

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 Gehrmann
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 Attention for Zero-Text-bias 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

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

Evaluating GPT-4 and ChatGPT on Japanese Medical Licensing Examinations

Jungo Kasai^{*} Yuhel 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 Gehrmke
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





_akhaliq

Following

Message



578 posts

334 followers

0 following

AK

POSTS

REELS

TAGGED

/visual Cortex and CortexBench

Website | Blog post | Paper



We're releasing CortexBench and our first Visual Cortex model, VC-1. CortexBench is a collection of 17 different AI tasks spanning locomotion, navigation, dexterous and mobile manipulation. We performed the largest and most comprehensive empirical study of pre-trained visual representations (PViTs) for Embedded AI (EAI), and find that none of the existing PViTs perform well across all tasks. Next, we trained VC-1 on a combination of over 4,000 ours of egocentric videos from 7 different sources and ImageNet, totaling over 5.6 million images. We show that when adapting VC-1 through task-specific losses or a small amount of in-domain data, VC-1 is competitive with or outperforms state-of-the-art on all benchmark tasks.

Procedure-Aware Pretraining for Instructional Video Understanding

Honglu Zhou^{1,2}, Roberto Martín-Martín^{1,3}, Mubbasir Kapadia³, Silvio Savarese³ and Juan Carlos Niebles¹
¹Salesforce Research, ²Rutgers University, ³UT Austin
{hzhou, mml353}@rutgers.edu, robertom@cs.utexas.edu, {ssavarese, jniebles}@salesforce.com

Abstract

Our goal is to learn a video representation that is useful for downstream procedure understanding tasks in instructional videos. Due to the small amount of available annotations, a key challenge in procedure understanding is to be able to extract from unlabeled videos the procedural knowledge such as the identity of the task (e.g., "make latte"), its steps (e.g., "pour milk"), or the potential next steps given partial progress in its execution. Our main insight is that instructional videos depict sequences of steps that repeat between instances of the same or different tasks, and that this structure can be well represented by a *Procedural Knowledge Graph* (PKG), where nodes are discrete steps and edges connect steps that occur sequentially in the instructional activities. This graph can then be used to generate pseudo labels to train a video representation that encodes the procedural knowledge in a more accessible form to generalize to multiple procedure understanding tasks. We build a PKG by combining information from a text-based procedural knowledge database and an unlabeled instructional video corpus and then use it to generate training pseudo labels with four novel pre-training objectives. We call this PKG-based pre-training procedure and the resulting model *PaperLika*. Procedure-Aware Pretraining for Instructional Knowledge Acquisition. We evaluate *PaperLika* on COIN and CrossTask for procedure understanding tasks such as task recognition, step recognition, and step forecasting. *PaperLika* yields a video representation that improves over the state of the art: up to 11.33% gains in accuracy in 12 evaluation settings. Implementation is available at <https://github.com/salesforce/paperlika>.

Figure 1: Training a video representation for procedure understanding with supervision from a procedural knowledge graph.

The figure shows two parts: (a) Procedure Knowledge Graph Building and (b) Pre-training. Part (a) shows a graph with nodes representing steps and edges representing sequential relationships. Part (b) shows the pre-training process where the graph is used to generate pseudo labels for training a video representation model.

∞-Diff: Infinite Resolution Diffusion with Subsampled Mollified States

Sam Bond-Taylor, Chris G. Willcocks
Department of Computer Science
Durham University
{samuel.e.bond-taylor, christopher.g.willcocks}@durham.ac.uk

Abstract

We introduce ∞-Diff, a generative diffusion model which directly operates on infinite resolution data. By randomly sampling subsets of coordinates during training and learning to denoise the content at those coordinates, a continuous function is learned that allows sampling at arbitrary resolutions. In contrast to other recent infinite resolution generative models, our approach operates directly on the raw data, not requiring latent vector compression for context, using hypernetworks nor relying on discrete components. As such, our approach achieves significantly higher sample quality, as evidenced by lower FID scores, as well as being able to effectively scale to higher resolutions than the training data while retaining detail.

Evaluating GPT-4 and ChatGPT on Japanese Medical Licensing Examinations

Jungo Kasai^{*}, Yubei 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 (禁選肢) 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 JGAKU QA as well as all model outputs and examination data. We hope that our results and benchmark will spur progress on more diverse applications.

SoftCLIP: Softer Cross-modal Alignment Makes CLIP Stronger

Yuting Gao^{1*}, Jinfeng Liu^{1,2,*}, Zihan Xu^{1,*}, Tong Wu¹, Wei Liu², Jie Yang², Ke Li¹, Xing Sun¹
¹Tencent YouTu Lab ²Shanghai Jiaotong University
{yutinggao, lanxxu}@tencent.com, ljf19991226@sjtu.edu.cn

Abstract

In the preceding biennium, vision-language pre-training has achieved noteworthy success on several downstream tasks. Nevertheless, acquiring high-quality image-text pairs, where the pairs are entirely exclusive of each other, remains a challenging task, and noise exists in the used datasets. To address this issue, we propose a novel approach that relaxes the strict one-to-one constraint and achieves a soft cross-modal alignment by using a softened target, which is generated from the self-intra-modal self-similarity. The intra-modal self-similarity is indicative to enable two pairs have some local similarities and model many-to-many relationships between modalities. Besides, since the positive still dominates the softened target distribution, we disentangle the self-similarity in the distribution to further boost the relation with the negatives in the cross-modal learning. Experiments demonstrate the effectiveness of SoftCLIP on ImageNet zero-shot classification and CC3M/CC12M as pre-training datasets. SoftCLIP achieves a top-1 accuracy improvement of 6.8%/7.2% CLIP baseline.

Language Models can Solve Computer Tasks

Guanyao Kim
University of California, Irvine
kgwu@uci.edu

Pierre Baldi
University of California, Irvine
pfbaldi@cs.uci.edu

Stephen McAleer^{*}
Carnegie Mellon University
smcaleer@cs.cmu.edu

Abstract

Agents capable of carrying out general tasks on a computer can improve efficiency and productivity by automating repetitive tasks and assisting in complex problem-solving. Ideally, such agents should be able to solve new computer tasks presented to them through natural language commands. However, previous approaches to this problem require large amounts of expert demonstrations and task-specific reward functions, both of which are impractical for new tasks. In this work, we show that a pre-trained large language model (LLM) agent can execute computer tasks guided by natural language using a simple prompting scheme where the agent Recursively Critiques and Improves its output (RCI). The RCI approach significantly outperforms existing LLM methods for automating computer tasks and surpasses supervised learning (SL) and reinforcement learning (RL) approaches on the MiniWoB++ benchmark. RCI is competitive with the state-of-the-art SL+RL method, using only a handful of demonstrations per task rather than tens of thousands, and without a task-specific reward function. Furthermore, we demonstrate RCI prompting's effectiveness in enhancing LLMs' reasoning abilities on a suite of natural language reasoning tasks, outperforming chain of thought (CoT) prompting. We find that RCI combined with CoT performs better than either separately.

Scaling Up Visual Speech Recognition With Synthetic Sup

Lakomkin², Konstantinos Vougioukas², Pingchuan Ma², Honglie Chen

SparseViT: Revisiting Activation Sparsity for Efficient High-Resolution Vision Transformer

Guanyao Chen^{1,2,*}, Zhijian Liu^{4,*}, Haotian Tang⁴, Li Yi^{1,3}, Hang Zhao^{1,3}, Song Han⁴

Ambiguity



Fine Details