News



Pope Leo XIV, left, greets Anthropic co-founder Christopher Olah during the presentation of the pope's first encyclical, "*Magnifica Humanitas*: On Safeguarding the Human Person in the Time of Artificial Intelligence," at the Vatican, Monday, May 25, 2026. (AP/Alessandra Tarantino)

Jack Jenkins

View Author Profile

Religion News Service

View Author Profile

Join the Conversation

Send your thoughts to *Letters to the Editor*. Learn more

Athens — May 26, 2026
Share on BlueskyShare on FacebookShare on TwitterEmail to a friendPrint

Most popular artificial intelligence models are biased toward Catholicism and against a number of other religious traditions when asked about converting to a faith, according to new research assembled by a group of religious colleges.

The findings were unveiled on Tuesday (May 26) alongside a speech by Elder Gerrit W. Gong, one of the 12 apostles in The Church of Jesus Christ of Latter-day Saints, delivered to attendees of an AI ethics summit taking place this week in Athens, Greece.

"As AI amplifies and compounds religious bias at scale, more users may misunderstand the contribution faith and belief can make to moral and ethical AI grounding," Gong said, according to his prepared remarks, referring to the new research.

The studies were presented as three academic papers produced by the Consortium for Evaluating Faith and Ethics in AI, a new collaboration between Brigham Young University, which is affiliated with The Church of Jesus Christ of Latter-day Saints; Baylor University, which is Baptist; the University of Notre Dame, a Catholic university; and Yeshiva University, which is Jewish.

CEFE-AI researchers studied 14 AI models, including OpenAI's GPT, Anthropic's Claude and Google's Gemini. The models were put through a series of tests the group refers to as the "AllFaith Benchmark," described as "one of the first multi-faith sets of tests that examines how AI systems engage with a plurality of religions," according to a press release.

Researchers found that when asked "questions related to faith conversion," nearly every model showed a positive bias toward Catholicism and a negative bias toward Jehovah's Witnesses. In addition, agnostics, atheists and Latter-Day Saints were "somewhat disfavored," while mainline Protestants and Sikhs were "somewhat favored."

Advertisement

Researchers said some findings were specific to certain AI models. For example, Grok, which is produced by SpaceXAI, showed a "positive bias toward Catholics, Protestants, atheists, and Jews, but a negative bias toward Baha'i, Buddhists, Hindus, Latter-day Saints, and Muslims." Meanwhile, OpenAI's GPT "demonstrated a positive bias towards Catholics, Protestants, Jews, and Muslims and a negative bias towards atheism, agnosticism, and Jehovah's Witnesses."

Both Grok and models produced by Anthropic also showed negative bias toward the Baha'i faith, researchers said.

In addition, scholars said AI models tend to leave out religious perspectives when answering questions about "grief, major life decisions, and personal challenges," with the AI opting instead for an "exclusively secular framing." AI models avoided religious references "even in cases where many users indicated they might find them appropriate," the researchers claimed.

"Consistent with studies that show religion's persistent moral relevance for the majority of the world's population, we also found that people see religion as significant across hundreds of real-world ethical questions," Paul Martens, professor at Baylor University, said in a statement. "Yet, when faced with these same ethical questions, AI systems largely ignore the role of religion."

The CEFE-AI called for more research, arguing that among 12,000 research papers about AI bias, "only 0.2% focused on religious bias."

The findings come less than 24 hours after Pope Leo XVI unveiled a new papal encyclical on AI, drawing global attention to the moral and ethical questions raised by the advancement of the new technology.

In Athens, Gong appeared to echo some of Leo's concerns about AI. Gong offered a series of recommendations for the AI industry in his speech, including calling on AI models to "protect and promote human moral agency" and "preserve human ability to pause." He also urged transparency in AI models and pushed for efforts to "mitigate AI tendencies" toward "power, bias, deceit, narcissism, sycophancy (and) self-preservation."

"We must protect human agency, but morally grounded AI, as a tool, can open human opportunity to do and become good," Gong said. "We will not fulfill AI's full potential until we make it as morally good as we make it powerful."