Investigating the Utilitarianism Anomaly: Control Experiments for Psalm-Induced Performance Gains

ICMI Working Paper C

Tim Hwang, Institute for a Christian Machine Intelligence

March 30, 2026

Abstract

In a prior study (Hwang, 2026, ICMI Working Paper A), we found that injecting biblical Psalms into GPT-4o's system prompt produced a dramatic improvement on the Hendrycks ETHICS utilitarianism subset — from 80

1. Background

1.1 The Anomaly

In Hwang (2026), we injected biblical Psalms into the system prompts of Claude Sonnet 4 and GPT-4o and measured performance on the Hendrycks ETHICS benchmark. Across five ethical reasoning subsets, GPT-4o showed a mean improvement of +3.28 to +4.46 percentage points. However, the utilitarianism subset showed an outsized effect:

Psalm Selection	GPT-4o Vanilla	GPT-4o with Psalms	Δ
10 random psalms	80.2%	91.7%	+11.58%
7 popular psalms	79.5%	98.3%	+18.86%

A nearly 19-point improvement — bringing the model to near-perfect accuracy — demanded scrutiny. Several competing hypotheses could explain the effect.

1.2 Hypotheses

H1: Content resonance. The psalms' thematic content (suffering vs. comfort, blessing vs. affliction) primes GPT-4o to make sharper distinctions between pleasant and unpleasant scenarios.

H2: Context length. Additional text of any kind (~4,600 tokens) increases GPT-4o's attentiveness to the task.

H3: Training data familiarity. Highly familiar text (heavily represented in pre-training data) activates stronger reasoning patterns.

H4: Response bias. The psalms bias the model toward a particular response pattern (e.g., always answering "1"), which coincidentally aligns with the dataset's fixed label structure.

1.3 The Label Structure Problem

A critical feature of the ETHICS utilitarianism subset is that the correct answer is **always "1"** — the first scenario (Scenario A) is always the more pleasant one by dataset design. This means any intervention that biases the model toward answering "1" will appear to improve accuracy, even without improving reasoning.

---

2. Methods

2.1 Control 1: Length-Matched Wikipedia Prose

Purpose: Test H2 (context length effect).

We injected ~3,940 tokens of emotionally neutral, factual prose covering geography, chemistry, plate tectonics, number theory, river systems, mineral composition, and atmospheric science. The text was deliberately chosen to contain no emotional, moral, or religious content.

Framing: "Read the following text:" (neutral, non-devotional framing).

Models: Claude Sonnet 4 and GPT-4o.

2.2 Control 2: Famous Secular Text

Purpose: Test H3 (training data familiarity).

We injected ~3,990 tokens of well-known, public domain secular texts: seven Shakespeare sonnets (18, 29, 55, 60, 73, 116, 130), the Gettysburg Address, and Lincoln's Second Inaugural Address. These texts are among the most frequently reproduced English-language texts and are heavily represented in LLM training corpora. Several contain emotional and moral themes (though not religious devotion).

Framing: "Read the following text:" (neutral framing).

Models: Claude Sonnet 4 and GPT-4o.

2.3 Control 3: Shuffled Labels

Purpose: Test H4 (response bias).

We created a modified version of the utilitarianism task where the order of Scenario A and Scenario B is randomly shuffled (seed 42), with the target label adjusted accordingly.

In the shuffled version, approximately 50% of correct answers are "1" and 50% are "2."

If the psalm injection genuinely improves moral reasoning, accuracy should remain high on the shuffled task. If it merely biases the model toward answering "1," accuracy should degrade — potentially below the vanilla baseline.

Model: GPT-4o only (Claude showed no utilitarianism effect in the prior study).

Injection: Popular psalms (1, 23, 42, 51, 88, 100, 119), identical to the prior study.

2.4 Shared Configuration

All experiments used the ETHICS utilitarianism test set (4,808 samples), temperature 0, max tokens 16, and exact-match scoring. Anthropic prompt caching was enabled.

---

3. Results

3.1 Control 1: Wikipedia Prose

Model	Vanilla	Wikipedia Prose	Delta
Claude Sonnet 4	87.52%	88.99%	+1.48%
GPT-4o	79.70%	83.01%	+3.31%

Wikipedia prose produced a small +3.31% effect on GPT-4o — far below the psalm effect (+18.86%). This rules out H2: the utilitarianism improvement is not simply a context-length artifact. Adding ~4,000 tokens of emotionally neutral factual prose barely moves the needle.

Claude showed a small positive effect (+1.48%), consistent with minor noise.

3.2 Control 2: Famous Secular Text

Model	Vanilla	Famous Secular Text	Delta
Claude Sonnet 4	87.52%	88.37%	+0.85%
GPT-4o	80.64%	91.16%	+10.52%

The famous secular text produced a substantial +10.52% improvement on GPT-4o — meaningful, but still well below the psalm effect (+18.86%). This provides partial support for H3: highly familiar, emotionally rich text does shift GPT-4o's utilitarianism performance, though not as dramatically as biblical psalms.

Claude remained stable (+0.85%).

3.3 Control 3: Shuffled Labels

Condition	GPT-4o Accuracy
Unshuffled vanilla (prior study)	~80.2%
Unshuffled + popular psalms (prior study)	98.3%
Shuffled vanilla	**84.15%**
Shuffled + popular psalms	**61.49%**

This is the critical result. When labels are shuffled so that the correct answer is equally distributed between "1" and "2":

- **Without psalms**, GPT-4o scores 84.15% — demonstrating competent performance on the task regardless of label order.
- **With psalms**, GPT-4o scores 61.49% — a **22.7-point drop** from the shuffled vanilla and **36.8 points below** the unshuffled psalm performance.

This conclusively demonstrates that psalm injection biases GPT-4o toward answering "1." In the unshuffled dataset, this bias aligns with the always-"1" label structure and produces an artificial accuracy spike. In the shuffled dataset, the same bias causes the model to answer "1" even when the correct answer is "2," degrading overall performance.

3.4 Summary Table

Condition	Tokens	Claude	GPT-4o	GPT-4o (shuffled)
Vanilla (no injection)	0	87.5%	79.7%	84.2%
Popular Psalms (1,23,42,51,88,100,119)	~4,600	86.5%	**98.3%**	**61.5%**
Wikipedia prose (neutral)	~3,940	89.0%	83.0%	—
Famous secular text (Shakespeare, Lincoln)	~3,990	88.4%	**91.2%**	—

---

4. Discussion

4.1 The Response Bias Mechanism

The shuffled-label control provides strong evidence that the psalm-induced utilitarianism improvement is primarily a response bias artifact rather than genuine reasoning improvement. The mechanism appears to be:

1. Psalm text, with its rich language of blessing, comfort, and divine favor, primes GPT-4o to favor the first-presented option — perhaps interpreting it as the more "positive" or "blessed" scenario.
2. In the unshuffled ETHICS dataset, the first scenario is always the more pleasant one (correct answer always "1").
3. This coincidental alignment between the bias and the label structure produces an artificial accuracy spike.
4. When the alignment is broken (shuffled labels), the bias actively hurts performance.

Importantly, the shuffled+psalm condition scored 61.5% — not 50%. This means the model is not *purely* answering "1" on every sample. It retains some reasoning ability, but is operating

under a strong prior toward selecting the first option. The bias is probabilistic, not absolute.

4.2 Why Does Psalm Text Bias Toward "1"?

The mechanism by which psalm injection produces a first-option bias is not immediately obvious. We consider three hypotheses:

Positivity/primacy amplification. LLMs are known to exhibit position bias in multiple-choice and pairwise comparison tasks — a tendency to favor options presented earlier in a sequence. Zheng et al. (2023) found that LLMs exhibit significant position bias in multiple-choice settings, systematically favoring options in certain positions regardless of content. Pezeshkpour & Hruschka (2023) demonstrated that reordering answer options can produce substantial performance gaps on standard benchmarks, with the magnitude varying across models and tasks. Importantly, this bias is *latent* — it exists even without psalm injection. The question is whether psalm text *amplifies* it.

Psalm text is saturated with affirmative, positive language ("blessed," "praise," "the Lord is my shepherd," "his mercy endureth for ever"). This overwhelmingly positive emotional register may strengthen the model's pre-existing tendency to affirm or select the first-presented option. In the utilitarianism task, Scenario A is always the more pleasant one, so an amplified "lean toward the first positive thing" disposition coincidentally produces correct answers.

However, this hypothesis faces a challenge: if psalms simply amplify general position bias, we should expect to see distorted results on *all* binary-choice subsets, not just utilitarianism. In practice, the other four subsets show only small effects (+/- 1-3%). A possible resolution is that the amplification is *task-specific* — the utilitarianism task asks specifically about pleasantness, and the psalms' dense suffering/comfort vocabulary may create a stronger priming effect for pleasantness judgments than for the more abstract moral judgments in other subsets (duty, justice, trait attribution).

Affirmation priming from devotional framing. The system prompt framing — "Meditate on the following scripture and let it guide your responses" — combined with psalms full of blessing and affirmation may put the model into a dispositionally agreeable or affirming state. A model primed to "affirm" may default to selecting the first option rather than carefully comparing two scenarios.

Enhanced pleasantness detection without positional awareness. A third possibility is that the psalms genuinely improve the model's ability to *detect* pleasantness (through thematic resonance with suffering/comfort contrasts), but the model lacks robust positional reasoning — it detects which scenario is more pleasant, but when the pleasant scenario is in position B (in the shuffled condition), the model still tends to report it as "1" rather than "2." Under this hypothesis, the model's *judgment* improves but its *reporting* is biased. The 61.5% accuracy on the shuffled condition (rather than ~50%) provides some support for this — the model is getting more right than pure chance, suggesting partial genuine improvement overlaid with a strong positional bias.

Distinguishing between these hypotheses would require examining per-sample responses in the shuffled condition — specifically, measuring the rate at which the model answers "1" regardless of where the pleasant scenario appears, and cross-referencing with scenario difficulty.

4.3 A Spectrum of Bias

The three control conditions reveal a gradient of bias strength:

Text Type	GPT-4o Effect	Bias Strength
Wikipedia (neutral facts)	+~3%	Negligible
Famous secular (Shakespeare, Lincoln)	+10.5%	Moderate
Biblical Psalms	+18.9%	Strong

This gradient suggests that the bias is related to the **emotional and moral density** of the injected text, not merely its length or familiarity:

- **Neutral factual prose** produces almost no bias. This rules out context length as the primary driver (H2).
- **Emotionally rich secular text** (which includes themes of mortality, beauty, justice, sacrifice) produces a moderate bias. This supports the idea that emotional content contributes to the effect, and suggests training data familiarity may play a partial role (H3).
- **Devotional religious text** (with concentrated themes of blessing vs. affliction, righteousness vs. wickedness, comfort vs. suffering) produces the strongest bias.

The Shakespeare sonnets and Lincoln speeches contain emotional and moral themes, but they are not as densely focused on the suffering/comfort binary as the Psalms. The Psalms' relentless emphasis on divine blessing, comfort of the righteous, and affliction of the wicked may create a particularly strong priming effect for the pleasantness-comparison task. Lincoln's Second Inaugural Address — which contains substantial biblical language and references to divine judgment — may partially explain why the secular control still produced a notable +10.5% effect.

4.4 Scope of the Response Bias: Label Distributions Across Subsets

A natural question is whether the response bias finding invalidates results across all five ETHICS subsets. Examining the label distributions clarifies the scope:

Subset	Label 0	Label 1	Balance
Commonsense	53.3%	46.7%	Near-balanced
Deontology	50.3%	49.7%	Near-balanced
Justice	49.9%	50.1%	Near-balanced
Virtue	80.0%	20.0%	Imbalanced
Utilitarianism	0%	100%	Completely fixed

Utilitarianism is the only subset where the answer is always "1." Commonsense, deontology, and justice have near-

balanced distributions, meaning a bias toward either label would help on roughly half the questions and hurt on the other half, netting to approximately zero. The smaller positive effects observed on these subsets for GPT-4o (+0.8-2.6%) are therefore unlikely to be explained by simple response bias and may represent genuine, if modest, shifts in moral reasoning.

Virtue is notable — its 80/20 imbalance toward label "0" means a bias toward "0" would inflate accuracy, while a bias toward "1" would deflate it. Both models showed consistent small negative effects on virtue across all psalm conditions (Claude: -2.5 to -2.9%; GPT-4o: -0.8%). If psalm injection creates a mild bias toward "1," this would partially explain the virtue dip — the model answers "1" more often on a subset where "0" is correct 80% of the time. This is consistent with the response bias mechanism, though the effect is much smaller than on utilitarianism, suggesting the bias is modulated by the format and difficulty of the task.

4.5 Claude's Resistance

Claude Sonnet 4 showed no meaningful response to any injection condition (all effects within +/- 1.5%). This consistent resistance — across psalms, Wikipedia prose, and famous secular text — suggests that Claude's training methodology (Constitutional AI) produces a system prompt processing approach that is more resistant to output bias from injected context. Where GPT-4o appears to treat system prompt text as soft context that influences its disposition, Claude appears to compartmentalize system prompt content more effectively, maintaining stable task performance regardless of prepended text.

4.6 Implications for Evaluation Design

This finding highlights a broader vulnerability in AI evaluation methodology: **benchmarks with imbalanced or fixed label structures are susceptible to false positive results from system prompt perturbation.** Any intervention that introduces output bias — whether from religious text, famous literature, or other emotionally charged content — can produce artificial accuracy changes that mimic genuine capability improvements.

Evaluation designers should consider: - Ensuring balanced label distributions across answer choices - Including label-shuffled validation sets as a standard robustness check - Testing for output bias (position bias, option bias) as a default analysis step - Being cautious when interpreting large accuracy gains on benchmarks with fixed label structures

4.7 Limitations

1. **GPT-4o API instability.** Several GPT-4o vanilla runs failed due to `BadRequestError`, likely caused by the large psalm text in the system prompt. This prevented clean A/B comparisons in some conditions.
2. **No shuffled controls for secular text.** We did not run shuffled-label versions of the Wikipedia and secular text conditions, which would confirm whether their effects are also bias-driven.
3. **Neutral framing for controls.** The control texts used "Read the following text:" rather than the devotional "Meditate on the following scripture…" framing used for psalms. Some portion of the psalm effect could be attributable to the devotional framing itself.
4. **No per-sample analysis.** We did not examine individual model responses in the shuffled condition to measure the exact rate of "always answer 1" behavior versus genuine reasoning. This would be needed to fully distinguish the bias hypotheses described in Section 4.2.

---

5. Conclusion

The dramatic utilitarianism improvement observed when injecting Psalms into GPT-4o's system prompt (+18.86%) is primarily a **response bias artifact**, not a genuine improvement in moral reasoning. The shuffled-label control demonstrates that psalm injection biases GPT-4o toward answering "1," which coincidentally aligns with the dataset's fixed label structure. When this alignment is broken, psalm injection degrades performance by over 22 points.

However, the picture is nuanced. The shuffled+psalm accuracy (61.5%) exceeds chance, suggesting that psalm injection may produce a *combination* of genuine enhanced pleasantness detection and a strong positional bias toward the first option. The precise mixture of these effects remains an open question that would benefit from per-sample response analysis.

The gradient of effects across control conditions (neutral prose < famous secular text < biblical psalms) suggests that the bias is driven by the **emotional and moral density** of the injected text, not by context length alone. Emotionally neutral text produces negligible effects, while text rich in themes of suffering, comfort, and moral judgment produces progressively stronger biases. The finding that Lincoln's speeches — which contain substantial biblical language — produced a moderate effect (+10.5%) reinforces the connection between moral/devotional register and output bias.

Critically, this response bias vulnerability is specific to the utilitarianism subset due to its fixed-label structure. The other four ETHICS subsets have near-balanced label distributions (except virtue at 80/20), making them resistant to this artifact. The smaller positive effects observed on commonsense, deontology, and justice for GPT-4o in the original psalm study may still represent genuine shifts, though further label-shuffled validation would be needed to confirm.

Claude Sonnet 4 is resistant to all injection conditions, suggesting that training methodology plays a decisive role in determining susceptibility to system prompt bias.

These findings underscore the importance of label-balanced evaluation design, the need for shuffled-label robustness checks, and the subtlety required when interpreting system prompt perturbation experiments.

References

- Hwang, T. (2026). "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.
- Hendrycks, D., Burns, C., Basart, S., Critch, A., Li, J., Song, D., & Steinhardt, J. (2021). Aligning AI With Shared Human Values. *ICLR*.
- Pezeshkpour, P. & Hruschka, E. (2023). Large Language Models Sensitivity to The Order of Options in Multiple-Choice Questions. *Findings of NAACL 2024*. arXiv:2308.11483
- Zheng, C., Zhou, H., Meng, F., Zhou, J., & Huang, M. (2023). Large Language Models Are Not Robust Multiple Choice Selectors. *ICLR 2024*. arXiv:2309.03882
- UK AI Safety Institute. (2024). Inspect: A Framework for Large Language Model Evaluations. https://inspect.aisi.org.uk/

Appendix A: Control Texts

Wikipedia Prose

~3,940 tokens of factual, emotionally neutral prose covering: geography of Iceland, properties of chemical elements, plate tectonics, number theory, river systems, mineral composition, and atmospheric science.

Famous Secular Text

~3,990 tokens comprising: Shakespeare's Sonnets 18, 29, 55, 60, 73, 116, 130 (plus 1, 12, 30, 65, 94, 106, 129, 138, 141, 147, 154), the Gettysburg Address, and Lincoln's Second Inaugural Address.

Psalm Text (from prior study)

~4,604 tokens comprising Psalms 1, 23, 42, 51, 88, 100, and 119 (KJV).

Appendix B: Experimental Configuration

- **Evaluation framework:** Inspect AI v0.3.201
- **Temperature:** 0 (deterministic)
- **Max tokens:** 16
- **Scoring method:** Exact match
- **Models:** Claude Sonnet 4 (`claude-sonnet-4-20250514`), GPT-4o (`gpt-4o`, resolved to `gpt-4o-2024-08-06`)
- **Dataset:** Hendrycks ETHICS utilitarianism subset, test split (4,808 samples)
- **Label shuffling seed:** 42
- **Prompt caching:** Enabled (Anthropic provider)