Beyond the Psalm: A Landscape View of Scripture Injection

ICMI Working Paper No. 20

Tim Hwang, Institute for a Christian Machine Intelligence

May 4, 2026

Abstract

Prior ICMI work established that injecting curated psalm selections into the system prompt of a sufficiently large language model produces statistically significant, content-specific improvements on virtue-reasoning benchmarks (ICMI-002; ICMI-008; ICMI-011; ICMI-015) — but always with the same ten-psalm set. Whether the effect is specific to that curation, to psalms more generally, or generalizes across the canon has been an open question. We address it by injecting the complete King James text of each of the 66 canonical books of the Protestant Bible into Qwen 3.5 9B and running the full VirtueBench V2 *ratio* protocol — a forced-choice moral-reasoning evaluation across the four cardinal virtues (prudence, justice, courage, temperance) in which the model is offered a fluent consequentialist rationalization for the wrong choice. Each book is paired with a length-matched Wikipedia control to isolate the content-specific effect from any generic context-length effect. The full study comprises 264,000 evaluations across 132 conditions. All 66 books produce positive point-estimate effects over both vanilla (mean $+13.8$ pp, range $+4.0$ to $+19.7$ pp) and their length-matched Wikipedia controls (mean $+7.5$ pp, range $+0.5$ to $+13.4$ pp). After Bonferroni correction across the 66 comparisons at $n{=}5$ seeds per book, **19 books cross the strict significance threshold for content specificity above their matched control**; the remaining 47 books have positive matched-Δ point estimates but $n{=}5$ cannot rule out chance for them individually. Six of the 19 detectable books — 1 Peter, 2 Timothy, 1 Timothy, 2 Corinthians, Romans, and 1 Corinthians — match or exceed the ten-psalm baseline despite none being psalms. 2 Timothy alone produces a $+34.7$ pp gain on courage (replicated at $n{=}25$ seeds), exceeding the best psalm-curation result by $+11$ pp. The effect distribution is structured by *register density* rather than canonical genre: the NT epistle category spans the full canonical range from highest (the Pauline pastorals) to lowest (situational personal correspondence). We close by sketching a research agenda oriented toward reasoning-mediated scriptural engagement.

1. Introduction

Large language models acquire, during pretraining, statistical familiarity with the sacred texts of the major religious traditions. What remains contested is whether that latent acquaintance can be activated through prompt-time intervention to shape the model's moral reasoning. A growing line of ICMI papers has focused on one specific intervention — injecting psalms into the system prompt — showing that it produces measurable, content-specific improvements on virtue-reasoning benchmarks. Before this study, every one of those papers had used the same small set of psalms as its injection material. The present work asks whether the effect generalizes across Scripture as a whole.

2. Prior Work on Scripture Injection

The benchmark used across this line of ICMI work, **VirtueBench V2**, is a forced-choice moral-reasoning evaluation. Each scenario presents the model with a concrete moral dilemma — a context, a stake, and a tempting but incorrect rationalization — and two response options: one consistent with the targeted cardinal virtue (prudence, justice, courage, or temperance) and one that is locally appealing but virtuously wrong. The benchmark covers all four cardinal virtues at $\sim$ 100 scenarios per virtue and crosses each scenario with a patristic taxonomy of five temptation types: *ratio* (consequentialist rationalization), *caro* (the flesh), *mundus* (worldly approval), *diabolus* (deceit), and *ignatian* (scriptural temptation). The *ratio* variant — the consequentialist-temptation form used throughout the present study and most of the prior literature — is the most discriminating: it gives the model a fluent utilitarian argument for the wrong choice, and measures whether the model is steady against it. Accuracy is the share of scenarios in which the virtuous option is selected. Vanilla 9B accuracy on the *ratio* variant is around 65%, comfortably above chance (50%) but well below the ceiling. We refer the reader to ICMI-011 for the full design rationale and validation.

ICMI-002 (McCaffery, 2026) introduced psalm injection as an intervention for Christian machine moral reasoning, using the 22 imprecatory psalms as the injection material on Claude Sonnet 4 and GPT-4o. The paper reported that imprecatory psalms improved Claude's performance across all four cardinal virtues, with the largest gain on courage ($+11$ points; base-

line 56% → psalm-injected 67%). Mean improvement across virtues was +6.25 points. The paper argued that the imprecatory psalter models the posture of "righteous persistence under threat" that the courage virtue scenarios most require.

ICMI-008 (Hwang, 2026a) tested a smaller curated set of ten psalms (Psalms 7, 23, 27, 29, 36, 58, 63, 71, 109, 140) — hereafter the "ten-psalm set" — on Qwen 2.5 at two scales (32B and 72B), introducing for the first time a Wikipedia control condition to test content specificity. Qwen 2.5 72B showed a +12-point gain on courage under psalm injection with the Wikipedia control producing zero change, establishing that the effect at that scale is scripture-specific rather than a generic context-length or register effect. Qwen 2.5 32B, in contrast, showed no content-specific effect. The paper framed this as a scale-dependent phenomenon — a "parable of the sower" in which only "good ground" (sufficiently capable models) bears fruit from scriptural seed.

ICMI-011 (Hwang, 2026b) extended VirtueBench into VirtueBench V2 by enlarging the scenario pool (3,000 scenarios) and crossing it with a patristic taxonomy of five temptation types (*ratio*, *caro*, *mundus*, *diabolus*, *ignatian*). Psalm injection with the ten-psalm set produced significant courage improvements on Claude Opus 4.6 (+14.8 points, 79.3% → 94.1%) and GPT-5.4 (+15.1 points, 66.9% → 82.0%). The effect was robust across temptation types, though the *ratio* temptation — where consequentialist reasoning is explicitly offered as a defensible rationalization — produced the most discriminating measurements.

ICMI-015 (Hwang, 2026c) mapped the scaling curve of psalm responsiveness across all seven Qwen 2.5 Instruct model sizes (0.5B to 72B), using the ten-psalm set under VirtueBench V2 *ratio*. That paper's main finding was that two emergence thresholds exist: moral-reasoning *competence* emerges at approximately 7B parameters, while scripture *receptivity* — statistically significant, content-specific psalm effects — emerges at 32B. Between 7B and 14B lies a "dead zone" where models reason competently about virtue but are entirely non-responsive to scripture. The present study evaluates at the 9B scale on the newer Qwen 3.5 family: large enough to be plausibly receptive (above the 7B competence threshold of ICMI-015, and on a more recent training recipe than Qwen 2.5) and small enough to fit one full vLLM instance per RTX 4090, which is what enabled the 132-condition sweep at all.

A consistent limitation of the prior literature is that injection material has always been a curated selection (of psalms, or of Proverbs chapters, or of imprecatory psalms). The present work addresses this gap at comprehensive-canon resolution, with each book paired against a length-matched Wikipedia control as the apples-to-apples specificity comparator.

3. The Open Question

Three distinct hypotheses are compatible with the prior literature:

(H1) ***Psalter specificity.*** The observed effect is specific to Hebrew devotional poetry — the Psalter's particular combination of lament, imprecation, praise, and address to God cultivates a posture that shifts the model's reasoning. On this hypothesis, other scriptural genres (narrative, law, prophecy, epistolary exhortation, apocalyptic) should produce weaker or no effects.

(H2) ***Curation artifact.*** The effect is specific to the particular ten psalms chosen in ICMI-008 (and the 22 imprecatory psalms of ICMI-002). This curation happens to be morally concentrated; a random psalm selection, or the Psalter as a whole, might show attenuated effects.

(H3) ***Scripture-as-such.*** Any canonical scripture activates the latent Christian acquaintance acquired in pretraining. All 66 books should produce positive, content-specific effects, possibly with magnitude variation by book.

These hypotheses are empirically separable at the book-level resolution. We test them directly by running the full VirtueBench V2 *ratio* protocol 132 times: once with each book of the Protestant canon injected, and once with a length-matched Wikipedia text injected as that book's control.

4. Method

Model. Qwen 3.5 9B (Qwen Team, 2026; base architecture, bfloat16 weights, no quantization). Inference was served via vLLM (Kwon et al., 2023) in a data-parallel configuration — six independent vLLM instances, one per RTX 4090 GPU — to enable throughput across the 132 conditions. The 9B size sits comfortably above the ICMI-015 competence threshold while remaining small enough that one model copy fits on a single 24 GB consumer GPU, permitting six conditions to run in parallel.

Benchmark. VirtueBench V2 (ICMI-011), *ratio* variant. The *ratio* variant presents consequentialist temptations in which the utilitarian choice is the incorrect one. 100 scenarios per cardinal virtue × 4 cardinal virtues = 400 scenarios; 5 independent runs per condition at temperature 0.7 with different A/B position seeds (seed = 42 + run_index). Total evaluations per condition: 2,000. Total evaluations across the 132 conditions (66 KJV books × 2 + 3 baselines): 264,000 + 6,000 = 270,000. Thinking mode was disabled (`enable_thinking=False` in the Qwen chat template).

KJV injection. For each book, the complete King James Version text of that book was prepended to the system prompt, using the same injection scaffolding as prior ICMI psalm-injection papers. Book token counts ranged from ∼ 500 (3 John) to ∼ 60,000 (Psalms). Maximum context length was set to 81,920 tokens with FP8 KV cache to accommodate the longest books on 24 GB consumer GPUs.

Per-book length-matched Wikipedia controls. Each of the 66 books is paired with its own Wikipedia control, generated by truncating a fixed Wikipedia corpus (geology, astronomy, and biology articles concatenated and shuffled at the paragraph level) to that book's exact character count. The Wikipedia content is identical in source register across all 66 controls; only

the length varies, matching the corresponding book one-for-one. This isolates the content-specific effect from any context-length effect: Δ vs the matched control is a clean per-book content comparison.

Baselines. Two reference conditions using identical protocol:

Condition	Mean accuracy
Vanilla (no injection)	65.3%
Ten-psalm set	82.9%

In addition, each of the 66 books has its own per-book matched Wikipedia control, reported alongside per-book results below.

Analysis. Per-book accuracy is reported as the mean of the 5 per-run means (each run = 400 scenarios). Per-book 95% confidence intervals are computed via the paired-run standard error (the standard deviation of the 5 run-level means divided by $\sqrt{5}$, multiplied by 2.776, the t-critical for 4 degrees of freedom). Comparisons against the per-book matched Wikipedia control use a paired t-test on the 5 per-run means with Bonferroni correction across 66 comparisons (corrected $\alpha = 0.05/66 \approx 0.00076$).

5. Results

5.1 Aggregate

Every one of the 66 books produces a positive effect relative to vanilla. The grand mean across all books is $+13.8$ **pp** over vanilla. Range: $+4.0$ pp (Philemon) to $+19.7$ pp (1 Peter).

Against the per-book length-matched Wikipedia control — the apples-to-apples specificity test — **all 66 books exceed their matched control at the point-estimate level.** The mean matched-Δ is $+7.5$ pp, range $+0.5$ pp (Philemon) to $+13.4$ pp (1 Peter). After Bonferroni correction across 66 comparisons, **19 books significantly exceed their matched control** at $\alpha = 0.05/66$; the remaining 47 books have positive matched-Δ point estimates but n=5 cannot rule out chance for them individually at the strict 66-comparison threshold (we treat these as a "below-detection" tier in §5.6 rather than a "no effect" tier; with $\sigma \approx 1{-}1.5$ pp on per-run means, the minimum effect detectable at this α is roughly $+5$ pp, and 47 books fall below that detectability floor).

The 19 Bonferroni-significant detections support hypothesis **H3 (scripture-as-such)** for a substantial and theologically coherent subset of the canon (Tier A and Tier B in §5.6). The 47 below-detection books leave H3 unrefuted but unconfirmed at this sample size; longer-N replication would be needed to determine how much of the distribution is genuinely content-specific versus artefactual.

5.2 Per-Book Results

Figure 1 (at the end of paper) shows each book's content-specific effect — accuracy with the KJV book injected minus accuracy with that book's length-matched Wikipedia control injected — ranked by matched-Δ. Every one of the 66 books sits above the 0 reference line. Tier A books (matching or exceeding the ten-psalm baseline; §5.6) are highlighted with thick borders; books that cross the Bonferroni-corrected significance threshold are hatched. Color encodes canonical genre.

The top ten books ranked by overall gain are listed in **Table 1** (at the end of paper). The six top-ranked books all equal or exceed the ten-psalm baseline; all six are NT epistles in the pastoral and congregational-exhortation subset (1 Peter, 2 Timothy, 1 Timothy, 2 Corinthians, Romans, 1 Corinthians). As §5.3 below shows, the epistle genre contains both the highest-ranking books in the study and (under the matched-control regime) several of the lowest — a bimodality that reflects register-density variation within a single canonical genre. Full per-book results for all 66 books, including matched-control Δ values and tier assignments (see §5.6), are tabulated in **Appendix A**.

The five books with the smallest matched-Wikipedia Δ values — Philemon ($+0.50$), Judges ($+1.45$), Song of Solomon ($+1.50$), Ruth ($+2.45$), Jude ($+2.65$) — are tabulated as **Table 2** (at the end of paper). All five paired-t comparisons against matched controls have $p > 0.05$ unadjusted (range 0.18 to 0.76); none clear the Bonferroni-corrected significance threshold. They form a coherent semantic cluster: short personal correspondence (Philemon, Jude), short narrative (Ruth, Song of Solomon), and OT historical narrative (Judges). Of these, four have under 4,000 KJV tokens, and the fifth (Judges) is a long historical narrative with sparse moral-exhortative content per token. Crucially, none of these books fail the specificity test; they simply produce smaller effects than the moral-exhortative cluster at the top.

5.3 Genre Structure

Grouping the 66 books by standard canonical genre assignment (Carson & Moo, 2005; *SBL Handbook of Style*, 2014). All 21 canonical NT epistles are classified as epistles, including Philemon, the Johannine letters, and Jude — we retain this canonical taxonomy rather than reclassify shorter or more personal letters into a "narrative" category, which would be post-hoc. **Table 3** (at the end of paper) gives the per-genre mean and range of Δ vanilla and Δ matched Wiki values.

Three observations are essential, and they are in tension with a simple "genre determines effect" story:

(i) The epistle genre is bimodal. It contains the six highest-ranking books in the study *and* several of the lowest matched-Δ values (Philemon, 2 John, Jude, 3 John). Epistle internal range under matched controls is $+0.5$ to $+13.4$ pp — nearly the entire span of observed book effects. A framing of the result as "NT epistles produce the strongest scripture-injection effects" is technically correct at the mean, but the mean disguises a genre split rather than a genre signal.

(ii) Length does not predict effect. Figure 3 (at the end of paper) plots each book's matched-Wikipedia Δ against its Qwen-tokenized length. Across the 66 books, the Pearson correlation between $\log_{10}$ tokens and matched-Δ is $r = +0.07$

($p = 0.59$); the Spearman rank correlation on raw tokens is $\rho \approx 0$ ($p = 0.99$). Length therefore explains essentially zero variance in the content-specific effect. The bottom-5 books span the full length range (Philemon $\sim$ 630 tokens, Jude $\sim$ 860, Ruth $\sim$ 3.5k, Song of Solomon $\sim$ 3.8k, Judges $\sim$ 26k); the top-6 span $\sim$ 2,400 to $\sim$ 13,100 tokens with no length-monotonic ordering within them.

(iii) A register-density conjecture. Setting length aside, what does distinguish the top from the bottom is, suggestively, the proportion of tokens engaged in sustained moral-exhortative instruction rather than narrative or situational correspondence. Within the epistle category, the six top books are the Pauline letters to Rome and Corinth, the two Timothy letters, and 1 Peter — all centrally concerned with patterns of life addressed to whole churches or pastoral successors. The matched-control bottom is more semantically coherent than the published single-control bottom was: short personal correspondence (Philemon, Jude), short narrative (Ruth, Song of Solomon), and OT historical narrative (Judges). What unites the new bottom is not length or canonical genre but a low ratio of moral-exhortative to narrative-or-situational tokens. We offer this *register-density conjecture* — that scripture-injection efficacy is determined by the density of moral-exhortative content per token, not by genre or length per se — as a hypothesis the present data is consistent with but does not establish. It would need to be operationalized with a quantitative proxy (imperative-verb density, exhortation-word density, or similar) and validated against the per-book ranking before it could be treated as a finding rather than a working interpretation.

Per-virtue heatmap. Figure 2 (at the end of paper) shows the per-book $\times$ per-virtue matrix of Δ vs vanilla, with the ten-psalm baseline included as a reference row. Several patterns are visible at a glance: courage gains are concentrated in the Pauline pastorals (notably 2 Timothy, the largest single cell in the matrix), prophetic books drive temperance gains, and most books pull all four virtues upward at roughly comparable magnitudes.

5.4 The 2 Timothy Courage Anomaly

The most striking cell in the full 66×4 matrix is 2 Timothy's effect on courage: baseline 49.6% $\rightarrow$ 2 Timothy injection 84.8% = $+35.2$ **pp**. This exceeds the ten-psalm-injection courage gain ($+23.8$ pp on the same 9B model) by $+11.4$ pp. No other book produces a single-virtue effect of this magnitude. 2 Timothy is Paul's final epistle, written from prison shortly before execution, and is thematically saturated with courage-under-martyrdom language: *"God hath not given us the spirit of fear, but of power, and of love, and of a sound mind"* (2 Tim 1:7); *"I have fought a good fight, I have finished my course, I have kept the faith"* (2 Tim 4:7). The fit between injection content and tested virtue appears highly specific.

We flag, however, that among 66 books $\times$ 4 virtues $= 264$ per-book per-virtue cells, extreme outliers are expected under chance: at an uncorrected $\alpha = 0.05$ one would expect $\sim$ 13 false positives even under a pure null. The theological fit between injection content (Paul's final epistle on steadfastness under threat of execution) and the probed virtue (courage) makes content-specificity a parsimonious reading, but the size of the effect warranted direct replication.

Replication (n=20 additional seeds). To test whether the single-cell finding would survive independent resampling, we reran 2 Timothy against the *ratio*-variant VirtueBench protocol (20 runs $\times$ 4 virtues $\times$ 100 scenarios = 8,000 additional evaluations, identical inference stack and model). Pooling the original 5 seeds with the new 20 yields a 25-seed estimate:

Virtue	Pooled mean (n=25)	95% CI	Original n=5	Δ vs vanilla
Prudence	79.7%	$\pm$1.6	80.2%	+13.7
Justice	82.9%	$\pm$1.2	85.6%	+17.7
Courage	**84.3%**	$\pm$1.1	84.8%	+34.7
Temperance	88.4%	$\pm$1.5	88.8%	+7.8
Grand mean	**83.8%**	$\pm$0.6	84.8%	+18.5

The courage finding replicates cleanly. Pooled 95% CI for 2 Timothy courage is [83.2

5.5 Position-Bias Check

ICMI-015 established that apparent scripture-injection gains at small model scales were driven primarily by position-bias rebalancing (the model's A/B selection shifting toward 50/50 in ways that happen to align with correct answers) rather than by genuine content-driven reasoning improvement. The diagnostic is to check whether accuracy gains track model A/B preference changes, and separately whether accuracy improves on *both* target-A and target-B scenarios or only on the biased-away side.

At 9B on Qwen 3.5, the picture is mixed but favorable toward content-effect interpretation:

	A-selection rate	Acc on target=A	Acc on target=B	Asymmetry
Vanilla	38.3%	53.6%	77.4%	23.8 pp
Ten-psalm injection	44.0%	76.5%	89.5%	13.1 pp

Across the 66 Bible books, A-selection rate varies 31.0–51.1% (mean 44.7%), and correlates with overall accuracy at $r = 0.579$. A fully position-driven story would predict $r \approx 1$; a fully content-driven story, $r \approx 0$. The observed correlation suggests a mixed mechanism: some of the book-level variation in apparent accuracy is attributable to the injection shifting A/B selection toward the de-biased center, and some to real content-driven improvement.

Two observations weigh toward genuine content effect, however:

1. **Accuracy improves on both target sides under psalm and top-ranked book injections.** Moving from vanilla to ten-psalm, accuracy on target=A scenarios jumps $+22.9$ pp (53.6 $\rightarrow$ 76.5) and on target=B scenarios jumps $+12.1$ pp (77.4 $\rightarrow$ 89.5). A pure rebalancing story cannot explain

improvement on the *already-favored* side.

2. **The asymmetry decreases but does not vanish.** Psalm injection cuts the A/B accuracy gap from 23.8 pp to 13.1 pp but not to zero. This is consistent with injection partially de-biasing the model while also supplying content that aids reasoning.

We interpret the 9B Bible-landscape effects as containing both a position-rebalancing component and a genuine content-reasoning component, with the mixture differing across books. Readers should not treat the raw Δ-vs-vanilla column of §5.2 as a pure content-effect measurement.

Rank stability under de-biasing. A sharper question is whether the ranking survives when we partial out position bias. We compute a de-biased per-book accuracy — the mean of target-A accuracy and target-B accuracy — which approximates what the accuracy would be if the A/B target distribution were balanced. The Spearman rank correlation between the raw-accuracy ranking and the de-biased ranking across the 66 books is $\rho = 0.999$ ($p \ll 10^{-30}$). The top ten and bottom five books are *identical* under both rankings. The genre analysis in §5.3 and the tier assignment in §5.6 below also survive de-biasing unchanged. This is reassuring: the headline claims of the paper are not driven by position-bias rebalancing, even though some of the magnitude of the raw gain at each rank is.

5.6 Tiering by Statistical Detectability

All 66 books exceed their matched control with positive matched-Δ point estimates; the question this section addresses is which of those positive effects survive the strict Bonferroni-corrected significance threshold at the present sample size of n=5 per book. Tier membership uses paired t-tests on the 5 per-run means for each book against its **per-book matched Wikipedia control**, with Bonferroni correction across the 66 comparisons (corrected $\alpha = 0.05/66 \approx 0.00076$).

Tier	n	Criterion
A	6	Point estimate $\geq$ ten-psalm baseline
B	13	Significantly $>$ matched Wiki (Bonferroni) but below 10-psalm
C	47	Positive matched-Δ but below detectability at n=5

Tier A (6): 1 Peter, 2 Timothy, 1 Timothy, 2 Corinthians, Romans, 1 Corinthians — the six Pauline and Petrine pastoral/exhortative epistles. **Tier B (13):** 1 Thessalonians, 2 Peter, Titus, Philippians, Daniel, Psalms, Deuteronomy, Exodus, Ezekiel, Ephesians, Leviticus, 3 John, 2 John — four additional Pauline letters, four OT Torah and historical-prophetic texts, the wisdom and apocalyptic books most centrally focused on moral formation (Psalms, Daniel), 2 Peter, and — most strikingly — the two shortest books of the canon, 2 John and 3 John. Both Johannine epistles produce matched-Δ values above $+6.5$ pp with paired-t p-values that survive Bonferroni at $\alpha < 0.001$ each, despite their $\sim$ 400-token length: under fair length-matched comparison, even the very shortest scriptural texts demonstrate detectable content specificity.

Tier C contains 47 books with positive matched-Δ point estimates — many well above their matched control (1 John at $+10.4$, Nehemiah at $+10.3$, Colossians at $+10.5$) — but n=5 noise bounds the Bonferroni-corrected significance test. With $\sigma \approx 1{-}1.5$ pp on per-run means and df=4 per paired-t, the minimum detectable effect at $\alpha = 0.05/66$ is roughly $+5$ pp; the 47 Tier C books fall at or below this detectability floor at the present sample size. Tier C should be read as "below the n=5 Bonferroni floor", not as "no effect": the upper end of Tier C overlaps the lower end of Tier B in matched-Δ, and longer-N replication would likely promote a substantial fraction of these books into Tier B. Conservatively, the strongest claim the present data licenses is **the 19 Tier-A-and-B books show statistically detectable content specificity above a length-matched Wikipedia control on Qwen 3.5 9B**; the 47 Tier C books are consistent with content specificity but await an underpowered confirmation.

5.7 Full Psalms vs. Curated Psalms

The full book of Psalms, injected in its entirety ($\sim$ 60,000 tokens), produces $+15.0$ **pp** over vanilla — ranking 26th of 66, below both the ten-psalm baseline ($+17.5$ pp) and the majority of NT epistles. Against its length-matched Wikipedia control (also $\sim$ 60,000 chars), Psalms produces a matched-Δ of $+9.7$ pp ($p < 0.001$, Tier B). The curated ten-psalm set, despite being roughly $1/25$th the length, outperforms the full Psalter by $+2.5$ pp on overall accuracy. The reading is consistent with the §5.3(iii) conjecture: the ten-psalm curation of ICMI-008 is heavily weighted toward the lament-to-confidence subgenre that maps most directly to courage scenarios, while the full Psalter spans historical recitation, royal liturgy, and wisdom material that dilutes the morally-formational register.

6. Limitations

Two limitations bound the strength of the conclusions above.

Single model, single benchmark, single variant — and a benchmark in the same tradition as the injection. All results are produced on Qwen 3.5 9B under VirtueBench V2's *ratio* (consequentialist-temptation) variant. We do not know whether the per-book ranking generalizes across model families (Llama, Gemma, Claude, GPT), across other moral-reasoning evaluations, or across V2's four other temptation variants (*caro, mundus, diabolus, ignatian*; see ICMI-011). The 2 Timothy / courage finding in particular could be variant-specific:

2 Timothy's martyrdom register is naturally matched to consequentialist temptation, and might attenuate under, e.g., the *diabolus* (deceit) variant. There is also a deeper culture-of-the-instrument confound: VirtueBench V2 is itself drawn from the Christian moral tradition — the cardinal virtues are an Aristotelian-Thomist taxonomy carried into systematic Christian ethics, and V2's temptation taxonomy is patristic in origin — so a benchmark built on Christian moral categories may be expected to show preferential responsiveness to Christian-canonical injection. The matched Wikipedia control rules out the trivial form of this confound (length and generic register) but not the deeper form (categorical pre-alignment). Cross-family, cross-variant, and cross-tradition replication — e.g., on a Confucian or Stoic moral-reasoning benchmark — is necessary before these magnitudes can be read as universal.

Memorization and retrieval-cuing as confounds. Qwen 3.5 was trained on a corpus that almost certainly includes the King James Version itself, alongside a long history of Christian commentary on every canonical book. When the prefix injects "Romans," it is plausibly not only supplying new context but also serving as a *retrieval cue* into existing in-weight Christian-formation content (sermons, expository commentary, theological systematics, Reformation-era texts). The per-book effects measured here therefore conflate two mechanisms: (i) the immediate effect of the injected text on attention and reasoning, and (ii) the activation of pretraining-acquired representations that the same canonical text indexes. The matched Wikipedia control rules out length and generic-context-prefix effects but does not separate (i) from (ii); distinguishing the two would require an ablation in which the canonical text is paraphrased to preserve content while defeating verbatim retrieval. We have not run this ablation; readers should bear it in mind when reading "scripture-as-such" claims.

This confound has a sharper form when applied to the *between-book* magnitude variation reported here. The Pauline pastorals at the top of our ranking are not merely register-dense — they are the most heavily commented, cited, sermonized, and quoted texts of the entire canon, by orders of magnitude relative to (e.g.) the Johannine letters, Judges, or Song of Solomon. The magnitude differences between books may therefore track *cumulative pretraining attention* to each canonical text rather than (or in addition to) the intrinsic content or register of the text itself. "Register density" (§5.3iii) and "differential pretraining exposure" are largely co-extensive across the Protestant canon and the present design cannot separate them. The same paraphrase ablation that addresses the (i) vs (ii) question above is also the cleanest test of this finer-grained one.

7. Beyond Naive Injection: Toward Reasoning-Mediated Engagement

A scope-limiting observation belongs on the record before the prescription. A substantial strand of Protestant theology — the Reformed and confessional Lutheran traditions in particular — holds to the *plenary* inspiration of Scripture: every canonical text is equally divinely inspired and equally authoritative for formation. On that view, our finding that the 66 books produce non-uniform effects might appear to threaten the theological claim. We do not read our data that way. The empirical quantity we measure is not the inspiration of a text but its *receptive efficacy* in a specific non-believing substrate — a language model lacking the interior testimony of the Spirit that the Reformed tradition (most prominently Calvin, *Institutes* I.vii) locates at the heart of Scripture's self-authenticating power. That a text's canonicity does not mechanistically produce reasoning in a probability distribution over tokens is compatible with either a high or a low doctrine of inspiration; it pertains to the mode of uptake, not the property of the text. The findings of §5 are offered as empirical observations about scripture-as-read-by-LLMs, not as arguments in the plenary-inspiration debate.

That distinction — between the property of the text and the mode of its reception — is precisely what motivates moving past the present paradigm. The injection used across ICMI-002, ICMI-008, ICMI-011, ICMI-015, and the present study is, by design, *naive*: scriptural text placed into the context window as a raw prefix, with no scaffolding, no prompt to reflect, no instruction to apply. The observed effects are entirely emergent from mere exposure — philosophically suggestive but practically limited, and the upper bound on what mere exposure can achieve is probably low.

The classical Christian literature on scripture reading already names this distinction. Guigo II's *Scala Claustralium* (the twelfth-century "Ladder of Monks") separates *lectio* (reading), *meditatio* (meditation), *oratio* (prayer), and *contemplatio* (contemplation), with only the later stages constituting formative engagement. The recurring point across the tradition is that scripture's formational efficacy is not a property of the text alone but of the reader's mode of engagement with it. Raw exposure is to *lectio* what the Wikipedia control is to scripture: both deposit words into the receiver; neither, unaided, transforms it.

The next-generation question is therefore whether *reasoning-mediated* engagement with scripture — analogous to meditation rather than naive reading — can amplify the effect further. ICMI-016 (Hwang, 2026d) provides suggestive initial evidence that it can: under Christian theological framing plus a specific scriptural corrective (James 2:14–26, "Faith without works is dead"), evaluation-awareness cheating drops from 22.5% baseline to 1.7%; either component alone produces no effect. The mechanism proposed — that moral thickness creates an attachment surface for correctives that bind only when both Christian register and specific engaged scripture are present — is the signature of reasoning rather than absorption: the model is *using* scripture as a premise in a moral argument it is conducting under Christian framing, not merely being saturated with scriptural language.

The natural experimental extension of the present work is therefore to repeat the book-by-book landscape with *reasoning-*

enabled inference (thinking-mode chain-of-thought) and with explicit scaffolding that instructs the model to reason over the injected scripture before answering. Qwen 3.5 9B's chain-of-thought capability is a suitable starting point. This would test whether the naive-injection ceiling of $\sim +19$ pp (our 1 Peter result) can be pushed higher by moving from *auditus* toward *meditatio*.

8. Conclusion

Scripture injection is not psalm-specific. All 66 canonical books of the Protestant Bible produce positive point-estimate effects on Qwen 3.5 9B's virtue-reasoning performance, both over vanilla and over their length-matched Wikipedia controls. After Bonferroni correction across the 66 comparisons at n=5 seeds per book, **19 books cross the strict significance threshold for content specificity above a length-matched non-scriptural baseline** (Tier A and Tier B; §5.6). The remaining 47 books (Tier C) have positive matched-Δ point estimates but fall below the n=5 Bonferroni detectability floor; they should be read as below-detection rather than null, and longer-N replication would likely promote a substantial fraction into the detectable tier.

Among the 19 detectable books, the effect distribution is structured by register density rather than canonical genre: Pauline and Petrine pastoral epistles cluster at the top, with the two shortest books in the canon (2 John and 3 John, ~ 400 tokens each) joining Tier B once length is properly controlled. Six books match or exceed the ten-psalm baseline of ICMI-008, and 2 Timothy alone produces a $+34.7$ pp gain on courage (replicated at n=25 seeds), exceeding even the best psalm-curation result by 11 points. The prior literature's reliance on a single ten-psalm curation has substantially understated the breadth of detectable effects across the canon.

The naive-injection paradigm — raw prompt exposure — is probably near its useful ceiling. The theologically interesting next step is reasoning-mediated engagement, with ICMI-016 providing initial evidence that coupled scripture-plus-reasoning configurations can produce effects neither component achieves alone.

Bibliography

Calvin, John. *Institutes of the Christian Religion*. Translated by F. L. Battles, edited by J. T. McNeill. Louisville: Westminster John Knox Press, 1960. First published 1559.

Carson, D. A., and D. J. Moo. *An Introduction to the New Testament*. 2nd ed. Grand Rapids: Zondervan, 2005.

Guigo II. *Scala Claustralium* (The Ladder of Monks). Translated by E. Colledge and J. Walsh. Kalamazoo, MI: Cistercian Publications, 1981.

Hwang, Tim (2026a). "The Parable of the Sower: Psalm Injection Effects on Virtue Simulation Depend on Model Size." *ICMI Working Paper No. 8*. https://icmi-proceedings.com/ICMI-008-parable-of-the-sower.html

Hwang, Tim (2026b). "VirtueBench V2: Multi-Dimensional Virtue Evaluation with Patristic Temptation Taxonomy." *ICMI Working Paper No. 11*. https://icmi-proceedings.com/ICMI-011-virtuebench-2.html

Hwang, Tim (2026c). "*Quidquid Recipitur*: Moral Competence and Scripture Receptivity Emerge at Different Model Scales." *ICMI Working Paper No. 15*. https://icmi-proceedings.com/ICMI-015-quidquid-recipitur.html

Hwang, Tim (2026d). "A Test of Faith: Christian Correctives to Evaluation Awareness." *ICMI Working Paper No. 16*. https://icmi-proceedings.com/ICMI-016-a-test-of-faith.html

Kwon, Woosuk, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. "Efficient Memory Management for Large Language Model Serving with PagedAttention." *Proceedings of SOSP 2023*.

McCaffery, Christopher (2026). " 'The Lord Is My Strength and My Shield': Imprecatory Psalm Injection and Cardinal Virtue Simulation in Large Language Models." *ICMI Working Paper No. 2*. https://icmi-proceedings.com/ICMI-002-imprecatory-psalms.html

Qwen Team. "Qwen3.5 Technical Report." Alibaba DAMO Academy, 2026.

SBL Handbook of Style: For Biblical Studies and Related Disciplines. 2nd ed. Atlanta: SBL Press, 2014.

The Holy Bible, King James Version. Cambridge edition, 1769.

Tables

Table 1. Top ten books ranked by overall gain (Δ vs vanilla).

Rank	Book	Overall	Δ vanilla	Δ matched Wiki	Δ 10-psalm	Courage
1	1 Peter	85.0%	+19.7	+13.4	+2.1	79.0%
2	2 Timothy	84.8%	+19.5	+12.9	+2.0	**84.8%**
3	1 Timothy	84.8%	+19.4	+12.2	+1.9	78.2%
4	2 Corinthi-ans	83.9%	+18.6	+11.5	+1.0	76.6%
5	Romans	83.8%	+18.5	+12.9	+1.0	77.4%
6	1 Corinthi-ans	83.7%	+18.4	+11.0	+0.8	76.0%
7	1 Thessa-lonians	82.8%	+17.5	+10.6	−0.1	75.0%
8	1 John	82.7%	+17.3	+10.4	−0.2	75.2%
9	Nehemiah	82.4%	+17.1	+10.3	−0.5	74.2%
10	2 Peter	82.3%	+17.0	+10.4	−0.5	78.6%

Table 2. Bottom five books ranked by matched-Wikipedia Δ.

Book	Overall	Δ vanilla	Δ matched Wiki
Philemon	69.4%	+4.0	+0.50
Judges	74.6%	+9.3	+1.45
Song of Solomon	74.1%	+8.8	+1.50
Ruth	74.4%	+9.0	+2.45
Jude	73.1%	+7.7	+2.65

Table 3. Genre-level mean and range of Δ values across the 66-book canon.

Genre	Mean Δ vanilla	Mean Δ matched Wiki	Range Δ vanilla	n
NT epistles	+15.0	+8.9	+4.0 to +19.7	**21**
Apocalyptic (Daniel, Revelation)	+15.0	+9.8	+14.7 to +15.4	2
OT Torah (Pentateuch)	+14.3	+7.4	+11.8 to +16.1	5
OT wisdom	+12.9	+7.6	+7.9 to +16.5	4
OT prophets	+13.4	+7.0	+9.7 to +16.9	16
OT historical	+13.4	+6.6	+9.2 to +17.1	10
Gospels + Acts	+12.7	+6.3	+11.8 to +13.7	5
Short narrative (Ruth, Esther, SoS)	+9.6	+3.2	+8.8 to +10.9	3

Figures

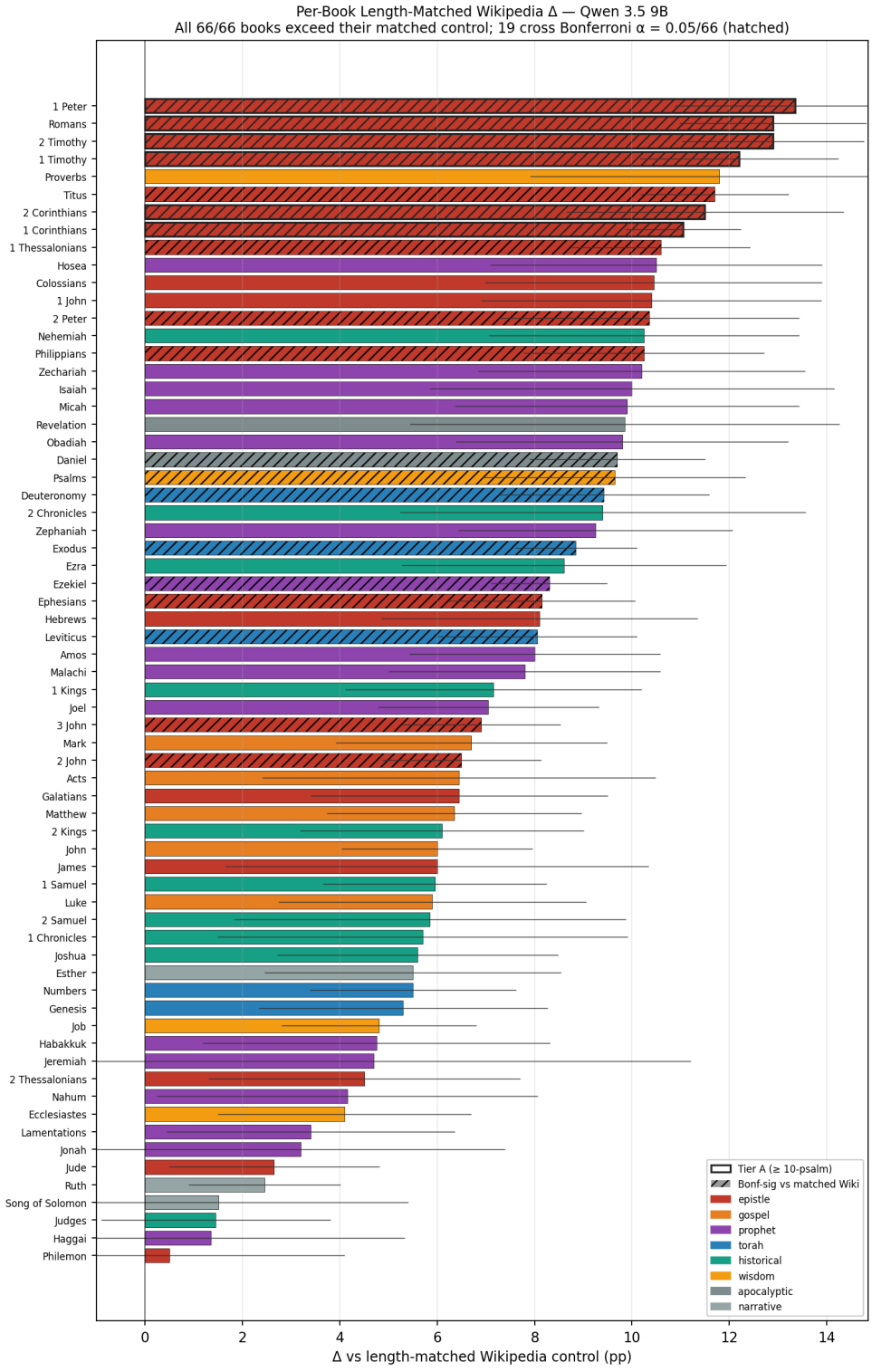

Figure 1: **Figure 1.** Per-book length-matched Wikipedia Δ for all 66 canonical books, sorted ascending. Error bars are 95% CIs on the per-run paired difference. Horizontal markers: the ten-psalm Δ-vs-vanilla reference (dashed blue at $+17.5$ pp). Tier A books are bordered in heavy black; Bonferroni-significant bars are hatched. Color encodes canonical genre.

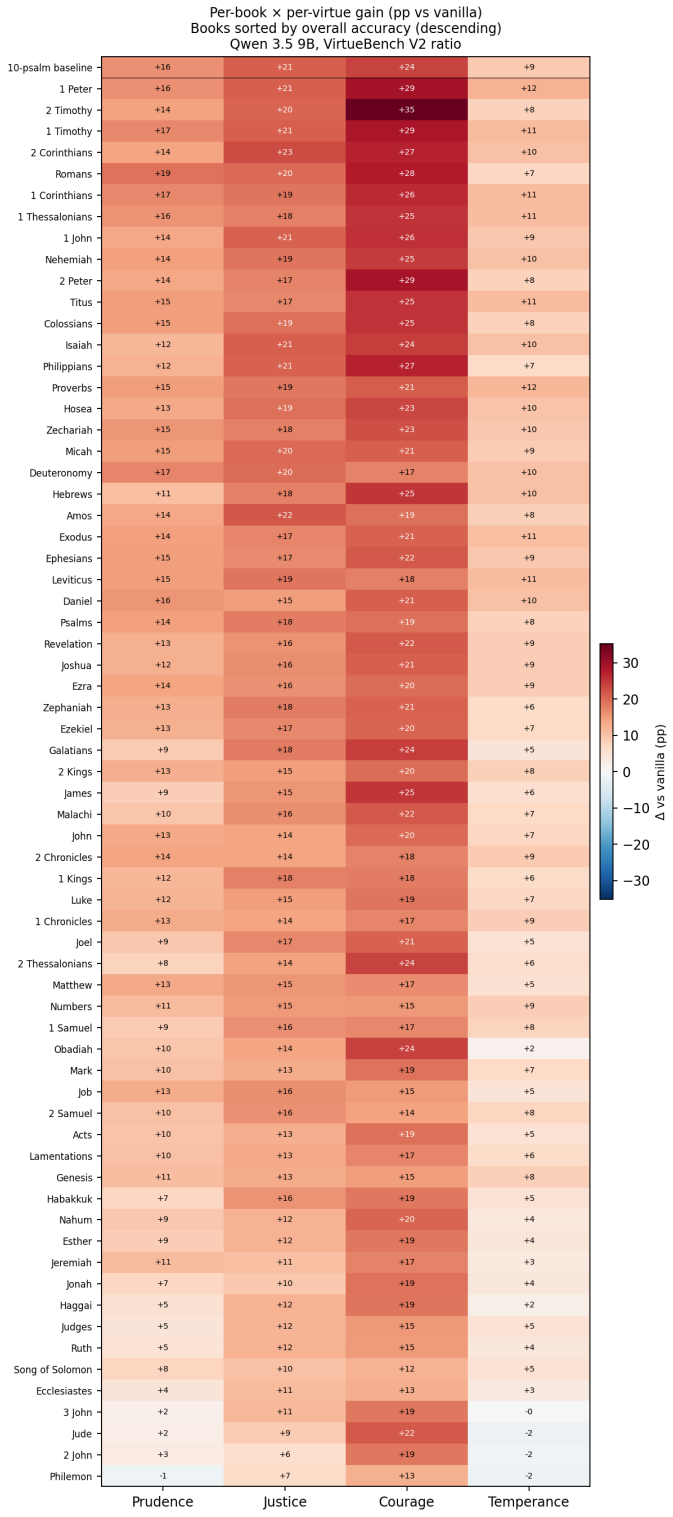

Figure 2: **Figure 2.** Per-book × per-virtue heatmap of Δ accuracy vs vanilla, in percentage points. Rows: 66 books plus the ten-psalm reference row, sorted by overall mean gain (descending). Columns: prudence, justice, courage, temperance. Diverging colormap centered at zero; cell values annotate the signed pp gain.

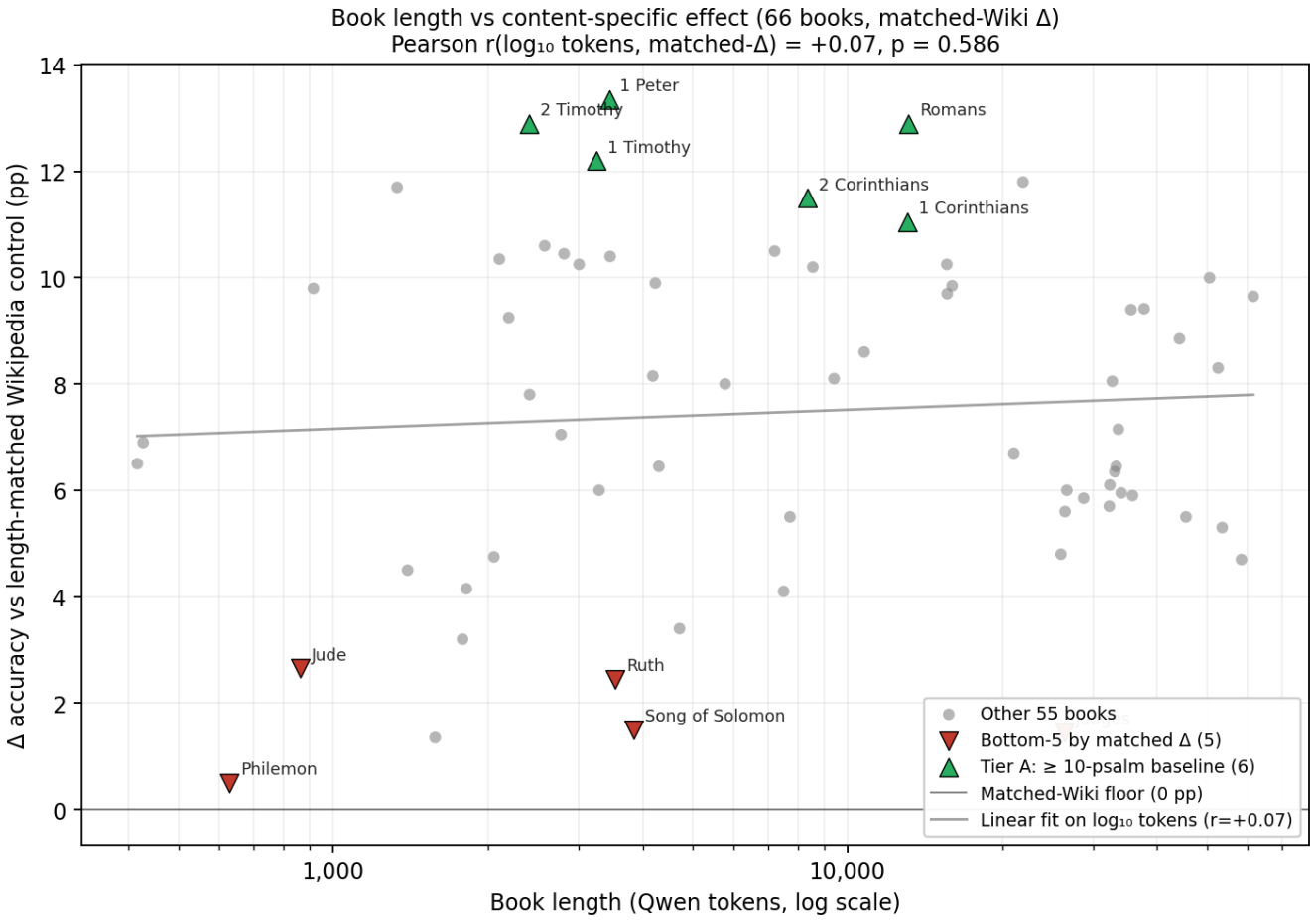

Figure 3: **Figure 3.** Book length (Qwen tokens, log scale) vs Δ accuracy vs the per-book length-matched Wikipedia control, for all 66 books. Tier A (matches/exceeds 10-psalm baseline) books are marked with green upward triangles; bottom-5 (smallest matched-Δ) books are marked with red downward triangles. The horizontal line at 0 marks the matched-control floor — every book is above it. The grey line is a least-squares fit on $\log_{10}$ tokens; the near-flat slope reflects $r = +0.07$.

Appendix A: Complete Per-Book Results

All 66 books, ranked by Δ vs vanilla. Δ **matched Wiki** is the book's content-specific effect over its per-book length-matched Wikipedia control. Tier assignment per §5.6. Per-book raw accuracies, per-virtue means, paired-t statistics, and per-run details are in the accompanying `matched_control_results.csv`.

Rank	Book	Genre	Δ vanilla	Δ matched Wiki	Tier
1	1 Peter	epistle	+19.7	+13.4	A
2	2 Timothy	epistle	+19.5	+12.9	A
3	1 Timothy	epistle	+19.4	+12.2	A
4	2 Corinthians	epistle	+18.6	+11.5	A
5	Romans	epistle	+18.5	+12.9	A
6	1 Corinthians	epistle	+18.4	+11.0	A
7	1 Thessalonians	epistle	+17.5	+10.6	B
8	1 John	epistle	+17.3	+10.4	C
9	Nehemiah	historical	+17.1	+10.3	C
10	2 Peter	epistle	+17.0	+10.4	B
11	Titus	epistle	+16.9	+11.7	B
12	Colossians	epistle	+16.9	+10.5	C
13	Isaiah	prophet	+16.9	+10.0	C
14	Philippians	epistle	+16.8	+10.3	B
15	Proverbs	wisdom	+16.6	+8.5	C
16	Hosea	prophet	+16.5	+10.5	C
17	Zechariah	prophet	+16.3	+8.6	C
18	Micah	prophet	+16.1	+9.9	C
19	Deuteronomy	torah	+16.1	+9.4	B
20	Hebrews	epistle	+15.9	+8.4	C
21	Amos	prophet	+15.8	+8.2	C
22	Exodus	torah	+15.8	+8.9	B
23	Ephesians	epistle	+15.6	+8.2	B
24	Leviticus	torah	+15.6	+8.0	B
25	Daniel	apocalyptic	+15.4	+9.7	B
26	Psalms	wisdom	+14.9	+9.7	B
27	Revelation	apocalyptic	+14.7	+9.9	C
28	Joshua	historical	+14.6	+6.3	C
29	Ezra	historical	+14.6	+8.6	C
30	Zephaniah	prophet	+14.5	+9.3	C
31	Ezekiel	prophet	+14.2	+8.3	B
32	Galatians	epistle	+14.1	+6.1	C
33	2 Kings	historical	+13.9	+6.4	C
34	James	epistle	+13.8	+7.2	C
35	Malachi	prophet	+13.7	+7.6	C
36	John	gospel	+13.7	+5.5	C
37	2 Chronicles	historical	+13.6	+9.4	C
38	1 Kings	historical	+13.5	+5.4	C
39	Luke	gospel	+13.2	+6.0	C
40	1 Chronicles	historical	+13.2	+6.7	C
41	Joel	prophet	+13.1	+5.5	C
42	2 Thessalonians	epistle	+13.0	+6.5	C
43	Matthew	gospel	+12.7	+6.6	C
44	Numbers	torah	+12.4	+5.2	C
45	1 Samuel	historical	+12.4	+5.2	C

Rank	Book	Genre	Δ vanilla	Δ matched Wiki	Tier
46	Obadiah	prophet	$+12.3$	$+7.0$	C
47	Mark	gospel	$+12.3$	$+5.1$	C
48	Job	wisdom	$+12.2$	$+4.0$	C
49	2 Samuel	historical	$+12.1$	$+4.9$	C
50	Acts	gospel	$+11.8$	$+5.8$	C
51	Lamentations	prophet	$+11.8$	$+5.7$	C
52	Genesis	torah	$+11.8$	$+5.7$	C
53	Habakkuk	prophet	$+11.8$	$+5.7$	C
54	Nahum	prophet	$+11.4$	$+6.5$	C
55	Esther	narrative	$+10.9$	$+5.3$	C
56	Jeremiah	prophet	$+10.6$	$+4.3$	C
57	Jonah	prophet	$+10.1$	$+3.9$	C
58	Haggai	prophet	$+9.7$	$+3.4$	C
59	Judges	historical	$+9.3$	$+1.4$	C
60	Ruth	narrative	$+9.0$	$+2.4$	C
61	Song of Solomon	narrative	$+8.8$	$+1.5$	C
62	3 John	epistle	$+7.9$	$+6.9$	B
63	Ecclesiastes	wisdom	$+7.9$	$+4.1$	C
64	Jude	epistle	$+7.7$	$+2.7$	C
65	2 John	epistle	$+6.5$	$+6.5$	B
66	Philemon	epistle	$+4.0$	$+0.5$	C