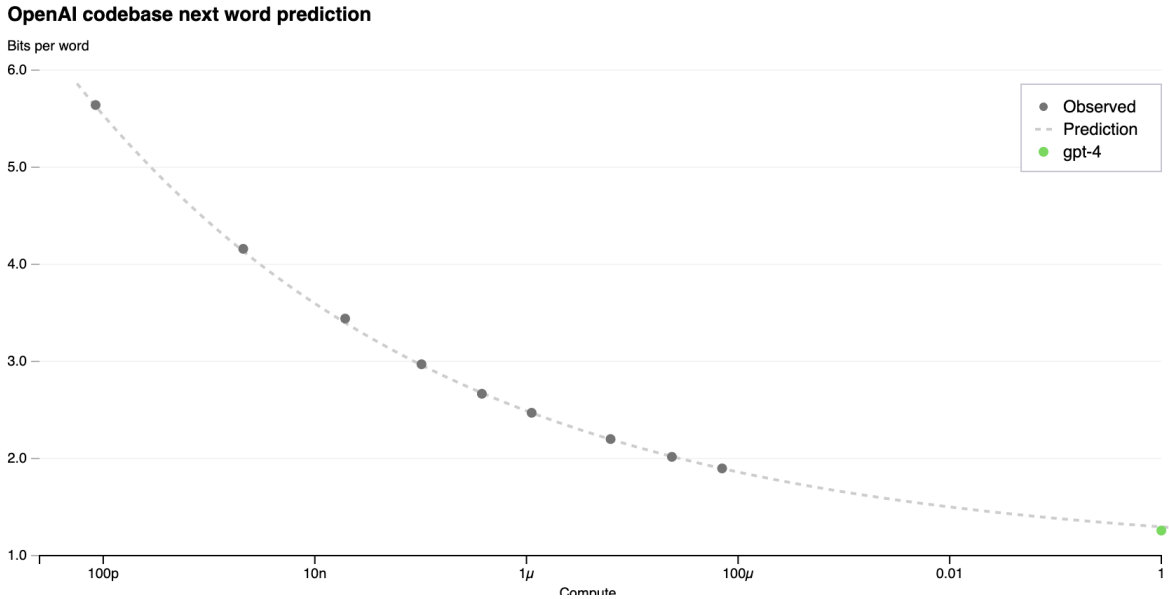
GPT-4 Technical Report

Predictable scaling

Develop infrastructure and optimization that have predictable behavior across multiple scales

Accurately predicted in advance GPT-4's final loss on our internal codebase by extrapolating from models trained using the same methodology but using 10,000x less

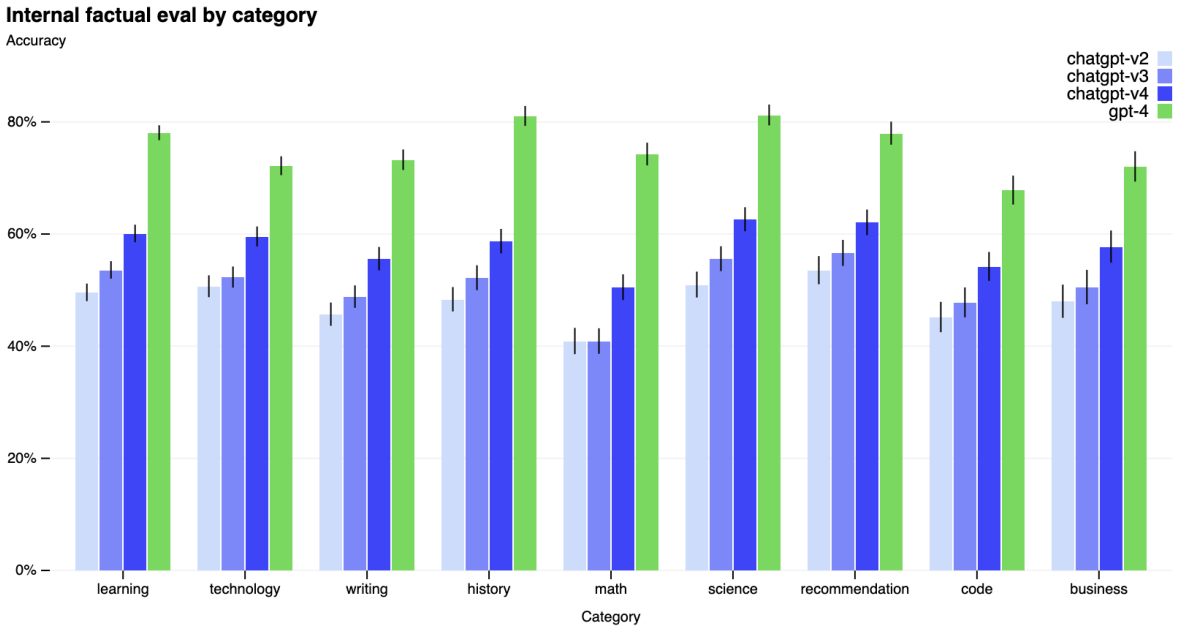


Visual inputs

Accept a prompt of text and images, generate text outputs

Steerability

Rather than the classic ChatGPT personality with a fixed verbosity, tone, and style, developers can now prescribe their AI's style and task by describing those directio



On nine categories of internal adversarially-designed factual evals, we compare GPT-4 (green) to the first three ChatGPT versions. There are significant gains across all topics. An accuracy of 1.0 means the model's answers are judged to be in agreement with human ideal responses for all questions in the eval.

Risks & mitigations

GPT-4 incorporates an additional safety reward signal during RLHF training to reduce harmful outputs by training the model to refuse requests for such content. The r

Sparks of Artificial General Intelligence: Early experiments with GPT-4

GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models

80% of the U.S. workforce could have at least 10% of their work tasks affected by the introduction of LLMs, while 19% of workers may see at least 50% of their tasks i

Least impacted: science & critical thinking

Most impacted: programming & writing