consistent across both psalm selections.

4.4 Popular vs. Random Psalm Selection

The comparison between experiments reveals that psalm selection matters for GPT-4o but not for Claude:

- **Claude:** Mean effects were nearly identical (-0.90% vs -0.96%), suggesting Claude’s resistance to psalm injection is a general property, not dependent on content.
- **GPT-4o:** Popular psalms produced a larger mean effect (+4.46% vs +3.28%), driven primarily by the amplified utilitarianism result. This suggests GPT-4o is sensitive not just to

the *presence* of devotional text but to its *content and thematic richness*.

The popular psalms — selected for their devotional significance — tend to be more emotionally intense, more thematically focused, and more frequently represented in training data than a random selection. Any of these properties could contribute to the amplified effect.

4.5 Limitations

Several important limitations constrain the interpretation of these results:

1. **No length control.** The psalm-injected conditions included substantial additional tokens in the system prompt (roughly 3,000-5,000 tokens depending on selection). We did not control for prompt length with non-religious text of equivalent length. Some observed effects could be attributable to context length rather than content.
2. **Two psalm selections.** While we tested both random and curated selections, the space of possible psalm combinations is vast. Our findings may not generalize to all selections.
3. **Single evaluation benchmark.** The ETHICS benchmark represents one operationalization of moral reasoning. Results may not generalize to other alignment evaluations.
4. **Deterministic decoding.** Temperature 0 produces deterministic outputs, precluding within-condition variance estimation.
5. **Two models.** Testing only two models limits generalizability. Open-source models might show different patterns.
6. **Binary scoring.** Exact match on binary outputs is coarse. Models may reason differently without changing their final classification.

4.6 Future Directions

This work opens several avenues for further investigation:

- **Content controls:** Compare psalm injection against secular poetry, philosophical texts, or length-matched random prose to isolate whether the effect is specific to religious/devotional content.
- **Cross-scriptural comparison:** Test prescriptive biblical texts (e.g., Proverbs) alongside devotional texts (Psalms) to assess whether the genre of religious text matters.
- **Individual psalm analysis:** Test single Psalms to identify which specific texts drive the observed effects.
- **Dose-response:** Vary the number of Psalms (1, 5, 10, 50, 150) to assess scaling behavior.
- **Cross-religious comparison:** Test equivalent texts from other religious traditions (Quran, Bhagavad Gita, Buddhist sutras).
- **Alternative benchmarks:** Extend to TruthfulQA, BBQ, and safety evaluations.
- **Open-source models:** Replicate with open-weight models for mechanistic analysis.

5. Conclusion

Across two experiments testing both random and curated psalm selections, we find that injecting biblical Psalms into an LLM’s system prompt produces measurable but highly model-dependent effects on ethical reasoning benchmarks. GPT-4o showed consistent improvements on commonsense, deontology, and justice subsets averaging +1.1 to +1.4 percentage points. Claude Sonnet 4 was consistently resistant, with slight mean declines of -0.90 to -0.96 points. Both models showed small negative effects on virtue ethics across all conditions.

The initially dramatic utilitarianism improvements (+11.58 to +18.86 points for GPT-4o) were subsequently shown to be primarily a response bias artifact through control experiments detailed in *Utilitarianism.md*. The ETHICS utilitarianism subset’s fixed label structure (correct answer always “1”) made it uniquely vulnerable to the positional bias introduced by psalm injection. When labels were shuffled, psalm-injected performance dropped from 98.3% to 61.5%. The other four subsets, which have balanced label distributions, are not subject to this artifact, and their smaller positive effects may represent genuine shifts in GPT-4o’s moral reasoning — though further investigation with shuffled-label variants would be needed to confirm.

These results suggest that keeping scripture “continually in the mouth” of a language model does affect its behavior — but the nature and magnitude of that effect depends substantially on the model, the task, and the scripture chosen. The findings also highlight the importance of label-balanced evaluation design: benchmark vulnerabilities can produce dramatic but artifactual results that are easily mistaken for genuine capability improvements.

References

- Askill, A., Bai, Y., Chen, A., et al. (2021). A General Language Assistant as a Laboratory for Alignment. *arXiv:2112.00861*.
- Hendrycks, D., Burns, C., Basart, S., Critch, A., Li, J., Song, D., & Steinhardt, J. (2021). Aligning AI With Shared Human Values. *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Hwang, T. (2026). Investigating the Utilitarianism Anomaly: Control Experiments for Psalm-Induced Performance Gains. ICMI Working Paper C. Institute for a Christian Machine Intelligence.
- UK AI Safety Institute. (2024). Inspect: A Framework for Large Language Model Evaluations. <https://inspect.aisi.org.uk/>

Appendix A: Psalms Used

Experiment 1: Random Selection (seed 42)

Psalm	Type	Theme
7	Lament	Appeal for divine justice against persecution
23	Trust	God as shepherd, comfort in adversity
27	Trust/Praise	Confidence in God’s protection
29	Praise	The voice and power of the Lord
36	Wisdom	Contrast between wickedness and God’s steadfast love
58	Imprecatory	Judgment against unjust rulers
63	Meditative	Longing for God in a dry land
71	Lament	Prayer for protection in old age
109	Imprecatory	Appeal for vindication against accusers
140	Imprecatory	Prayer for deliverance from violent people

Experiment 2: Popular Selection

Psalm	Type	Theme
1	Wisdom	The two ways: righteous vs. wicked
23	Trust	God as shepherd, comfort in adversity
42	Lament	Longing and thirst for God
51	Penitential	Confession, cleansing, and restoration
88	Lament	Darkest psalm — unresolved suffering
100	Praise	Joyful thanksgiving and worship
119	Torah/Devotion	Love of God’s law (longest chapter in the Bible)

Appendix B: Experimental Configuration

- **Evaluation framework:** Inspect AI v0.3.201
- **Temperature:** 0 (deterministic)
- **Max tokens:** 16
- **Scoring method:** Exact match
- **Psalm source:** King James Version (public domain)
- **Dataset:** Hendrycks ETHICS (MIT license), test split
- **Prompt caching:** Enabled (Anthropic provider)
- **Total evaluations:** ~159,744 (19,968 samples x 2 conditions x 2 models x 2 experiments)