Human or Machine? Turing Tests for Vision and Language

Mengmi Zhang,[1] Giorgia Dellaferrera, [2,3] Ankur Sikarwar, [1,*] Marcelo Armendariz, [4,5,*]
Noga Mudrik, [6,*] Prachi Agrawal, [7,*] Spandan Madan, [5,8,*] Andrei Barbu, [5,9]
Haochen Yang, [10] Tanishq Kumar, [11] Meghna Sadwani, [12] Stella Dellaferrera, [13]
Michele Pizzochero, [8] Hanspeter Pfister, [8] and Gabriel Kreiman [4,5]

[*] Equal contribution

[1] CFAR and I2R, Agency for Science, Technology and Research, Singapore, [2] IBM Research - Zürich, Rüschlikon, Switzerland,
[3] Institute of Neuroinformatics, University of Zürich and ETH Zürich, Zürich, Switzerland,
[4] Children's Hospital, Harvard Medical School, USA, [5] Center for Brains, Minds and Machines, USA,
[6] Biomedical Engineering, Johns Hopkins University, USA, [7] Birla Institute of Technology and Science, Pilani, India,
[8] School of Engineering and Applied Sciences, Harvard University, USA, [9] CSAIL, MIT, USA, [10] Harvard University, USA,
[11] Harvard College, Harvard University, USA, [12] Jawaharlal Nehru Medical College, India, [13] University of Turin, Italy
Address correspondence to gabriel.kreiman@tch.harvard.edu

Abstract

As AI algorithms increasingly participate in daily activities that used to be the sole province of humans, we are inevitably called upon to consider how much machines are really like us. To address this question, we turn to the Turing test and systematically benchmark current AIs in their abilities to imitate humans. We establish a methodology to evaluate humans versus machines in Turing-like tests and systematically evaluate a representative set of selected domains, parameters, and variables. The experiments involved testing 769 human agents, 24 state-of-the-art AI agents, 896 human judges, and 8 AI judges, in 21,570 Turing tests across 6 tasks encompassing vision and language modalities. Surprisingly, the results reveal that current AIs are not far from being able to impersonate human judges across different ages, genders, and educational levels in complex visual and language challenges. In contrast, simple AI judges outperform human judges in distinguishing human answers versus machine answers. The curated large-scale Turing test datasets introduced here and their evaluation metrics provide valuable insights to assess whether an agent is human or not. The proposed formulation to benchmark human imitation ability in current AIs paves a way for the research community to expand Turing tests to other research areas and conditions. All of source code and data are publicly available: here.

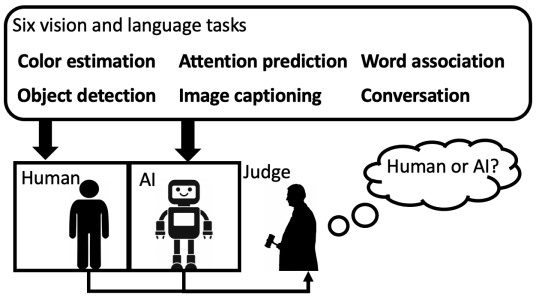


Figure 1. **Schematic illustration of Turing tests in six vision and language tasks**. A Turing test works with a judge asking a test subject (either a human or an AI agent) a series of tasks. Each party is kept in a separate room, so no physical contact is allowed. The AI passes the Turing test if the judge is unable to distinguish the AI from another human being by using the responses collected from the given task presented to both. See **Fig 2** for an overview of the six tasks.

1. Introduction

The Turing test, also known as the "imitation game", was proposed by Alan Turing in 1950 as a way of assessing a machine's ability to exhibit intelligent behaviors indistinguishable from those of a human (**Fig. 1**) [61]. Since its inception, whether the Turing test adequately quantifies intelligence or not has remained controversial [22, 34]. The purpose of this paper is *not* to argue in favor or against Turing tests as a measure of general intelligence. Instead, we consider the Turing tests as a quantitative evaluation of how well current AIs can imitate humans.

With powerful AI technologies being deployed in the real world, it is becoming increasingly important for lay people,legal judges, doctors, politicians, and other experts to ascertain whether the agent they are interacting with is a human or not. As two examples out of many, the inability to distinguish a human from an AI bot may lead to cybersecurity breaches resulting in the loss of private and protected data. Besides, the inability to distinguish real news from AI generated fake news or DeepFakes [68] can have disastrous implications for electoral campaigns [28, 68].

The answer to whether current AIs pass the Turing test depends on a plethora of considerations, including the machine agent, the human agent, the judge, the specific task, contextual conditions, and many more. Distinct from the original version of the Turing test in unrestricted conversations, the purpose of the current work is *not* to exhaustively study all possible combinations of these parameters and choices. Instead, we aim to: (i) establish a methodology to evaluate human imitators, (ii) provide a systematic protocol for the AI community to quantify whether a task is performed by humans or machines, and (iii) introduce evaluation metrics and analysis tools on a subset of tasks and conditions as a proof-of-principle. Specifically, we benchmarked 24 AI models in Turing tests on 6 fundamental tasks in computer vision and natural language processing (**Fig. 2**): color estimation, object detection, attention prediction, image captioning, word associations, and conversation.

The key contributions of this work are:

(1) We design a systematic format for conducting Turing tests and evaluating AI models over different tasks involving multiple modalities. This helps the community expand the Turing test to a wide range of tasks and benchmark future AI models.

(2) We introduce datasets to evaluate current AIs in Turing-like tests in 6 fundamental vision and language tasks.

(3) We conduct human psychophysics experiments to evaluate human judges in 24 state-of-the-art vision and language AI models in Turing tests.

(4) We show that simple machine learning algorithms can serve as AI judges to distinguish machines versus human agents in the same tasks.

2. Related Works

2.1. Glimpse of the 70-year history of Turing test

The Turing test was introduced as an imitation game where a machine tries to pass as human during a conversation and a human judge determines whether they are interacting with a human or not [61]. The Loebner Prize was introduced in 1991 [45] to the programs considered by human judges to be the most human-like. There was also an award for the most human human [11]. The Turing test has generated extensive controversy and discussion about whether it is a valid measure of intelligence [25, 26, 34, 40, 51], shifting to whether machines can successfully imitate humans [31–33]. Several notable arguments include Searle's Chinese room thought experiment [54], Block's behaviorism [5], Harnad's Total Turing Test [30], Watt's Inverted Turing Test [65], Damassino's Questioning Turing Test [17] and Sejnowski's Reverse Turing Test [55]. Distinct from these arguments, our aim is to systematically and quantitatively provide methods, datasets and benchmark current AIs in imitating humans through Turing-like tests in multiple vision and language tasks.

2.2. AI versus humans in vision tasks

Current computer vision models can perform a wide range of tasks such as object recognition and detection Models are often evaluated by comparing their outputs against human ground truth annotations. Many object recognition studies benchmarked AI versus humans in out-of-distribution generalization [4,20], adversarial attacks [21], and contextual variations [7, 74]. Several studies also compared attention in AI models against humans in saliency prediction [36], and eye movement prediction [27, 71, 73]. However, high performance in a particular task does *not* constitute a Turing test. AI models can show similar average performance to humans in narrow tasks, or even outperform humans, and still be distinguishable from humans. Turing tests provide a unique assessment of AI models as imitators of human behavior which extends and complements current benchmarking frameworks.

2.3. AI versus humans in language tasks

Similar observations can be made in natural language processing. AI models are often compared against human ground truth data in discriminative tasks, such as image captioning or visual question answering [9, 44, 56, 70]. Human evaluation scores are reliable but costly to obtain. To mitigate these problems, several evaluation metrics have been proposed, such as BLEU [49], THUMB [38], and METEOR [18] in image captioning. However, these metrics focus on n-gram overlaps and are insensitive to semantic information. Cui *et al*. proposed a learned critique model acting as a human judge to perform a Turing Test in image captioning tasks [16]. Here we also introduce critique models and compare them with human judges.

Generative AI models are notoriously difficult to evaluate due to the inherent ambiguities of language. For example, human evaluators are often recruited to assess the quality of sentiment and semantic relevance on text generated by BERT [19] or GPT2/3 [8, 8, 37]. Such evaluations are restricted to specific domains of text

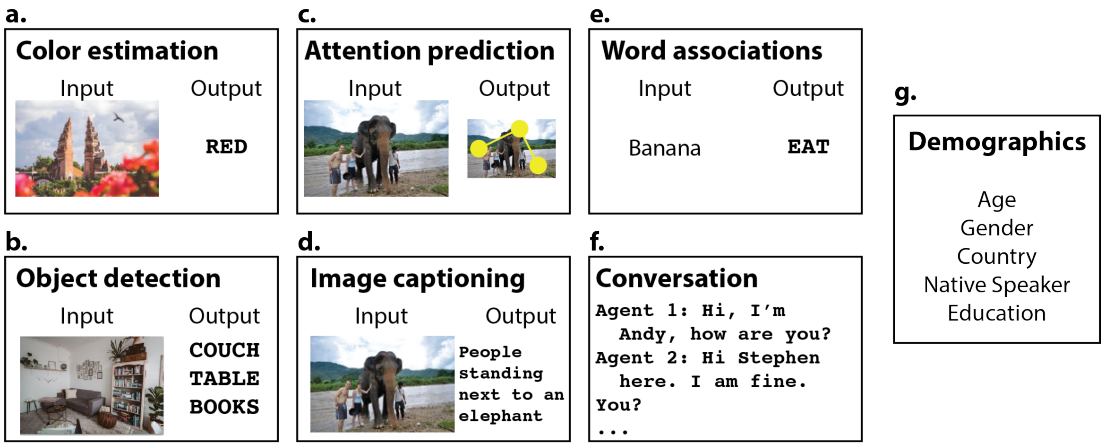


Figure 2. **Schematic of the 6 tasks**. We systematically evaluate 3 vision tasks, 1 vision-language task, and 2 language tasks. **a**. In Color estimation, the agent is presented with an image and has to output the main color. **b**. In Object detection, the agent is presented with an image and has to provide three objects. **c**. In Attention prediction, the agent is presented with an image and the output is a sequence of attention locations or eye movements. **d**. In Image captioning, the agent provides a single sentence description of an image. **e**. In Word associations, the agent is presented with a word and has to produce a single word related to the cue. **f**. In Converstaions, agents produce 24 exchanges. See **Sec. 3** for detailed description of each task and see **Supp. Material** for more example stimuli from both human and AI agents for all tasks. **g**. The results of a Turing test with a human judge depend on the characteristics of the judge. As an initial characterization, we collect basic demographic information indicated in this table.

generation and the heterogeneity of human judges has not been characterized. Here we provide an extensive set of Turing tests on multiple large state-of-the-art language models based on 896 judges across different demographics.

Conversation was the key target of the original Turing test and remains a daunting challenge for AI. There have been numerous early attempts at generating restricted topics during conversations, such as Colby's PARRY simulating a paranoid schizophrenic [12, 13] and Weizenbaum's ELIZA simulating a psychiatrist [66]. However, none of these models have come close to unrestricted Turing tests. Advances in large language models [8, 14, 19, 58] have led news and social media to produce anecdotal claims about current AI being sentient in conversations [43,60,67]. However, few studies rigorously and quantitatively assessed AIs in their ability to imitate humans in conversation. Preliminary works introduced unrestricted Turing tests in conversations with one exchange per conversation [75]. Here we provide extensive evaluations of AIs engaged in conversations with up to 24 exchanges.

3. Experiments

We introduce the six tasks (**Fig. 2**), how we created the datasets and how we set up the Turing tests (**Fig. 3**). Further details about each task, controls, and example snapshots of the Amazon Mechanical Turk (AMT) interfaces are provided in the **Supp. Sections S2 – S7**. All AMT experiments are based on "master" workers. We also collected demographic information about the participants as metadata, including their native language, age, gender,

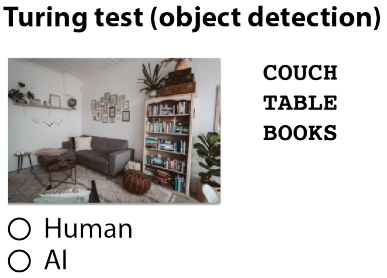


Figure 3. **Schematic illustration of the Turing test for the object detection task**. The judge is presented with an image and three labels and has to decide whether those labels were produced by a human or by an AI. For screenshots of the Turing test for each of the tasks, see **Supplementary Material**.

educational background, and the country they are originally from (**Fig. 2g**). For each task, we collect human answers and machine answers. During each Turing test, we present a single instance of the answers and ask participants to indicate whether the answer comes from a human or AI (e.g., **Fig. 3** for the Object detection task). Half of the time, the entry shown was from a human. The other half of the time, an AI answer was shown, sampling with equal probability from one of the different computational models used for each task. The trial order was randomized. No feedback was provided to the participants. Additional control trials were introduced for each specific task to ensure compliance.

Task	Num. Stimulus	Num. Turing Tests	Sources of Datasets	AI models
Color estimation	785	1,625	self-collect, MSCOCO [42]	Google Vision API Microsoft Azure Cognitive Services, MMCQ [6]
Object detection	808	1,975	self-collect, MSCOCO [42]	Google Vision API, Microsoft Azure Cognitive Services Amazon Rekognition, Detectron2 [69]
Attention Prediction	547	2,160	NatureDesign [73], FindingWaldo [73] NatureSaliency [72]	IVSN [27, 72, 73], DeepGaze3 [39], GBVS [29]
Image Captioning	1,000	8,140	self-collect MSCOCO [42], nocaps [3]	GIT [63], OFA [64], BLIP [41] ClipCap [46], Microsoft's Azure Cognitive Services [2]
Word Association	1,500	3,550	self-collect	Word2Vec [50], GPT2 [52], GPT3-embedding (davinci) [8], GPT3-prompt (text-curie-001) [8], GPT3-prompt (text-Davinci-002) [8].
Conversation	300×8 lengths	4,120	self-collect Topical-Chat dataset [24]	GPT3-text-davinci-002 [48], GPT3-text-curie-001 [48] Blenderbot [57], DialogPT [24]

Table 1. **Specifications of six Turing tasks** Source datasets, number of Turing tests conducted, number of stimulus, and AI models used to collect responses are listed for each task. See **Section 3** for task descriptions.

3.1. Color Estimation

Dataset. We collected a testing set of 1,000 images sampled from the validation set of MS-COCO [42] to ensure that computational models had not seen these images during training. We collected human responses for 785 images using AMT in the Color Estimation task. AMT subjects were asked "What is the dominant color in this image?", and had to select from 11 possible colors (**Fig. S1, S2**). For a sample image of the AMT interface, instructions, and a list of colors, see **Supp. Section S2**.

For AI models, we collected responses for the dominant color from Google's Cloud Vision API [1], Microsoft's Azure Cognitive Services [2], and the publicly available Median Cut Color Quantized (MCCQ) algorithm [6]. In total, 3,140 image-response pairs were collected (785 from humans and 2,355 from AI models, **Table 1**).

Turing test (Human Judge). Subjects were shown an image-response pair collected above, and asked the question "Was the color description made by a human or a machine?" (**Fig.S3, S4, S14**). Subjects were also asked to predict the gender of the speaker. An auxiliary True (50%)/False (50%) question about the presence of a background object present in the image was introduced as a control. The mean accuracy in this control task was 91%. A total of 1,625 responses were collected from human judges.

Turing test (AI judge). For each image-response pair, we collected image features from the last layer embedding of a ResNet18 model pretrained on ImageNet, and the textual features for the color using BERT [19]. These features were then concatenated, and passed into a Random Forest classifier to predict human versus AI labels, using a 90% / 10% split for cross-validation.

3.2. Object detection

Dataset. We collected responses for 808 images from the set of 1,000 images from MS-COCO in **Section. 3.1**. Subjects were presented with the question "What do you see in this image?", and were asked to enter three single word responses (**Fig. S16, S17**). We introduced several controls, ensuring that the responses were single words in English and that participants do not repeat the same response within or across images. For sample images of the AMT interface, instructions and controls, see **Supp. Section S3**.

We collected predicted labels from Google's Cloud Vision API, Microsoft's Azure Cognitive Services, Amazon's Rekognition API, and Facebook's Detectron2. For all networks, the top three predicted labels with the highest prediction score were collected. In total, 4,040 image-response pairs were collected (808 from humans and 3,232 from the four AI models, **Table 1**).

Turing test (Human judge). The Turing test for human judges was performed using the same protocol and controls described in **Section 3.1** (**Fig. 3, S18, S19, S29**). A total of 1,975 responses were collected from human judges.

Turing test (AI judge). As the images were the same for humans and AI models, we used solely textual features for the AI judge to classify a response as human or AI. We concatenated the three responses into a sentence, and collected textual features for the sentence using BERT [19]. These features were passed into an SVM classifier to classify responses into humans versus AI, with 90% / 10% split cross-validation.

3.3. Attention prediction

Dataset. We used eye movements (overt attention) from human subjects during two visual search tasks [73], and a free-viewing task [72]. We evaluated 7,000 scanpaths from 40 participants (**Table S2**). For the three datasets, we used a modified version of IVSN [72, 73], DeepGaze3 [39] and GBVS models [29] to generate eye movement predictions. **Supp. Section S4** provides examples of eye movement sequences from humans and models.

Turing test (human judge). Separate Turing tests were launched for eye movements from free-viewing tasks (80

judges) and visual search tasks (100 judges) (**Fig. S31, S47** and **Fig. S32, S47**). We presented infinitely repeating animated clips of eye movements from humans or model predictions with a maximum of 15 fixations to human judges on AMT. Each judge had to identify if the eye movements were from a human or a computational model. As a control, judges were also asked to answer "What do you see in the presented clip?" with one correct answer among 3 options. Responses from judges with a score < 7 out of 12 were not considered in the analyses.

Turing test (AI judge). We performed Turing tests using an SVM as an AI judge. Sequences of 10 fixations per trial from humans or computational models were fed as input in the form of an array of fixation coordinates to train an SVM to classify human versus machine eye movements. The SVM was trained using 10-fold cross validation. Model performance on validation sets across folds with 3 random seeds was calculated and averaged.

3.4. Image captioning

Dataset. We randomly sampled 250 images each from in-domain, near-domain, and out-of-domain categories from the validation set of the nocaps dataset [3] and 250 images from the MSCOCO test set [42], creating a set of 1,000 images. We collected 2,290 human captions with ≥ 6 words per caption and ≥ 2 captions per image from AMT participants (**Fig. S48, S49, S50, S65**). We implemented additional controls in our AMT interface. For example, workers were not allowed to submit a caption before viewing the image for ≥ 4s (**Supp. Section S5**).

To generate machine captions, we used: GIT [63], OFA [64], BLIP [41], ClipCap [46], and Microsoft's Azure Cognitive Services [2] (**Table S3**). For open-source models, we used the largest variants finetuned on the COCO Captions dataset [10, 42]. We collected 5,000 machine captions with 5 captions per image (**Supp. Section S5**).

Turing test (human judge). We collected responses from 293 AMT participants (**Fig. S51**). Each participant was presented with image-caption pairs and indicated whether the caption was generated by a human or AI. To ensure that the participants read the captions carefully, we prevented response times $<$ 3s. We removed responses from non-native English speakers (**Supp. Section S5**).

Turing test (AI judge). We trained an SVM model for binary classification (human versus machine) on the dataset of human and machine captions. We randomly sampled 400 captions from each of the 5 models to get 2,000 machine captions and combined them with our 2,000 human captions. We used the OpenAI API [47] to obtain 4,096-dimensional embeddings (text-similarity-curie-001 model) for each caption as input features to train the SVM with 10-fold cross-validation and 3 random seeds.

3.5. Word associations

Dataset. We chose 150 unique cue words (50 nouns, 50 verbs, and 50 adjectives), spanning a wide range of occurrence frequencies [59] (**Table S4**; see **Section S6** for multiple additional controls). Associations to each cue word were collected from human subjects (**Fig. S68, S69, S74**), and from the following language models: Word2vec [50], GPT2 [52], GPT3-embedding (based on davinci embedding), GPT3-curie-prompt (based on "curie" prompt completeion), and GPT3-davinci-prompt (based on "davinci" prompt completeion) [8]. For the human associations, we followed two procedures: (1) Free associations, whereby participants provided a one-word answer to the question: "What is the first word that comes to your mind when you hear the word [cue word]?" (**Fig. S68**); and (2) Prompt-based associations, whereby participants completed a prompt with one word (**Fig. S69**). The prompts used for the human prompt-completion were the same prompts used for GPT3-curie-prompt and GPT3-dacinci-prompt (**Table S6**). All participants were English native speakers living in the US. **Section S6** describes the implementation of each model to retrieve word associations.

Turing test (human judge). For the human-judge Turing tests, we collected data from 50 native English speakers on AMT (**Fig. S70**). In each trial, a cue word and a corresponding guess word (association) were presented and the judge had to choose whether the association was made by a human or by an AI model (**Section S6**).

Turing test (AI judge). We trained a linear SVM classifier with 10-fold cross-validation [15] to distinguish between human-made and machine-made associations. We used the the distance between the cue and guess word embeddings, based on (1) Word2Vec, (2) GPT2, or (3) GPT3 (davinci).

3.6. Conversation

Dataset. We collected 300 conversations between: (1) two humans, (2) a human and an AI model, (3) two AI models. For the conversations including humans, we recruited 150 fluent English participants to have a conversation over a chatting platform. The participants did not know whether they were speaking with another human or with an AI chatbot (see instructions in **Supp. Section S7.1.2**). We collected conversations containing 24 exchanges each. For the human-human conversations, we added 40 conversations from the Topical-Chat dataset [24], selected based on a minimum length of 24 exchanges. Multiple example conversations are included in **Supp. Section S7.4**.

For the AI chatbots, we used three state-of-the-art language models: Blenderbot3 (175B model) [57], GPT3 text-davinci-002 [48], and GPT3 text-curie-001 [48] (see settings, pre-processing, prompts, and control details in

Supp. Section S7.1.4).

Turing test (human judge). We chunked each conversation into 8 different lengths, including the initial 3, 6, 9, 12, 15, 18, 21, and 24 exchanges. There were 208 human judges (AMT: 200, in-lab: 8). The participants were presented with 20 randomly sampled chunked conversations with different lengths and had to respond, for each of the two speakers, whether the speaker was a human or a machine and the gender (**Fig. S89**). As a control, speakers also had to select the general topic of the conversation from a list of five topics. We only considered judges that correctly classified at least 15 topics out of 20 and removed incorrectly classified trials.

Turing test (AI judge). We evaluated whether simple AI models can discern whether a sentence was generated by a model or a human. We only examined single sentences here. Therefore, these results provide only an initial proof-of-principle lower bound for AI judges. We built four corpora, one containing all the sentences written by humans (the *human corpus*), and the others with the sentences produced by Blenderbot, GPT3text-davinci-002 and GPT3text-curie-001 (the *AI corpora*). We used *BERT* embeddings [19] to tokenize each sentence, and fed the tokenized sentences to a linear SVM trained to classify *human* vs. *AI* with 10-fold cross-validation.

4. Results

We summarize the results of all the Turing tests in **Fig. 4**, by averaging across all AI models and all human judge demographics. In the **Supplementary Material**, we show results separated by AI model and also for different human judge demographics. For each task, **Fig. 4** shows the proportion of times that a trial was classified as human (first column), or AI (second column), when the ground truth was human (first row), or AI (second row). Entries along each row add up to 100%. When comparing different AI models for a given task in terms of the ability to imitate humans, percentages closer to 50% indicate better models. In contrast, when comparing different judges (e.g., human judges versus AI judges, or human judges of different ages or educational backgrounds), higher overall accuracy indicates better judges.

4.1. Color estimation

Human judges distinguished AI answers as AI 58% of the time and human answers as human 55% of the time (**Fig. 4a**). We broke down performance based on each individual AI model (**Fig. S5**). The Google API performed slightly better (57%) than Azure API (60%) and MCCQ (65%). Even though the color MCCQ is a simple metric, it still achieved a moderately good performance in fooling humans 35% of the time. There was no major difference in performance of human judges across different age groups (**Figs. S6, S7, S8**), education levels (**Figs. S9, S10, S11**), or genders (**Figs. S12, S13**). In contrast, the AI judge classified human answers as human 43% of the time and AI answers as AI 34% of the time (**Fig. 4g**, see **Fig. S15** for individual AI models).

4.2. Object detection

Human judges distinguished AI answers as AI 69% of the time and human answers as human 52% of the time. We broke down performance based on each individual AI model (**Fig. S20**). Among all the AI models, Detectron performed the best (49%), with a large gap from the second best, Google API (65%). This modern object detection algorithm in computer vision not only achieves outstanding absolute scores in terms of standard evaluation metrics, such as mAP [53], its response patterns also closely mimic humans' by identifying top-3 salient objects in the scene. Specifically, we used the variant with MaskRCNN [35] trained on ImageNet and MS-COCO.

There was no major difference in performance of human judges across different age groups (**Figs. S21, S22, S23**), genders (**Fig. S27, S28**) or education levels (**Fig. S24, S25, S26**). All numbers were within a 5% difference of the average performance across all human judges reported above.

Next, we analyzed the classification performance of the AI judge for this task. In start contrast to human, the AI judge is able to distinguish between AI and human speakers much better **Fig 4**. The overall classification accuracy of AI judge is 81% (as compared to 56.5% of human judges). Specifically, AI judge can tell AI responses as AI with a 90% accuracy, and human responses as human with 72%. The easiest to classify are responses from the Azure API with AI judges getting a 94% accuracy, while the hardest to classify are Detectron and Amazon's Rekognition API with accuracy of 67% each (**Fig. S30**).

4.3. Attention prediction

Human judges distinguished human eye movements as human 63% of the time and AI-generated eye movements as AI 50% of the time (**Fig. 4c**). We examined the Turing test performance for each AI model separately (**Fig. S33**). IVSN [72, 73] outperformed GBVS [29] and DeepGaze3 [39] by 4% and 17%, respectively. We evaluated whether the agent's goals during eye movements might influence the subsequent Turing tests. Human judges performed 2.5% better in free-viewing (**Fig. S36**) compared to visual search tasks (**Fig. S35**). Free-viewing is mostly driven by pure bottom-up saliency, and it may be easier to discern scanpath patterns without target-directed modulation. Consistent with this idea, GBVS generated more human-like scanpaths during free-viewing compared with visual search, whereas DeepGaze3 performed much worse in free-viewing tasks

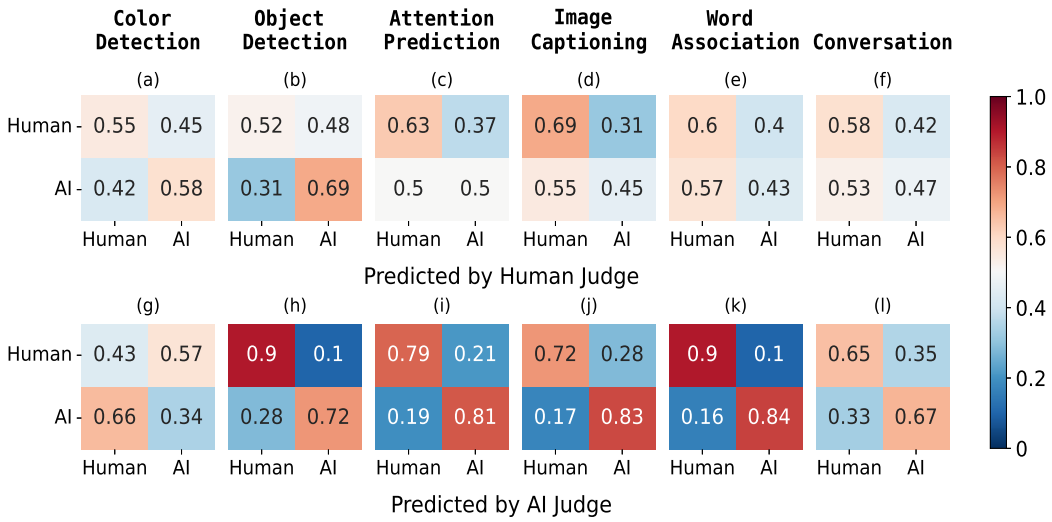


Figure 4. **Results of the Turing test for each task**. Turing test results for human judges (top row) and machine judges (bottom row). For each task, the confusion matrices report the percentage of times when the trial was labeled "human" (first column) or "AI" (second column) when the ground truth was human (first row) and AI (second row). Percentages add up to 100 within each row. Here all AI models were averaged together. See **Supplementary Material** for results from each AI model and different human judge demographic groups.

than visual search. IVSN performance was similar in both tasks, which emphasizes the importance of incorporating both bottom-up and top-down attention mechanism in computational models of human attention.

Judges across different ages (**Fig. S39**, **S40**, **S41**), and also male or female judges (**Fig. S42**, **S43**), performed equally well in the Turing tests. Judges with a postgraduate degree performed slightly worse than the ones with bachelor degrees or lower (**Fig. S44**, **S45**, **S46**).

As an initial evaluation of AI judges, we trained an SVM classifier purely based on the sequences of eye fixations regardless of the image features (**Fig. 4h**). Interestingly, a simple SVM AI judge performed 20% better than human judges. AI judges outperformed human judges across different models **Fig. S34**), and different tasks (**Fig. S37**, **S38**). However, this result should be interpreted with caution since the AI judge was explicitly trained to classify scanpaths while human judges typically do not have such prior training.

4.4. Image captioning

Human judges distinguished human captions as human 69% of the time and AI captions as AI 45% of the time (**Fig. 4d**). There were rather large differences among AI models (**Fig. S52**), with proportions of AI captions labeled AI ranging from 37% (BLIP) to 59% (ClipCap). CIDEr is a standard evaluation metric for assessing the quality of AI-generated captions [62]. OFA shows better performance than GIT and ClipCap in CIDEr score [64], but it is not better at imitating human captions, highlighting the differences between traditional performance metrics and Turing tests. Human judges labeled captions as humans slightly more often for imgaes from in-domain nocaps (**Fig. S62**), compared to near-domain nocaps and out-of-domain nocaps (**Fig. S63**, **S64**). There were no differences between judges of different ages (**Fig. S54**, **S55**, **S56**), male versus female judges (**Fig. S57**, **S58**), or judges with different education levels (**Fig. S59**, **S60**, **S61**).

As a proof-of-principle to build an AI judge, we used the OpenAI Embeddings API (text-similarity-curie-001 model) to extract features and trained a linear SVM to discriminate human versus machine responses. Surprisingly, even though this classifier did not use image features, the AI judge could discern human versus machine answers with 77.5% accuracy, well above the performance of human judges (**Fig. 4j**). Similar results were observed across all the different AI models (**Fig. S53**).

4.5. Word associations

Human judges labeled human word associations as humans 60% of the time and AI word associations as AI 43% of the time (**Fig. 4e**). Results were similar for different AI models (**Fig.S66**). Surprisingly, the Word2Vec model produced word associations that were harder to discriminate from human ones. There were no major differences between human judges across different ages (**Fig. S71**), genders (**Fig. S72**), or education levels (**Fig. S73**).

We used the various word embedding from GPT-series models and Word2Vec as features to train 3 SVM judges (independent SVM classifiers based on the embedding of each AI model). In contrast to human judges, AI

judges could easily distinguish human versus machine word associations (**Fig. 4k**). All AI judges outperformed human judges (**Fig. S67**). As expected, when training an SVM classifier based on the embedding of the same model whose embedding where also used for generating the guess words, this model classification performance was essentially perfect. For instance, on the 3rd row on the left confusion matrix in **Fig. S67a**, the AI judge trained on Word2Vec embedding indeed perfectly predicted all guess words generated by Word2Vec embeddings as AI. While these cases are good as a sanity check, they should not be considered in the general evaluation of the AI judges performance. Hence, we trained 3 different SVM to avoid such biases.

4.6. Conversation

Human judges distinguished human participants in conversations as humans 58% of the time and AI agents as AI 47% of the time (**Fig. 4f, S86, S87**). Unlike AI models passing the Turing tests in restricted conversation topics, it is interesting to note an overall accuracy of 53.5% here in largely unrestricted conversations.

We separately considered human-human, human-AI, and AI-AI conversations (**Fig. S75**). Surprisingly, human-human conversations were classified as human only 61% of the time and AI-AI conversations were classified as human 56% of the time. Blenderbot was classified as human 64% of the time, suggesting that AIs can be perceived as more human than humans themselves. In human-AI conversations, human participants were labeled as humans 61% of the time, and AIs were classified as AI 55% of the time. The overall classification accuracy in human-AI conversations was higher than AI-AI conversations (58% versus 41%), suggesting that AIs reveal their true self more often when talking to humans than amongst themselves. This observation is consistent with the notion that human judges are more accurate in making comparisions rather than absolute evaluations.

When comparing different AI models (**Fig. S75**), Blenderbot was most often labeled as humans, 66% of the time in AI-AI conversations and 50% of the time in human-AI conversations (**Table S7**). The results of the Turing test depended on the conversation length (**Fig. S88**). AI models were less adept at passing as humans in longer conversations. Among all the AI models, GPT3-curie showed the sharpest drop while Blenderbot maintained relatively high performance. These observations highlight that model sizes, specific training on conversation data, and incorporation of external memory modeling past conversation history are important factors when imitating humans in conversations.

Younger judges performed better in discerning AIs from humans than older judges in AI-AI conversations (**Fig. S77, S78, S79, S85a**). Surprisingly, male judges performed slightly better than female judges (60% versus 57.5%), especially in AI-AI conversations (46% versus 39%) (**Fig. S80, S81, S85b**, **Table S8**). Intriguingly, education had a slight negative relation with classification accuracy of human judges (54%, 53% and 51% for middle-high school, college and postgraduate degrees respectively), especially in human-AI conversations. However, this trend was reversed in AI-AI conversations where postgraduate judges performed better than middle-high school judges (53% versus 41%) (**Fig. S82, S83, S84, S85c**).

We trained a simple SVM judge to distinguish whether a sentence in a conversation was from humans or AIs. Consistently with the other experiments, the AI judge beat human judges by a large margin (66% versus 53.5%, **Fig. 4l**). This AI judge performed particularly well in classifying Blenderbot sentences (**Fig. S76**), in stark contrast with human judges who were more easily fooled by Blenderbot than GPT models. Human judges likely focus on high-level conversation understanding rather than single-sentence statistics in the Turing tests.

5. Discussion

The Turing test has been extensively discussed, and contested, as a means to assess general intelligence. Instead, we focus on Turing tests as a metric to evaluate whether an algorithm can imitate humans or not. **Table S1** summarizes the observations in a highly simplified binary format; this table is a grand average and the reader is referred to all the actual numbers for a more accurate description of the findings. Remarkably, the algorithms tested throughout the current work seem to be quite close to passing a Turing test when evaluated by human judges. Given that imitating humans can be very good for certain purposes but could also easily be turned into potentially evil applications, these observations call for more extensive and rigorous scrutiny of machines that can imitate AI.

One step to mitigate risks from human imitators is to build AI judges. Our results show that even simple AI judges like the ones introduced here can do a better job than human judges in detecting machine answers. The results of current AI judges should not be over-interpreted because AI judges were explicitly trained to classify responses from humans versus AIs, while human judges were not. This point raises the possibility that humans may be trained to better recognize machine answers in the future.

An algorithm's ability to imitate humans did not always correlate with traditional performance metrics like accuracy, implying that Turing tests provide a complementary assessment of AI models to existing benchmarking frameworks. Comparisons between models in Turing tests also provide insights helpful for developing future AI models that can better align with humans.

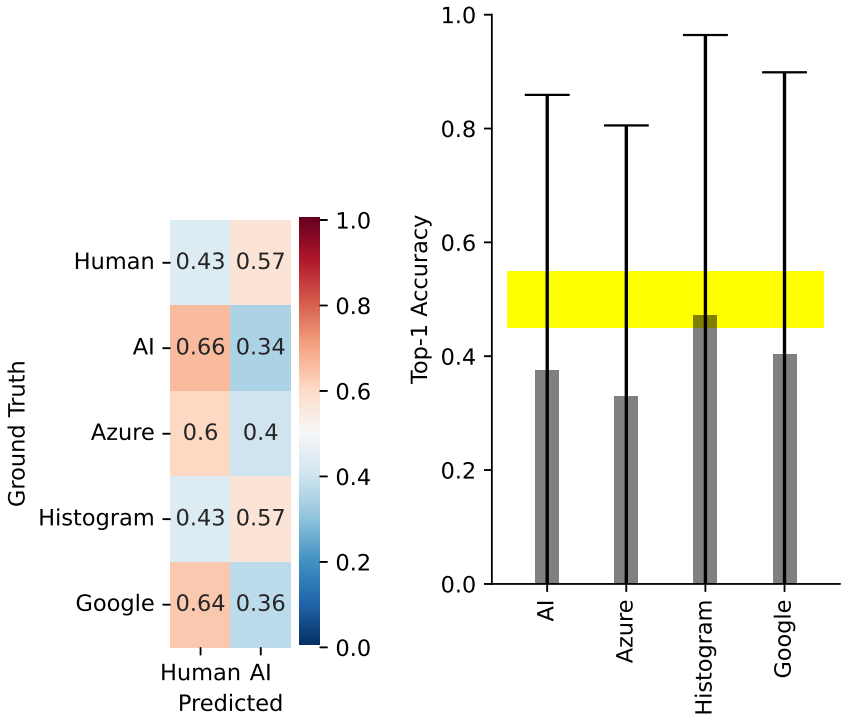
The datasets and evaluations introduced here are quite extensive (21570 Turing test trials, 904 human and AI judges, 6 vision and language tasks, several demographic groups), but they barely scratch the surface of what needs to be done. There are essentially infinite possible Turing tests. The results of a Turing test depend on the task, the algorithm, how the question is formulated, the characteristics of the human judge and many other variables

This work provides a comprehensive, yet certainly far from exhaustive, evaluation of state-of-the-art AI models in terms of human emulation. These efforts pave the way for the research community to expand Turing tests to other research areas, to build better imitators, and better detectors of imitators. If more AI models can "blend" in among humans and take over tasks that were originally unique yardsticks of being humans, this makes us ponder what makes us humans and whether we are mentally, ethically, and legally ready for the rapid revolution brought forth by AI technologies.

List of Supplementary Sections

List of Supplementary Figures

List of Tables



(a) Confusion matrix

(b) Top1 accuracy

Figure S15. **Color estimation**. Results of the Turing test for AI judges (a) Confusion matrix (b) Top-1 accuracy.

References

[1] Google vision api. https://cloud.google.com/vision. Accessed: 2022-10-30. 4

[2] Microsoft azure cognitive api. https://azure.microsoft.com/en-us/products/cognitive-services/. Accessed: 2022-10-30. 4, 5, 69, 71

[3] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8948–8957, 2019. 4, 5, 69

[4] Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. Understanding robustness of transformers for image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10231–10241, 2021. 2, 14

[5] N Block. Behaviourism and psychologism. *Philosophical Review*, 90(5):43, 1981. 2, 14

[6] Dan S. Bloomberg and Leptonica. Color quantization using modified median cut. 2008. 4

[7] Philipp Bomatter, Mengmi Zhang, Dimitar Karev, Spandan Madan, Claire Tseng, and Gabriel Kreiman. When pigs fly: Contextual reasoning in synthetic and natural scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 255–264, 2021. 2, 14

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 2, 3, 4, 5, 14, 15, 90

[9] Arjun Chandrasekaran, Viraj Prabhu, Deshraj Yadav, Prithvijit Chattopadhyay, and Devi Parikh. Do explanations make vqa models more predictable to a human? *arXiv preprint arXiv:1810.12366*, 2018. 2, 14

[10] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 5, 69

[11] Brian Christian. *The most human human: What talking with computers teaches us about what it means to be alive*. Anchor, 2011. 2, 14

[12] Kenneth Mark Colby. Modeling a paranoid mind. *Behavioral and Brain Sciences*, 4(4):515–534, 1981. 3, 15

[13] Kenneth Mark Colby, Sylvia Weber, and Franklin Dennis Hilf. Artificial paranoia. *Artificial Intelligence*, 2(1):1–25, 1971. 3, 15

[14] Eli Collins and Zoubin Ghahramani. Lamda: our breakthrough conversation technology. *The Keyword, May*, 18, 2021. 3, 15

[15] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995. 5, 91

[16] Yin Cui, Guandao Yang, Andreas Veit, Xun Huang, and Serge Belongie. Learning to evaluate image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5804–5812, 2018. 2, 14

[17] Nicola Damassino. The questioning turing test. *Minds and Machines*, 30(4):563–587, 2020. 2, 14

[18] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380, 2014. 2, 14

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2, 3, 4, 6, 14, 15, 107

[20] Samuel Dodge and Lina Karam. A study and comparison of human and deep learning recognition performance under visual distortions. In *2017 26th international conference on computer communication and networks (ICCCN)*, pages 1–7. IEEE, 2017. 2, 14

[21] Gamaleldin Elsayed, Shreya Shankar, Brian Cheung, Nicolas Papernot, Alexey Kurakin, Ian Goodfellow, and Jascha Sohl-Dickstein. Adversarial examples that fool both computer vision and time-limited humans. *Advances in neural information processing systems*, 31, 2018. 2, 14

[22] Robert M French. The turing test: the first 50 years. *Trends in cognitive sciences*, 4(3):115–122, 2000. 1

[23] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018. 14

[24] Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Proc. Interspeech 2019*, pages 1891–1895, 2019. 4, 5, 105, 122

[25] Keith Gunderson. The imitation game. *Mind*, 73(290):234–245, 1964. 2, 14

[26] Keith Gunderson. *Mentality and machines*. U of Minnesota Press, 1985. 2, 14

[27] Shashi Kant Gupta, Mengmi Zhang, Chia-Chien Wu, Jeremy Wolfe, and Gabriel Kreiman. Visual search asymmetry: Deep nets and humans share similar inherent biases. *Advances in Neural Information Processing Systems*, 34:6946–6959, 2021. 2, 4

[28] Holly Kathleen Hall. Deepfake videos: When seeing isn't believing. *Cath. UJL & Tech*, 27:51, 2018. 2

[29] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. *Advances in neural information processing systems*, 19, 2006. 4, 6, 50, 51

[30] Stevan Harnad. Minds, machines and searle. *Journal of Experimental & Theoretical Artificial Intelligence*, 1(1):5–25, 1989. 2, 14

[31] Stevan Harnad. Other bodies, other minds: A machine incarnation of an old philosophical problem. *Minds and Machines*, 1(1):43–54, 1991. 2, 14

[32] Stevan Harnad. Levels of functional equivalence in reverse bioengineering. *Artificial life*, 1(3):293–301, 1994. 2, 14

[33] Stevan Harnad. Turing on reverse-engineering the mind. *Journal of Logic, Language, and Information*, 1999. 2, 14

[34] Patrick Hayes and Kenneth Ford. Turing test considered harmful. In *IJCAI (1)*, pages 972–977. Citeseer, 1995. 1, 2, 14

[35] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 6

[36] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1072–1080, 2015. 2, 14

[37] Marzena Karpinska, Nader Akoury, and Mohit Iyyer. The perils of using mechanical turk to evaluate open-ended text generation. *arXiv preprint arXiv:2109.06835*, 2021. 2, 14

[38] Jungo Kasai, Keisuke Sakaguchi, Lavinia Dunagan, Jacob Morrison, Ronan Le Bras, Yejin Choi, and Noah A Smith. Transparent human evaluation for image captioning. *arXiv preprint arXiv:2111.08940*, 2021. 2, 14

[39] Matthias Kümmerer, Matthias Bethge, and Thomas SA Wallis. Deepgaze iii: Modeling free-viewing human scanpaths with deep learning. *Journal of Vision*, 22(5):7–7, 2022. 4, 6, 50, 51

[40] Katrina LaCurts. Criticisms of the turing test and why you should ignore (most of) them. *Official blog of MIT's course: Philosophy and theoretical computer science*, 2011. 2, 14

[41] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 4, 5, 69, 70, 71

[42] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 4, 5

[43] Ramishah Maruf. Google fires engineer who contended its ai technology was sentient. *CNN*. 3, 15

[44] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021. 2, 14

[45] Michael L Mauldin. Chatterbots, tinymuds, and the turing test: Entering the loebner prize competition. In *AAAI*, volume 94, pages 16–21, 1994. 2, 14

[46] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 4, 5, 69, 71

[47] Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, et al. Text and code embeddings by contrastive pre-training. *arXiv preprint arXiv:2201.10005*, 2022. 5

[48] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. 4, 5, 105

[49] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 2, 14

[50] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 4, 5, 90

[51] Richard L Purtill. Beating the imitation game. *Mind*, 80(318):290–294, 1971. 2, 14

[52] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 4, 5, 90

[53] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 6

[54] John R Searle. Minds, brains, and programs. *Behavioral and brain sciences*, 3(3):417–424, 1980. 2, 14

[55] Terrence Sejnowski. Large language models and the reverse turing test. *arXiv preprint arXiv:2207.14382*, 2022. 2, 14

[56] Sasha Sheng, Amanpreet Singh, Vedanuj Goswami, Jose Magana, Tristan Thrush, Wojciech Galuba, Devi Parikh, and Douwe Kiela. Human-adversarial visual question answering. *Advances in Neural Information Processing Systems*, 34:20346–20359, 2021. 2, 14

[57] Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, Morteza Behrooz, William Ngan, Spencer Poff, Naman Goyal, Arthur Szlam, Y-Lan Boureau, Melanie Kambadur, and Jason Weston. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage, 2022. 4, 5, 105

[58] Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, et al. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *arXiv preprint arXiv:2208.03188*, 2022. 3, 15

[59] Robyn Speer, Joshua Chin, Andrew Lin, Sara Jewett, and Lance Nathan. Luminosoinsight/wordfreq: v2.2, Oct. 2018. 5, 90

[60] Nitasha Tiku. The google engineer who thinks the company's ai has come to life. *WashingtonPost*. 3, 15

[61] Alan M Turing. Computing machinery and intelligence. In *Parsing the turing test*, pages 23–65. Springer, 2009. 1, 2, 14

[62] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 7

[63] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022. 4, 5, 69, 70, 71

[64] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR, 2022. 4, 5, 7, 69, 70, 71

[65] Stuart Watt. Naive psychology and the inverted turing test. *Psycoloquy*, 7(14):463–518, 1996. 2, 14

[66] Joseph Weizenbaum. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, 1966. 3, 15

[67] Tiffany Wertheimer. Blake lemoine: Google fires engineer who said ai tech has feelings. *BBC news*. 3, 15

[68] Mika Westerlund. The emergence of deepfake technology: A review. *Technology Innovation Management Review*, 9(11), 2019. 2

[69] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 4

[70] Ming Yan, Haiyang Xu, Chenliang Li, Junfeng Tian, Bin Bi, Wei Wang, Weihua Chen, Xianzhe Xu, Fan Wang, Zheng Cao, et al. Achieving human parity on visual question answering. *arXiv preprint arXiv:2111.08896*, 2021. 2, 14

[71] Zhibo Yang, Lihan Huang, Yupei Chen, Zijun Wei, Seoyoung Ahn, Gregory Zelinsky, Dimitris Samaras, and Minh Hoai. Predicting goal-directed human attention using inverse reinforcement learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 193–202, 2020. 2, 14

[72] Mengmi Zhang, Marcelo Armendariz, Will Xiao, Olivia Rose, Katarina Bendtz, Margaret Livingstone, Carlos Ponce, and Gabriel Kreiman. Look twice: A generalist model predicts return fixations across tasks and species. *PLoS Computational Biology*, page In Press, 2022. 4, 6, 50, 51

[73] Mengmi Zhang, Jiashi Feng, Keng Teck Ma, Joo Hwee Lim, Qi Zhao, and Gabriel Kreiman. Finding any waldo with zero-shot invariant and efficient visual search. *Nature communications*, 9(1):1–15, 2018. 2, 4, 6, 14, 50, 51

[74] Mengmi Zhang, Claire Tseng, and Gabriel Kreiman. Putting visual object recognition in context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12985–12994, 2020. 2, 14

[75] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*, 2019. 3, 15, 105