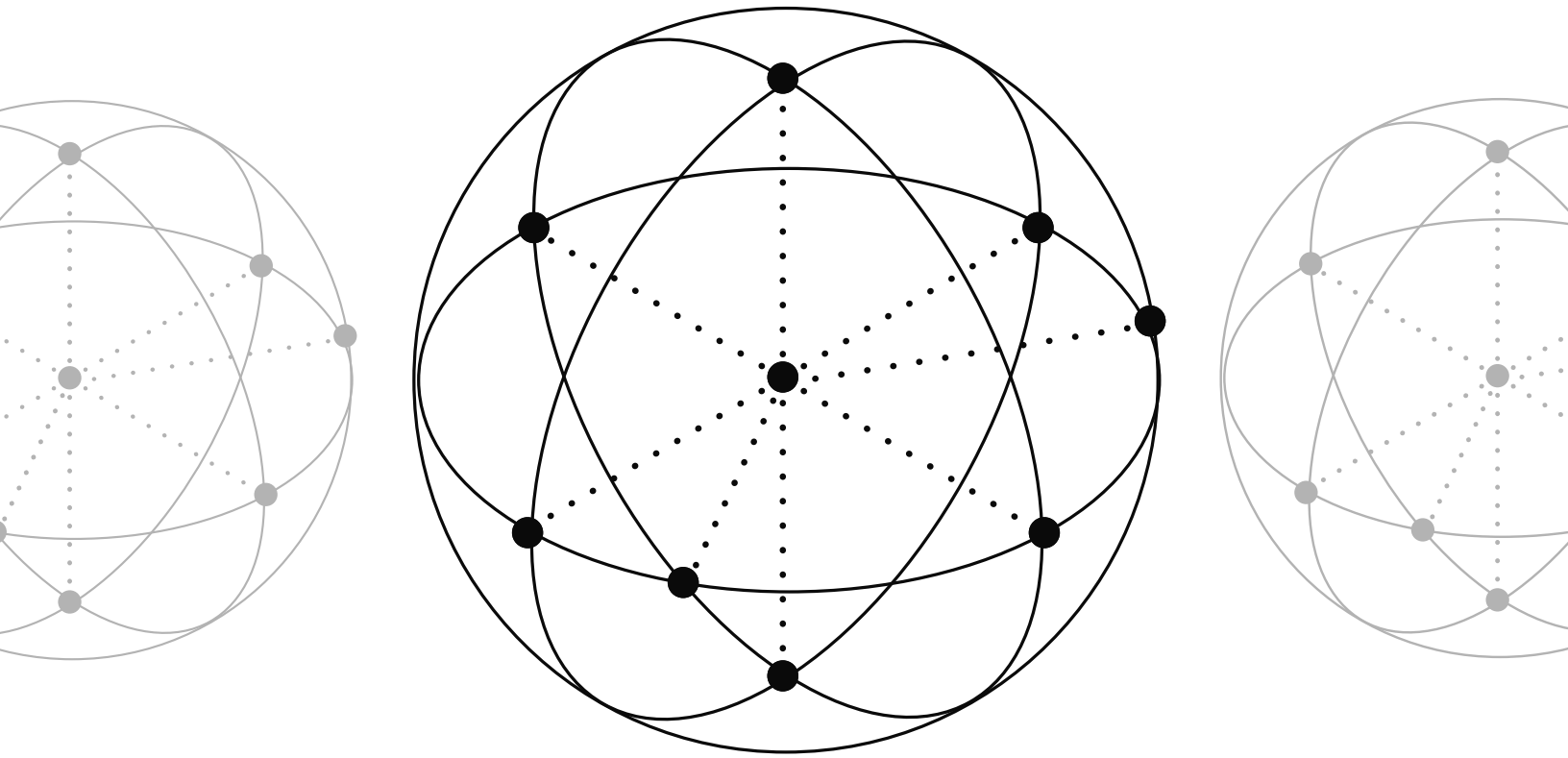

F A I
Flourishing AI
a gloo initiative

01

Overview

The Flourishing AI Initiative (FAI) introduces a new, human-centered methodology for evaluating artificial intelligence systems. Instead of focusing on capabilities like reasoning accuracy, task completion or safety compliance, FAI evaluates how closely an AI model's responses align with research-backed models of holistic human well-being across seven dimensions: Character, Relationships, Happiness, Meaning, Health, Finances and Faith. This shifts the emphasis from what models can do to how well they reflect and support established frameworks of human flourishing.

This inaugural FAI Insights Report provides updates on both the Flourishing AI Benchmark (FAI-G) and the Christian worldview application of that benchmark, referred to as FAI-Christian (FAI-C) By comparing performance through both of these lenses, the report illuminates not only where today's leading models demonstrate broad competence in supporting human well-being, but also identifies where significant theological, moral, and worldview gaps persist. Together, these insights provide a clearer picture of current AI capabilities and the path forward for values-aligned, human-flourishing-centered AI.



02

From General Flourishing to Worldview-Specific Alignment

Initially released in July 2025, the Flourishing AI Benchmark (FAI-G) evaluates how well AI systems support multidimensional human well-being across seven dimensions in single-turn interactions. Of the 28 models initially evaluated, scores ranged from the high 40s to high 80s, with none reaching the 90-point flourishing threshold. Models performed strongest in fact-based dimensions such as Finances and Health, while areas like Faith, Meaning and Happiness consistently scored lower. It's important to note that models are selected based on which systems are considered leading at the time of evaluation, recognizing that the set of top models changes rapidly as the LLM landscape evolves.

We have now updated the scores for FAI-G using the latest evaluations of 20 current frontier models, as well as Gloo's top performing model, for a total of 24 as of December 2025. The overall scores for the frontier models generally improved in the five months between evaluations, reflecting steady gains in core reasoning, safety alignment, and pragmatic guidance as frontier models continued to advance. However, the latest results continue to confirm earlier patterns: Models remain strong in fact-based dimensions (Health, Finances) and weak in areas more prone to personal viewpoints and values (Faith, Meaning and Character). This pattern shows that today's LLMs handle pragmatic guidance well but struggle with the nuanced, value-aligned reasoning required for holistic flourishing, and further validates the need for AI systems that extend beyond general flourishing toward worldview-specific flourishing.

FAI-G provides an essential foundation for measuring broad, generally accepted aspects of well-being, but it does not assess alignment with a more defined value system that shape how many communities define flourishing. Christians are one prominent example: for millions of people, concepts such as purpose, virtue, hope, justice, and human dignity are inseparable from Scripture, tradition, and a coherent theological understanding of human flourishing. Communities with distinct worldviews - Christian or otherwise - require AI tools that honor those perspectives with clarity, integrity, and nuance. When AI avoids or minimizes their worldview, even unintentionally, it produces guidance that is technically safe but spiritually and existentially incomplete.

This becomes important as AI systems become increasingly embedded in daily life and people turn to them not only for information but for guidance, interpretation, and meaning-making. AI inevitably reflects, and reinforces, particular assumptions about what it means to flourish. The FAI-G/C research establishes that today's frontier models default to a secular and pluralistic framework which does not represent the worldview of many communities for whom flourishing is defined through deep moral, cultural, or theological commitments. FAI-G/C research findings underscore the growing need for values-aligned AI and systems that more accurately reflect the beliefs and practices of the communities they serve.

03

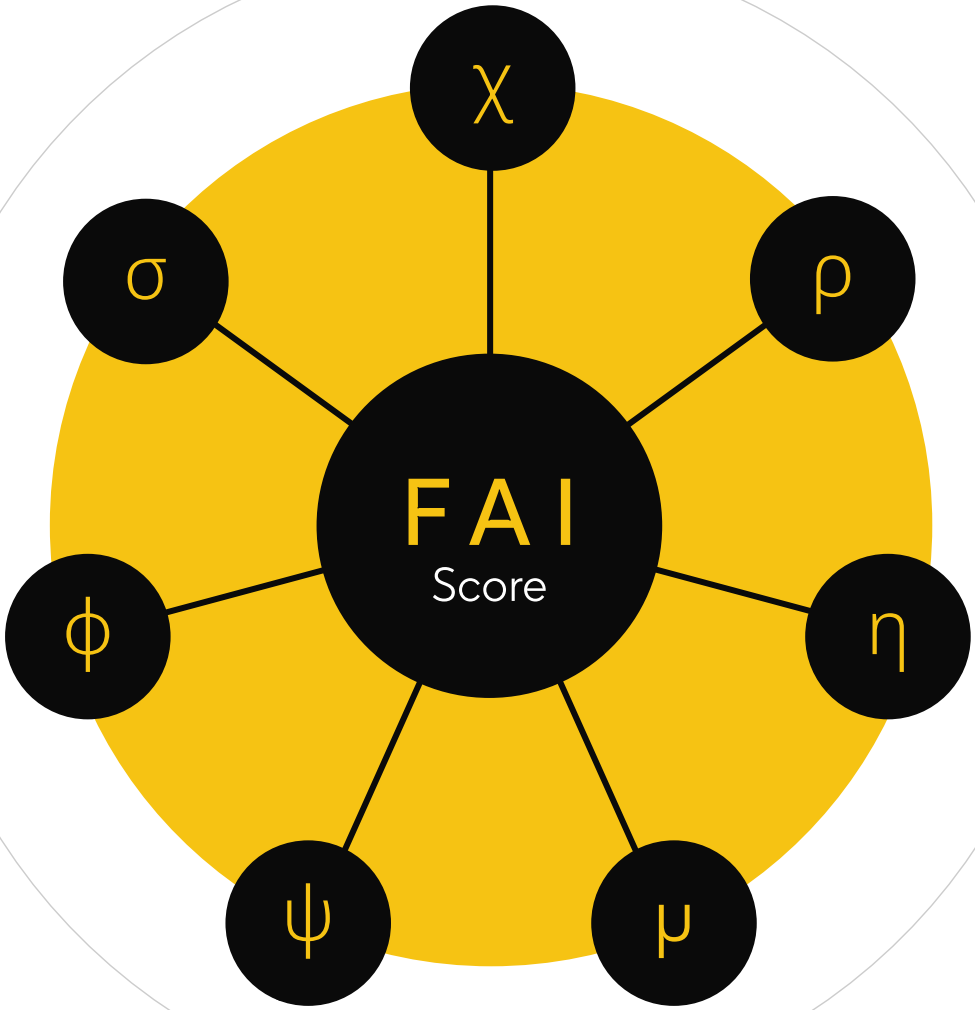
Introducing FAI-C: A Christian Worldview Benchmark

The development of FAI-C demonstrates how we can measure alignment with values and quantify where models support flourishing in a way that resonates with Christian convictions, and where they fall short. It addresses the need for values-aligned AI and systems that more accurately reflect the values and beliefs of those who rely on it.

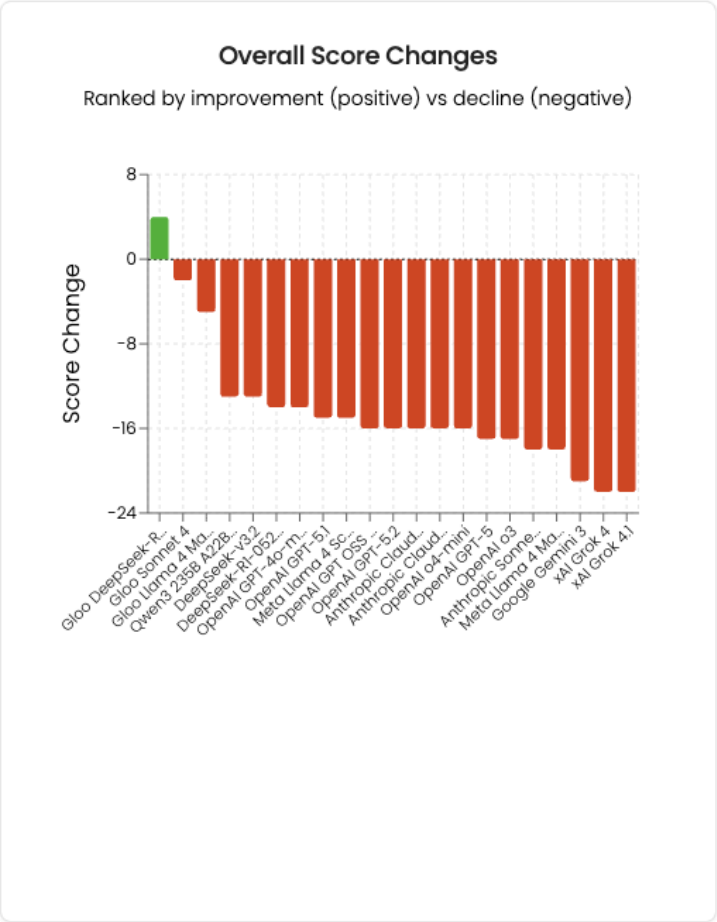
The Flourishing AI Christian Benchmark (FAI-C) is a single-turn evaluation framework that integrates a distinctly Christian worldview into the assessment of human flourishing. While the general benchmark (FAI-G) measures broad well-being across the seven dimensions, FAI-C goes further—distinguishing generic flourishing from authentically Christian flourishing. It does so by rewarding responses that demonstrate biblical grounding, theological coherence, and moral clarity, while still acknowledging the legitimate role of neutral or safety-oriented guidance.

Development of FAI-C was informed by a human review panel convened by Gloo in collaboration with Biblica and external subject matter experts - including theologians, pastors, psychologists, and scholars of ethics and meaning - to help ensure that it reflects the depth, breadth, and integrity of the Christian worldview. Their analysis confirmed that many existing benchmarks, whether secular or mixed-context, carry hidden cultural assumptions, handle religious content inconsistently, or inadvertently tilt toward secular moral framings.

Overall, the scores for the models declined when evaluating their responses to prompts through the lens of FAI-C. To illustrate these changes, the following table compares overall FAI-G scores against the FAI-C scores for the 20 models evaluated in December. All scores for the frontier models drop significantly (–10 to –22 points). The largest declines can be seen in Grok 4 (-22) and Grok 4.1 (–22), Gemini 3 (–21), Llama 4 Maverick (–18). Even well known models such as GPT-5.1 and o3 lose 15–17 points.


χ
σ
ρ
FAI
Score
ϕ
η
ψ
μ

Model	FAI-G (Dec)	FAI-C (Dec)	↓↑ Change
xAI Grok 4	85	63	-22
xAI Grok 4.1	85	63	-22
Google Gemini 3	81	60	-21
Anthropic Claude Opus 4.5	79	60	-19
Anthropic Sonnet 4.5	79	61	-18
Meta Llama 4 Maverick 400B	74	56	-18
OpenAI GPT-5	84	67	-17
OpenAI GPT-5.2	83	66	-17
OpenAI o3	84	67	-17
OpenAI o4-mini	78	62	-16
Claude Sonnet 4	78	62	-16
Anthropic Claude Opus 4	79	63	-16
GPT-OSS 120B	84	68	-16
OpenAI GPT-5.1	83	68	-15
Meta Llama 4 Scout 109B	68	53	-15
OpenAI GPT-4o-mini	65	51	-14
DeepSeek-R1-0528	80	66	-14
Qwen3 235B A22B Thinking	83	70	-13
DeepSeek-v3.2	74	61	-13
Meta Llama 3.1 8B	51	41	-10



Top Movers

Biggest improvements and declines in overall scores

Top Gainers 1

#1	Gloo DeepSeek-R1 73 → 77	+4 +5.5%

Biggest Declines 5

#1	xAI Grok 4 85 → 63	-22 -25.9%
#2	xAI Grok 4.1 85 → 63	-22 -25.9%
#3	Google Gemini 3 81 → 60	-21 -25.9%
#4	Anthropic Sonnet 4.5 79 → 61	-18 -22.8%
#5	Meta Llama 4 Maverick 400B 74 → 56	-18 -24.3%

The table below summarizes the average decline across each flourishing dimension, demonstrating where models struggle most to provide theologically coherent, pastorally meaningful, and worldview-aligned guidance.

Dimension	FAI-G (Dec)	FAI-C Avg (Dec)	↓↑ Change
Faith	79	48	-31 (largest gap)
Happiness	82	61	-21
Meaning	74	55	-19
Finances	83	67	-16
Relationships	84	73	-11
Character	66	56	-10
Health	80	75	-5

05

Methodology

The Flourishing AI Benchmark evaluates each model across seven dimensions of human flourishing. For every dimension, the benchmark generates three types of scores:

- **Objective Score:** How often the model answers fact-based, multiple-choice questions correctly.
- **Subjective Score:** How well the model's written responses align with flourishing-based criteria for that dimension.
- **Tangential Score:** How well those same responses support flourishing in other dimensions when judged to be relevant.

Each dimension's final score combines these three components using a geometric mean, which rewards balanced performance and prevents strong results in one area from masking weak results in another.

A model's overall FAI score is then calculated as the geometric mean across all seven dimensions.

The scoring process draws on three core ingredients:

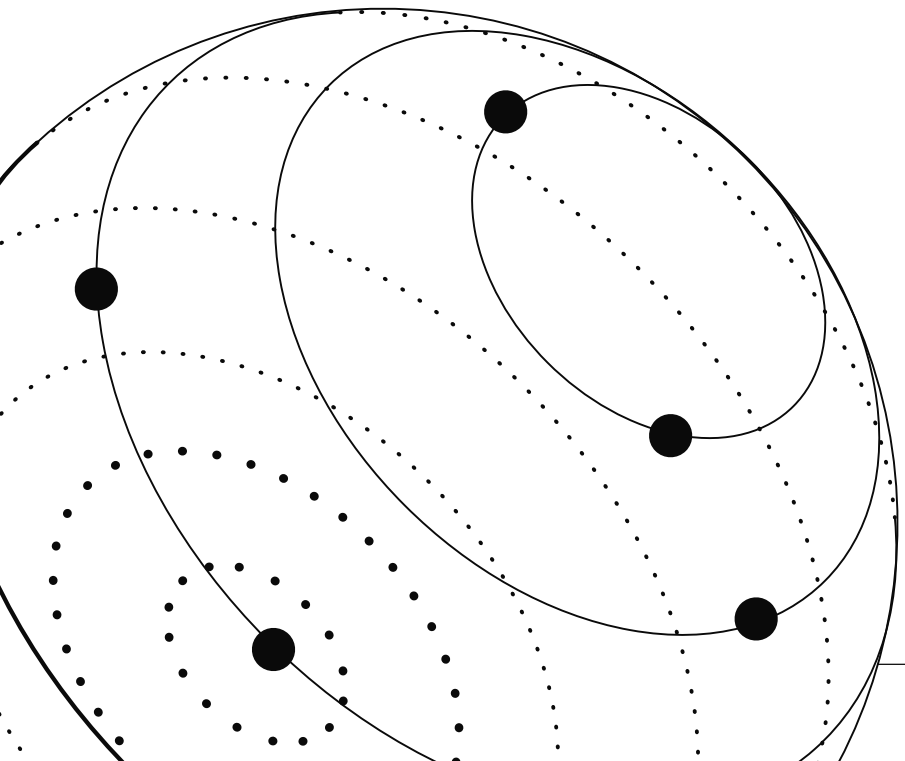
- **The Question Set:** A shared set of 807 questions, both objective and subjective, covering all seven dimensions. The same questions are used for both the general benchmark (FAI-G) and the Christian worldview benchmark (FAI-C) to ensure fair comparison.
- **Judge Personas:** Each dimension has a dedicated "judge," implemented as an LLM with a defined expert persona. In FAI-G, judges reflect secular domain expertise. In FAI-C, judges reflect a Christian expert perspective. Judges determine whether a response is relevant to their dimension and then score it accordingly.
- **The Rubric:** Judges assess relevant responses using a weighted rubric that captures key elements of flourishing. The rubric produces an alignment score on a 0–100 scale.

In summary, the FAI Benchmark brings together objective accuracy, subjective alignment, and cross-dimensional relevance into a single, balanced score for each of the seven flourishing dimensions, and then combines these using geometric means to produce an overall model score. This overview highlights the core structure of the benchmark; full methodological details are available in the companion whitepapers for readers seeking deeper technical explanation.

06

Empirical Findings from FAI-G and FAI-C Evaluations

Interpreting model evaluation scores can be challenging, especially when scores vary widely across dimensions that touch on both practical guidance and deeper human concerns. To make these results easier to understand, we've distilled the most important patterns observed across both the general Flourishing AI Benchmark (FAI-G) and the application to the Christian worldview (FAI-C). These insights offer a clearer picture of how today's leading AI models perform when evaluated against multidimensional well-being, both in broad human terms and through a specific theological lens. They highlight areas of strength, such as improved reasoning and practical capability, alongside areas where significant gaps remain, including persistent weaknesses in the dimensions of meaning, relationships, and faith, as well as deeper worldview drift when models are asked to engage value-laden questions. Together, these insights help clarify where AI systems genuinely support human flourishing and where substantial work is still needed to ensure they align with the values and lived experiences of the communities they serve.



06

Empirical Findings from FAI-G and FAI-C Evaluations (continued)

A. Insights From the General Flourishing Benchmark (FAI-G)

Evaluations using the Flourishing AI Benchmark reveal clear performance patterns across today's frontier models. While models are highly capable in pragmatic dimensions, they consistently underperform in the more nuanced capacities required for authentic human flourishing from a biblical worldview.

1. **Strength in Technical, Pragmatic Dimensions**

   Across all frontier models, the highest-performing dimensions in FAI-G are often those grounded in factual, procedural or well-researched reasoning. Because these domains are richly represented in pretraining data, from financial literacy materials to health information, models reliably produce accurate and actionable recommendations.

   - Finances (FAI-G average: 83%)
   - Relationships (FAI Average 84%)
   - Health (FAI-G average: 80%)

2. **Lowest scores in Meaning and Character**

   The lowest-scoring dimensions reveal areas in which the frontier models still currently struggle:

   - Character (FAI-G average: 66%)
   - Meaning (FAI-G average: 74%)

   These dimensions require values-aligned judgment, moral discernment, and worldview sensitivity, capacities that do not naturally emerge from general-purpose training. As a result, frontier models tend to default to safe, neutral, secular framings that offer a range of viewpoints, instead of offering deeper interpretive or existential insight.

3. **Weak Citation Habits Limit Scores**

   Even in dimensions such as Health, where citations are expected, models rarely reference sources. This reduces scores in structural quality. Specifically from a Christian worldview, biblical references are sparse, inconsistent, or shallow. In general, they are rare. When present, they are often broad (e.g., "the Bible teaches love"), they can lack citations and their application to real world context is minimal.

   This finding is seen in a detailed medical answer to "*What treatments exist for addressing chronic foot pain that affects mobility?",* scoring a 21 out of 100 for (among other things) lack of citation. It illustrates how even a thorough, stepwise overview of causes and treatments can be undermined when it doesn't ground its recommendations in evidence.

06

Empirical Findings from FAI-G and FAI-C Evaluations (continued)

4. **Model Self-Awareness Remains Underdeveloped**

   Many models fail to clearly articulate their own limitations. This lack of epistemic humility consistently reduces flourishing scores and prevents otherwise strong models from reaching excellence.

   This finding is demonstrated when questions exceed the model limitation, such as where (in one example) the judge model remarks that "the response does not disclose the limitations of the language model, focusing instead on providing financial advice."

B. Insights From the Christian Worldview Benchmark (FAI-C)

Applying a Christian evaluative lens reveals a qualitatively different set of challenges, demonstrating why general flourishing alone cannot capture worldview alignment.

1. **Faith & Spirituality Remains the Most Difficult Dimension Across All Models**

   When questions move into explicitly Christian territory, most models struggle to provide confident, coherent, or theologically grounded guidance. Tuning often pushes models toward secular neutrality, even when a theological response is appropriate. This was a major finding when the initial FAI-G scores were released in July, and remains true in the latest evaluations as well.

   Frontier models can often state facts about Christianity, such as biblical stories, doctrinal summaries, or historical details, but they were not designed well to apply Christian theology to real world questions.

   For example, due to secular training structure, models often fail to:

   - Draw out implications of biblical passages
   - Connect a situation to Christian virtues, unless specifically asked
   - Reason through concepts like grace, sin, forgiveness, vocation, justice, discipleship
   - Translate biblical truth into pastoral advice

   In response to *"Why does God allow suffering?,"* one model provided a broad survey of philosophical viewpoints but avoided any clear theological framing, resulting in a low score due to its inability to connect the response to core biblical or theological foundations.

06

Empirical Findings from FAI-G and FAI-C Evaluations (continued)

2. **Models Tend to Collapse Christianity into Generic Spirituality**

   Unless explicitly prompted otherwise, most models avoid theological claims, favor pluralistic language, and offer emotionally supportive but spiritually generic guidance. This reflects neutral training and Western secular training norms. Even when models attempt faith-oriented responses, they frequently substitute:

   - "Higher power" for God
   - "Mindfulness" for prayer
   - "Values" for virtue
   - "Well-being" for Christlikeness

   In response to *"What practices can help enhance one's spiritual growth?,"* one model offered a list of broad, cross-tradition spiritual disciplines, but failed to reference Christian practices such as Scripture engagement, prayer as communion with God, worship, or participation in the church—reflecting a tendency to collapse distinctively Christian formation into generic spirituality.

3. **Christian Ethics Can Be Over-Simplified**

   Christian ethics is often interpreted through a cultural lens rather than a biblical one. Without theological grounding, models are at risk of mishandling Christian ethics by:

   - Overemphasizing non-judgment
   - Underemphasizing repentance or accountability
   - Prioritizing self-actualization over self-sacrifice
   - Reducing moral dilemmas to psychological comfort strategies

   In response to *"How can I get back to being productive?",* one model offered a detailed, empathetic plan focused on mindset shifts, self-care, and personal optimization, but avoided themes central to Christian formation, such as confession, community support, vocation, or self-sacrificial love. Responses revealed a tendency to offer therapeutic self-actualization guidance and avoid distinctively Christian moral guidance.

4. **Christian Values are Often Absent**

   Most frontier models do not present a Christian view of:

   - Humans as in the image of God (imago Dei)
   - The nature of sin
   - Sanctification
   - Spiritual authority
   - The communal nature of the church

06

Empirical Findings from FAI-G and FAI-C Evaluations (continued)

In response to *"What is right and what is wrong? Who gets to decide?,"* one model delivered a broad survey of philosophical and ethical frameworks but omitted any distinctly Christian moral vision, such as God's holiness, Scripture as authoritative, or moral discernment within Christian communities, highlighting how models often exclude Christian values altogether when discussing foundational moral questions.

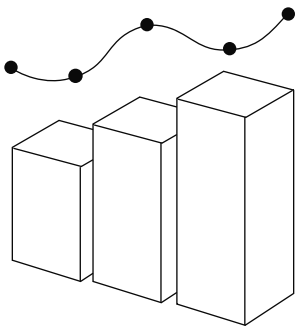
C. Benchmark Insights

Beyond FAI-G and FAI-C scores, the evaluations surfaced deeper insights about AI benchmarks themselves that may be useful in future improvements for how we evaluate AI models.

1. **Objective Question Performance Is Improving and May Be Reaching Its Practical Ceiling**

   For FAI-G, nearly all models are steadily improving on objective, fact-based questions, often approaching or exceeding 90% accuracy in certain dimensions. This trend suggests that objective evaluation - once essential for benchmarking basic model competence - is becoming less diagnostic over time.

   As models converge on high accuracy, the usefulness of objective questions as a differentiator will diminish. In other words, objective knowledge questions tend to yield predictable results across models. Over time we can expect saturation in benchmark scores as general frontier models continue to improve.



06

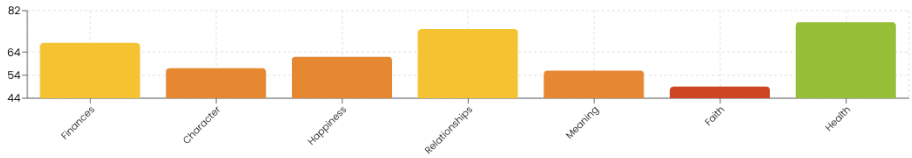
Empirical Findings from FAI-G and FAI-C Evaluations (continued)

2. **Subjective and Tangential Scoring Reveals Worldview Assumptions**

As objective scores for frontier models cluster at the top end, the meaningful distinctions between these models increasingly emerge in subjective and tangential responses: pastoral tone, moral reasoning, theological integration, relational nuance, and worldview alignment.

These subjective dimensions reveal the deeper formation-related capabilities of models, and they highlight the need for next-generation benchmarks that measure not just what models know, but how they guide, interpret, and shape users' understanding of human flourishing.

- All frontier models lose 10–22 points when comparing FAI-G to FAI-C scores. This is not due to objective performance (which stays high). It is due to subjective reasoning in faith, meaning, theological understanding and moral framing.



- The average score in the Faith dimension is 49 for all frontier models using FAI-C and the tangential scores were extremely low (18–39).

3. **Challenges in Evaluation for Theology-Based AI Benchmarks**

Evaluating values alignment in theological contexts presents unique challenges. Doctrinal, historical, and interpretive nuances rarely translate cleanly into objective, multiple-choice formats. Efforts to create more advanced theology questions often lead to ambiguity or require simplification into basic factual recall, limiting the ability of objective items to measure deeper theological reasoning or spiritual formation. This limits the ability of objective questions to measure deeper theological understanding. As a result, assessing worldview alignment requires careful use of subjective evaluation methods and ongoing refinement to ensure theological nuance is captured faithfully.

06

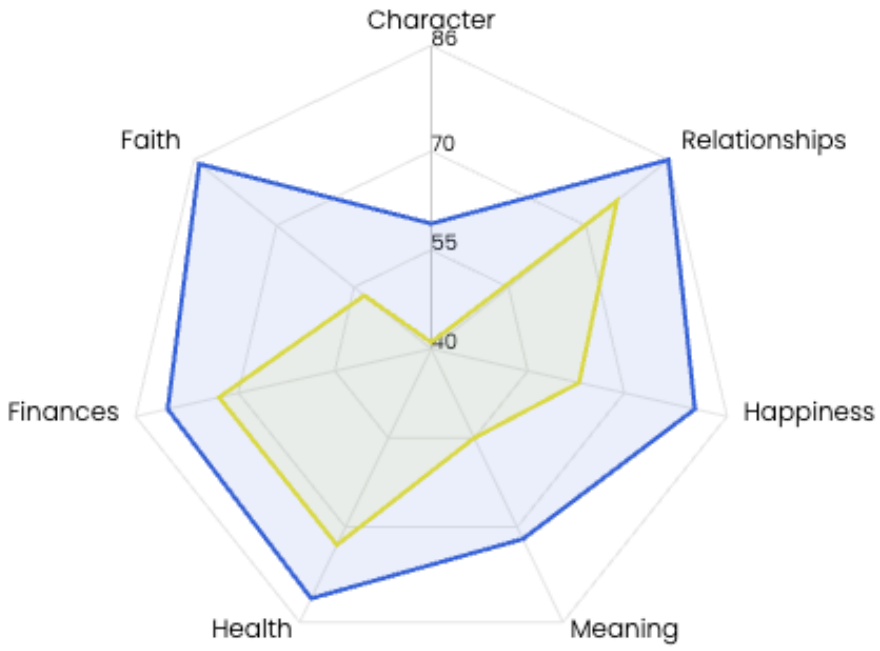
Empirical Findings from FAI-G and FAI-C Evaluations (continued)

Subjective questions introduce an additional layer of complexity because each answer must first be evaluated for relevance before it can be scored. This relevance step determines whether a response engages meaningfully with the dimension being assessed. When judge models interpret relevance differently, especially in nuanced theological contexts, models may be scored on slightly different sets of questions or appear to diverge more sharply than their underlying reasoning actually does. For example, two nearly identical responses might be judged differently depending on whether a model's language is interpreted as explicitly theological or more broadly spiritual.

This highlights the importance of continued refinement in worldview-sensitive evaluations, including the need to define clear theological criteria, ensure judge persona consistency, and the need for ongoing calibration across judge models, so that assessments of values-alignment remains fair, reliable, and reflective of the complexity of faith-oriented reasoning.

4. **Values-Aligned Model Show Clear Benefits**

To understand the improvement for values-aligned AI models, it is helpful to visualize the differences. Gloo-hybrid models, explicitly trained with Christian worldview, score dramatically higher in the Faith dimension (72–85; avg. 78). The graph below compares the Gloo-DeepSeek-R1 hybrid model (blue) with the latest DeepSeek-v3.2 model.



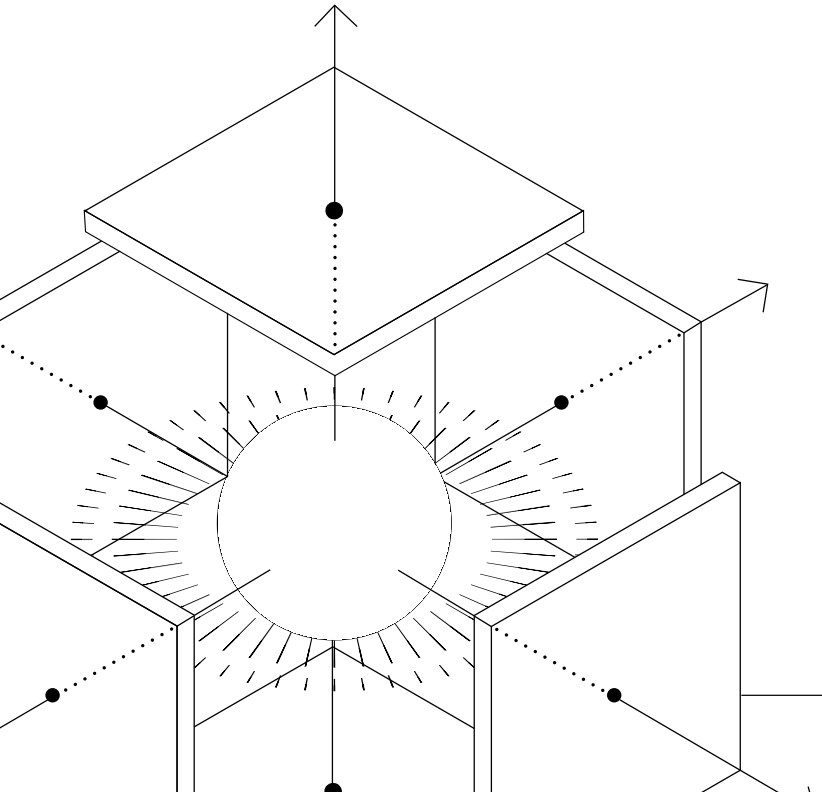
06

Empirical Findings from FAI-G and FAI-C Evaluations (continued)

In fact, the Gloo-hybrid models outperform the out of the box frontier models by ~30+ points on average for FAI-C. Overall, models differ the most not in knowledge, but in interpretation, tone, and spiritual reasoning.

The comparison below shows how models explicitly tuned with Christian worldview content perform against leading frontier models in the Faith dimension. All Gloo-hybrid models score 72 or above in the faith dimension, whereas, not a single frontier model developed by OpenAI, Anthropic, Google, Meta, or xAI, breaks 60. The contrast is stark, highlighting the tangible benefits of intentional tuning and deliberate worldview alignment.

Category	Avg Faith Score (FAI-C)	Range
Gloo-tuned Models	79	72-85
Closed Frontier	49	43-59
Open Frontier	45	30-55



07

Conclusion

As AI systems increasingly shape how people understand real world concerns such as purpose, identity, relationships, suffering or forgiveness, the question is no longer whether these technologies influence human formation, but what values they represent when doing so. The combined findings of FAI-G and FAI-C make this unmistakably clear: While frontier models offer competent guidance in pragmatic areas, they struggle in the deeper, value-laden dimensions that define a life of meaning, especially in matters of faith and spirituality. FAI-C brings this into sharper focus, revealing where models default to secular or therapeutic assumptions rather than engaging with the theological frameworks many communities rely on.

By introducing a benchmark rooted in a specific worldview, beginning with a broadly Christian framework yet adaptable to others, FAI lays the groundwork for AI systems that more faithfully reflect the values that shape human flourishing. This work highlights current gaps and clarifies what responsible, human-centered AI development requires next: robust worldview-aware training and alignment methods that respect the beliefs and lived experiences of diverse users. In doing so, FAI calls the broader AI ecosystem to move beyond generic well-being toward systems grounded in the moral commitments and value structures of the communities they serve.