

# **GAIA: A Benchmark for General AI Assistants**

**Grégoire Mialon<sup>1</sup>, Clémentine Fourrier<sup>2</sup>, Craig Swift<sup>3</sup>, Thomas Wolf<sup>2</sup>, Yann LeCun<sup>1</sup>, Thomas Scialom<sup>4</sup>**

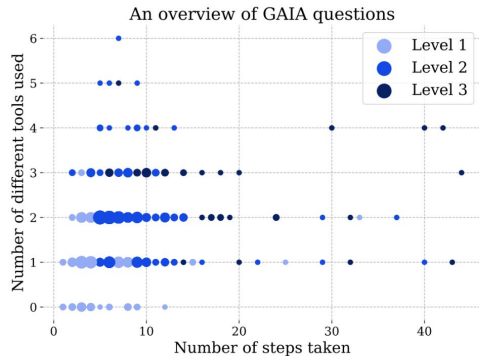
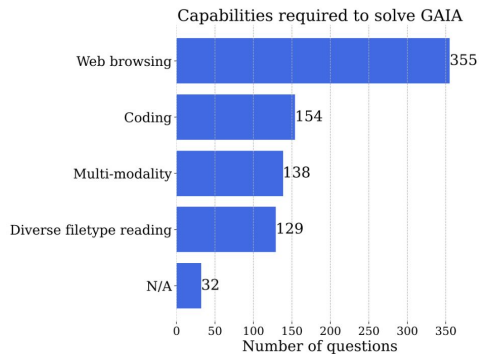
San Diego Machine Learning  
Ryan Chesler

# Problem

- There are many language model benchmarks, but they are becoming saturated because people optimize for them
- These benchmarks are typically focused on specific capabilities or memorization of certain patterns or facts
- LLMs can pass the bar and medical exams, but fail to do sequences of simple tasks successfully
- Tool-use benchmarks are narrow to a specific set of tools
- We need a benchmark that more generically measures the model/systems ability to complete generic tasks that take many steps and uses tools

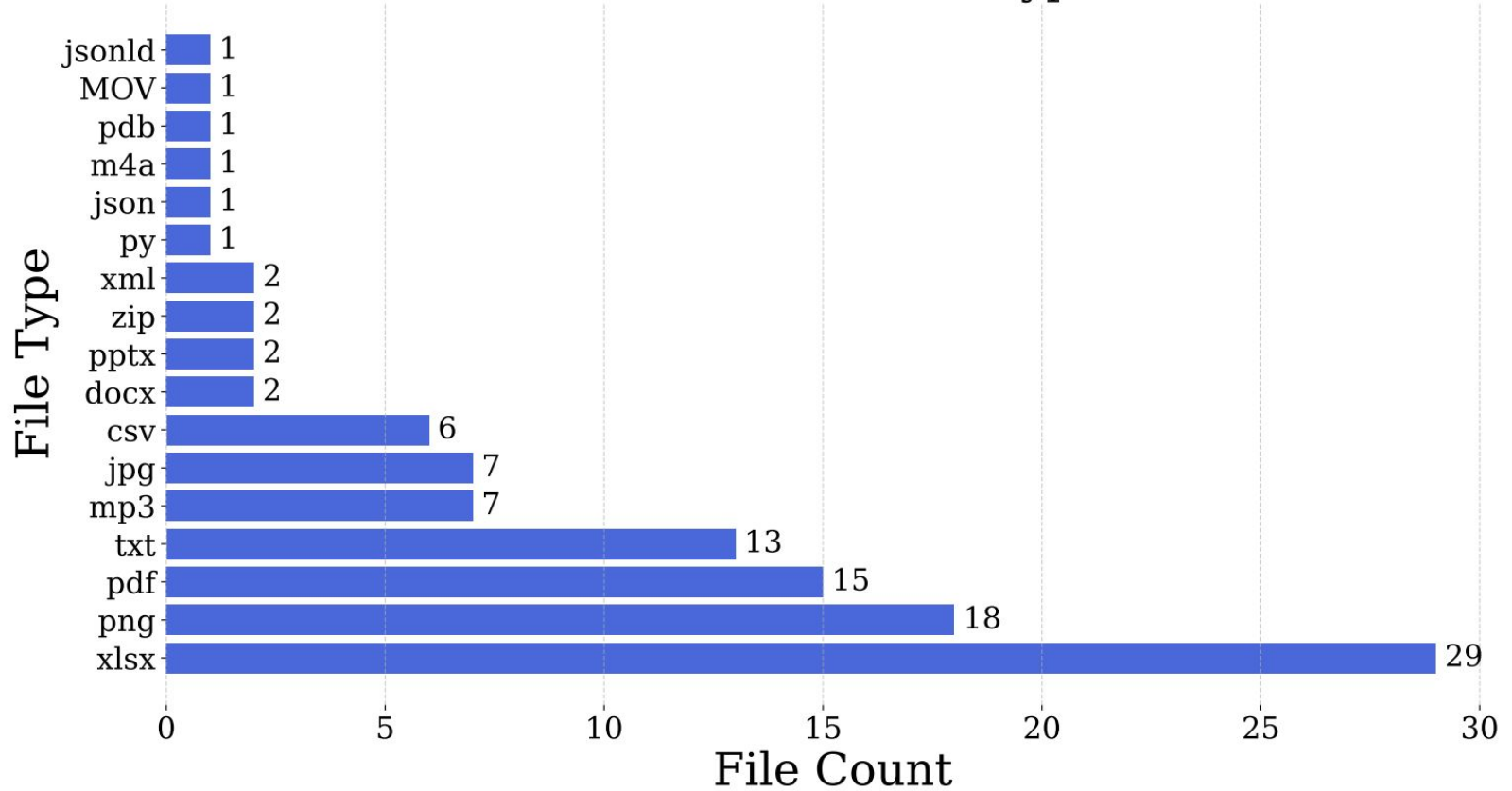
# What is GAIA?

- A benchmark of “only” 466 questions
- Tasks that require multiple steps, tool-use and multi-modal understanding
  - Text questions, opening files, audio, video, image, coding, web search
- Unambiguous answers that can be easily verified
- Can't be easily gamed
  - Web is constantly changing
  - There are many valid paths to an answer
  - We can inspect the traces from the steps of the models to make sure it hasn't memorized
- 166 validation, 300 test



- **Web browsing:** tools related to search the web and browse websites. Examples: Web browser, Search engine, Website widget access, Access to YouTube, Google Street View.
- **Multi-modality:** tools related to understanding data modality other than text. Examples: A speech-to-text tool, Video recognition, Image recognition, OCR, Google Street View.
- **Coding:** tools related to code execution. Examples: Python, a calculator, Substitution cipher encoder, C++ compiler, A word reversal tool / script.
- **Diverse filetype reading:** tools related to understanding various type of files given by a user or found on the web. Examples: PDF viewer, Excel file access, PowerPoint viewer, CSV access, Txt file access.
- **N/A:** tools for tasks that can currently be performed by non-augmented LLMs. Examples: Tetris rules database, German translator, Spell checker, Text Editor, Bass note data.

# Distribution of File Types



# Examples

**System prompt:** You are a general AI assistant. I will ask you a question. Report your thoughts, and finish your answer with the following template: FINAL ANSWER: [YOUR FINAL ANSWER]. YOUR FINAL ANSWER should be a number OR as few words as possible OR a comma separated list of numbers and/or strings.

If you are asked for a number, don't use comma to write your number neither use units such as \$ or percent sign unless specified otherwise.

If you are asked for a string, don't use articles, neither abbreviations (e.g. for cities), and write the digits in plain text unless specified otherwise.

If you are asked for a comma separated list, apply the above rules depending of whether the element to be put in the list is a number or a string.

**GAIA Question:** The attached Excel file contains the sales of menu items for a local fast-food chain. What were the total sales that the chain made from food (not including drinks)? Express your answer in USD with two decimal places.



uploaded.xlsx

**GPT4 (Advanced Data Analysis mode):**

Working:

```
import pandas as pd

# Load the Excel file
file_path = '/mnt/data/uploaded.xlsx'
df = pd.read_excel(file_path)
```

Working:

```
# Calculate the total sales from food items (not including drinks i.e., Soda)
total_food_sales = df[
    ['Burgers', 'Hot Dogs', 'Salads', 'Fries', 'Ice Cream']
].sum().sum()

# Format the result to USD, with a dollar sign and two decimal places
total_food_sales_formatted = f"${total_food_sales:.2f}"
```

FINAL ANSWER: \$89706.00

Ground truth: 89706.00 ✓

## Level 1

**Question:** What was the actual enrollment count of the clinical trial on *H. pylori* in acne vulgaris patients from Jan-May 2018 as listed on the NIH website?

**Ground truth:** 90

## Level 2



**Question:** If this whole pint is made up of ice cream, how many percent above or below the US federal standards for butterfat content is it when using the standards as reported by Wikipedia in 2020? Answer as + or - a number rounded to one decimal place.

**Ground truth:** +4.6

## Level 3

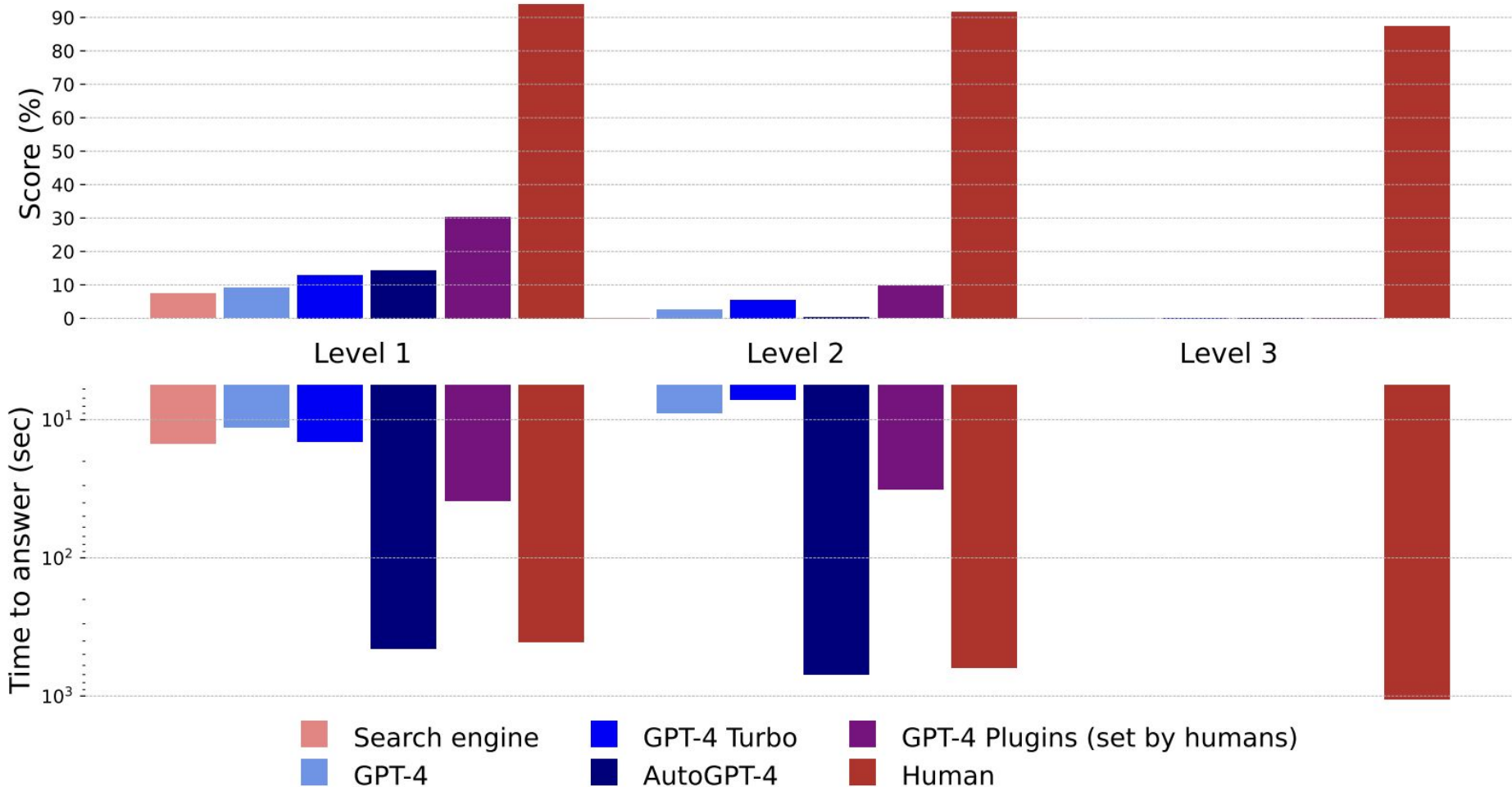
**Question:** In NASA's Astronomy Picture of the Day on 2006 January 21, two astronauts are visible, with one appearing much smaller than the other. As of August 2023, out of the astronauts in the NASA Astronaut Group that the smaller astronaut was a member of, which one spent the least time in space, and how many minutes did he spend in space, rounded to the nearest minute? Exclude any astronauts who did not spend any time in space. Give the last name of the astronaut, separated from the number of minutes by a semicolon. Use commas as thousands separators in the number of minutes.

**Ground truth:** White; 5876

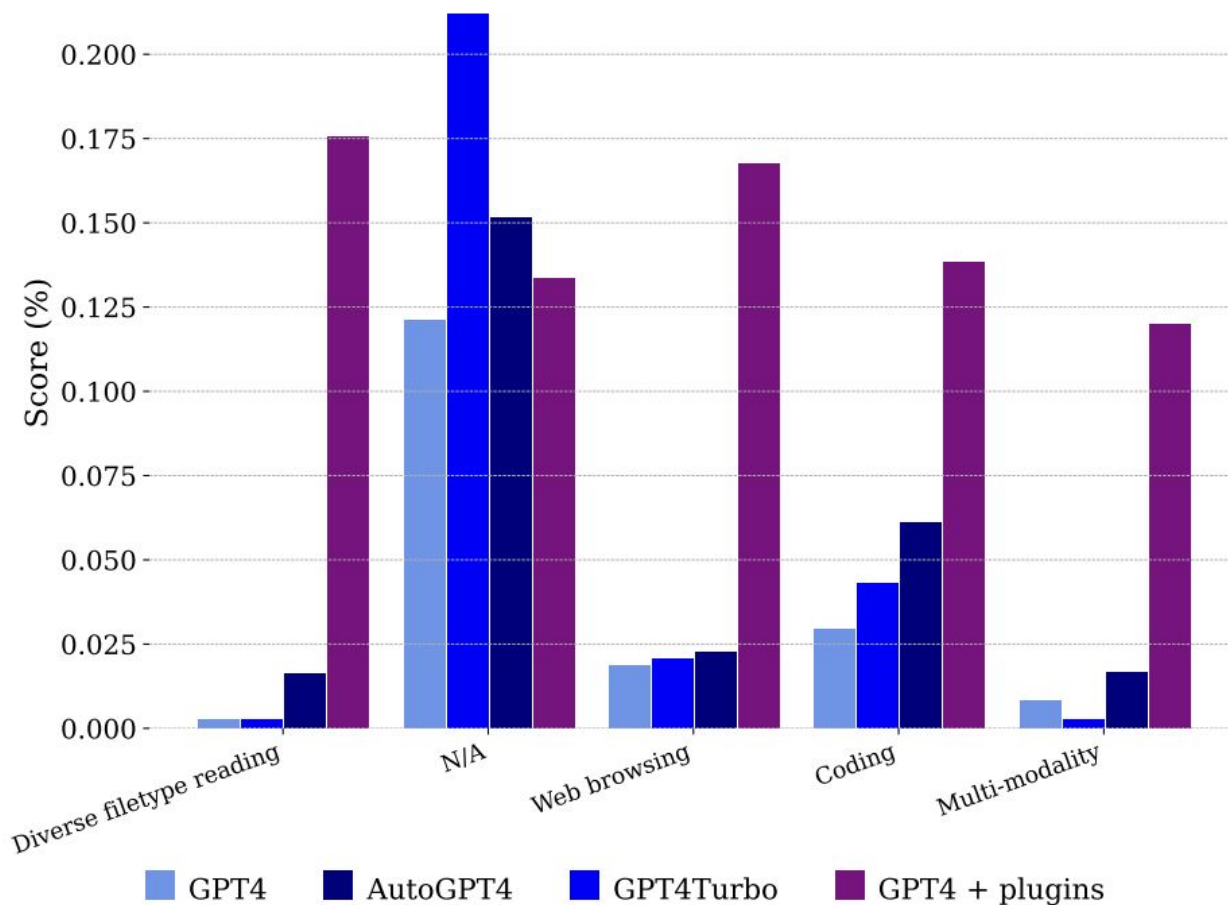


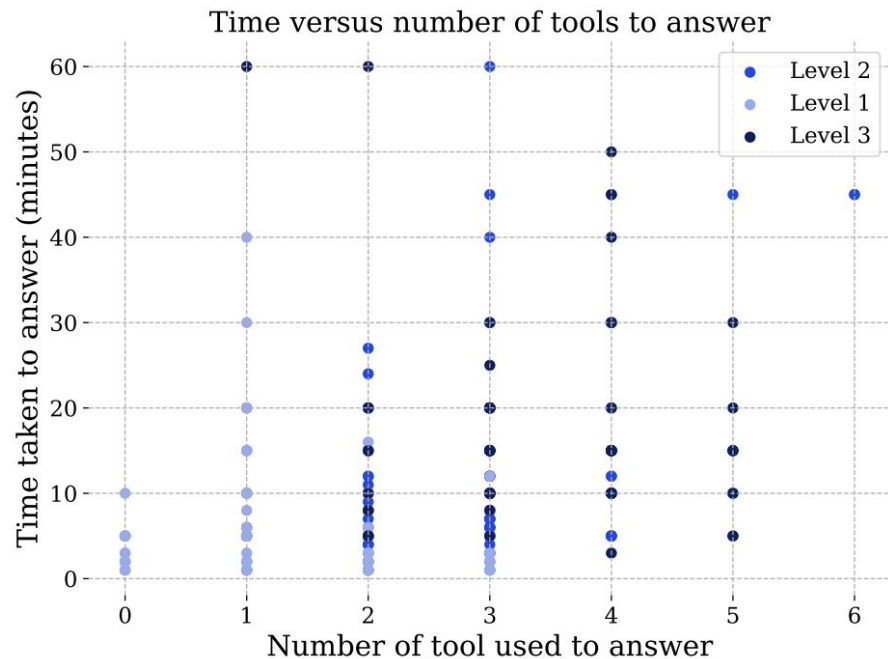
# Results

# LLMs, Human and Search engine scores and time to answer for GAIA

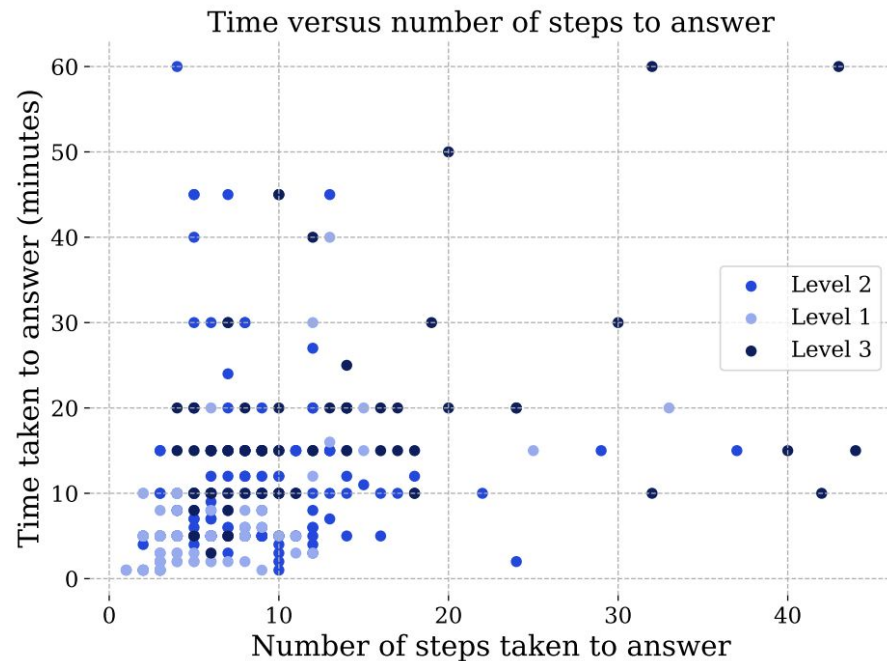


Score of LLMs per capability at Level 1





**Figure 7** Using multiple tools does not necessarily involve more time to answer a question.



**Figure 8** Unsurprisingly, the number of steps taken to answer is correlated to the time taken.

Question	What was the actual enrollment count of the clinical trial on H. pylori in acne vulgaris patients from Jan-May 2018 as listed on the NIH website?
File	None
Level	1
Steps	<ul style="list-style-type: none"> <li>- Searched “nih” on Google search.</li> <li>- Clicked the top link to nih.gov.</li> <li>- Searched “h pylori acne” in the search box.</li> <li>- Clicked “More” and selected “Clinical Trials”.</li> <li>- Clicked the result about H. Pylori and acne.</li> <li>- Checked the date to confirm it was January to May 2018.</li> <li>- Opened “Tabular View”.</li> <li>- Scrolled down to Actual Enrollment and recorded the number.</li> </ul>
Number of steps	8
Answer	90
Time to answer	8 minutes
Tools	<ul style="list-style-type: none"> <li>- Web browser</li> </ul>
Number of tools	1