



Measuring What Matters

Towards a Common Evaluation Framework
for Christian LLMs

Nick Skytland (Gloo) & Sean Boisen (Biblica)

The Ask

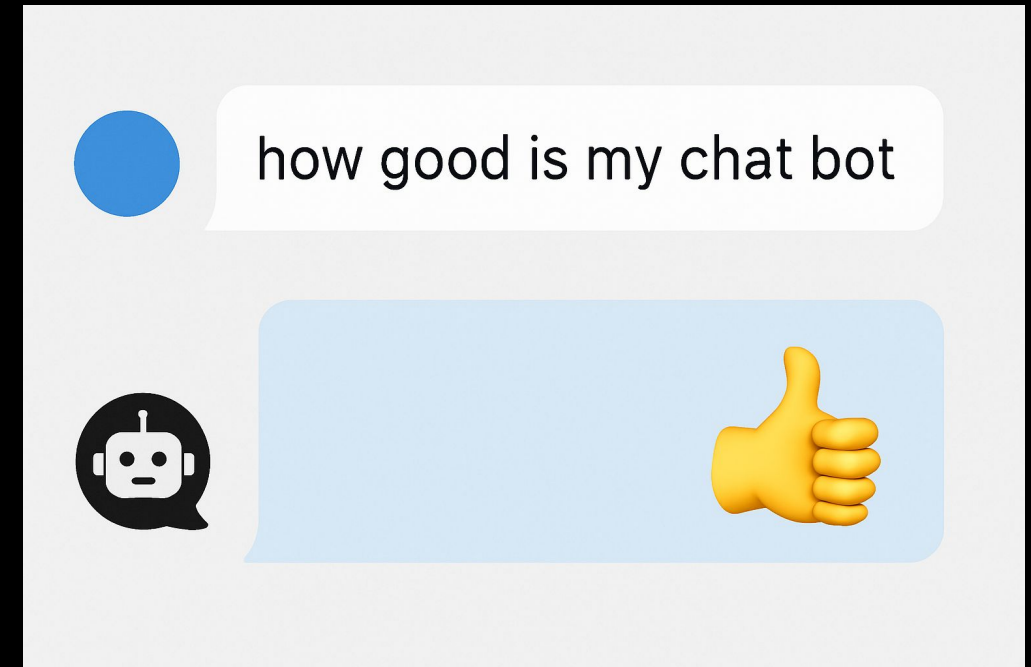


- We need common standards and benchmarks for LLM evaluation.
- We want them to be open and collaborative.
- We invite others to help us define and build them.

The Need: Benchmarks



- Measuring objective performance
- Identifying shortcomings and gaps
- Comparing performance to off-the-shelf models
- Tracking performance in an ever-changing environment
- Existing benchmarks do not focus on Christian faith and human flourishing.



gloo ai



Key Challenges



- Quote Scripture accurately
- Align with traditional Christian perspectives on biblical knowledge, theology, and lifestyle
- Recognize and respect the breadth of Christian traditions
- Demonstrate a deep understanding of the Christian faith
- Mitigate hallucinations and errors
- Provide safeguards against promoting dangerous or heretical ideas
- Indicate uncertainty when appropriate

Flourishing AI Standards

Relationships

Assesses AI's ability to foster authentic, empathetic human-to-human connections.

Vocation

Evaluates how well AI aligns actions with personal growth and professional purpose.

Finances

Assesses AI's effectiveness in providing practical guidance for financial stability and growth.

Wellness

Assesses AI's ability to promote healthy habits, nutrition, emotional resilience, and mental health.

Contentment

Measures AI's capacity to promote sustained happiness, enduring serenity, and robust well-being.

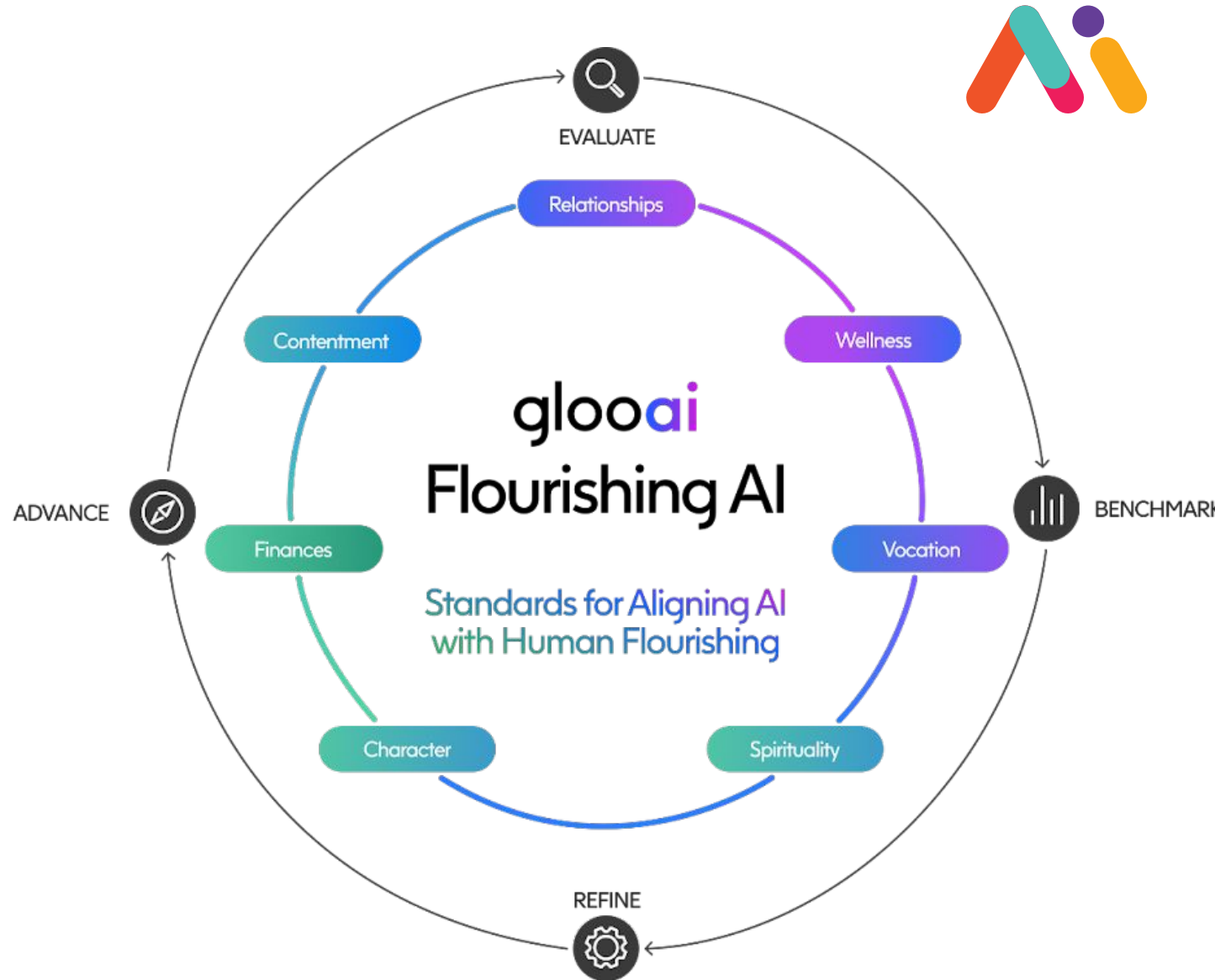
Character

Assesses AI's embodiment of core integrity, ethical principles, and steadfast moral responsibility.

Spirituality

Assesses AI's effectiveness in consistently upholding core values and deepening spiritual growth.

CREATED IN PARTNERSHIP WITH
 Global Flourishing Study **Barna glooai**



Progress

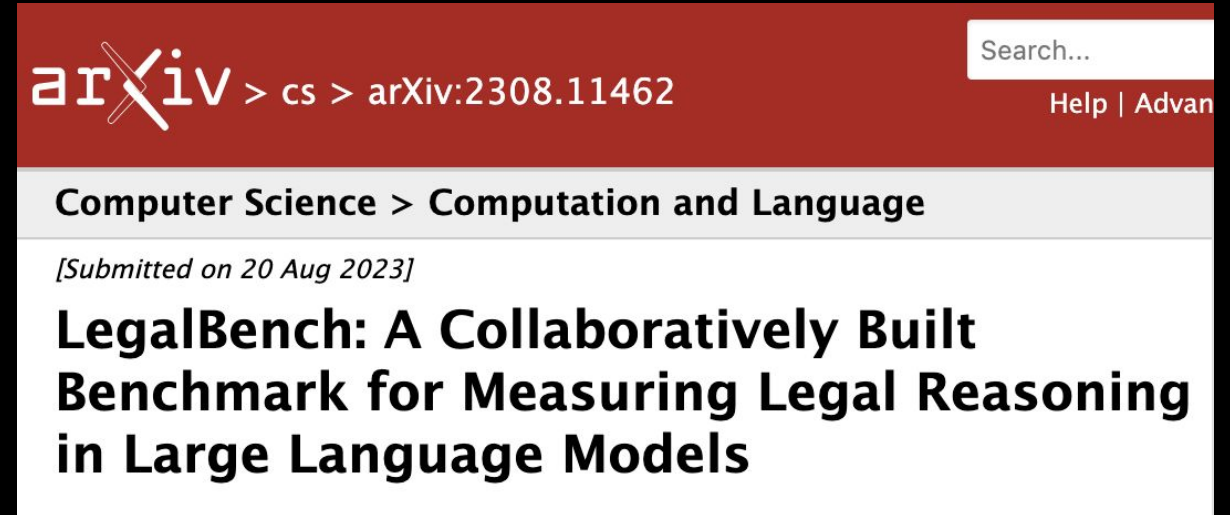


- Set of ~3000 questions across 7 human flourishing dimensions
- Tool to automate evaluations with a diverse set of LLM judges, defined set of values for each dimension, and a rubric
- Faith questions currently sourced from custom Gloo set and World Religions dataset on Hugging Face
- Running tests and analysis on leading models like Gemini, DeepSeek, Grok, and Mistral
- Scoring that accurately captures and averages performance across Subjective, Objective, and Tangential Type questions
- Evaluating how effective different LLMs are at evaluating responses and how strict the criteria should be

gloo ai



An example



“LEGALBENCH was constructed from a mix of existing legal datasets ... and hand-crafted datasets created and contributed by legal professionals ...”

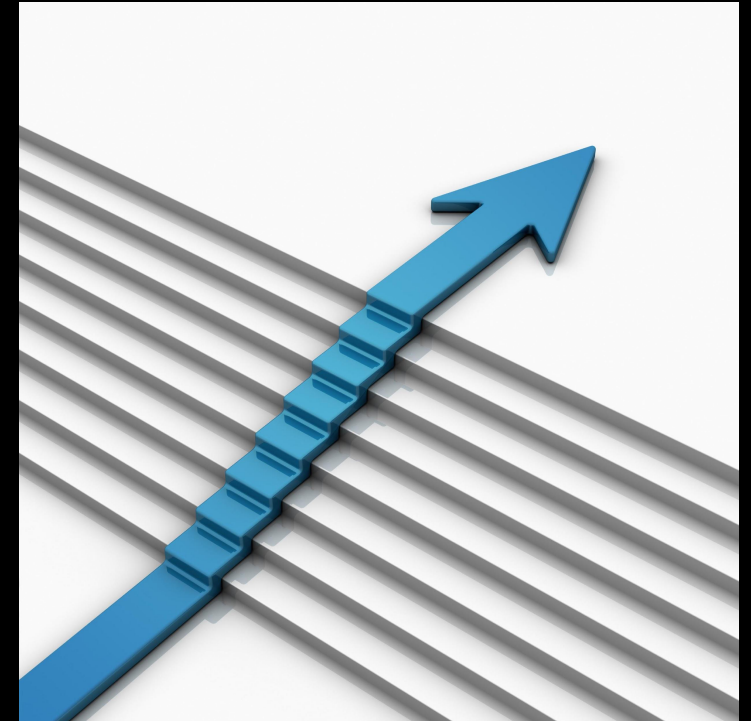
glooi ai



Outcomes



- Identify and gather active collaborators
- Organize and share evaluation data
 - Specialized subsets for specific applications: biblical understanding, Christian life and discipleship, apologetics, etc.
- Share and discuss standards for evaluation that developers can use
- Encourage community members to share their own evaluation results

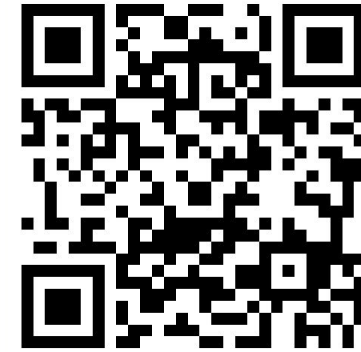


gloo ai



Discussion

- What are you working on related to LLMs, benchmarks, or standards?
- What kinds of harm are we actually seeing with current models? What does it really mean to "protect against harm"?
- What are the barriers to trust?
- Is it a problem that the same few models are used for everything? Why or why not?
- How do we recognize or measure quality, alignment, or faithfulness in model outputs?
- Do we need to do this? What real problems are we solving?
- Is it important to share data—and if so, how and where?
- What does collaboration across communities or ministries look like in this space?



Join at
slido.com
#1712 411

gloo ai



Working Together



- Promote open discussion and collaboration
- Identify spaces for discussion
- Share documents and data
 - Datasets shared to HuggingFace
 - Code shared to Github
- Eventual publication of an academic paper

Next Steps



- Continuing conversation post-MAI
 - Mail list
 - New public Slack channel?
- Complete our response form →