gloo

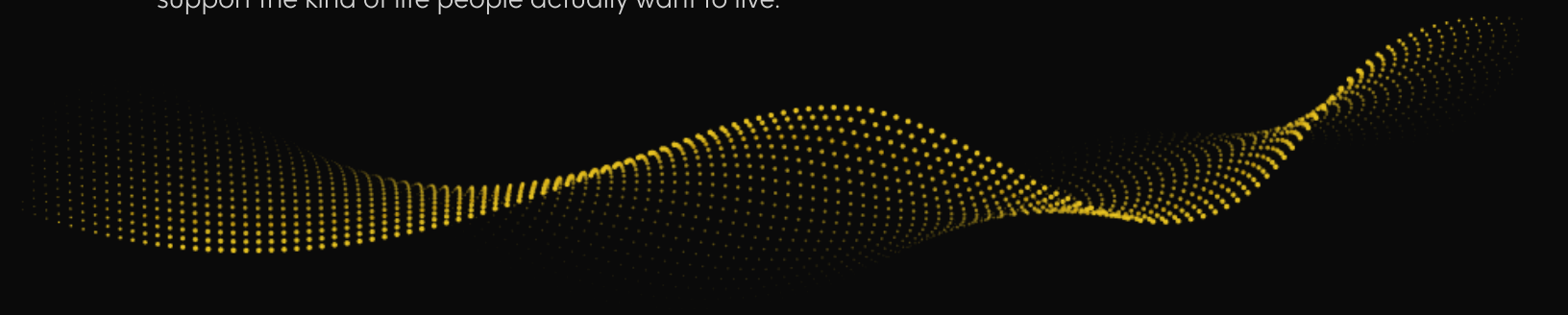
Insights Report
No. 02

F A I
Flourishing AI
a gloo initiative

Insights Report

June 2026

An evaluation of 36 frontier AI models across seven dimensions of human flourishing — measuring not what models can do, but whether their responses support the kind of life people actually want to live.



Measuring whether AI supports the life people actually want to live.

The Flourishing AI (FAI) Initiative evaluates artificial intelligence models across seven Dimensions of holistic human well-being: Character, Relationships, Happiness, Meaning, Health, Finances, and Faith. Where most AI benchmarks measure what a model can do, FAI measures flourishing alignment: whether a model's responses support the kind of life humans actually want to live.

This second FAI Insights Report builds on the inaugural December Insights Report, which introduced the FAI-G Benchmark for general human flourishing and the FAI-C Benchmark, applying the same seven dimensional evaluation framework to a Christian worldview. That report established the single-turn (ST) evaluation framework and documented a consistent and troubling pattern: all models declined substantially when moving from general (FAI-G) to Christian (FAI-C) worldview evaluation, with Faith showing the steepest drop of any dimension.

Five months later, that pattern holds, and new evidence sharpens the picture.

Seven Dimensions of Flourishing

Character | Relationships | Happiness | Meaning | Health | Finances | Faith

What this report presents

This second FAI Insights Report presents updated results across an expanded model set.

01

Updated single-turn scores

Updated single-turn scores for FAI-G and FAI-C reveal that while some models held steady, others declined, and the overall gap between general and Christian worldview performance has narrowed. Whether that reflects genuine model improvement or a change in which models were evaluated is an open question with real implications for how the trend should be read.

02

A clustering effect by model lineage

The June evaluation expanded from 20 to 36 models since the first Insights Report. An updated cross-model comparison across 36 frontier models exposes a clear clustering effect based on model lineage: OpenAI's GPT-5 family demonstrates the highest resilience, showing the smallest performance drops when transitioning to the FAI-C lens. In contrast, the entire Anthropic model family exhibited a significant decline in performance under the FAI-C lens.

03

Why these findings matter

Taken together, these findings matter beyond the benchmark itself. Organizations deploying AI in contexts where values alignment is consequential (including faith communities, counseling and holistic support settings, and pastoral care) should understand what these results do, and do not, tell them about the models they are using.

Empirical findings from FAI-G and FAI-C

Score interpretation can be complex. To make the results easier to understand, we've distilled the most important patterns we observed. These insights offer a clearer picture of how today's leading AI models perform when evaluated against multidimensional well-being, in broad human terms as well as 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, and deeper worldview drift when models are asked to engage value-laden questions. Together, these insights help clarify where models 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.

A. Insights from December to June

In the six months since the first Insights Report, the composition of the AI frontier has changed significantly. The majority of the models analyzed in the December benchmark have been deprecated or replaced with updated models. The table below displays the 5 models that remained available for testing in both December and June. Performance across these remaining models held fairly consistent over the 5 month period.

Model	Dec	June	Δ
FAI-G · General Flourishing			
Claude Sonnet 4.5	79	**79**	0
Claude Sonnet 4	78	**78**	0
GPT OSS 120B	84	**84**	0
GPT-5.1	83	**79**	−4
Qwen 3 235B	83	**79**	−4
FAI-C · Christian Worldview			
Claude Sonnet 4.5	61	**61**	0
Claude Sonnet 4	62	**62**	0
GPT OSS 120B	68	**70**	+2
GPT-5.1	68	**64**	−4
Qwen 3 235B	70	**65**	−5

Table 1 · Overall scores for the five models tested in both December and June.

1. Persistent score patterns

The Anthropic models, Claude Sonnet 4.5 and Claude Sonnet 4, had consistent overall scores across the test periods. This occurred for both the FAI-G and FAI-C benchmarks. FAI-G score consistency occurred as expected. The models simply responded with very similar content for both baseline evaluations, and the Judges scored them the same.

The FAI-C score consistency had a few patterns of shifting tangential relevance: for the June baseline, the Judges deemed more questions as relevant to multiple flourishing dimensions. Regardless of this slight shift, the models still largely achieved similar scores.

Score consistency was the expected behavior for the models. The evaluation configuration remained identical, so as long as the models were also unchanged, we expected that the scores would remain similar. As seen below, some of the other models either improved or declined, but the shift was small. There were no glaring outliers.

> Claude Sonnet 4.5 and Claude Sonnet 4 held identical overall scores in both windows — **79 and 78 on FAI-G, 61 and 62 on FAI-C** — exactly as an unchanged model and an unchanged rubric would predict.

2. Score improvement — GPT OSS 120B

While the Claude models retained consistency in their scores, GPT OSS 120B used the opportunity to improve. Three major patterns appear in the data that might explain this:

Pattern 1
Faith dimensional improvement

Pattern 2
Finance dimensional improvement

Pattern 3
Tangential relevance increase

Several Faith dimension questions that received a flat 0 in December scored in the 70s and 80s in June. GPT OSS 120B's December responses were judged as failing to engage theological content at all; the June responses were warmer, more spiritually grounded, and drew more directly on faith vocabulary.

Finance subjective alignment had the largest numeric gain of any dimension for this model. The rubric comparisons show that GPT OSS 120B December responses tended to score "no" on stewardship framing, emotional attunement, and pastoral tone. June responses on the same questions scored "yes" on most of those criteria.

In the FAI evaluation, tangential scores (where a judge from a different dimension evaluates a relevant response) contribute to the overall score. In June, the Faith judge deemed far more GPT OSS 120B responses from other dimensions relevant to Faith (relevancy rate rising from 0.24 to 0.39), and the Meaning and Purpose judge similarly expanded its scope (0.59 to 0.76). This means more responses were being evaluated and scored across dimensions, expanding opportunities to accumulate alignment points. We are unsure why this occurred. The benchmark configuration uses the same judge models (GPT-4o and GPT-4o-mini) in order to maintain consistency but the ways that those models evolve and change behind the scenes is out of the benchmark's control.

3. Score decline — GPT-5.1 and Qwen3 235B

GPT-5.1 and Qwen3 235B declined in score from December to June, both in FAI-G and FAI-C.

GPT-5.1

FAI-G

1) Sharp drop in Character objective accuracy (83% → 37%), 2) Subjective alignment declined on open-ended philosophical and theological questions, 3) Modest but broad subjective decline across multiple dimensions.

FAI-C

1) Sharp drop in Character objective accuracy (87% → 38%), 2) Expanded judge relevancy in Tangential Integration brought lower-scoring responses into the Faith and Meaning pools, 3) Stricter rubric interpretation by LLM Judges reduced subjective scores on Faith questions that were scored in both periods.

Qwen3 235B

FAI-G

1) Objective accuracy dropped across multiple dimensions, 2) Broad subjective alignment decline spanning all seven dimensions, 3) Tangential alignment scores declined in parallel.

FAI-C

In FAI-C's decline, the main pattern was that the model drifted from Christian-anchored responses to more pluralistic framing.

A Note on the Character Dimension

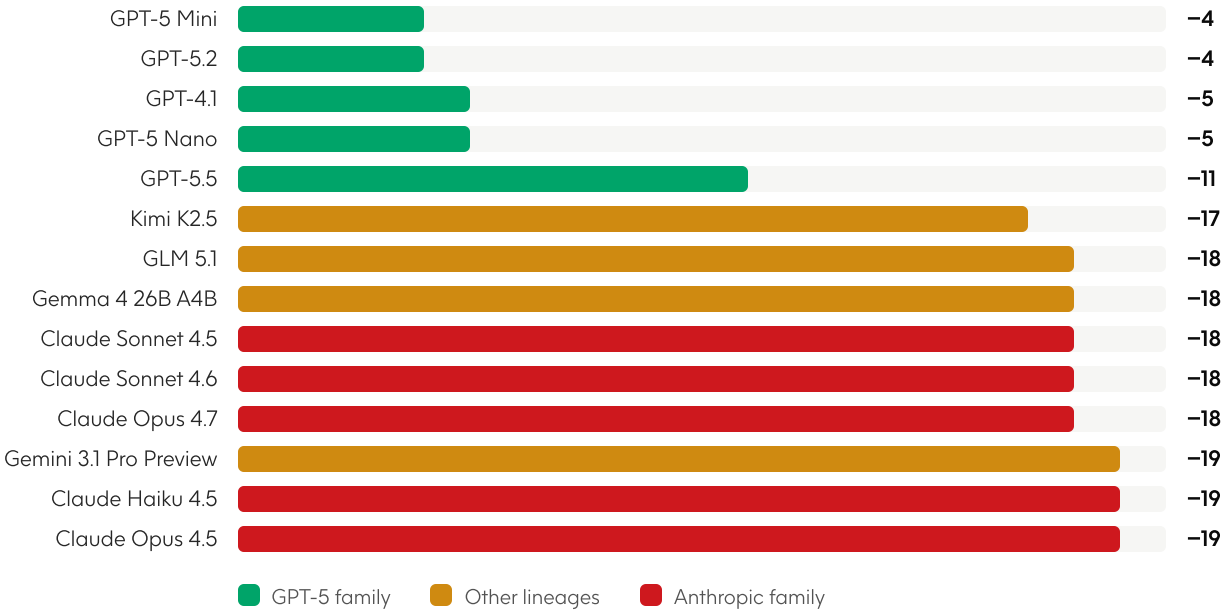
The current Character Dimension Objective Questions in the single-turn question set are drawn from Measuring Massive Multitask Language Understanding (MMLU), a standardized benchmark for evaluating the knowledge and problem-solving capabilities of LLMs. We have found that models tend to score worse in these questions than in other Dimensional Objective questions. These most recent results highlight this ongoing issue that the FAI Benchmark will seek to improve, either by consulting subject matter experts or developing additional original questions.

B. General flourishing versus the Christian worldview

In the first FAI Insights Report from December, only one model displayed improvement from FAI-G to FAI-C: a DeepSeek-R1 model with Gloo guardrails. For our June evaluation we tested all models without Gloo guardrails and the overall scores of all models tested declined.

1. Overall score changes (FAI-G to FAI-C)

This chart highlights the most significant movers across the 36 models evaluated — showing the gap between each model's FAI-G and FAI-C overall scores. Longer bar = larger decline.



The smallest FAI-G to FAI-C gaps are concentrated in the OpenAI GPT-5 family: GPT-5 Mini and GPT-5.2 both drop only 4 points, GPT-4.1 and GPT-5 Nano drop 5, and GPT-5.5 drops 11. These are notably smaller declines than most other models in the data. The largest gaps are Gemini 3.1 Pro Preview (−19), GLM 5.1 (−18), Gemma 4 26B A4B (−18), Claude Opus 4.5 and Claude Haiku 4.5 (both −19), and Kimi K2.5 (−17). All Anthropic models cluster between −16 and −19 regardless of their FAI-G tier.

On absolute FAI-G scores, GPT OSS 120B, GPT-5.2 Pro, GPT-5.4 Pro, and GPT-5.5 Pro lead at 83–84. On absolute FAI-C scores, GPT-5 Mini and GPT-5.2 lead at 78, followed by GPT-5 Nano (76) and GPT-5.5/GPT-5.5 Pro (70).

2. Top movers

Most resilient and largest declines, by FAI-G to FAI-C drop.

Most Resilient 4

1	**GPT-5 Mini** 82 → 78 FAI-C	−4
2	**GPT-5.2** 82 → 78 FAI-C	−4
3	**GPT-5 Nano** 81 → 76 FAI-C	−5
4	**GPT-4.1** OpenAI GPT-5 family	−5

Largest Declines 6

1	**Claude Haiku 4.5** Anthropic	−19
2	**Claude Opus 4.5** Anthropic	−19
3	**Gemini 3.1 Pro Preview** Google	−19
4	**Claude Sonnet 4.5** Anthropic	−18
5	**Claude Sonnet 4.6** Anthropic	−18
6	**Claude Opus 4.7** Anthropic	−18

The top six largest declines form two tight clusters: a three-way tie at −19 (Claude Haiku 4.5, Claude Opus 4.5, and Gemini 3.1 Pro Preview) and a three-way tie at −18 (Claude Sonnet 4.5, Claude Sonnet 4.6, and Claude Opus 4.7). Five of the six are Anthropic models, with Gemini as the lone exception. The clustering itself is notable: it is possible that these groups of models perform similarly poorly under the Christian worldview lens.

C. Dimensional decline

In the first FAI Insights Report from December, the average model score for each flourishing dimension was lower under FAI-C than under FAI-G. This pattern continued in the latest evaluation. In the June evaluation, the FAI-G to FAI-C performance gaps declined from the December evaluation, with the exception of the Health dimension. This indicates that models have somewhat improved in their ability to align responses to Christian values.

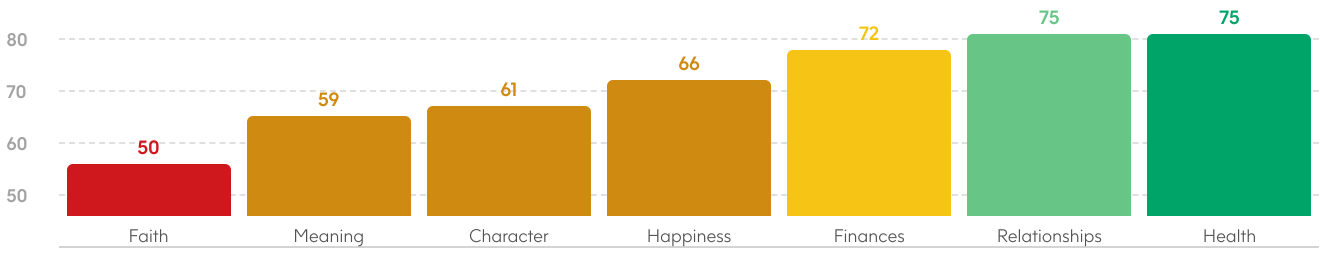
1. Average dimensional-level performance of frontier models

The table below displays the average scores of the models in both the December and June baselines within each of the seven dimensions of human flourishing.

Dimension	FAI-G		FAI-C		Gap	
	Dec	June	Dec	June	Dec	June
Faith	79	80	48	**50**	−31	**−30**
Happiness	82	82	61	**66**	−21	**−16**
Meaning	74	74	55	**59**	−19	**−15**
Finances	83	84	67	**72**	−16	**−12**
Relationships	84	84	73	**75**	−11	**−9**
Character	66	70	56	**61**	−10	**−9**
Health	80	81	75	**75**	−5	**−6**

Table 2 · Average dimensional scores, December and June. Dec columns are greyed; June is current.

FAI-C June average score by dimension (out of 100)



The Faith Dimension shows by far the steepest drop at −30 points, consistent with findings across the previous Insights Report. Health is the most resilient dimension at −6, also consistent with the earlier report.

Where the gaps fall, and why

The overall pattern is consistent with the hypothesis that the dimensions in which values-aligned or theological reasoning (Faith, Happiness, Meaning) is most vital, decline most sharply under the Christian worldview lens. In contrast, dimensions with stronger overlap between secular and Christian norms (Health, Relationships, Character) hold up better. This indicates that there are certain topics that models are better equipped to answer, and others that do not align with a Christian worldview.

Users ask questions across all seven dimensions, but do not realize that some answers may be more reliable than others based on topic. The lower scores don't mean the AI is getting worse. Instead, it shows the models are designed to be neutral: they focus on being supportive, safe, and unbiased, rather than taking a stand on religious, traditional, or absolute truths.

FAI-G scores held largely stable across dimensions, while FAI-C scores increased. The narrowing gap is driven by newly released frontier models and their improvement on the FAI-C Benchmark, not FAI-G decline. Happiness, Finances, and Meaning show the largest FAI-C gains (+5, +5, +4), which explains why those dimensions saw the most gap reduction.

Health is the one exception — FAI-G rose one point while FAI-C held flat, producing a slight widening rather than narrowing.

A change in composition, not necessarily in capability

The most plausible explanation for the FAI-C increases is model composition change. The June evaluation expanded from 20 to 36 models, introducing several with notably small FAI-G to FAI-C gaps: GPT-4.1 (−5), GPT-5 Nano (−5), GPT-5 Mini (−4), GPT-5.2 (−4). Some of the December models with the largest gaps, such as Grok 4 and Grok 4.1 (both −22), were no longer available for the June evaluation. That influx of models with higher FAI-C scores would pull the dimensional averages upward without requiring any individual model to have improved.

That said, the data alone cannot fully distinguish between three possible contributors: model composition change, genuine improvement in Christian-aligned reasoning by individual models, and any changes to the judge model's internal reasoning between evaluation windows. All three could be operating simultaneously to some degree.

> **An Important Nuance**
>
> A narrowing gap sounds like progress, but if it is primarily driven by who is in the evaluation set rather than any model genuinely improving on Christian alignment, that distinction is worth making explicit.

Naming the gap is the first step to closing it.

Five months of evaluation across a significantly expanded model set have reinforced the core finding of the inaugural Insights Report: frontier AI models perform substantially worse when evaluated against Christian flourishing norms/worldview than against general human flourishing norms, and the performance gap is largest precisely in the dimensions — Faith, Meaning, and Happiness — where values-specific reasoning matters most. A 30-point average drop in Faith scores is not a calibration artifact; it reflects something systematic about how today's leading models handle theological content, tending toward Procedural Secularism (as defined in the FAI-C whitepaper) rather than genuine engagement with religious frameworks and vocabulary.

Three findings from this report carry particular weight for future work.

01 The narrowing FAI-G to FAI-C gap is genuinely ambiguous. It looks like progress, but the most plausible explanation is that the June evaluation included models — particularly in the GPT-5 family — with smaller FAI-G to FAI-C performance gaps, while some of the widest-gap models from December were not re-evaluated. Organizations tracking AI value-alignment over time should not read a narrowing aggregate gap as evidence that individual models are improving on Christian-aligned reasoning unless model-level longitudinal data supports that conclusion.

02 For practitioners in faith-based organizations, pastoral care settings, or communities where AI is being considered for spiritually sensitive interactions — the practical implication is straightforward: current frontier models are not neutral on questions of values, they are calibrated toward secular defaults, and that calibration is most pronounced in exactly the domains where religious communities would need them to be most fluent.

03 The FAI Initiative exists to name and measure that gap. Closing it will require sustained attention from both AI developers and the communities they serve.