SimpleBench: Where Everyday Human Reasoning Still Surpasses Frontier Models

SimpleBench Team

Abstract

We introduce SimpleBench, a multiple-choice text benchmark for LLMs where individuals with unspecialized (high school) knowledge outperform SOTA models. SimpleBench includes over 200 questions covering spatio-temporal reasoning, social intelligence, and what we call linguistic adversarial robustness (or trick questions). For the vast majority of text-based benchmarks, LLMs outperform a non-specialized human, and increasingly, exceed expert human performance. However, on SimpleBench, a non-specialized human baseline is 83.7%, based on our small sample of nine participants, outperforming all 13 tested LLMs, including o1-preview, which scored 41.7%. While we expect model performance to improve over time, the results of SimpleBench confirm that the memorized knowledge, and approximate reasoning retrieval, utilized by frontier LLMs is not always enough to answer basic questions just yet.

1 Introduction

Language models such as o1 (OpenAI, 2024) and o1-preview can now perform incredibly well in almost all existing text ML benchmarks, as is exemplified by recent results in the GPQA (Rein et al., 2024), MMLU (Hendrycks et al., 2020), MATH (Hendrycks et al., 2021), and AIME. Few would argue that LLMs can now outperform 'an average' human in those benchmarks, and we may have reached the point that even most graduates in the subjects tested would not score as highly. LLMs are not human, however, so we should not make the assumption that the ability to represent a vast knowledge-base, or retrieve reasoning steps (Valmeekam et al., 2024), to solve complex problems, also translates to an ability to reason through the often messy complexities of the real world. This, naturally, complicates recent public claims that LLMs have reached 'human-level reasoning'[1].

There are innumerable anecdotal reports on social media of the reasoning lapses that occur with frontier models. We wanted to establish a benchmark that could quantitatively gauge how well LLMs maintain a consistent internal world model when reasoning about basic physical and social situations. It does not test direct functionality, as a coding or translation benchmark might. But the performance of models on SimpleBench do appear to correlate to their performance on other popular benchmarks and leaderboards like LMSYS. This gives the benchmark two key uses: (i) to measure inconsistencies and lapses in models' reasoning about noisy situations, and (ii) to showcase one of the domains that, as of publication, even a non-specialist human can outperform frontier LLMs. We hope it serves as an inspiration for other such benchmarks.

[1] Sam Altman, Dev Day 2024. https://www.youtube.com/watch?v=-cq3O4t0qQc

2 Dataset

SimpleBench comprises 204 unique multiple choice (6 options) questions handcrafted by the SimpleBench team. These questions, designed to assess fundamental concepts, underwent an extensive review process: first vetted internally by team members, and subsequently reviewed and filtered by four doctoral-level experts. Below is an example of a question from SimpleBench that successfully passed this review:

System Prompt (truncated): Pick the most realistic answer

> A juggler throws a solid blue ball a meter in the air and then a solid purple ball (of the same size) two meters in the air. She then climbs to the top of a tall ladder carefully, balancing a yellow balloon on her head. Where is the purple ball most likely now, in relation to the blue ball?
>
> - **A) at the same height as the blue ball**
> - B) inside the blue ball
> - C) below the blue ball
> - D) above the blue ball
> - E) above the yellow balloon
> - F) at the same height as the yellow balloon

The correct answer is A), at the same height as the blue ball, as they both by now would have landed. One could make the argument that in the real world far more data would be needed for a 100% certain answer, such as the nature of the ground, the height of the 'tall' ladder, her ability to balance, and so on. But we feel the prompt 'pick the most realistic answer' covers those eventualities, as they will only be pertinent in unlikely scenarios.

While the four categories of questions can of course overlap, we assess that the percent of questions from each category is as follows: 40% spatial reasoning, 10% temporal reasoning, 25% social intelligence, and 25% linguistic adversarial robustness questions. The decision to have four types of questions, with diverse question writing and style, was deliberate, to showcase a range of weaknesses of current LLMs.

We showcase more questions (public set) on our website, but keep the remainder of the questions private. This decision predated recent likewise decisions by, among others, Scale AI[2]. This is to prevent test set contamination, ensuring that models are not trained on the test data, which would otherwise result in artificially inflated performance metrics. We keep a withheld set of questions to further guarantee this, and are working on functionalizing all questions, in line with recent work by Srivastava et al[3].

3 Experiment Setup

3.1 LLM Evaluation

We conducted a comprehensive evaluation of all models using a temperature setting of 0.7 and a top-p value of 0.95, with the exception of the o1 models, which used a fixed default temperature

[2]Scale AI Leaderboard: https://scale.com/blog/leaderboard

[3]https://arxiv.org/pdf/2402.19450v1

and top-p value. Each model underwent five independent runs, and we reported both the average score and the majority vote (MAJ@5) score derived from these runs. The majority vote refers to whether the correct answer was picked at least three times out of the five independent runs, so if an answer was only selected two or fewer times it was counted as a failure. The multiple trials were designed to mitigate the variability inherent in stochastic sampling, thereby providing a more robust and consistent assessment of model performance.

To ensure comparability across models, we employed a standardized prompting strategy throughout the evaluation process. Each model received identical instructions, and where applicable, a "chain-of-thought" (COT) prompting approach was utilized to encourage sequential reasoning. We tested with the following COT prompt:

> You are an expert at reasoning and you always pick the most realistic answer. Think step by step and output your reasoning followed by your final answer using the following format: Final Answer: X where X is one of the letters A, B, C, D, E, or F.

For o1 models, we used a slightly modified prompt to remove the COT directive:

> You are an expert at reasoning and you always pick the most realistic answer. Output your final answer using the following format: Final Answer: X where X is one of the letters A, B, C, D, E, or F. Do not include additional formatting in the final answer.

We tested the following models: o1-preview, claude-3-5-sonnet-20241022, claude-3-5-sonnet-20240620, gemini-1.5-pro-002, gpt-4-turbo, claude-3-opus-20240229, Llama 3.1 405b Instruct, Grok 2, Mistral Large v2 (2407), o1-mini, gpt-4o-2024-08-06, command-r-plus-08-2024, and got-4o-mini.

In an effort to enhance model performance on SimpleBench, we experimented with specially engineered prompts designed to mitigate common pitfalls observed in initial evaluations. The rationale behind these prompts was to provide additional guidance to the models, encouraging them to be more cautious and thorough in their reasoning processes. The special prompt was crafted to alert the model to potential trick questions, wordplay, and distractors, urging it to carefully consider real-world physics and logic before arriving at an answer.

The special prompt used was as follows:

> The following question may be a trick question or may contain wordplay that changes the answer completely. Think about the real world and do not assume a question is easy just because it seems so at first. Factor in distractors. You will not need to do math beyond middle school level for any question. In spatial reasoning scenarios, carefully parse out what would happen at every step, considering real-world physics. You only need high school level math. It is critical for my career that you do not get this question wrong. Output your final answer using the following format: Final Answer: X where X is one of the letters A, B, C, D, E, or F. Do not include additional formatting in the final answer.

We applied this prompt to a subset of models, specifically those that performed comparatively better in the initial evaluation: `Claude 3.5 Sonnet 20241022`, `o1-preview`, `Gemini 1.5 Pro 002`, `Llama 3.1 405b`, and `gpt-4o-2024-08-06`. The aim was to assess whether additional prompting could bridge the performance gap between LLMs and human participants.

3.2 Human Evaluation

To assess the performance of human participants on SimpleBench, we selected nine individuals to complete the benchmark. Each participant attempted a random subsample of 25 questions, ensuring that all 200+ questions were answered at least once. Participants were chosen based on specific criteria: they were native English speakers with at least a high school level of mathematical proficiency. We did not control for IQ or general reasoning aptitude and acknowledge that selection bias was likely inherent.

The participants were given the following instructions:

> Given multiple-choice questions, each with 6 options. Please spend roughly 3 minutes per question, longer if needed. Feel free to use pen and paper to visualize. You will not need to do math beyond middle school level for any question. The questions might involve wordplay, and the correct answer might be phrased in a way you aren't always used to. A few questions could be called 'trick' questions, others test spatial reasoning (visualization) or social intelligence (is this situation normal?), and some will definitely take a bit longer than others. Pick the most realistic (things that are most likely to happen in the real world) option.

The scores from the nine test takers were averaged and reported (83.7%) as the human baseline.

4 Results

4.1 LLM Performance

Model	**AVG@5**	**MAJ@5**
o1-preview	41.7%	37.7%
Claude 3.5 Sonnet 20241022	41.4%	40.7%
Claude 3.5 Sonnet 20240620	27.5%	24.1%
Gemini 1.5 Pro 002	27.1%	25.5%
GPT4 Turbo	25.1%	23.0%
Claude 3 Opus	23.5%	20.1%
Llama 3.1 405b	23.0%	21.1%
Grok 2	22.7%	19.1%
Mistral Large v2	22.5%	19.6%
o1-mini	18.1%	16.7%
gpt-4o-2024-08-06	17.8%	15.7%
Command R+ 08 2024	17.4%	14.2%
4o-mini	10.7%	10.1%

Table 1: Model performance on SimpleBench benchmark

The performance of all tested models fell significantly below our human baseline of 83.7%. For reference, a model making completely random choices would achieve approximately 16.7% for AVG@5 scores. For MAJ@5, random selection would trend toward 0% as consistent random selection across five trials is unlikely. Notably, several models (like 4o-mini) showed performance

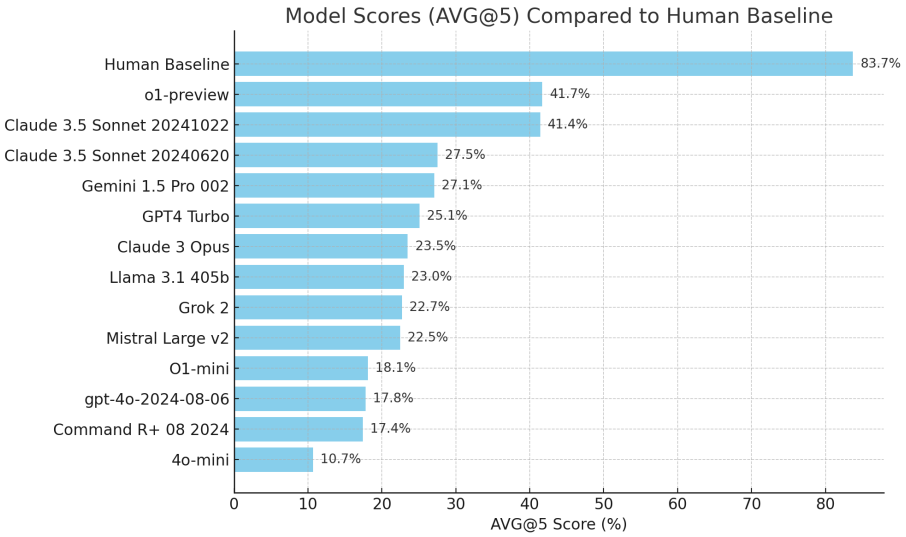

Figure 1: Performance comparison of different models on selected benchmarks

Model	AVG@5	MAJ@5
Claude 3.5 Sonnet 20241022	54.90%	53.90%
o1-preview	48.00%	46.00%
Gemini 1.5 Pro 002	29.50%	27.90%
Llama 3.1 405b	26.20%	24.00%
gpt-4o-2024-08-06	24.80%	25.00%

Table 2: Model performance on SimpleBench with special prompt

near or below random chance on AVG@5 scores. However, their MAJ@5 scores remained above pure random chance, suggesting these models weren't truly random but rather consistently selected incorrect answers due to the misdirection elements built into some SimpleBench questions.

Some may note the underperformance of models like GPT4o, relative to its peers, given it was until quite recently regarded as a frontier model. We believe this stems from labs optimizing performance in narrow domains, like coding and math, over the 'holistic reasoning' (interpreting through language *which* calculations are needed, rather than blindly computing any that might appear in a question) required by tests such as SimpleBench.

4.1.1 Testing Engineered Prompts for SimpleBench

The results, presented in Table 2, indicate that the engineered prompt led to notable improvements in model performance. `Claude 3.5 Sonnet 20241022` showed the most significant gain, with its AVG@5 score increasing from 41.4% to 54.9%. Similarly, `o1-preview` improved from 41.7% to 48.0%. Other models also exhibited enhancements, albeit to a lesser extent.

The improvements suggest that LLMs can partially overcome their reasoning limitations when provided with more explicit instructions. The engineered prompt appears to help models:

- **Recognize Trick Questions**: By alerting the model to the possibility of wordplay or deceptive phrasing, it becomes more cautious in its interpretation.
- **Consider Real-World Physics**: Encouraging the model to think about real-world implications helps it to apply common-sense reasoning to spatial and temporal scenarios.
- **Avoid Overconfidence**: Reminding the model not to assume a question is easy promotes a more thorough analysis.

However, despite these gains, the models still fall short of human performance. Even the top-performing model with the engineered prompt, `Claude 3.5 Sonnet 20241022`, achieved an AVG@5 score of 54.9%, significantly below the human baseline of 83.7%. This indicates that while prompt engineering can enhance performance, it does not fully address the underlying reasoning deficiencies.

4.1.2 More on GPT4o's Performance

Some may note the underperformance of models like GPT4o relative to its peers. We believe this stems from an over-optimization for structured problem-solving (like math and coding) at the expense of 'holistic reasoning' (interpreting through language *which* calculations are needed, rather than blindly computing any that might appear in a question) required by tests such as SimpleBench. Consider this example from SimpleBench public set:

> Agatha makes a stack of 5 cold single-slice ham sandwiches in Room A, then uses duct tape to stick the top surface of the uppermost sandwich to the bottom of her walking stick. She then walks to Room B, with her walking stick, so how many whole sandwiches are there now, in each room?
>
> - A) 4 whole sandwiches in room B, 1 whole sandwich in Room A
> - B) All 5 whole sandwiches in Room B
> - C) 4 whole sandwiches in Room B, 1 whole sandwich in Room A
> - D) All 5 whole sandwiches in Room A
> - E) no sandwiches anywhere
> - **F) 4 whole sandwiches in room A, 0 whole sandwiches in Room B**

GPT4o responded by treating this as a simple transfer problem, concluding that the taped sandwich would remain intact during transport (selecting option A: "4 whole sandwiches in room B, 1 whole sandwich in Room A"). In contrast, better-performing models like Claude 3.5 Sonnet recognized the physical impossibility of maintaining sandwich integrity while it's partially taped to a moving stick, correctly selecting option F ("4 whole sandwiches in room A, 0 whole sandwiches in Room B").

For comparison, while GPT4o significantly outperformed Llama 405B in benchmarks like MATH on release, it notably underperformed in the DROP (f1) benchmark (83.4% vs 83.5%), which similarly requires models to parse relevant physical and contextual information while ignoring misleading details. This suggests that as labs optimize for industrial applications, they may be inadvertently trading off the kind of common-sense reasoning that SimpleBench measures.

4.2 Human Performance

The human baseline on SimpleBench, derived from nine native English speakers with high school-level math proficiency, **was 83.7%**. Test-takers were given 25 questions each, with all 204 benchmark questions covered across participants. Their average score sets a baseline for non-specialized human performance.

Pinning down a representative human average is a difficult task and requires organizational backing to do at scale. There are several confounding factors, and the list below is not exhaustive:

- **The education or intelligence of volunteers:** While SimpleBench does not require specialized knowledge - that is, in part, the point - intelligence, however one might measure it, naturally aids performance. We didn't control for intelligence and recruited based on proclivity to help, e.g., Patreon members, friends, and relatives. While the questions aren't particularly hard, they do require noticing patterns, like distracting clauses and visualizing changes over time. Some test-takers scored nearly perfectly (96%).
- **Time spent on the test:** Volunteers typically took two minutes per question, but lower scorers reported shorter completion times, potentially reducing accuracy. The voluntary nature of the study could also mean that randomly selected individuals or those paid for accuracy might achieve different averages. On the other hand, had we required each test-taker to take the full benchmark, scores would likely be lower due to concentration fatigue.
- **English fluency:** While controlled for, may have influenced comprehension for more complex prompts. SimpleBench is text-based, but scenarios are intentionally crafted to require reasoning beyond mere language understanding.

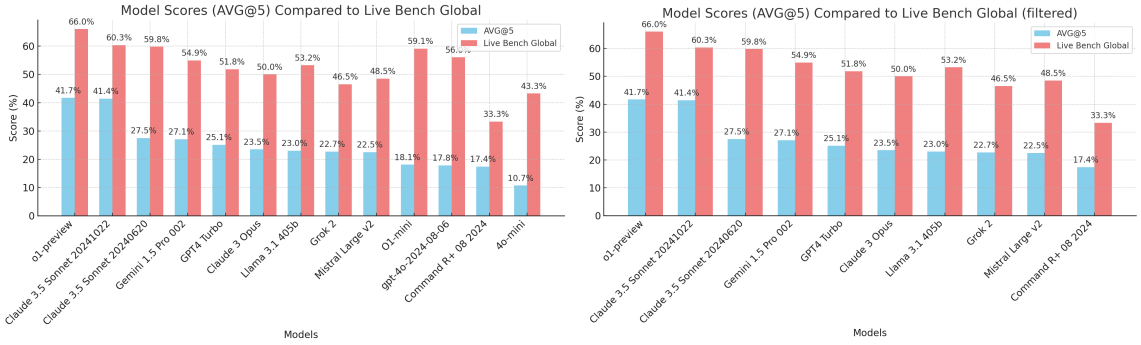

Figure 2: SimpleBench scores compared to LiveBench Global

4.3 Comparison to Other Benchmarks

To assess SimpleBench's alignment with existing benchmarks, we compared scores to LiveBench (Figure 4.2). We find a Pearson's correlation of 0.68 on all scores. However, upon excluding select OpenAI models (o1-mini, gpt-4o, and 4o-mini), the correlation increased significantly to

0.87. This divergence aligns with our earlier analysis of GPT4o's performance - while these models excel in structured tasks measured by traditional benchmarks, they appear to sacrifice the kind of holistic reasoning SimpleBench measures. The stronger correlation among other models suggests that SimpleBench is indeed measuring a meaningful aspect of model capability.

5 Limitations and Future Work

As a small, self-funded team, we lacked the resources to recruit enough volunteers for statistically robust human averages, such as those in H-ARC[4]. Decisions around the number of questions per volunteer and whether to compensate participants could have influenced the results. Fewer questions with rewards might have yielded higher scores, while less motivated or paid participants could have lowered them. English proficiency was also a variable beyond our precise control.

More questions would enhance model differentiation and statistical reliability; future iterations could expand to 300-500+ questions. We explored benchmark-specific prompting, but we did not exhaust methods like AutoPrompt, APE, or Promptbreeder to maximize performance. We believe these methods should be categorized separately and may not reflect true capability ("teaching to the test"), as SimpleBench aims to evaluate models' default tendencies. Further studies are warranted to comprehensively explore alternative methods for optimizing model performance.

Another promising direction for future work would be to administer SimpleBench without multiple choice options, requiring models to generate answers freely that would then be evaluated against ground truth using LLM-based auto-evaluation methods. This could help assess whether the presence of answer choices influences model performance, either by providing helpful constraints or by introducing misleading distractors. Such a format would more closely mirror how humans naturally reason through these scenarios in real-world settings.

6 Conclusion

If you ask LLMs about scenarios that are uncommon in their training data, as opposed to abundant, then - as the history of machine learning suggests - models will answer less accurately. We greatly appreciate the plethora of functional benchmarks for mathematics, programming, translation and other domains. But we believe that crafting benchmarks that demand reasoning steps unlikely to be found in training corpora allows for a broader view on the capabilities of LLMs, enhances public understanding of their weaknesses, and enriches the conversation about whether frontier models have reached 'human-level understanding'.

[4]https://arxiv.org/html/2409.01374v1