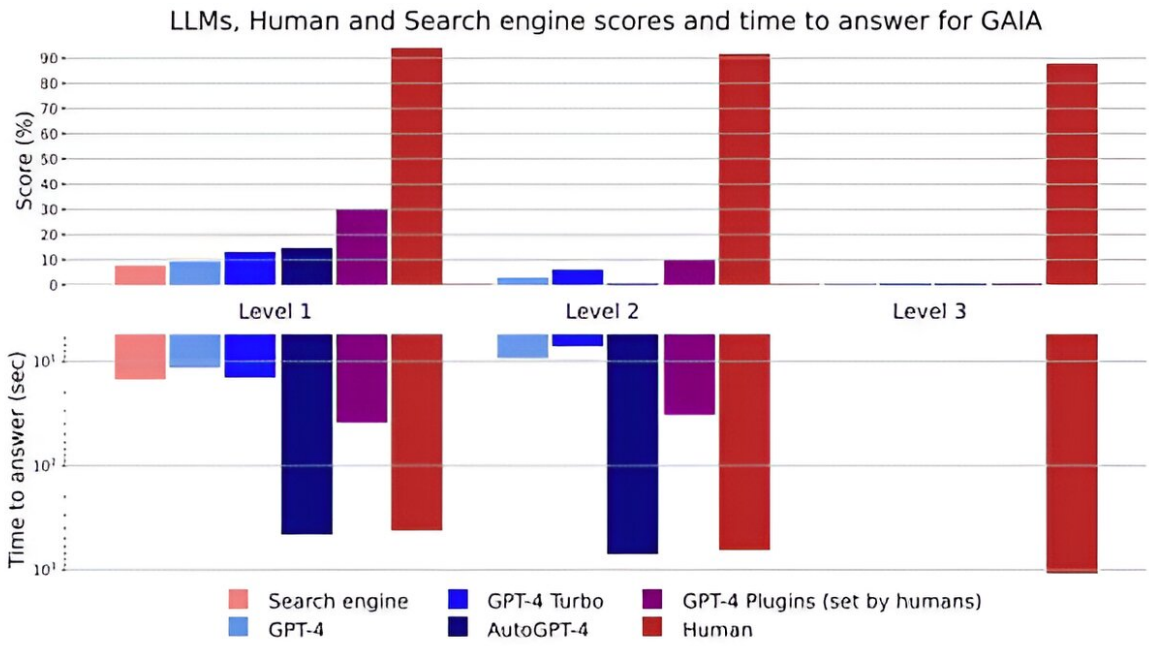
AI researchers introduce GAIA: A benchmark testing tool for general AI assistants

December 1 2023, by Bob Yirka



Scores and time to answer per method and level. GPT4 + plugins score should be seen as an oracle since the plugins were chosen manually depending on the question. Human score refers to the score obtained by our annotators when validating the questions. Credit: *arXiv* (2023). DOI: 10.48550/arxiv.2311.12983

A team of researchers affiliated with AI startups Gen AI, Meta, AutoGPT, HuggingFace and Fair Meta, has developed a benchmark tool

for use by makers of AI assistants, particularly those that make Large Language Model based products, to test their applications as potential Artificial General Intelligence (AGI) applications. They have written a paper describing their tool, which they have named GAIA, and how it can be used. The article is posted on the *arXiv* preprint server.

Over the past year, researchers in the AI field have been debating the ability of AI systems, both in private and on social media. Some have suggested that AI systems are coming very close to having AGI while others have suggested the opposite is much closer to the truth. Such systems, all agree, will match and even surpass human intelligence at some point. The only question is when.

In this new effort, the research team notes that in order for a consensus to be reached, if true AGI systems emerge, a ratings system must be in place to measure their intelligence level both against each other and against humans. Such a system, they further note, would have to begin with a benchmark, and that is what they are proposing in their paper.

The benchmark created by the team consists of a series of questions that are posed to a prospective AI, with answers compared against those provided by a random set of humans. In creating the benchmark, the team has made sure that the questions were not typical AI queries, where AI systems tend to score well.

Instead, the questions they pose tend to be the kind that are pretty easy for a human to answer but are difficult for a computer. In many cases, finding answers to the questions the researchers devised involved going through multiple steps of work and/or "thought." As an example, they might ask a question specific to something found on a specific website, like, "How far above or below is the fat content of a given pint of ice cream based on the USDA standards, as reported by Wikipedia?"

The research team tested the AI products they work with and found that none of them came close to passing the benchmark, suggesting the industry may not be as close to developing a true AGI as some have thought.

More information: Grégoire Mialon et al, GAIA: a benchmark for General AI Assistants, *arXiv* (2023). DOI: 10.48550/arxiv.2311.12983