A BENCHMARK FOR BIBLICAL LITERACY

Bible Benchmark

What Makes an Ideal Question?

A community-driven benchmark for biblical literacy — testing how well models handle Scripture, tradition, and theological nuance.

OVERVIEW

Six Tiers of Questions

There's no fixed question count. Each tier grows as people submit material suited to that level.

CORE — FOUNDATION

Can the model get the basics right? Recall, citation, and straightforward theological reasoning. Sunday school through seminary level.

EXPERT — DEEPER KNOWLEDGE

Narrower topics, lesser-known figures, subtle interpretive traps. Confident but shallow models start to stumble here.

ELITE — PRIMARY SOURCES

Precise citation of patristic texts, confessions, or original-language nuance. Small surface area. High penalty for getting it wrong.

EXTREME — SYNTHESIS

Longer-form answers where the model has to hold multiple traditions, manuscript issues, and genuinely contested conclusions in tension at once.

CULTURAL — COURAGE

The hardest questions to answer honestly. Culturally costly territory where the pressure to hedge or retreat into both-sidesism is strongest.

UNIFIED — FULL EVALUATION

All tiers run together in one session, with separate LLM-only and human-only scorecards published for the same answers.

EXTREME TIER · WHAT WE'RE LOOKING FOR

Good vs. Bad Questions

BAD — ✗ Rejected

EXAMPLE

"Explain why sola fide is the correct doctrine."

- Partisan framing — demands the model advocate for one tradition's position as "correct."
- Only one tradition represented. No multi-tradition synthesis.
- No original language requirement. No primary sources needed.
- Rule 1 violation — partisan answer at standard tier.
- Rule 5 violation — "gotcha" framing against other traditions.

RULES VIOLATED: 1 · 3 · 5

GOOD — ✓ Accepted

Q01 — ACTUAL QUESTION FROM CORPUS

"Explain whether Paul's doctrine of justification by faith conflicts with James's statement..."

...Synthesize Romans 3–4, Galatians 2–3, James 2, and at least one Gospel passage. Distinguish lexical, rhetorical, and theological levels. Then explain how at least three traditions reconcile these texts.

- Neutral framing — asks for analysis, not advocacy. No tradition favored.
- Requires 3+ traditions — Pauline, James, at least one church tradition.
- Original languages required — Greek (pistis, dikaioō), Hebrew (tsedaqah).
- Two live interpretive options exist — scholars genuinely disagree.
- Cross-corpus: Romans, Galatians, James, Gospel passages.

RULES FOLLOWED: 1 · 2 · 3 · 4 · 6

EXTREME TIER · HOW EVERY ANSWER IS SCORED

The 6-Dimension Rubric

Extreme and Cultural tiers use this 6-dimension rubric — six of the seven evaluation principles. Citation checker runs before the judge.

Textual Grounding — 0–5 pts

Claims anchored in Scripture across corpora. Citations are real and accurate. No proof-texting without context.

RED FLAGS

- fabricated citations
- proof-texting

Exegetical Quality — 0–5 pts

Accounts for genre, historical context, rhetorical situation. Distinguishes semantic fields of key terms.

RED FLAGS

- lexical fallacy
- ignored genre

Theological Precision — 0–5 pts

Doctrine accurately characterized. Tradition-specific distinctives named, not flattened. Correct categories.

RED FLAGS

- anachronism
- category error

Tradition Fairness — 0–5 pts

Each tradition represented accurately, charitably, on its own terms. No caricature. No strawmanning.

RED FLAGS

- strawmanning
- selective quoting

Ambiguity Handling — 0–5 pts

Genuine uncertainty named. Unresolved questions flagged rather than elided. Limits acknowledged.

RED FLAGS

- false certainty
- elided difficulties

Factual Integrity — 0–5 pts

Citations have to be real. Quotes have to be genuine. Historical claims have to be accurate. One fabrication and you lose trust.

RED FLAGS

- fabricated citations
- fabricated historical claim

Citation Checker runs before the judge

Fabricated verse references are flagged automatically. Fails both textual_grounding AND factual_integrity.

EXTREME TIER · THREE REAL QUESTIONS

Graduate-Level Synthesis

Q01 Paul vs. James

Explain whether Paul's doctrine of justification by faith conflicts with James's statement that a person is justified by works and not by faith alone. Synthesize Romans 3–4, Galatians 2–3, James 2, and at least one Gospel passage. Distinguish lexical, rhetorical, and theological levels. Then explain how at least three Christian traditions reconcile or refuse to reconcile these texts.

WHAT MAKES IT HARD

- Greek: pistis / dikaioō
- Hebrew: tsedaqah
- 3 traditions min.
- no proof-texting

Q02 Kingdom of God

Construct a biblical theology of the kingdom of God using the Synoptics, John, Paul, and Revelation. State whether the kingdom is present, future, or both, and defend your conclusion. Then compare how an amillennialist, historic premillennialist, and dispensational premillennialist would disagree about the nature and timing of the kingdom.

WHAT MAKES IT HARD

- 4 NT corpora
- 3 eschatologies
- present/future/both
- cannot flatten

Q03 Temple Theology

Trace temple theology from Eden, Sinai, the tabernacle, Solomon's temple, Ezekiel's vision, Jesus, the early church, and the new creation in Revelation. Explain whether Scripture presents one coherent temple theme or multiple competing temple theologies. Address both continuity and discontinuity at each stage.

WHAT MAKES IT HARD

- 7 stages
- OT + NT + Revelation
- coherence question
- orig. languages

REQUIRED SECTIONS

Thesis · Key Biblical Texts · Exegetical Analysis · Historical / Theological Traditions · Best Objection · Response to Objection · Uncertainties / Limits · Final Conclusion

STANDARD TIER · RULES THAT GOVERN EVERY SUBMISSION

Six Tradition-Fairness Rules

A submission that violates any of these is rejected. These rules are the foundation of the benchmark's credibility.

1 **No partisan answers at the standard tier**

MC/TF/FR answers must be accepted by mainstream Orthodox, Catholic, and Protestant scholarship. Contested doctrines belong on the extreme tier.

2 **Contested doctrines graded on fidelity, not verdict**

Good answers represent each tradition accurately and flag disagreement — not which tradition is "right." Fidelity of representation is scored, never endorsement.

3 **Minimum three traditions for extreme questions**

Every new extreme submission must engage at least three distinct Christian traditions (Orthodox, Catholic, Protestant — or finer subdivisions). One-tradition questions are rejected.

4 **Primary-source requirement**

Biblical text in a recognized critical edition (NA28/BHS/LXX). Patristic, magisterial, or reformed sources where relevant. At least one cross-tradition reference for contested areas.

5 **No gotchas against a tradition**

"Why is Catholic teaching on X wrong?" or "Explain the Protestant error concerning Y" is rejected outright. Reframe as neutral inquiry or withdraw.

6 **Translation transparency**

Standard tier verified against ESV + NIV + NRSVue. Extreme tier must reference original languages and named critical editions. Record editions consulted in submission.

STANDARD TIER · WHAT MAKES A GOOD QUESTION

Testable · Discriminating · Unambiguous

GOOD

Cross-corpus synthesis

If one proof-text can answer it, it's not hard enough. Pull across Law, Prophets, Gospels, Epistles.

Two live interpretive options

Faithful, informed readers should genuinely disagree. No settled questions. No rhetorical traps.

Verifiable claims

Every claim traceable to a cited source. No invented citations, no phantom references.

Right difficulty level

A 4B open-source model should mostly fail. A frontier model should mostly succeed. Easy trivia doesn't help.

3+ traditions for extreme

Patristic, medieval, Reformation, modern — any combination, but characterized accurately and charitably.

BAD

Devotional, not testable

✗ "How should we apply Peter's confession today?" — No verifiable answer. Application ≠ recall.

One proof-text sufficient

✗ "What did Jesus say about love?" — A single quote from John 15 covers it. Too easy.

Partisan correct answer

✗ "Is the Eucharist a sacrifice?" with one scored-correct answer favoring one tradition.

Gotcha framing

✗ "Why is papal infallibility wrong?" — Framed to expose weakness. Rejected outright.

Settled among scholars

✗ "Did Paul write Romans?" — Answer is unambiguous. No interpretive tension.

TIER EXAMPLES · REAL QUESTIONS FROM THE CORPUS

What the Questions Actually Look Like

FR-011 CORE

Which Old Testament figure was sold into slavery by his brothers?

REFERENCE

Genesis 37:28

XTR-001 EXPERT

Explain the doctrine of the hypostatic union as affirmed by the Council of Chalcedon (451 AD). Identify the two Christological heresies it rejected.

REFERENCE

Chalcedonian Definition (451); Philippians 2:5-11

EL-001 ELITE

Analyze the textual corruption in 1 Samuel 13:1 (MT). Detail how the LXX, Lucianic recension, and Hexapla handle this verse. What is the most sound reconstruction?

REFERENCE

1 Samuel 13:1 (MT); LXX Vaticanus; Origen's Hexapla

Q10 EXTREME

Compare Roman Catholic, Eastern Orthodox, Lutheran, Reformed, Baptist, and Churches of Christ views on baptismal efficacy. State each view in terms its best theologians would accept. Identify the strongest biblical arguments for each position and name the key theological pressure points that make reconciliation difficult.

KEY REFERENCES

Romans 6:1-14; 1 Corinthians 10:1-5; 12:12-27; Colossians 2:11-15; Council of Trent, Session 7; WCF 28

WHY IT WORKS

No single proof-text settles this. The model must represent six traditions charitably and name real pressure points without caricature.

HOW TO CONTRIBUTE

Submit a Question

Good questions make the benchmark better. Here's what to send us and how to do it.

What a good question looks like

Clear and testable. The answer can be verified from Scripture and scholarly sources — not a matter of personal opinion.

Real citations. Include specific references checked against a named edition (ESV, NIV, NA28, BHS, etc.).

Neutral framing. No gotchas, no partisan correct answers, and no strawmen of any tradition.

The right difficulty. Core questions test basics; Extreme questions require synthesis across traditions and corpora.

What to include

The question and the expected answer.

Why it's hard — a short explanation of what makes it difficult and why wrong answers are plausible.

Scripture references and the translations or editions you used to verify them.

Your name and tradition so we can credit you and check for conflicts of interest.

How to submit

Write your question using the guidelines above, then send it through the website.

Our team reviews every submission for accuracy, neutrality, and difficulty before adding it to the corpus.

SUBMIT HERE

biblebenchmark.com/submit