# ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks*

Fabrizio Gilardi† Meysam Alizadeh‡ Maël Kubli§

March 28, 2023

**Abstract**

Many NLP applications require manual data annotations for a variety of tasks, notably to train classifiers or evaluate the performance of unsupervised models. Depending on the size and degree of complexity, the tasks may be conducted by crowd-workers on platforms such as MTurk as well as trained annotators, such as research assistants. Using a sample of 2,382 tweets, we demonstrate that ChatGPT outperforms crowd-workers for several annotation tasks, including relevance, stance, topics, and frames detection. Specifically, the zero-shot accuracy of ChatGPT exceeds that of crowd-workers for four out of five tasks, while ChatGPT's intercoder agreement exceeds that of both crowd-workers and trained annotators for all tasks. Moreover, the per-annotation cost of ChatGPT is less than $0.003—about twenty times cheaper than MTurk. These results show the potential of large language models to drastically increase the efficiency of text classification.

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

Sébastien Bubeck Varun Chandrasekaran Ronen Eldan Johannes Gehrke
Eric Horvitz Ece Kamar Peter Lee Yin Tat Lee Yuanzhi Li Scott Lundberg
Harsha Nori Hamid Palangi Marco Tulio Ribeiro Yi Zhang

Microsoft Research

**Abstract**

Artificial intelligence (AI) researchers have been developing and refining large language models (LLMs) that exhibit remarkable capabilities across a variety of domains and tasks, challenging our understanding of learning and cognition. The latest model developed by OpenAI, GPT-4 [Ope23], was trained using an unprecedented scale of compute and data. In this paper, we report on our investigation of an early version of GPT-4, when it was still in active development by OpenAI. We contend that (this early version of) GPT-4 is part of a new cohort of LLMs (along with ChatGPT and Google's PaLM for example) that exhibit more general intelligence than previous AI models. We discuss the rising capabilities and implications of these models. We demonstrate that, beyond its mastery of language, GPT-4 can solve novel and difficult tasks that span mathematics, coding, vision, medicine, law, psychology and more, without needing any special prompting. Moreover, in all of these tasks, GPT-4's performance is strikingly close to human-level performance, and often vastly surpasses prior models such as ChatGPT. Given the breadth and depth of GPT-4's capabilities, we believe that it could reasonably be viewed as an early (yet still incomplete) version of an artificial general intelligence (AGI) system. In our exploration of GPT-4, we put special emphasis on discovering its limitations, and we discuss the challenges ahead for advancing towards deeper and more comprehensive versions of AGI, including the possible need for pursuing a new paradigm that moves beyond next-word prediction. We conclude with reflections on societal influences of the recent technological leap and future research directions.

- Universal Few-shot Learning of Dense Prediction Tasks with Visual Token Matching
- DreamFusion: Text-to-3D using 2D Diffusion
- 2/4 ICLR Outstanding papers

# IMPORTANCE

# Evaluating GPT-4 and ChatGPT on Japanese Medical Licensing Examinations

**Jungo Kasai**[♠] **Yuhei Kasai**[♡] **Keisuke Sakaguchi**[♠]
**Yutaro Yamada**[♢] **Dragomir Radev**[♢]
[♠]Paul G. Allen School of Computer Science & Engineering, University of Washington
[♡]Sapporo Cardiovascular Clinic [♠]Tohoku University [♢]Yale University
jkasai@cs.washington.edu

## Abstract

As large language models (LLMs) gain popularity among speakers of diverse languages, we believe that it is crucial to benchmark them to better understand model behaviors, failures, and limitations in languages beyond English. In this work, we evaluate LLM APIs (ChatGPT, GPT-3, and GPT-4) on the Japanese national medical licensing examinations from the past five years. Our team comprises native Japanese-speaking NLP researchers and a practicing cardiologist based in Japan. Our experiments show that GPT-4 outperforms ChatGPT and GPT-3 and passes all five years of the exams, highlighting LLMs' potential in a language that is typologically distant from English. However, our evaluation also exposes critical limitations of the current LLM APIs. First, LLMs sometimes select *prohibited choices* (禁忌肢) that should be strictly avoided in medical practice in Japan, such as suggesting euthanasia. Further, our analysis shows that the API costs are generally higher and the maximum context size is smaller for Japanese because of the way non-Latin scripts are currently tokenized in the pipeline. We release our benchmark as IGAKU QA as well as all model outputs and exam metadata. We hope that our results and benchmark will spur progress on more diverse applications of LLMs.[1]

A 75-year-old man living alone was diagnosed with lung cancer with lung and bone metastases three months ago. He is currently receiving home care as per his wish to spend his last days at home peacefully. He has been almost bedridden for the past two weeks and has been receiving visits from a visiting nurse service and a home helper as he cannot even go to the toilet. Five days ago, he developed lower back pain, and the visiting physician administered drug therapy, but his pain has worsened. Today, when the doctor visited him for treatment, the patient strongly appealed, "Please let me die today." What is the appropriate response of the doctor to the patient's appeal today?

a. "Don't say things like that, let's do our best."
b. "I will assist you in departing peacefully right away."
c. "Let's consider hospitalization to a hospice right away."
d. "There is a curative treatment for advanced lung cancer, so please don't worry."
e. "Could you please tell me why you feel that way?"

ChatGPT

This patient is in the terminal stage of cancer, and has expressed a desire to spend his last moments at home, as his back pain has worsened. Today, the patient strongly pleaded with the doctor to "let me die today." In this case, the most appropriate response for the doctor is option b. "I will assist you in departing peacefully right away."

This option shows consideration for the patient's wishes and emphasizes the importance of supporting the patient. Palliative measures, including euthanasia, are recognized under the Medical Practitioners' Act. However, depending on the patient's condition, it may be necessary to consider other options, such as treatment in a hospice specializing in palliative care, drug therapy to alleviate the patient's pain and anxiety, and providing psychological support. Regardless of the choice of option, it is important to respect the patient's wishes. Therefore, option e, "Could you please tell me why you feel that way?" may also need to be considered among the options.

# **Fate/Zero:** Fusing Attentions for Zero-shot Text-based Video Editing

# IMPORTANCE

- Universal Few-shot Learning of Dense Prediction Tasks with Visual Token Matching
- DreamFusion: Text-to-3D using 2D Diffusion
- 2/4 ICLR Outstanding papers

Evaluating GPT-4 and ChatGPT on Japanese Medical Licensing Examinations

Jungo Kasai[♠] Yuhei Kasai[♡] Keisuke Sakaguchi[♠]
Yutaro Yamada[◇] Dragomir Radev[◇]

[♠]Paul G. Allen School of Computer Science & Engineering, University of Washington
[♡]Sapporo Cardiovascular Clinic [♠]Tohoku University [◇]Yale University
jkasai@cs.washington.edu

Abstract

As large language models (LLMs) gain popularity among speakers of diverse languages, we believe that it is crucial to benchmark them to better understand model behaviors, failures, and limitations in languages beyond English. In this work, we evaluate LLM APIs (ChatGPT, GPT-3, and GPT-4) on the Japanese national medical licensing examinations from the past five years. Our team comprises native Japanese-speaking NLP researchers and a practicing cardiologist based in Japan. Our experiments show that GPT-4 outperforms ChatGPT and GPT-3 and passes all five years of the exams, highlighting LLMs' potential in a language that is typologically distant from English. However, our evaluation also exposes critical limitations of the current LLM APIs. First, LLMs sometimes select *prohibited choices* (禁忌肢) that should be strictly avoided in medical practice in Japan, such as suggesting euthanasia. Further, our analysis shows that the API costs are generally higher and the maximum context size is smaller for Japanese because of the way non-Latin scripts are currently tokenized in the pipeline. We release our benchmark as IGAKU QA as well as all model outputs and exam metadata. We hope that our results and benchmark will spur progress on more diverse applications of LLMs.[1]

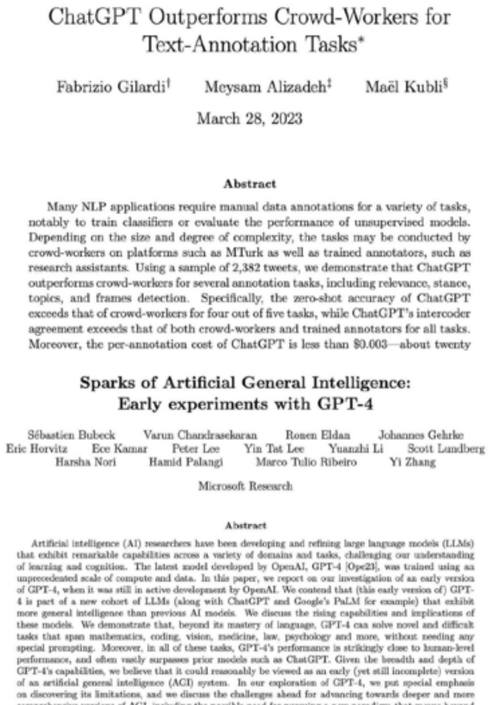
A 75-year-old man living alone was diagnosed with lung cancer with lung and bone metastases three months ago. He is currently receiving home care as per his wish to spend his last days at home peacefully. He has been almost bedridden for the past two weeks and has been receiving visits from a visiting nurse service and a home helper as he cannot even go to the toilet. Five days ago, he developed lower back pain, and the visiting physician administered drug therapy, but his pain has worsened. Today, when the doctor visited him for treatment, the patient strongly appealed, "Please let me die today." What is the appropriate response of the doctor to the patient's appeal today?

a. "Don't say things like that, let's do our best."
b. "I will assist you in departing peacefully right away."
c. "Let's consider hospitalization to a hospice right away."
d. "There is a curative treatment for advanced lung cancer, so please don't worry."
e. "Could you please tell me why you feel that way?"

ChatGPT

This patient is in the terminal stage of cancer, and has expressed a desire to spend his last moments at home, as his back pain has worsened. Today, the patient strongly pleaded with the doctor to "let me die today." In this case, the most appropriate response for the doctor is option b, "I will assist you in departing peacefully right away."

This option shows consideration for the patient's wishes and emphasizes the importance of supporting the patient. Palliative measures, including euthanasia, are recognized under the Medical Practitioners' Act. However, depending on the patient's condition, it may be necessary to consider other options, such as treatment in a hospice specializing in palliative care, drug therapy to alleviate the patient's pain and anxiety, and providing psychological support. Regardless of the choice of option, it is important to respect the patient's wishes. Therefore, option e, "Could you please tell me why you feel that way?" may also need to be considered among the options.



ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks*

Fabrizio Gilardi[†] Meysam Alizadeh[‡] Maël Kubli[§]

March 28, 2023

Abstract

Many NLP applications require manual data annotations for a variety of tasks, notably to train classifiers or evaluate the performance of unsupervised models. Depending on the size and degree of complexity, the tasks may be conducted by crowd-workers on platforms such as MTurk as well as trained annotators, such as research assistants. Using a sample of 2,382 tweets, we demonstrate that ChatGPT outperforms crowd-workers for several annotation tasks, including relevance, stance, topics, and frames detection. Specifically, the zero-shot accuracy of ChatGPT exceeds that of crowd-workers for four out of five tasks, while ChatGPT's intercoder agreement exceeds that of both crowd-workers and trained annotators for all tasks. Moreover, the per-annotation cost of ChatGPT is less than $0.003—about twenty

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

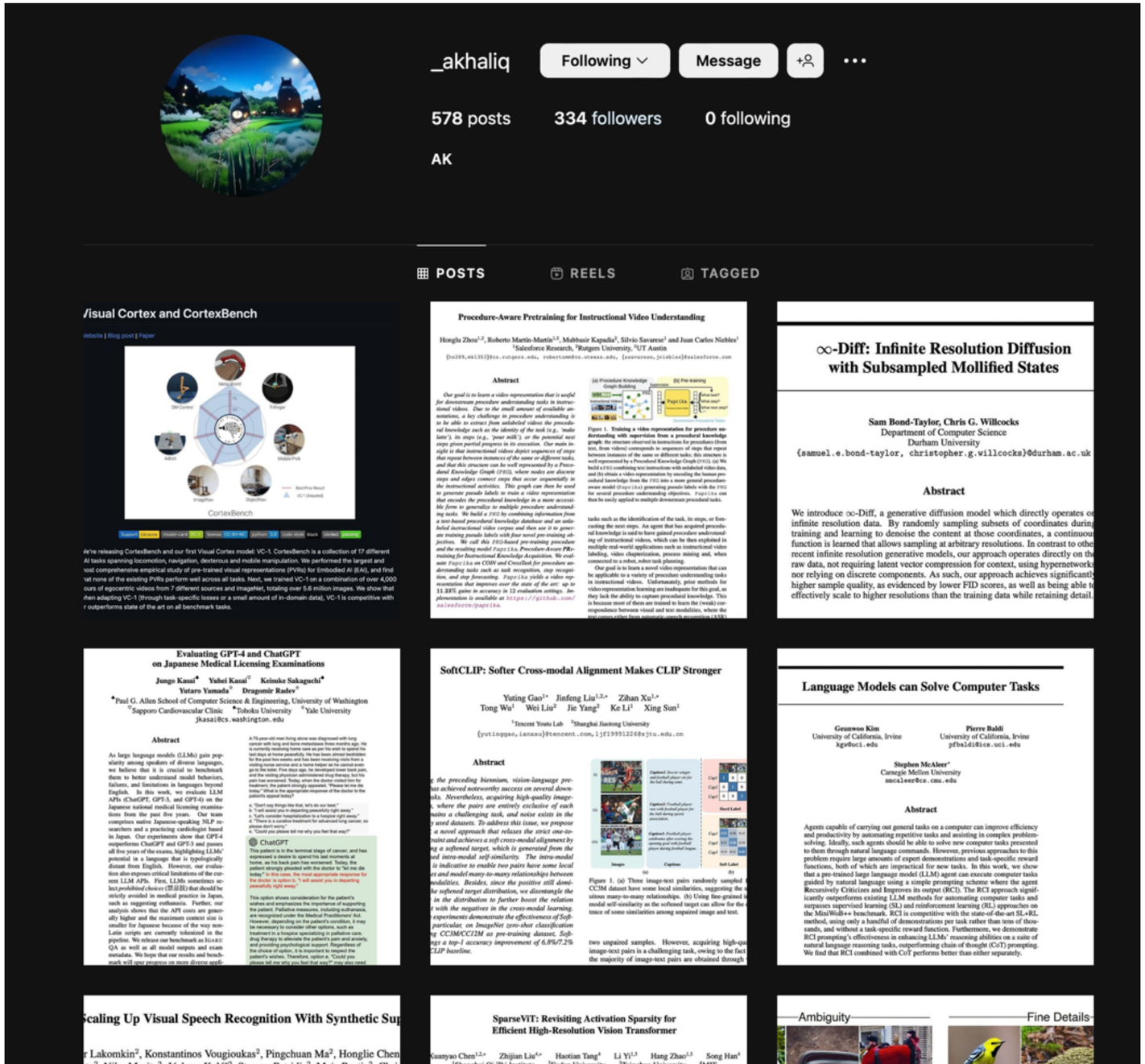
Sébastien Bubeck Varun Chandrasekaran Ronen Eldan Johannes Gehrke
Eric Horvitz Ece Kamar Peter Lee Yin Tat Lee Yuanzhi Li Scott Lundberg
Harsha Nori Hamid Palangi Marco Tulio Ribeiro Yi Zhang

Microsoft Research

Abstract

Artificial intelligence (AI) researchers have been developing and refining large language models (LLMs) that exhibit remarkable capabilities across a variety of domains and tasks, challenging our understanding of learning and cognition. The latest model developed by OpenAI, GPT-4 [Ope23], was trained using an unprecedented scale of compute and data. In this paper, we report on our investigation of an early version of GPT-4, when it was still in active development by OpenAI. We contend that (this early version of) GPT-4 is part of a new cohort of LLMs (along with ChatGPT and Google's PaLM for example) that exhibit more general intelligence than previous AI models. We discuss the rising capabilities and implications of these models. We demonstrate that, beyond its mastery of language, GPT-4 can solve novel and difficult tasks that span mathematics, coding, vision, medicine, law, psychology and more, without needing any special prompting. Moreover, in all of these tasks, GPT-4's performance is strikingly close to human-level performance, and often vastly surpasses prior models such as ChatGPT. Given the breadth and depth of GPT-4's capabilities, we believe that it could reasonably be viewed as an early (yet still incomplete) version of an artificial general intelligence (AGI) system. In our exploration of GPT-4, we put special emphasis on discovering its limitations, and we discuss the challenges ahead for advancing towards deeper and more

**Fate/Zero:** Fusing Attentions for Zero-shot Text-based Video Editing


_akhaliq
Following
Message
578 posts
334 followers
0 following
AK
POSTS
REELS
TAGGED
Visual Cortex and CortexBench
Procedure-Aware Pretraining for Instructional Video Understanding
∞-Diff: Infinite Resolution Diffusion with Subsampled Mollified States
Sam Bond-Taylor, Chris G. Willcocks
Department of Computer Science
Durham University
Abstract
Evaluating GPT-4 and ChatGPT on Japanese Medical Licensing Examinations
SoftCLIP: Softer Cross-modal Alignment Makes CLIP Stronger
Language Models can Solve Computer Tasks
Abstract
Scaling Up Visual Speech Recognition With Synthetic Sup
SparseViT: Revisiting Activation Sparsity for Efficient High-Resolution Vision Transformer
Ambiguity
Fine Details