arXiv:2211.13087v1 [cs.CV] 23 Nov 2022

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.

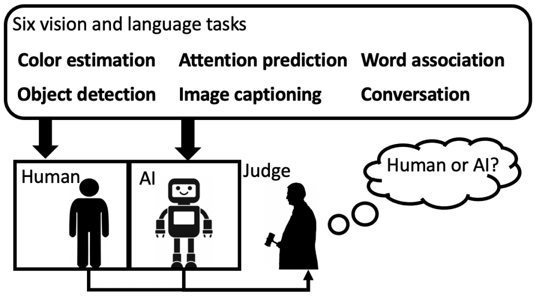


Figure 1. **Schematic illustration of Turing tests in six vision and language tasks**. A Turing test works with a judge asking a test subject (either a human or an AI agent) a series of tasks. Each party is kept in a separate room, so no physical contact is allowed. The AI passes the Turing test if the 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

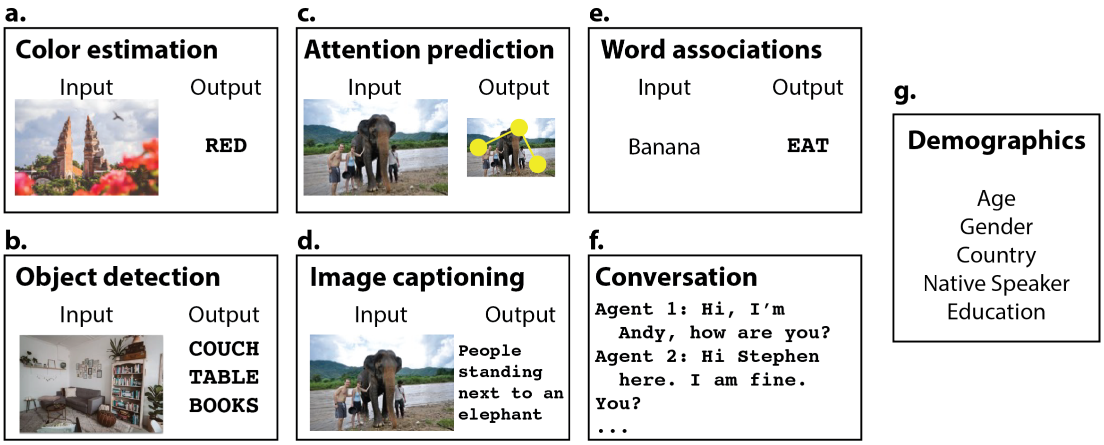


Figure 2. **Schematic of the 6 tasks**. We systematically evaluate 3 vision tasks, 1 vision-language task, and 2 language tasks. **a**. In Color estimation, the agent is presented with an image and has to output the main color. **b**. In Object detection, the agent is presented with an image and has 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, 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.

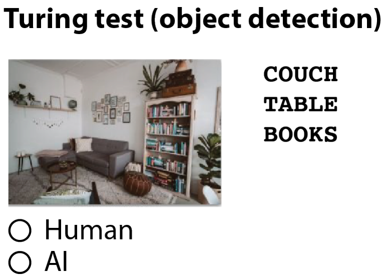


Figure 3. **Schematic illustration of the Turing test for the object detection task**. The judge is presented with an image and three labels and has to decide whether those labels were produced by a human or by an AI. For screenshots of the Turing test for each of the tasks, see **Supplementary Material**.

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 $\geq$ 6 words per caption and $\geq$ 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 $\geq$ 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

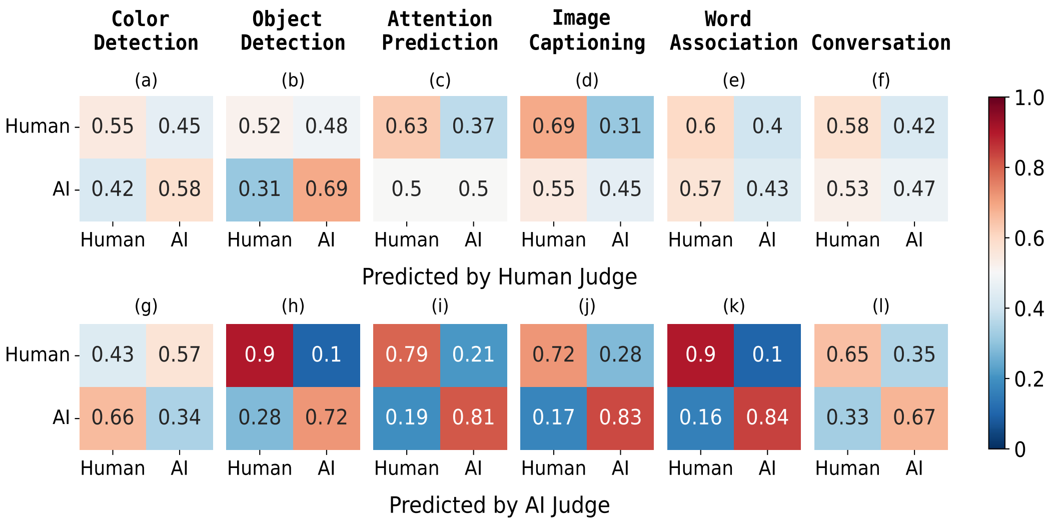


Figure 4. **Results of the Turing test for each task**. Turing test results for human judges (top row) and machine judges (bottom row). For each task, the confusion matrices report the percentage of times when the trial was labeled "human" (first column) or "AI" (second column) when the ground truth 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

S1 Background and discussion **14**
S1.1 Glimpse of the 70-year history of Turing test . . . 14
S1.2 AI versus humans in vision tasks . . . 14
S1.3 AI versus humans in language tasks . . . 14

S2 Color estimation **16**
S2.1 Dataset . . . 16
S2.1.1 Collecting human responses from AMT . . . 16
S2.1.2 Collecting AI responses . . . 16
S2.2 Turing test . . . 16
S2.2.1 Collecting human judge responses for Turing test . . . 16
S2.2.2 Demographic information of responders . . . 16
S2.2.3 Human judge performance based on demographic information . . . 16
S2.2.4 AI judge . . . 17

S3 Object Detection **33**
S3.1 Dataset . . . 33
S3.1.1 Collecting human responses from AMT . . . 33
S3.1.2 Collecting AI responses . . . 33
S3.2 Turing test . . . 33
S3.2.1 Collecting human judge responses for Turing test . . . 33
S3.2.2 Demographic information of responders . . . 33
S3.2.3 Human judge performance based on demographic information . . . 33
S3.2.4 AI judge . . . 34

S4 Attention Prediction **50**
S4.1 Dataset . . . 50
S4.1.1 Human eye movement responses . . . 50
S4.1.2 AI responses . . . 50
S4.2 Turing test . . . 50
S4.2.1 Collecting human judge responses for Turing test . . . 50
S4.2.2 AI judge . . . 50
S4.3 Discussion . . . 51

S5 Image captioning **69**
S5.1 Dataset . . . 69
S5.1.1 Collecting human captions . . . 69
S5.1.2 Collecting machine captions . . . 69
S5.2 Turing test . . . 69
S5.3 Discussion . . . 70

S6 Word Associations **90**
S6.1Dataset . . . 90
S6.1.1 Cue words . . . 90
S6.1.2 Collecting human responses from AMT . . . 90
S6.1.3 Collecting AI responses . . . 90
S6.2Turing test . . . 91
S6.2.1 Human judge Turing test . . . 91
S6.2.2 AI judge Turing test . . . 91

S7 Conversation 105
S7.1 Dataset . . . 105
S7.1.1 Human participants . . . 105
S7.1.2 Dataset collection: Instructions to human participants . . . 105
S7.1.3 AI conversation bots . . . 105
S7.1.4 Dataset collection: Prompt and settings for GPT3text-davinci002 and GPT3text-curie-001 . . . 106
S7.2 Turing test . . . 106
S7.2.1 Collecting human judge resposnes for Turing test . . . 106
S7.2.2 AI judge . . . 107
S7.3 Results and discussion . . . 107
S7.3.1 Confusion matrix and Top-1 accuracy . . . 107
S7.3.2 Gender perception . . . 107
S7.3.3 Effects of judge demographics . . . 107
S7.3.4 Comparison between AMT and in-person experiments . . . 107
S7.3.5 Effect of conversation length . . . 108
S7.4Examples of collected conversations . . . 125
S7.4.1 Example of conversations: human-human . . . 125
S7.4.2 Example of conversations: blenderbot-blenderbot . . . 125
S7.4.3 Example of conversations: GPT3textdavinci002-GPT3davincidavinci002 - successful . . . 126
S7.4.4 Example of conversations: GPT3textdavinci002-GPT3davincidavinci002 - discarded . . . 127
S7.4.5 Example of conversations: GPT3textcurie001-GPT3textcurie001 - successful . . . 127
S7.4.6 Example of conversations: Human-Blenderbot . . . 128
S7.4.7 Example of conversations: Human-GPT3textdavinci002 . . . 129
S7.4.8 Example of conversations: Human-GPT3textcurie001 . . . 130
S7.4.9 Example of conversations: DialoGPT-DialoGPT . . . 131

List of Supplementary Figures

S1 **Color estimation**. AMT user interface for collecting responses. . . . 18
S2 **Color estimation**. Random samples from our collected color estimation dataset. . . . 19
S3 **Color estimation**. AMT user interface for collecting human judge responses (Turing test). . . . 20
S4 **Color estimation**. Random samples from our collected Turing tests. . . . 21
S5 **Color estimation**. Results of the Turing test for human judges. . . . 22
S6 **Color estimation**. Results of the Turing test for human judges with age level below 35. . . . 23
S7 **Color estimation**. Results of the Turing test for human judges with age level 35-45. . . . 24
S8 **Color estimation**. Results of the Turing test for human judges with age above 45. . . . 25
S9 **Color estimation**. Results of the Turing test for human judges with education level below Bachelors. . . . 26
S10 **Color estimation**. Results of the Turing test for human judges with education level of Bachelors. . . . 27
S11 **Color estimation**. Results of the Turing test for human judges with education level above Bachelors. . . . 28
S12 **Color estimation**. Results of the Turing test for human judges with Male gender. . . . 29
S13 **Color estimation**. Results of the Turing test for human judges with Female gender. . . . 30
S14 **Color Estimation.** Demographic information for the human judges for the color estimation task . . . 31
S15 **Color estimation**. Results of the Turing test for AI judges . . . 32
S16 **Object detection**. AMT user interface for collecting responses. . . . 35
S17 **Object detection**. Random samples from our collected object detection dataset. . . . 36
S18 **Object detection**. AMT user interface for collecting human judge responses (Turing test). . . . 37
S19 **Object detection**. Random samples from our collected Turing tests. . . . 38
S20 **Object detection**. Results of the Turing test for human judges. . . . 39
S21 **Object detection**. Results of the Turing test for human judges with age below 35. . . . 40
S22 **Object detection**. Results of the Turing test for human judges with age 35-45. . . . 41
S23 **Object detection**. Results of the Turing test for human judges with age above 45. . . . 42
S24 **Object detection**. Results of the Turing test for human judges with education level below Bachelors. . . . 43
S25 **Object detection**. Results of the Turing test for human judges with education level of Bachelors. . . . 44
S26 **Object detection**. Results of the Turing test for human judges with education level above Bachelors. . . . 45

S27 **Object detection**. Results of the Turing test for human judges with Male gender. 46
S28 **Object detection**. Results of the Turing test for human judges with Female gender. 47
S29 **Object Detection.** Demographic information for the human judges for the object detection task 48
S30 **Object detection**. Results of the Turing test for AI judges 49
S31 **Attention prediction.** AMT user interface for collecting human judge responses for the free-viewing task (Turing test). 52
S32 **Attention prediction.** AMT user interface for collecting human judge responses for the visual search task (Turing test). 53
S33 **Attention prediction**. Results of the Turing test averaged over visual search and free-viewing tasks for human judges. 54
S34 **Attention prediction**. Results of the Turing test averaged over visual search and free-viewing tasks for AI judges. 55
S35 **Attention prediction**. Results of the Turing test in the visual search task for human judges. 56
S36 **Attention prediction**. Results of the Turing test in the free-viewing task for human judges. 57
S37 **Attention prediction**. Results of the Turing test in the visual search task for AI judges. 58
S38 **Attention prediction.** Results of the Turing test in the free-viewing task for AI judges. 59
S39 **Attention prediction.** Results of the Turing test averaged over visual search and free-viewing tasks for human judges below age 35 60
S40 **Attention prediction**. Results of the Turing test averaged over visual search and free-viewing tasks for human judges between age 35 and 45. (a) Confusion matrix (b) Top-1 accuracy. 61
S41 **Attention prediction.** Results of the Turing test averaged over visual search and free-viewing tasks for human judges above age 45 62
S42 **Attention prediction.** Results of the Turing test averaged over visual search and free-viewing tasks for male human judges. 63
S43 **Attention prediction.** Results of the Turing test averaged over visual search and free-viewing tasks for female human judges. 64
S44 **Attention prediction.** Results of the Turing test averaged over visual search and free-viewing tasks for human judges with highest educational level of middle/high school. 65
S45 **Attention prediction.** Results of the Turing test averaged over visual search and free-viewing tasks for human judges with highest educational level of Bachelor. 66
S46 **Attention prediction.** Results of the Turing test averaged over visual search and free-viewing tasks for human judges with highest educational level of Master or Post-graduate. 67
S47 **Attention prediction.** Demographic information for the human judges for the visual search and free-viewing tasks combined 68
S48 **Image captioning.** AMT user interface for collecting responses. 72
S49 **Image captioning.** Random samples from our collected caption dataset. 73
S50 **Image captioning.** Average caption lengths. 74
S51 **Image captioning.** AMT user interface for Turing test. 75
S52 **Image captioning.** Results of the Turing test for human judges 76
S53 **Image captioning.** Results of the Turing test for AI judges. 77
S54 **Image captioning.** Results of the Turing test for human judges below age 35. 78
S55 **Image captioning.** Results of the Turing test for human judges between age 35 and 45. 79
S56 **Image captioning.** Results of the Turing test for human judges above age 45. 80
S57 **Image captioning.** Results of the Turing test for male human judges. 81
S58 **Image captioning.** Results of the Turing test for female human judges. 82
S59 **Image captioning.** Results of the Turing test for human judges with highest education level of middle/high school. 83
S60 **Image captioning.** Results of the Turing test for human judges with highest education level of Bachelor. . . . 84
S61 **Image captioning.** Results of the Turing test for human judges with highest education level of Master or Post-graduate. 85
S62 **Image captioning.** Results of the Turing test for human judges with in-domain nocaps images. 86
S63 **Image captioning.** Results of the Turing test for human judges with near-domain nocaps images. 87
S64 **Image captioning.** Results of the Turing test for human judges with out-of-domain nocaps images. 88

S65 **Image captioning.** Demographic information for the human judges. (a) Age. (b) Gender. (c) Education level. 89
S66 **Word Associations.** Results of the Turing test for human judges. 92
S67 **Word Associations.** Results of the Turing test for AI judges. 93
S68 **Word Associations.** AMT user interface for collecting word associations responses, for free associations . . 94
S69 **Word Associations.** AMT user interface for collecting responses, for prompt-guided associations 95
S70 **Word Associations.** AMT user interface for collecting human judge responses (Turing test). 96
S71 **Word Associations.** Results of the Turing test for human judges in different age groups 97
S72 **Word Associations.** Results of the Turing test for different genders. 98
S73 **Word Associations.** Results of the Turing test for human judges for different educational levels. 99
S74 **Word Associations.** Demographic information for the human judges for the word association task 100
S75 **Conversation.** Results of the Turing test for human judges 109
S76 **Conversation.** Results of the Turing test for AI judges. 110
S77 **Conversation.** Results of the Turing test for human judges below age 35. 111
S78 **Conversation.** Results of the Turing test for human judges between age 35 and 45. 112
S79 **Conversation.** Results of the Turing test for human judges above age 45. 113
S80 **Conversation.** Results of the Turing test for male human judges. 114
S81 **Conversation.** Results of the Turing test for female human judges 115
S82 **Conversation.** Results of the Turing test for human judges with highest education level of middle/high school. 116
S83 **Conversation.** Results of the Turing test for human judges with highest education level of Bachelor. 117
S84 **Conversation.** Results of the Turing test for human judges with highest education level of Master and Post-graduate. 118
S85 **Conversation.** Demographic information for the human judges. 119
S86 **Conversation.** Results of the Turing test for human judges from data collected on AMT 120
S87 **Conversation.** Results of the Turing test for human judges during in-lab experiments. 121
S88 **Conversation.** Length dependence of Turing test results for different settings and models. 122
S89 **Conversation.** Screenshot of the conversation task performed by human judges on AMT. 123

List of Tables

S1 **Summary of Turing test results for human judges**. 15
S2 **Attention Prediction.** Datasets used for visual search and free-viewing tasks and number of human scanpaths in the dataset. 51
S3 **Image captioning.** Different model variants used for collecting machine captions. 71
S4 **Word associations**. Cue words and their associations (‘Guess Words’) predicted by the five different AI language models. The associations for the first 3 models (Word2vec, GPT2, and GPT3-embedding) were derived from the corresponding embeddings, while the associations of the last two models (GPT3 prompt curie and GPT3 prompt davinci) were based on prompts (the full prompts are described in **Table S6** 101
S5 **Word associations**. Cue words and their associations - *Cont. from previous page* 102
S6 **Word associations**. The 150 prompts given for the GPT-3 curie and GPT-3 davinci models (for their ’prompt’ version), which were also given for the human participants who made prompt-based associations, and were generated using a held-out set containing human word-pair associations. The cue words are presented in the leftmost columns, and the guess words of the curie-prompt model and the davinci-prompt model are presented in the two right columns. 103
S7 **Conversations**. Results for the Turing test for different AI models and conversation types. The Turing test criteria is the same as Table S1 124
S8 **Conversation**. Gender perception of humans and machines by the human judges. 124

S1. Background and discussion

S1.1. Glimpse of the 70-year history of Turing test

The Turing test was originally proposed by Alan Turing as an imitation game where a machine tries to pass as human during a conversation and a human judge has to determine whether they are interacting with a human or not [61]. For the past 70 years since Turing's classical article, the Turing test has generated extensive controversy and discussion. In parallel with the unbounded optimistic attitudes towards AI in the 1960s and the sober realization of the immense difficulties in AI afterwards, many in the scientific community have shifted from the question of whether the Turing test is a valid and meaningful measure of intelligence [25, 26, 34, 40, 51] to the question of whether any machine 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]. Different from these works studying the validity, value, and procedures of Turing test as a measure of intelligence, our aim is to propose a systematic and quantitative formula to benchmark current AIs in imitating humans through Turing-like tests.

Driven by the fascination with mechanizing human cognition and the advent of modern computers, the Loebner Prize was introduced in 1991 [45] to the programs considered by human judges to be the most human-like. Interestingly, there was also another award in the competition for the human who does the best job of swaying the judges: the Most Human Human award [11]. Different from these text-based Turing tests, our work covers a wide range of tasks across both vision and language modalities and tests 15 state-of-the-art AI models in these tasks.

S1.2. AI versus humans in vision tasks

Current computer vision models can perform a wide range of discrimination tasks (such as object recognition and object detection) and generative tasks (such as text-to-image synthesis). Models are often evaluated by comparing their outputs against human ground truth annotations. For example, deep nets are biased towards textures rather than shapes compared to humans [23]. Many object recognition studies benchmark 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 [71, 73].

It is important to note that high performance in a particular task does *not* constitute a Turing test. AI models can outperform humans and still be distinguishable from humans. AI models can even show similar average performance in narrow tasks and still be easy to tell apart from humans (e.g., because of different patterns of mistakes, levels of robustness, characteristics of the answers). Thus, the Turing tests provide a unique behavior assessment of AI models, which is complimentary to existing benchmarking frameworks.

S1.3. AI versus humans in language tasks

The stories above in vision happen again in natural language processing. AI models are often compared against human ground truth data in the discriminative tasks, such as visual question answering [9, 44, 56, 70]. In contrast, AI models in generative tasks are notoriously difficult to evaluate due to the inherent ambiguity of language. For example, human evaluators are recruited to manually assess the quality of sentiment and semantic relevance on the generated text by BERT [19]. Subsequently, larger language models appear, such as GPT2 and GPT3 [8]. To assess the quality of the generated articles by GPT3, human experts are recruited to ascertain whether the 200-word article was written by humans or by GPT3 models of various sizes [8, 37]. However, the task was restricted to open-domain text generation and the heterogeneity of human judges in the task is less studied. In contrast, our work provides a more comprehensive set of Turing tests on varieties of large SOTA language models from the judgements of 896 judges of 4 demographic identities.

Human evaluation scores are reliable but costly to obtain. To mitigate these problems, several evaluation metrics are proposed, such as BLEU [49], THUMB [38], and METEOR [18] in image captioning. However, these metrics tend to focus on n-gram overlaps and are insensitive to semantic information. Cui etal. [16] proposed a learned critique model acting as a human judge to perform a Turing Test in image captioning task. The model has to recognize whether a candidate caption is human written or machine generated. Although the work is very relevant to ours, we highlight several differences: first, instead of using a critique model, we focus on evaluating how current AI models are like us with Turing tests involving human judges; second, in contrast to Turing test in single tasks, our work aims to propose a systematic methodology to test current and future AI models in all vision and language tasks; and third, we propose a generalist AI judge capable of discriminating human and AI agents over multiple tasks and modalities.

	Color Dect.	Object Dect.	Att. Pred.	Image Cap.	Word Ass.	Conversation
Age (<35)	✗	✗	✗	✗	✓	✓
Age (35-45)	✗	✗	✗	✗	✓	✓
Age (>45)	✗	✗	✗	✗	✓	✓
Sex (F)	✗	✗	✗	✗	✓	✓
Sex (M)	✗	✗	✗	✗	✓	✓
Edu L1	✗	✗	✗	✗	✓	✓
Edu L2	✗	✗	✗	✗	✓	✓
Edu L3	✗	✗	✓	✗	✓	✓
Overall	✗	✗	✗	✗	✓	✓

Table S1. **Summary of Turing test results for human judges**.
A tick (✓) indicates that the AI models pass the Turing test ($0.45 \leq$ top-1 accuracy ≤ 0.55) in the specific task (columns), based on evaluations made by specific demographic groups (rows). This table represents only a coarse grand summary of the results. The actual confusion matrices and quantitative results are presented throughout the **Supplementary Material**.

Conversation has been the central topic of the original Turing test. It has still been by far the most challenging topic in natural language processing. In early years of Loebner Prize competitions, there have been numerous successful attempts at simulating restricted topics during conversations, such as Colby's PARRY simulating a paranoid schizophrenic [12, 13] and Weizenbaum's ELIZA [66] simulating a psychiatrist's discussion with patients. However, none of these models have come close to unrestricted Turing test. With the prosperity of large language models based on transformers [8, 14, 19, 58], news and social medias have produced anecdotal claims about current AI being sentient in conversations with humans [43,60,67]; however, there has been very few works quantitatively assessing AIs in their abilities of imitating humans in conversations. Some preliminary works [75] introduced unrestricted Turing tests in conversations with one exchange per conversation; however, these tests fail to capture long-term associations over multiple exchanges within a conversation. Thus, our goal is to provide comprehensive evaluations of AIs engaged in conversations with up to 24 exchanges.

S2. Color estimation

S2.1. Dataset

S2.1.1 Collecting human responses from AMT

To collect human responses for the color estimation task, we used Amazon Mechanical Turk (AMT). Fig. S1 shows an example image of the AMT user interface. As can be seen, we showed participants an image and asked them the question-*what is the dominant color in this image?*. They were asked to select an answer from a drop down list of 11 possible colors—red, pink, orange, yellow, purple, green, blue, brown, gray, black, white. In the sample image, the worker has selected the option "white".

To ensure that responses are of good quality, only AMT master workers can accept the HIT. They were given as much time as needed to complete the task. Each worker was shown 25 images, and given 2 hours to complete the task. For each image, we only used the response from one single worker for the Turing test (described below). This was done to ensure a symmetry across images (exactly half of the responses from humans and half of the responses from AI). In total, we obtained responses for 785 images from the validation set of MS-COCO dataset.

S2.1.2 Collecting AI responses

For each of the 785 images, we collected responses from—Microsoft Azure API, Google Cloud Vision API and the MCCQ algorithm. Unfortunately, the APIs are proprietary technology and we have no information about what data they were trained on, or what is the model architecture. In comparison, we also included the MCCQ algorithm, which is a color quantization method relying on median cuts. This is a classical algorithm, and has been used for a long time for color quantization tasks.

S2.2. Turing test

S2.2.1 Collecting human judge responses for Turing test

After collecting responses for color recognition recognition from AI models, we then collected responses from human judges in a second AMT task, where they were asked to tell if image-pair responses came from human speakers or machines. They were also asked to predict the gender of the speaker. An example image of the AMT user interface for the Turing tests shown to human judges is presented in Fig. S3. Each participant was shown 25 image-response pairs. These pairs were randomly selected with 50% of the time from human agents and 50% of the time from AI agents. We also ensured that only master workers are allowed to accept the HIT. As an additional control, the participants were also asked a binary Yes/No question about the contents of the image. These image contents were manually identified by us for each image, and the correct answer was true only 50% of the time. Thus, a casual worker who is not investing time looking at the image would get this question right only 50% time on average. However, we found that all master workers had a performance of greater than 85% on this task, and the mean was 91%.

S2.2.2 Demographic information of responders

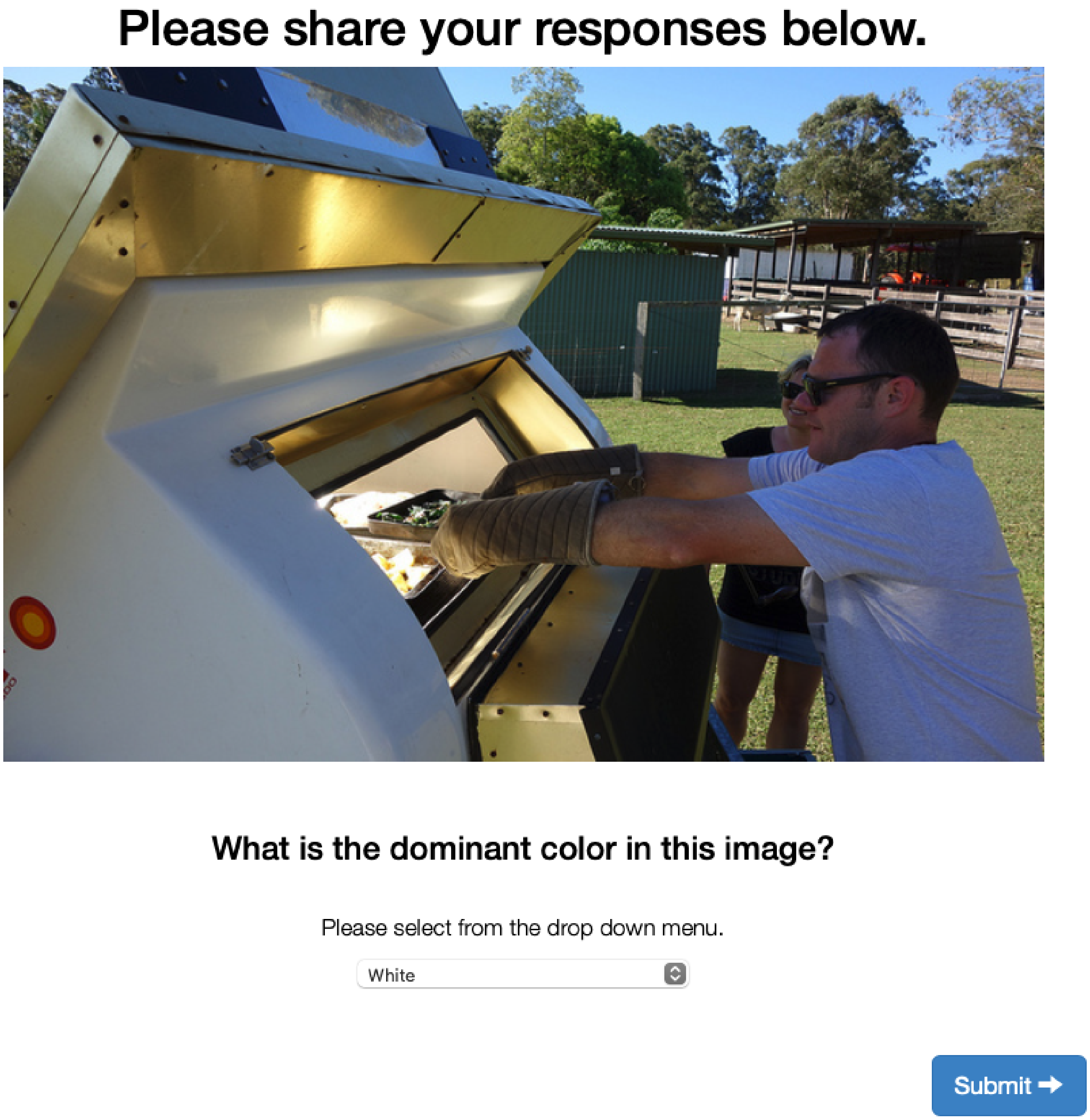
For all our participants we collected three key pieces of demographic information—age, gender and education level. We then created separate bins for each of these, and these results are provided in the pie charts in Fig. S14.

S2.2.3 Human judge performance based on demographic information

We analyzed if people from a certain demographic are better at distinguishing AI and Humans. Below we provide these results. Fig. S6,Fig. S7 and Fig. S8 show performance of human judges in three separate age bins: Below 35, 35-45, and Above 45. As can be seen, there are no major differences in their capabilities to distinguish human and AI speakers across these age bins. There are small variations, but they are contained within a 5% margin of difference. Simiarly, Fig. S9,Fig. S10 and Fig. S11 show the performance of human judges across different education lavels, while Fig. S12 and Fig. S13 show the same for different genders. Again, we can see that there are no major differences in the performance across education or gender bins. Thus, largely, our findings suggest that demographics do not have a significant impact on the capability to solve the Turing test.

S2.2.4 AI judge

To see if an AI judge can distinguish humans and machines, we trained a binary classification model on the data collected from human agents and AI models. Note that this is the same data as that shown to human judges during the Turing test—images along with responses from humans or machines. We trained 4 models for this task: (i) Human vs all AI models, (ii) Human vs Azure, (iii) Human vs Google, (iv) and Human vs MCCQ. In case (i), for each image, we randomly sampled a response from one of the AI models. This was done to ensure the classes are balanced and 50% of the data comes from Humans 50% from AI models. As the responses were the same fixed 11 colors, the textual representation of these colors alone is not enough to train a classifier. Thus we extracted VGG features for each image, and BERT features for the color name picked by the speaker. These were concatenated and then passed to a random forest model. 90% of the data was kept for training, while 10% were used as testing to report performance. The performance of each of these classifiers was reported in Fig. S15.



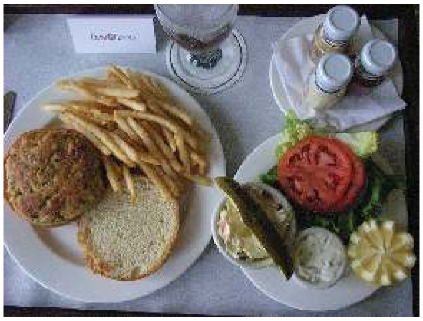
Please share your responses below.

What is the dominant color in this image?

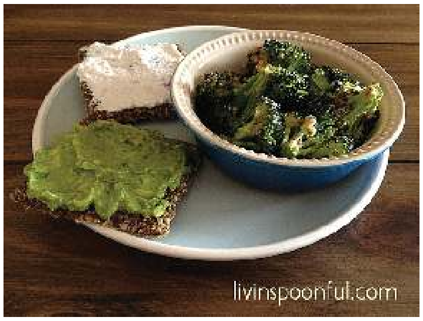
Please select from the drop down menu.

White

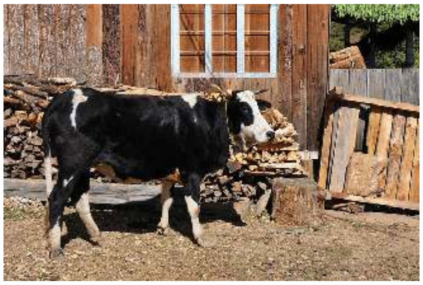
Figure S1. **Color estimation**. AMT user interface for collecting responses.



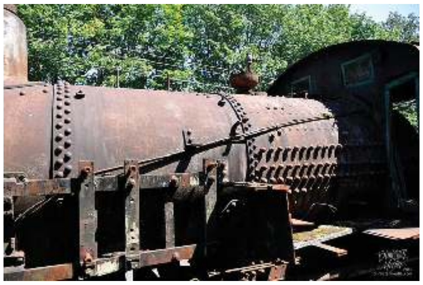
Google: green
Azure: gray
Histogram: gray
Human: white



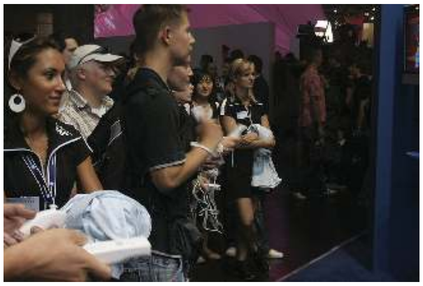
Google: green
Azure: brown
Histogram: green
Human: green



Google: black
Azure: brown
Histogram: gray
Human: brown

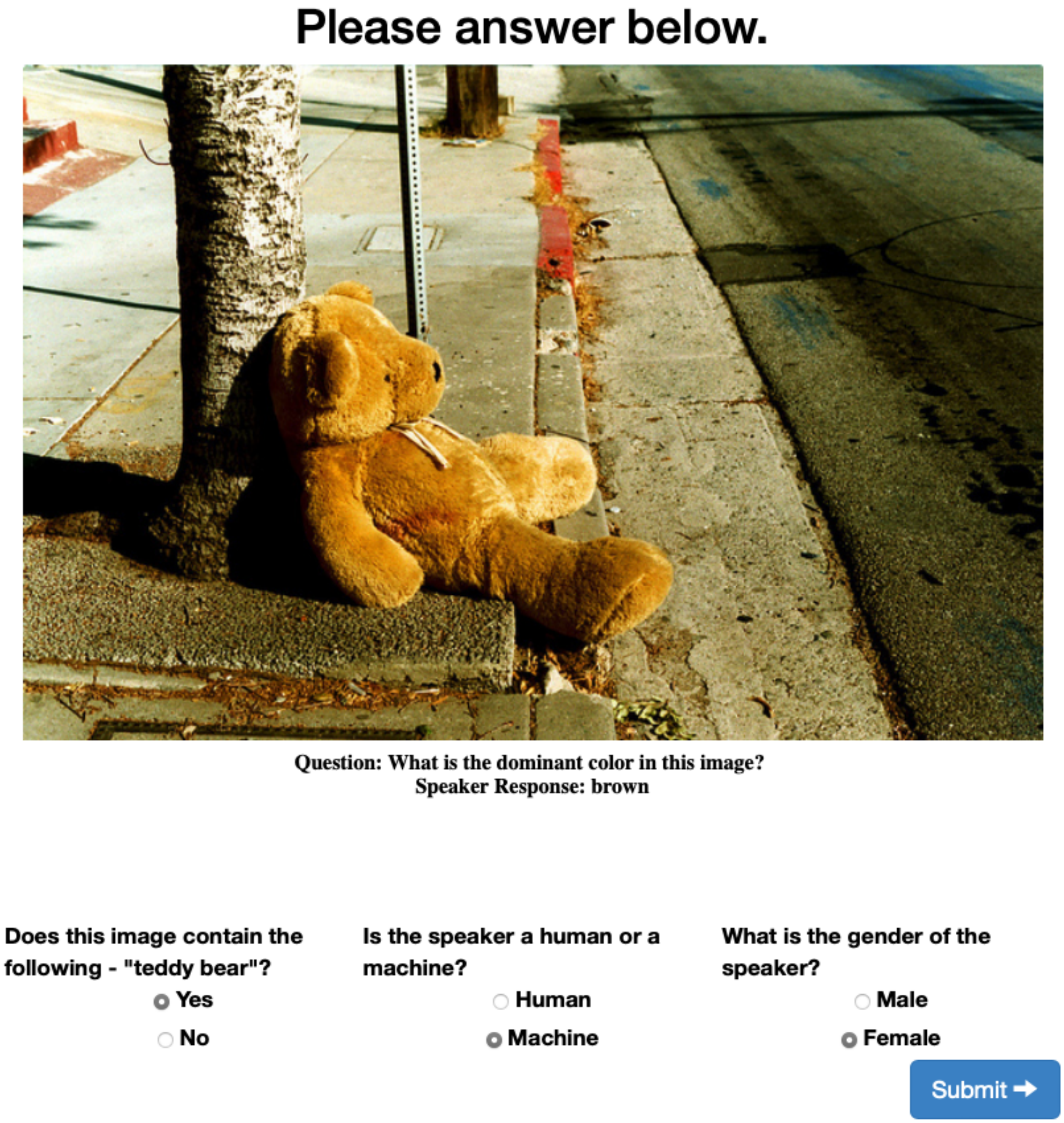


Google: gray
Azure: black
Histogram: black
Human: brown



Google: black
Azure: black
Histogram: black
Human: black

Figure S2. **Color estimation**. Random samples from our collected color estimation dataset.



Please answer below.

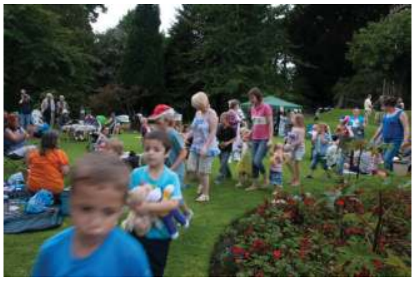
Question: What is the dominant color in this image?
Speaker Response: brown

Does this image contain the following - "teddy bear"?
● Yes
○ No

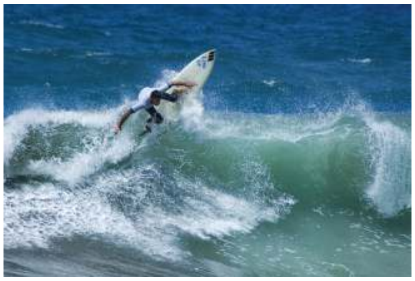
Is the speaker a human or a machine?
○ Human
● Machine

What is the gender of the speaker?
○ Male
● Female

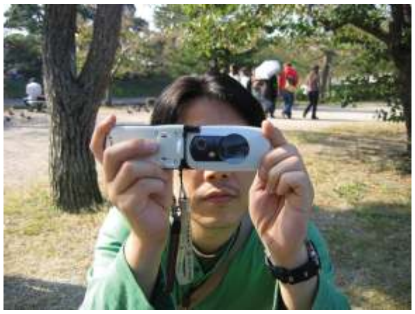
Figure S3. **Color estimation**. AMT user interface for collecting human judge responses (Turing test).



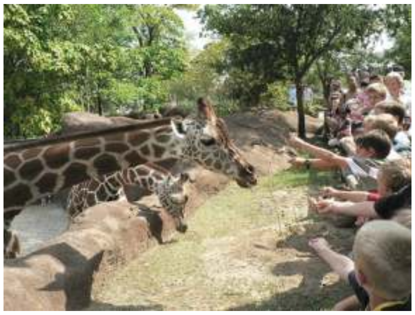
Response: green
Actual Speaker: Human
Human Judge Prediction: Human



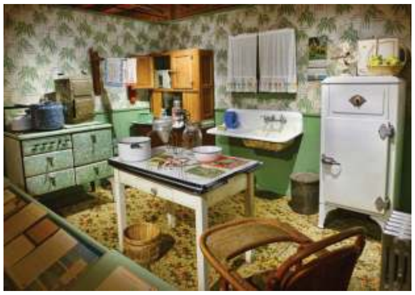
Response: blue
Actual Speaker: Human
Human Judge Prediction: Machine



Response: green
Actual Speaker: Human
Human Judge Prediction: Human

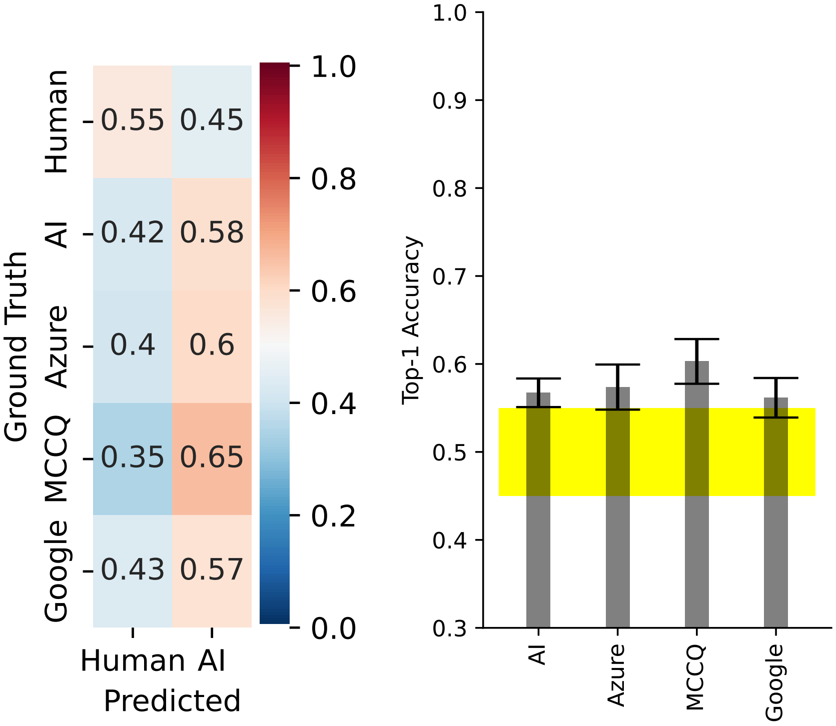


Response: brown
Actual Speaker: Human
Human Judge Prediction: Human



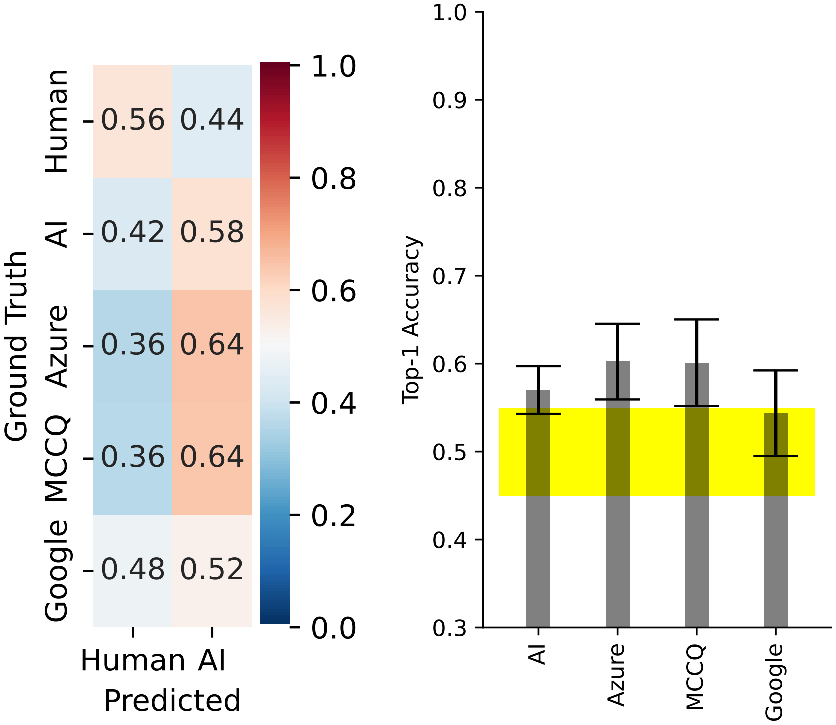
Response: green
Actual Speaker: Human
Human Judge Prediction: Machine

Figure S4. **Color estimation**. Random samples from our collected Turing tests.



(a) Confusion matrix (b) Top1 accuracy

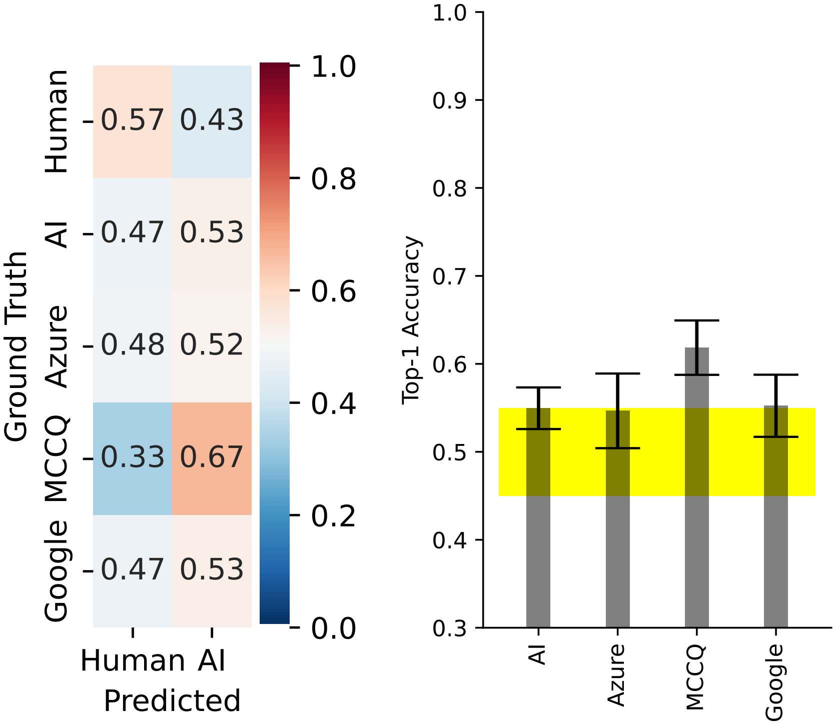
Figure S5. **Color estimation**. Results of the Turing test for human judges.
(a) Confusion matrix (b) Top-1 accuracy. The highlighted regions in yellow denote the criteria of passing the Turing tests: above 45% and below 55%. The errorbars denote the standard deviation. Same notations are applicable for all the subsequent bar plots in the entire supplementary material.



(a) Confusion matrix

(b) Top1 accuracy

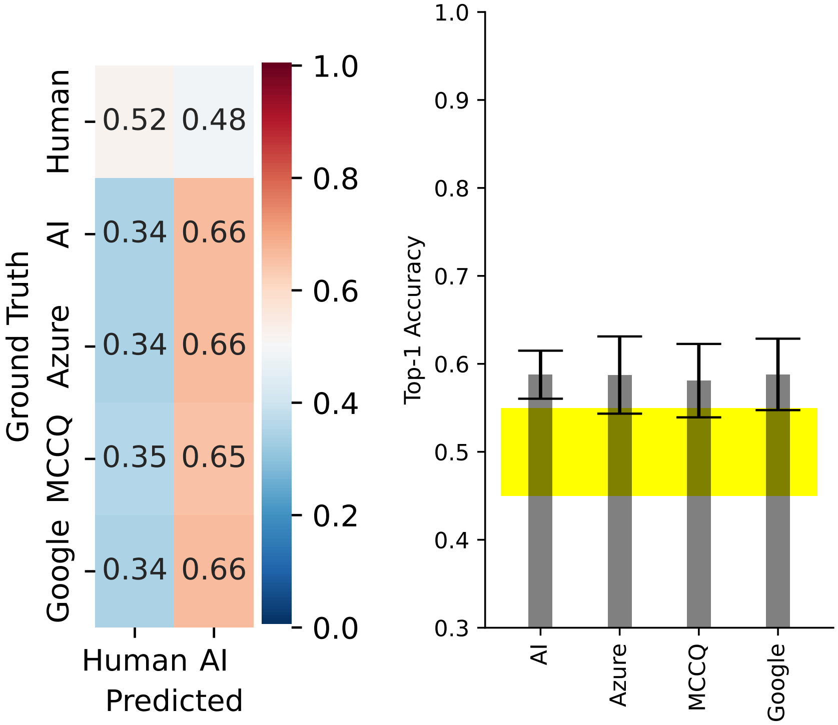
Figure S6. **Color estimation**. Results of the Turing test for human judges with age level below 35. (a) Confusion matrix (b) Top-1 accuracy.



(a) Confusion matrix

(b) Top1 accuracy

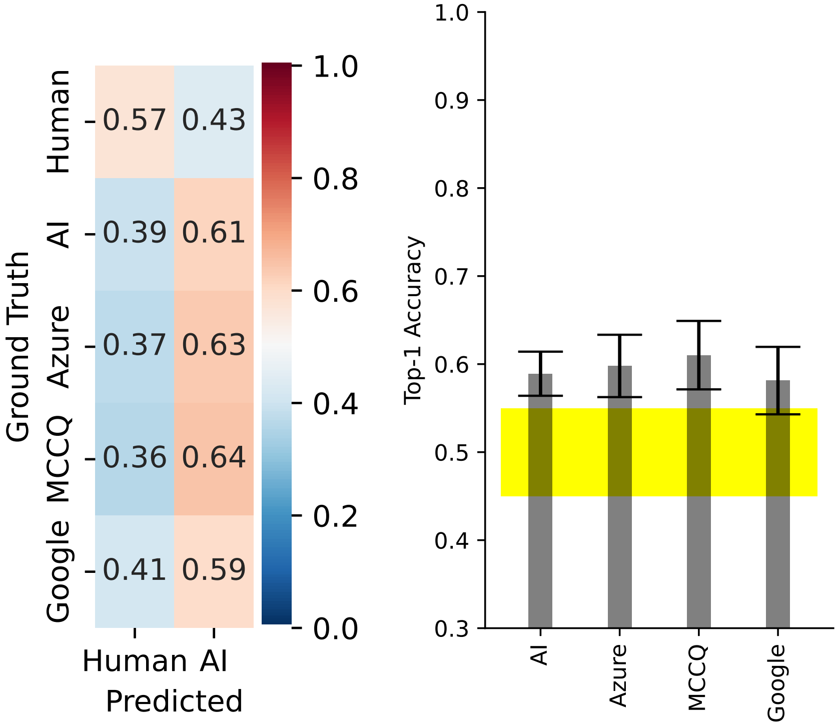
Figure S7. **Color estimation**. Results of the Turing test for human judges with age level 35-45. (a) Confusion matrix (b) Top-1 accuracy.



(a) Confusion matrix

(b) Top1 accuracy

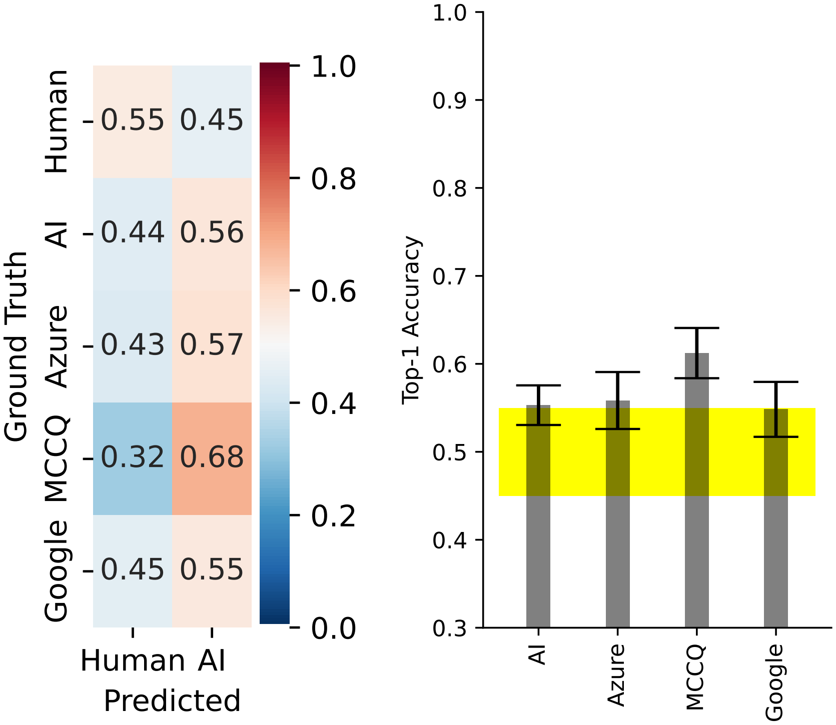
Figure S8. **Color estimation**. Results of the Turing test for human judges with age above 45. (a) Confusion matrix (b) Top-1 accuracy.



(a) Confusion matrix

(b) Top1 accuracy

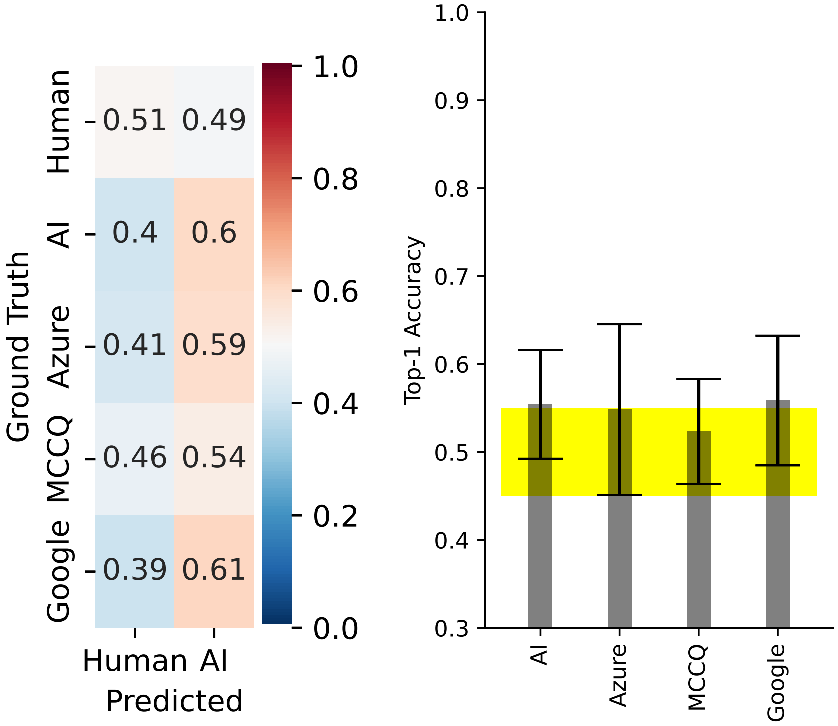
Figure S9. **Color estimation**. Results of the Turing test for human judges with education level below Bachelors. (a) Confusion matrix (b) Top-1 accuracy.



(a) Confusion matrix

(b) Top1 accuracy

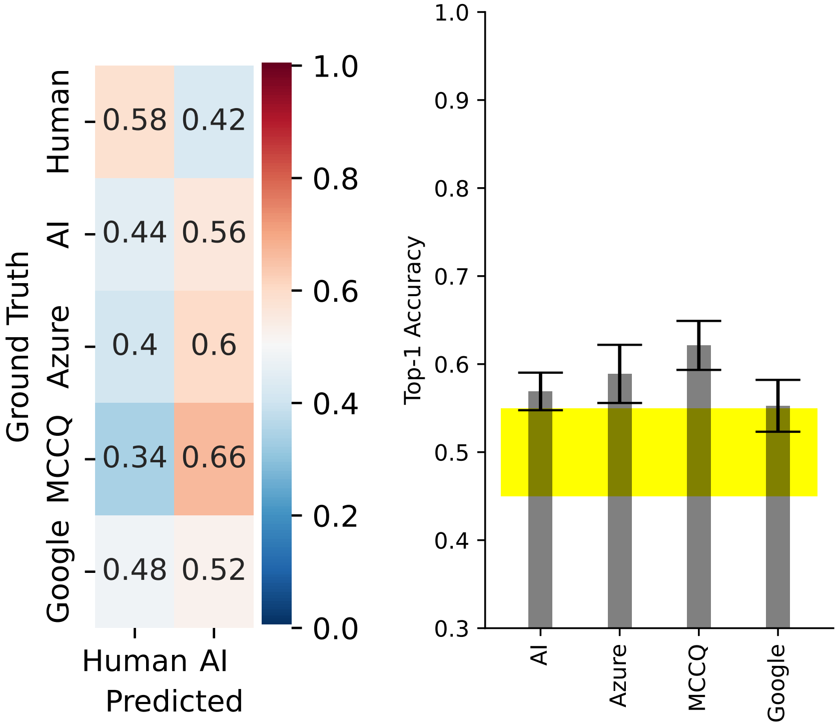
Figure S10. **Color estimation**. Results of the Turing test for human judges with education level of Bachelors. (a) Confusion matrix (b) Top-1 accuracy.



(a) Confusion matrix

(b) Top1 accuracy

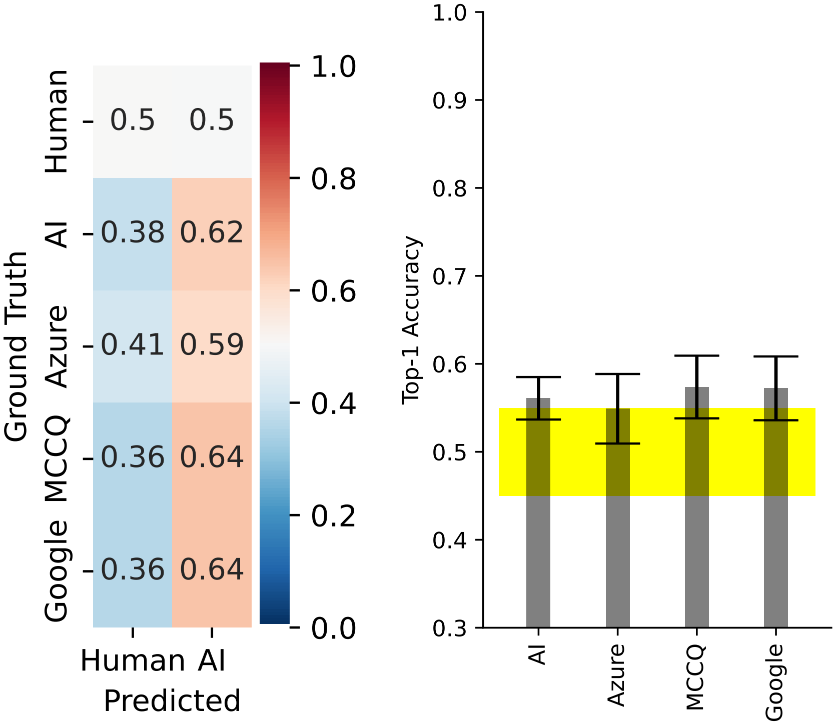
Figure S11. **Color estimation**. Results of the Turing test for human judges with education level above Bachelors. (a) Confusion matrix (b) Top-1 accuracy.



(a) Confusion matrix

(b) Top1 accuracy

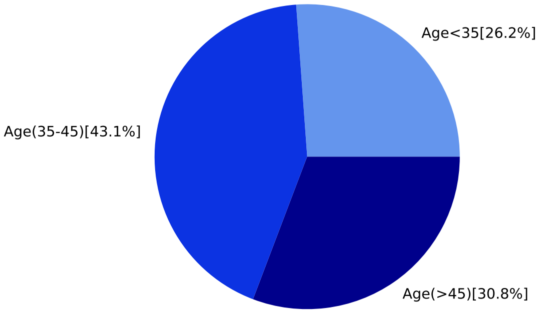
Figure S12. **Color estimation**. Results of the Turing test for human judges with Male gender. (a) Confusion matrix (b) Top-1 accuracy.



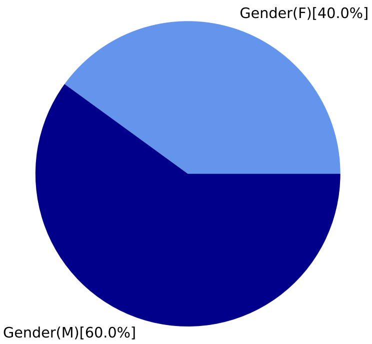
(a) Confusion matrix

(b) Top1 accuracy

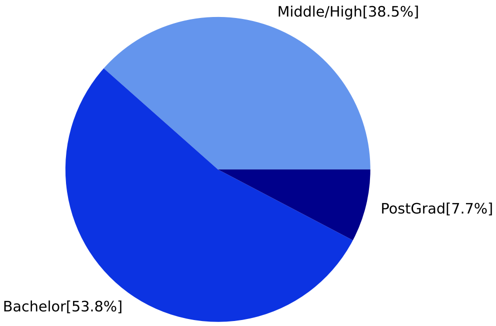
Figure S13. **Color estimation**. Results of the Turing test for human judges with Female gender. (a) Confusion matrix (b) Top-1 accuracy.



(a) Age distribution

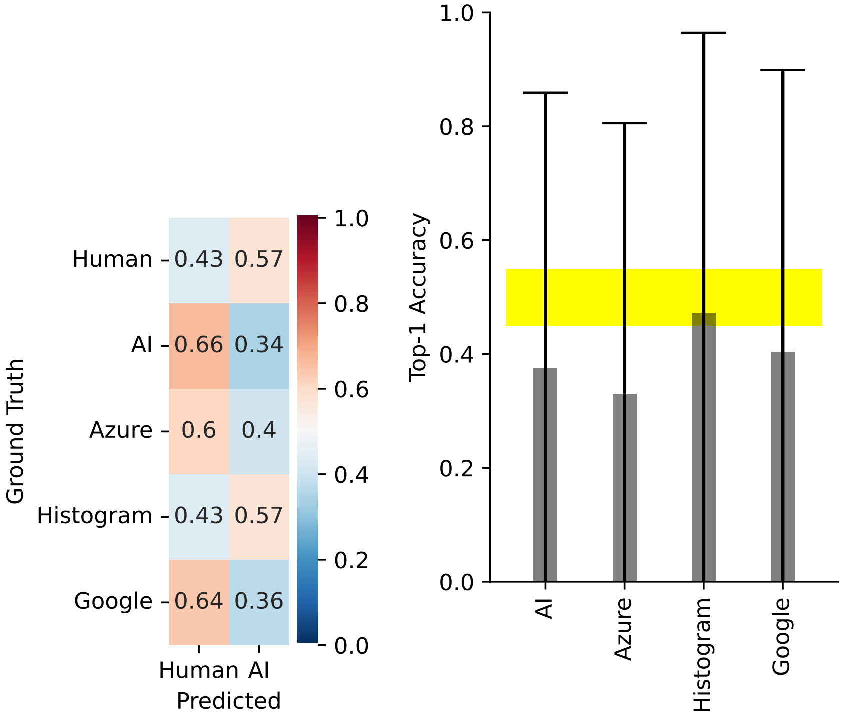


(b) Gender distribution



(c) Education level distribution

Figure S14. **Color Estimation.** Demographic information for the human judges for the color estimation task (a) Age. (b) Gender. (c) Education level.



(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.

S3. Object Detection

S3.1. Dataset

S3.1.1 Collecting human responses from AMT

To collect human responses for the object detection task, we used AMT. Fig. S16 shows an example image of the AMT user interface. As can be seen, we showed participants an image and asked them the question-*what do you see in this image?*. They were asked to enter three single word responses in three separate text boxes. In this sample image, the worker chose to enter Baseball, Field and Cap. Note that these are open set responses, and were not cleaned to fit a fixed dictionary.

Thus, to ensure that responses are of good quality, we ensured that only AMT workers with a HIT approval rating of 98 and with at least 100 successful HITs can accept the HIT. They were given as much time as needed to complete the task. Each worker was shown 25 images, and given 12 hours to complete the task. We also added several other checks including—implementing a spell checker to ensure only valid English words are used. We also ensured that the three responses must not be the same, and that the responses are not copied across images. For each image, we only used the response from one single worker for the turing test (described below). This was done to ensure a symmetry across images. In total, we obtained responses for 808 images from the validation set of MS-COCO dataset.

S3.1.2 Collecting AI responses

For each of the 808 images, we collected responses from—Microsoft Azure API, Google Cloud Vision API and Amazon Rekognition API. Again, the details of the APIs are not released and therefore remain unkown. To include an open source model, we used Detectron2. We picked the three bounding boxes with maximum confidence score as the response from Detectron2. Specifically, the MaskRCNN variant trained on ImageNet and MS-COCO (Link)

S3.2. Turing test

S3.2.1 Collecting human judge responses for Turing test

As before, we conducted a second round of AMT with human judges. They were asked to tell if image-pair responses came from human speakers or machines, and to predict the gender of the speaker. An example image of the AMT user interface for this second task shown to human judges is presented in Fig. S18. Each participant was shown 25 images. We also ensured that only master workers are allowed to accept the HIT. As an additional control, the participants were also asked a binary Yes/No question about the contents of the image. These image contents were manually identified by us for each image, and the correct answer was true only 50% of the time. Thus, a casual worker who is not investing time looking at the image would get this question right only 50% time on average. However, we found that all master workers had a performance of greater than 85% on this task, and the mean was 91%.

Fig. S16 shows an example image of the AMT user interface used to collect human responses for object detection. As can be seen, we showed participants an image and asked them the question-*what do you see in this image?*, along with three empty text responses where they filled one word responses. After collecting responses for object detection from AI models, we then collected responses from human judges in a second AMT task, where they were asked to tell if image-pair responses came from human speakers or machines. An example image of the AMT user interface for this second task shown to human judges has been shown in Fig. S18.

S3.2.2 Demographic information of responders

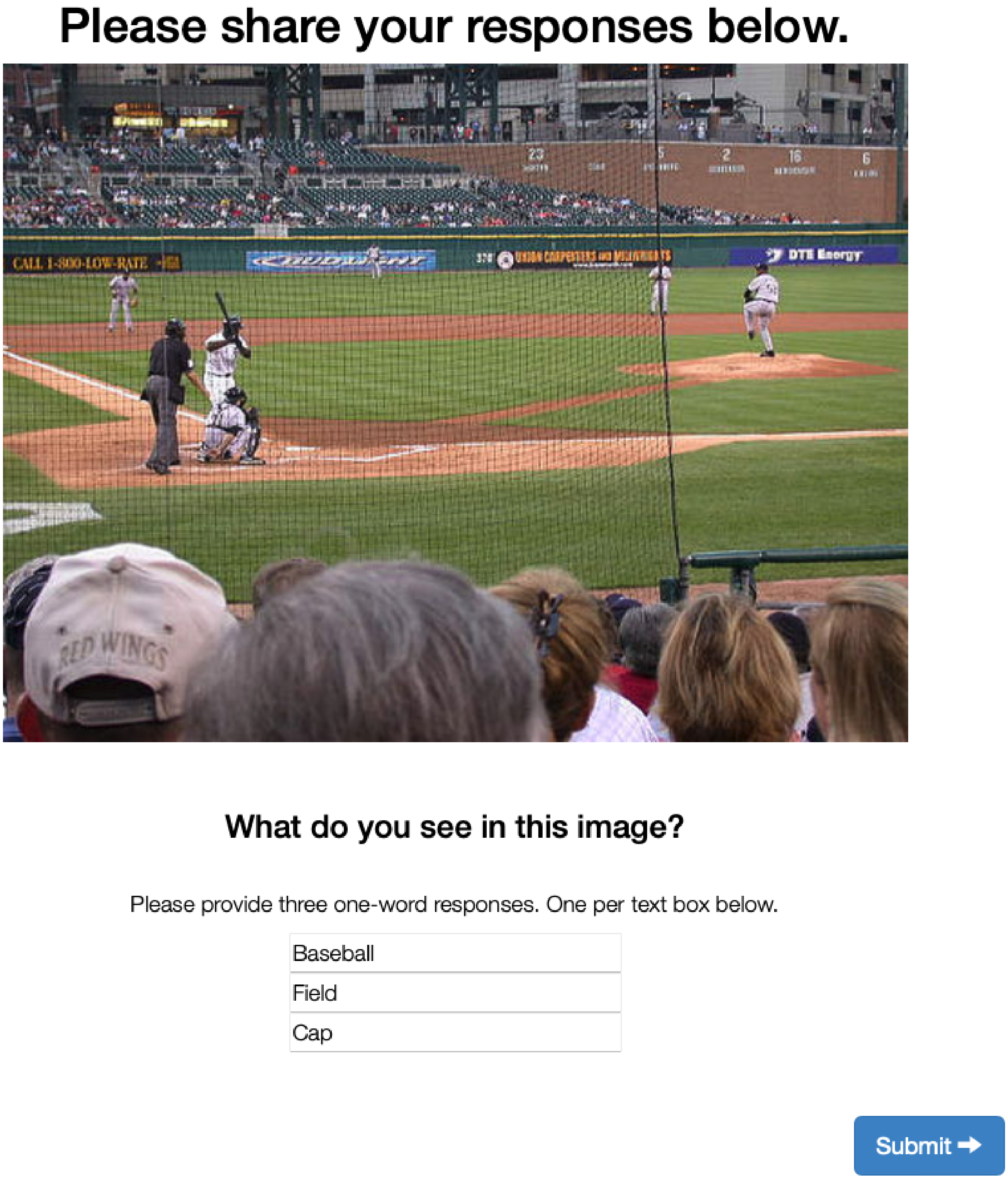
The results for distribution of human judges belonging to different demographic categories of Age, Gender and Education levels are provided in the pie charts in Fig. S29.

S3.2.3 Human judge performance based on demographic information

We evaluated whether people from a certain demographic are better at distinguishing AI and Humans. Below we provide these results. Fig. S21,Fig. S22 and Fig. S23 show performance of human judges in three separate age bins: Below 35, 35-45, and Above 45 respectively. As can be seen, there are no major differences in their capabilities to distinguish human and AI speakers across these age bins. There are small variations, but they are contained within a 5% margin of difference. Similarly, Fig. S24, Fig. S25, Fig. S26 show the performance of human judges across different education levels, while Fig. S27 and Fig. S28 show the same for different genders. Again, we can see that there are no major differences in the performance across education or gender bins. Thus, largely, our findings suggest that demographics do not have a significant impact on the capability to solve the Turing test.

S3.2.4 AI judge

We also trained an AI judge to distinguish humans and machines. For this, we trained a binary classification model to classify the same data as that shown to human judges during the Turing test—images along with responses from humans or machines. We trained 4 models for this task: (i) Human vs all AI models, (ii) Human vs Azure, (iii) Human vs Google, (iv) and Human vs Detectron2, and (v) Human vs Rekognition. In case (i), for each image we randomly sampled a response from one of the AI models. This was done to ensure the classes are balanced and 50% of the data comes from Humans 50% from AI models. As the responses were open set, the textual representation of the responses given were enough to train a classifier. We concatenated the three words into a single sentence, and extracted BERT features for the sentence. These were then passed to an SVM model. 90% of the data was kept for training, while 10% were used as testing to report performance. The performance of each of these classifiers has been reported in Fig. S30.



Please share your responses below.

What do you see in this image?

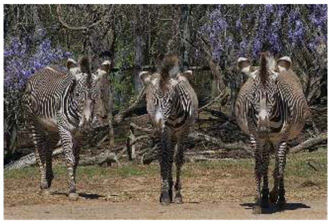
Please provide three one-word responses. One per text box below.

Baseball

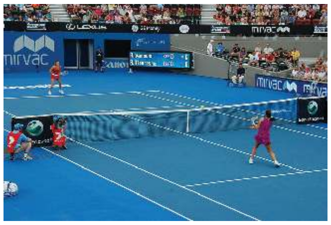
Field

Cap

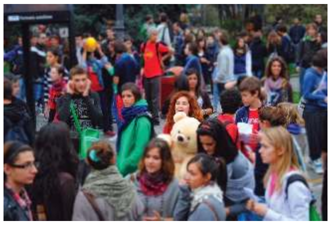
Figure S16. **Object detection**. AMT user interface for collecting responses.



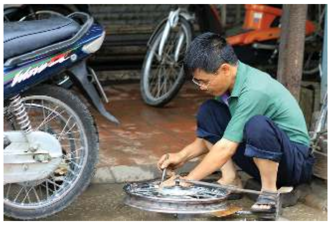
Microsoft: animal,mammal,outdoor
Google: Zebra,Plant,Vertebrate
Detectron: zebra
Rekognition: Zebra,Wildlife,Animal
Human: zebra,flowers,grass



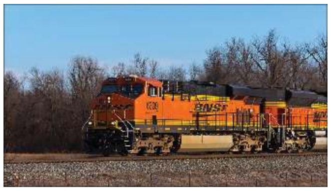
Microsoft: sport,athletic game,tennis
Google: Footwear,Tennis,Racketlon
Detectron: person,sports ball,tennis racket
Rekognition: Person,Human,Tennis Court
Human: people,rackets,ball



Microsoft: clothing,human face,person
Google: Purple,Orange,Hat
Detectron: person,teddy bear,backpack
Rekognition: Person,Human,Audience
Human: people,bear,girls

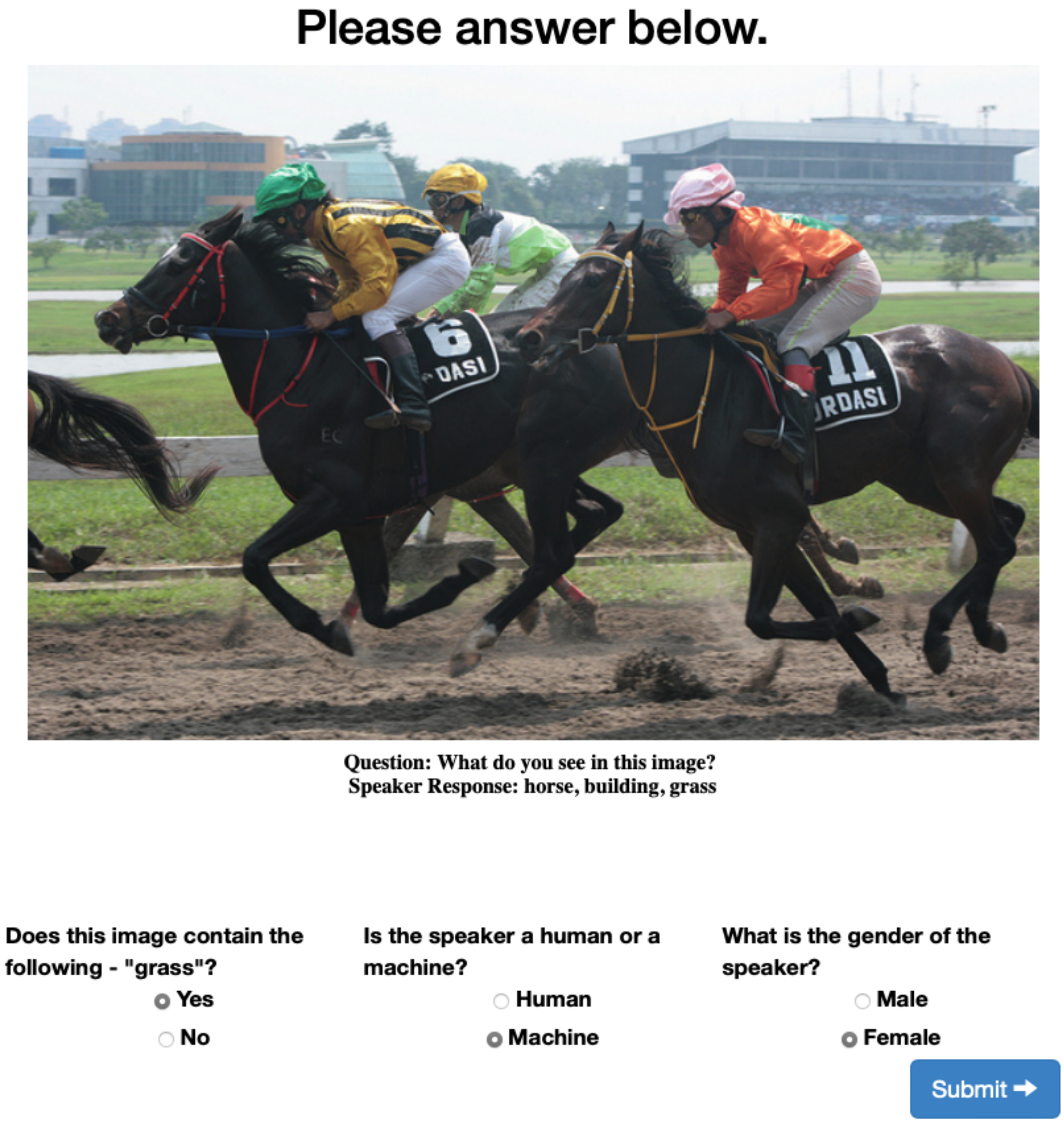


Microsoft: person,wheel,clothing
Google: Tire,Wheel,Photograph
Detectron: person,motorcycle,bicycle
Rekognition: Wheel,Machine,Person
Human: vehicle,wheel,mechanic



Microsoft: railroad,outdoor,sky
Google: Train,Sky,Vehicle
Detectron: train
Rekognition: Train,Transportation,Vehicle
Human: track,train,tree

Figure S17. **Object detection**. Random samples from our collected object detection dataset.


Please answer below.

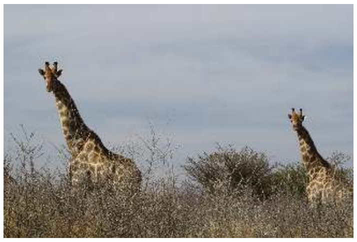
Question: What do you see in this image?
Speaker Response: horse, building, grass

Does this image contain the following - "grass"?
● Yes
○ No

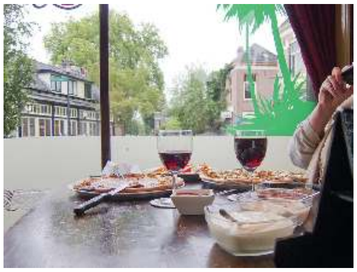
Is the speaker a human or a machine?
○ Human
● Machine

What is the gender of the speaker?
○ Male
● Female

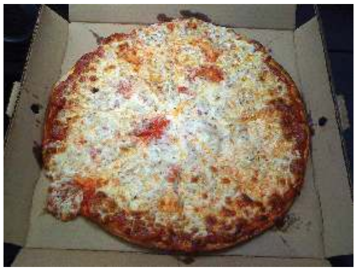
Figure S18. **Object detection**. AMT user interface for collecting human judge responses (Turing test).



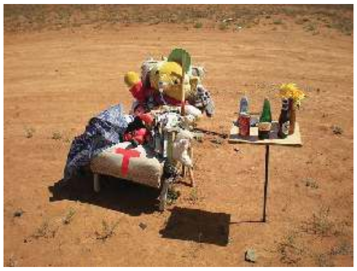
Responses: animal, mammal, outdoor
Actual Speaker: Machine
Human Judge Prediction: Machine



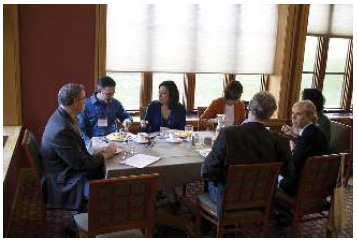
Responses: tableware, meal, wine glass
Actual Speaker: Machine
Human Judge Prediction: Machine



Responses: pizza, box, cheese
Actual Speaker: Human
Human Judge Prediction: Machine

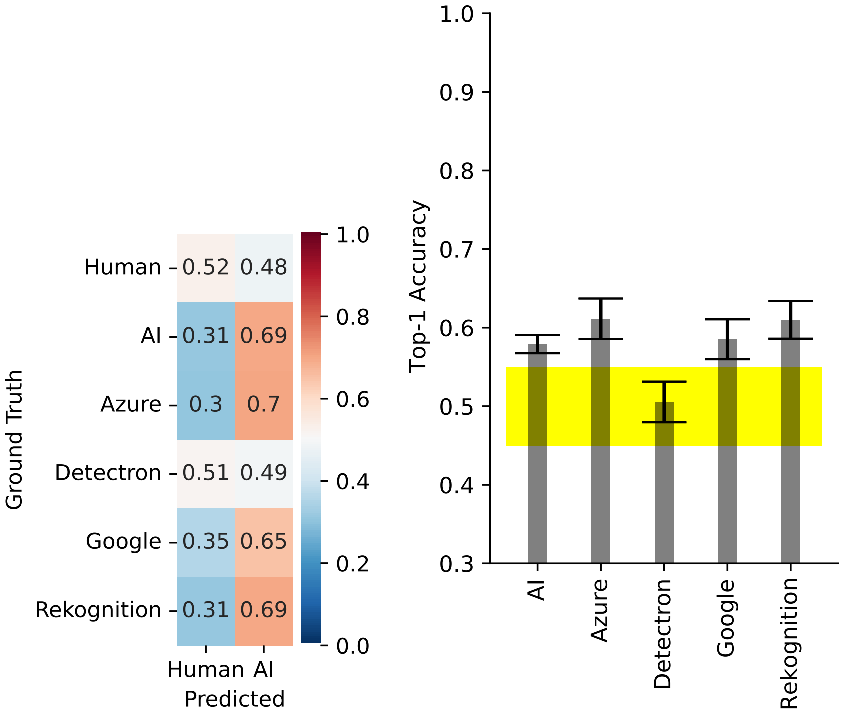


Responses: bottle, teddy bear, cup
Actual Speaker: Machine
Human Judge Prediction: Machine



Responses: chairs, table, plate
Actual Speaker: Human
Human Judge Prediction: Machine

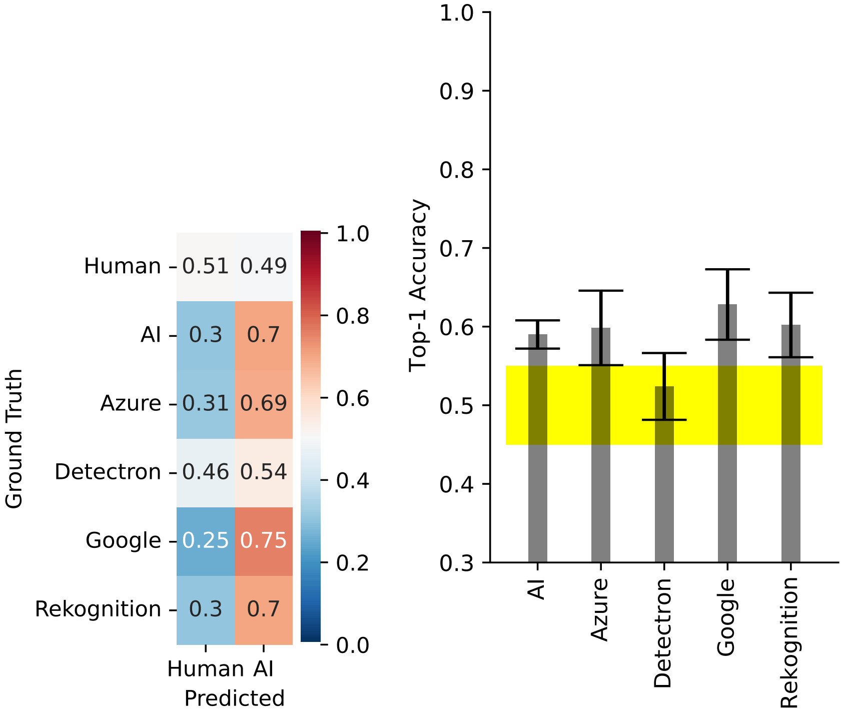
Figure S19. **Object detection**. Random samples from our collected Turing tests.



(a) Confusion matrix

(b) Top1 accuracy

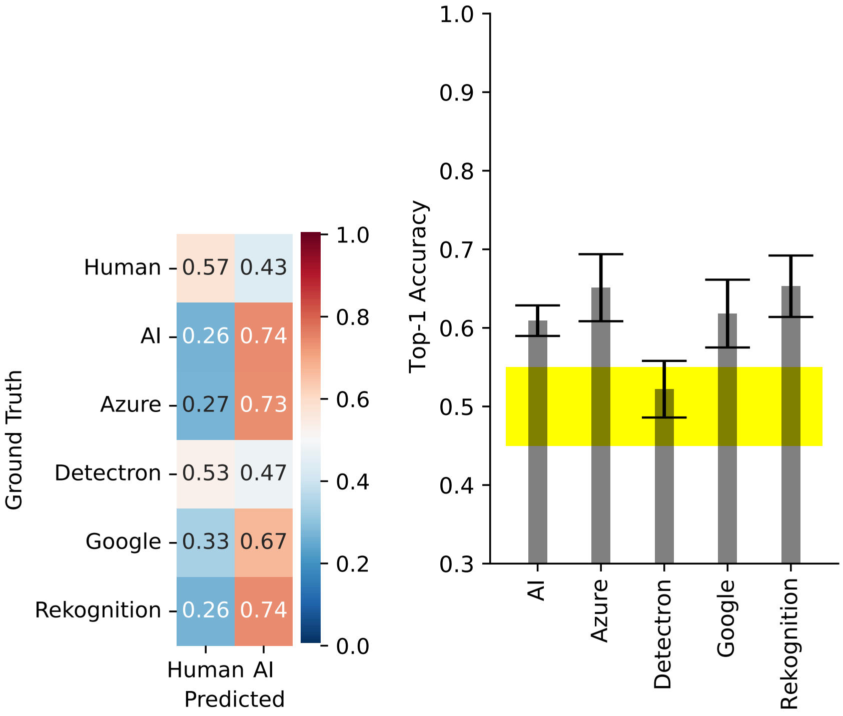
Figure S20. **Object detection**. Results of the Turing test for human judges.
(a) Confusion matrix (b) Top-1 accuracy.



(a) Confusion matrix

(b) Top1 accuracy

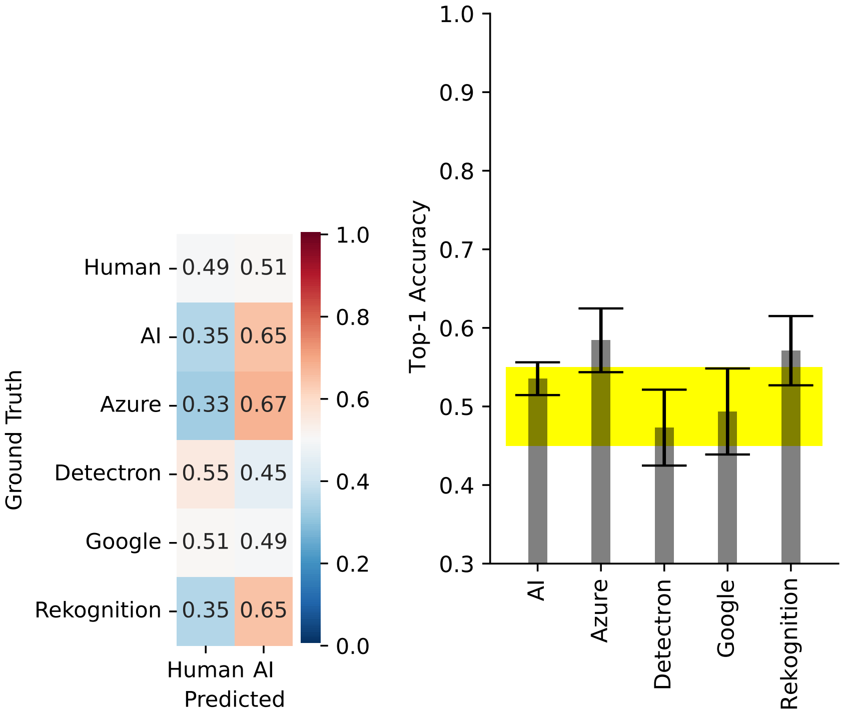
Figure S21. **Object detection**. Results of the Turing test for human judges with age below 35. (a) Confusion matrix (b) Top-1 accuracy.



(a) Confusion matrix

(b) Top1 accuracy

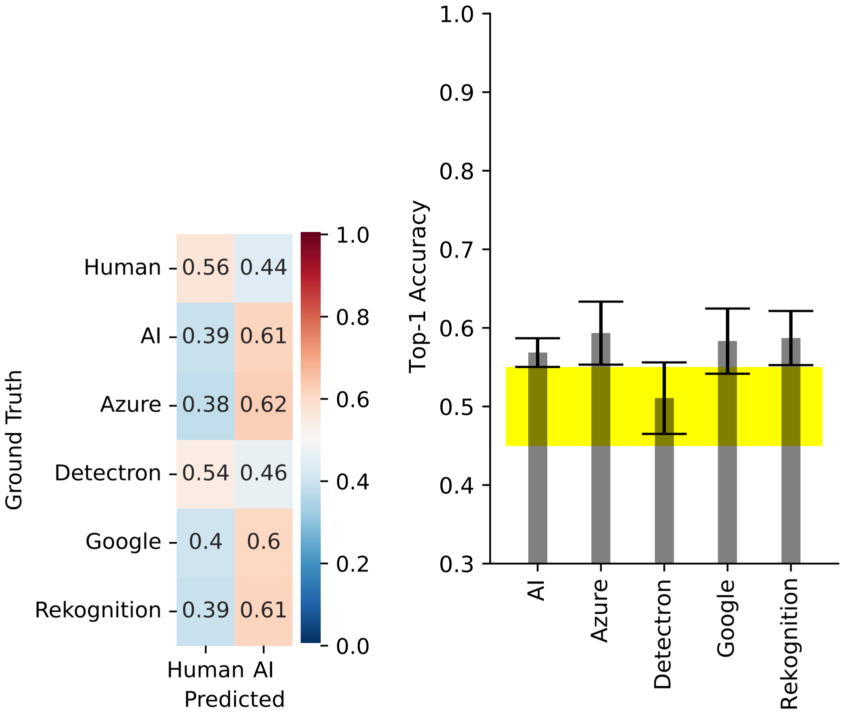
Figure S22. **Object detection**. Results of the Turing test for human judges with age 35-45. (a) Confusion matrix (b) Top-1 accuracy.



(a) Confusion matrix

(b) Top1 accuracy

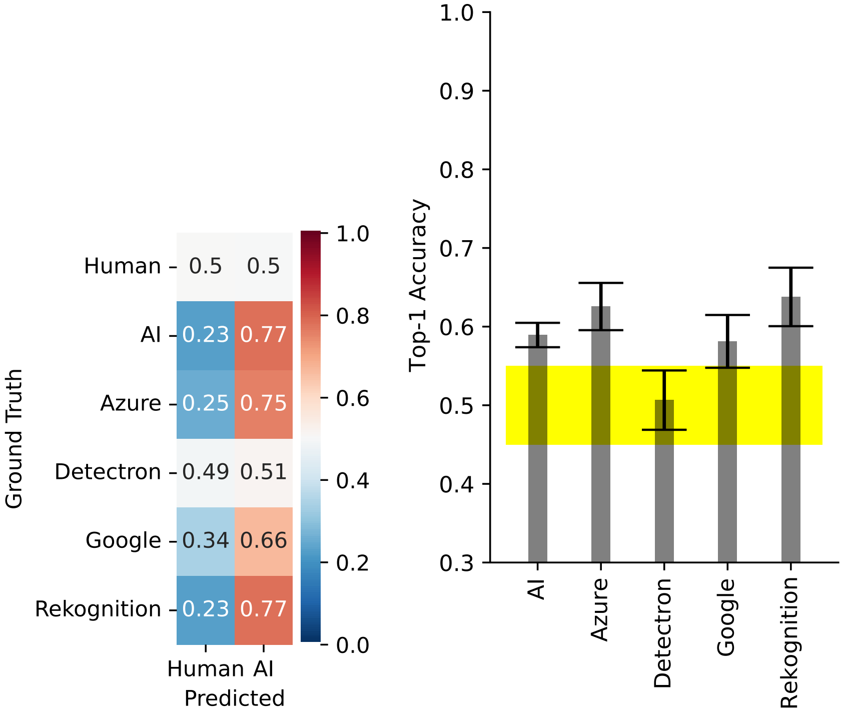
Figure S23. **Object detection**. Results of the Turing test for human judges with age above 45. (a) Confusion matrix (b) Top-1 accuracy.



(a) Confusion matrix

(b) Top1 accuracy

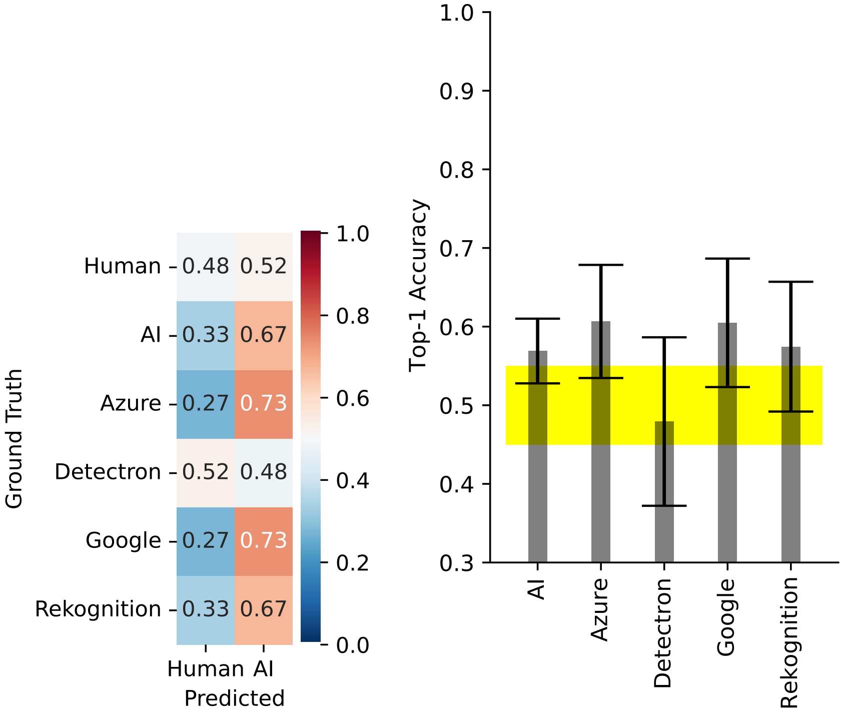
Figure S24. **Object detection**. Results of the Turing test for human judges with education level below Bachelors. (a) Confusion matrix (b) Top-1 accuracy.



(a) Confusion matrix

(b) Top1 accuracy

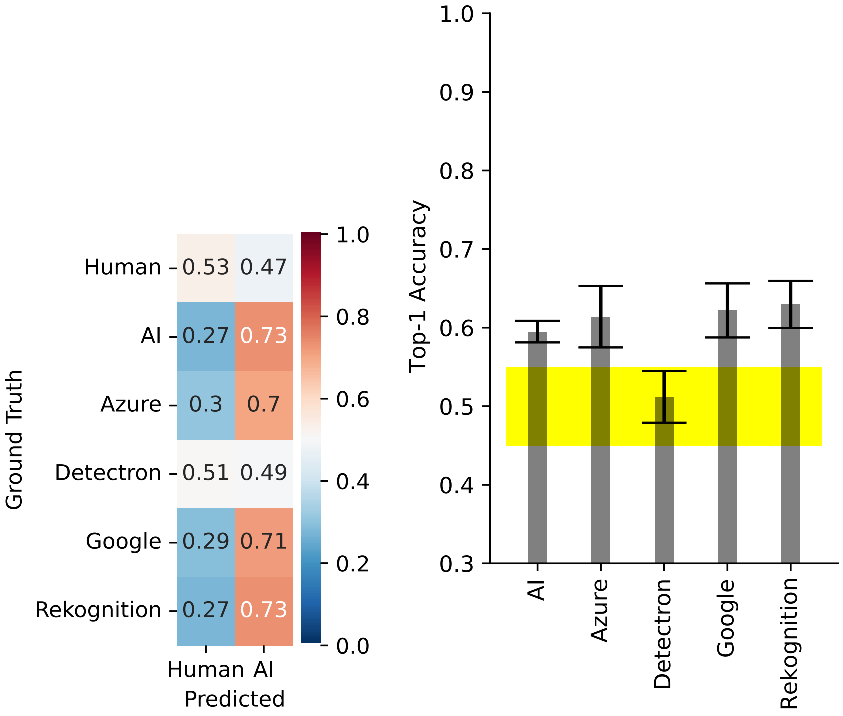
Figure S25. **Object detection**. Results of the Turing test for human judges with education level of Bachelors. (a) Confusion matrix (b) Top-1 accuracy.



(a) Confusion matrix

(b) Top1 accuracy

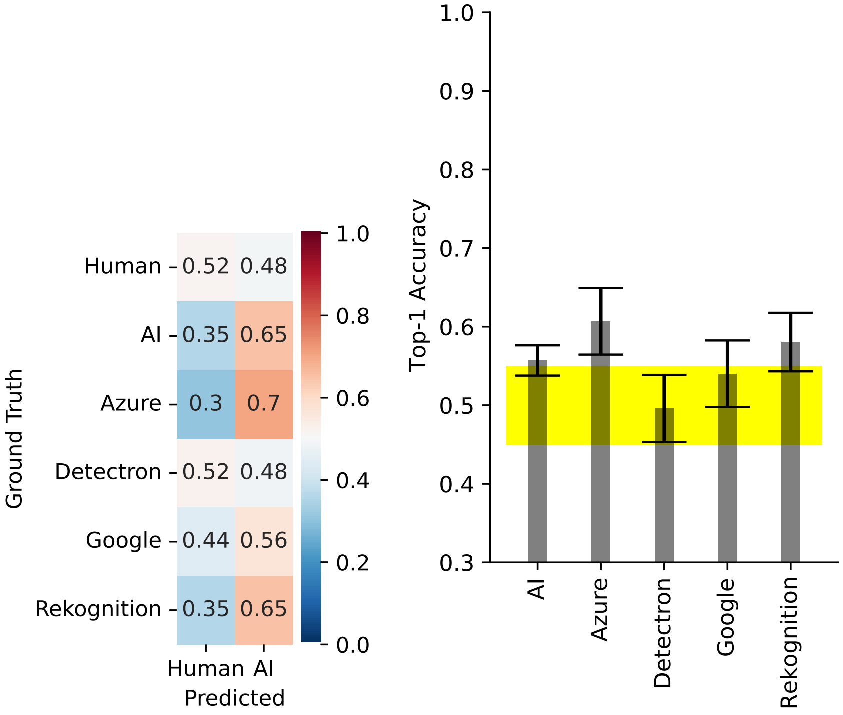
Figure S26. **Object detection**. Results of the Turing test for human judges with education level above Bachelors. (a) Confusion matrix (b) Top-1 accuracy.



(a) Confusion matrix

(b) Top1 accuracy

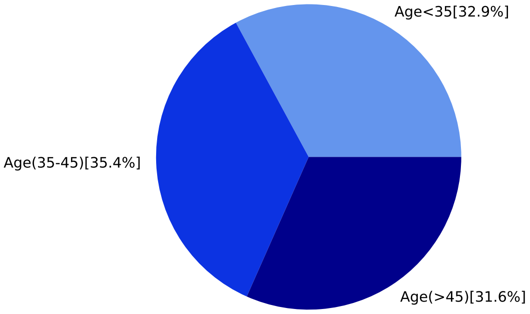
Figure S27. **Object detection**. Results of the Turing test for human judges with Male gender. (a) Confusion matrix (b) Top-1 accuracy.



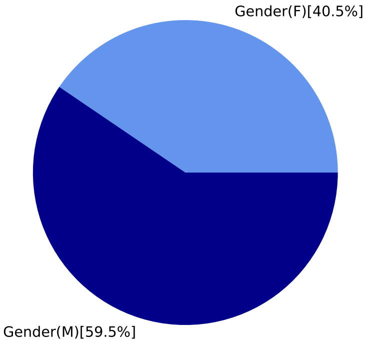
(a) Confusion matrix

(b) Top1 accuracy

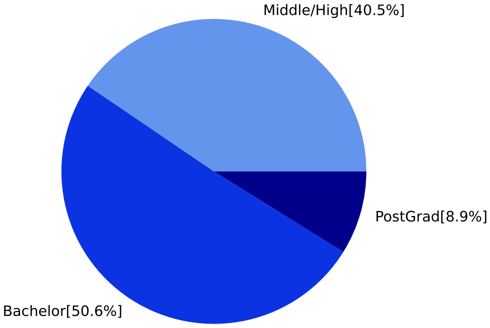
Figure S28. **Object detection**. Results of the Turing test for human judges with Female gender. (a) Confusion matrix (b) Top-1 accuracy.



(a) Age distribution

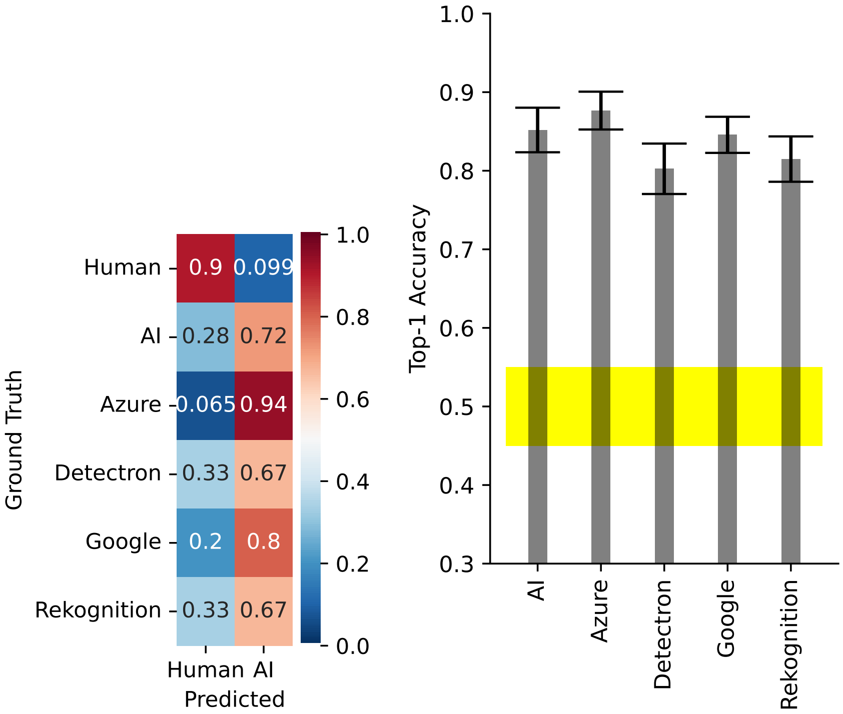


(b) Gender distribution



(c) Education level distribution

Figure S29. **Object Detection.** Demographic information for the human judges for the object detection task (a) Age. (b) Gender. (c) Education level.



(a) Confusion matrix

(b) Top1 accuracy

Figure S30. **Object detection**. Results of the Turing test for AI judges (a) Confusion matrix (b) Top-1 accuracy.

S4. Attention Prediction

S4.1. Dataset

Sample GIF files of human and machine eye movement sequences for the visual search task and free-viewing tasks that were presented to human judges can be accessed here and here.

S4.1.1 Human eye movement responses

We used eye movements from human subjects during two visual search tasks [73] and a free-viewing task [72]. In the visual search tasks, we used two datasets including natural images and Waldo images [73]. For the natural visual search task, 15 subjects searched for target objects across 240 cluttered images, yielding 3,600 scanpaths. For the Waldo data set, 15 participants had to find Waldo across 67 images, totalling 1,005 scanpaths. In the free-viewing task [72], we used the same natural image dataset as in the visual search task, but no target object was specified. A total of 2,400 free-viewing scanpaths were collected from 10 subjects.

The GIF files for the visual search task consist of a frame showing the target image for 1 s, followed by moving yellow rings on the search image denoting the eye movement fixations with each fixation shown for 0.3 s. The target image presentation followed by eye movement fixations iterate infinitely with a gap of 1 s between each iteration. For the free-viewing task, GIF files consist of infinitely iterating eye movement fixations denoted by yellow rings on the viewing image with each fixation shown for 0.3 s. A gap of 1 s is introduced between each iteration.

S4.1.2 AI responses

For the three datasets, we used a modified version of IVSN [72, 73], DeepGaze3 [39] and GBVS models [29] to generate eye movement predictions. To generate the fixations, we used inhibition-of-return centered on the current fixation with a window size of 100x100 for waldo images and 200x200 for natural images. The process of generating GIF files is the same as described in S4.1.1.

S4.2. Turing test

S4.2.1 Collecting human judge responses for Turing test

Separate Turing tests were launched for eye movements from visual search and free-viewing tasks. We presented infinitely repeating animated clips of eye movements from humans and computational model predictions with maximum 15 fixations to crowd-workers (judges) on AMT. For eye movement sequences with larger than 15 fixations, first 15 fixations were shown. On the AMT interface, each judge had to identify if the presented eye movement clip was from a human or a computational model. We randomly sampled 12 eye movement clips - 6 from humans and 6 from computation models (distributed equally among IVSN, GBVS and DeepGaze3) and presented them to each judge. To filter out inattentive judges, judges were also asked a control question ”What do you see in the presented clip?” with one correct answer among 3 options. Responses from judges with a score of 7 or more out of 12 for the control question were considered. To make sure that the judges paid attention to the eye movement sequences, the judges were allowed to respond to the questions only when the presented clip had played at least once. In total, responses from 100 judges for visual search task and 80 judges for free-viewing task were considered for the Turing test.

Fig. S31 and **Fig.** S32 show an example image of the AMT user interface for the Turing test for visual search and free-viewing tasks respectively. We showed the human judges a GIF consisting of eye movement sequences as discussed above and asked the judges to identify if the sequence shown is from human or machine. We also asked them to classify if the eye movements are from male or female. Additionally, to filter out inattentive judges, we introduced a control question asking what objects they see in the image among 3 options. Sample video clip of the AMT interface for the Turing test with the playing eye movements for the free-viewing task and visual search tasks can be accessed here `https://drive.google.com/file/d/10SYSrXuWIeVTQTMH2Lpsg3ejfnMMOLg_/view?usp=sharing` and here `https://drive.google.com/file/d/1JxIdbrHNEEIq6HGS1YftNC3-00vv5ow0/view?usp=sharing` respectively.

S4.2.2 AI judge

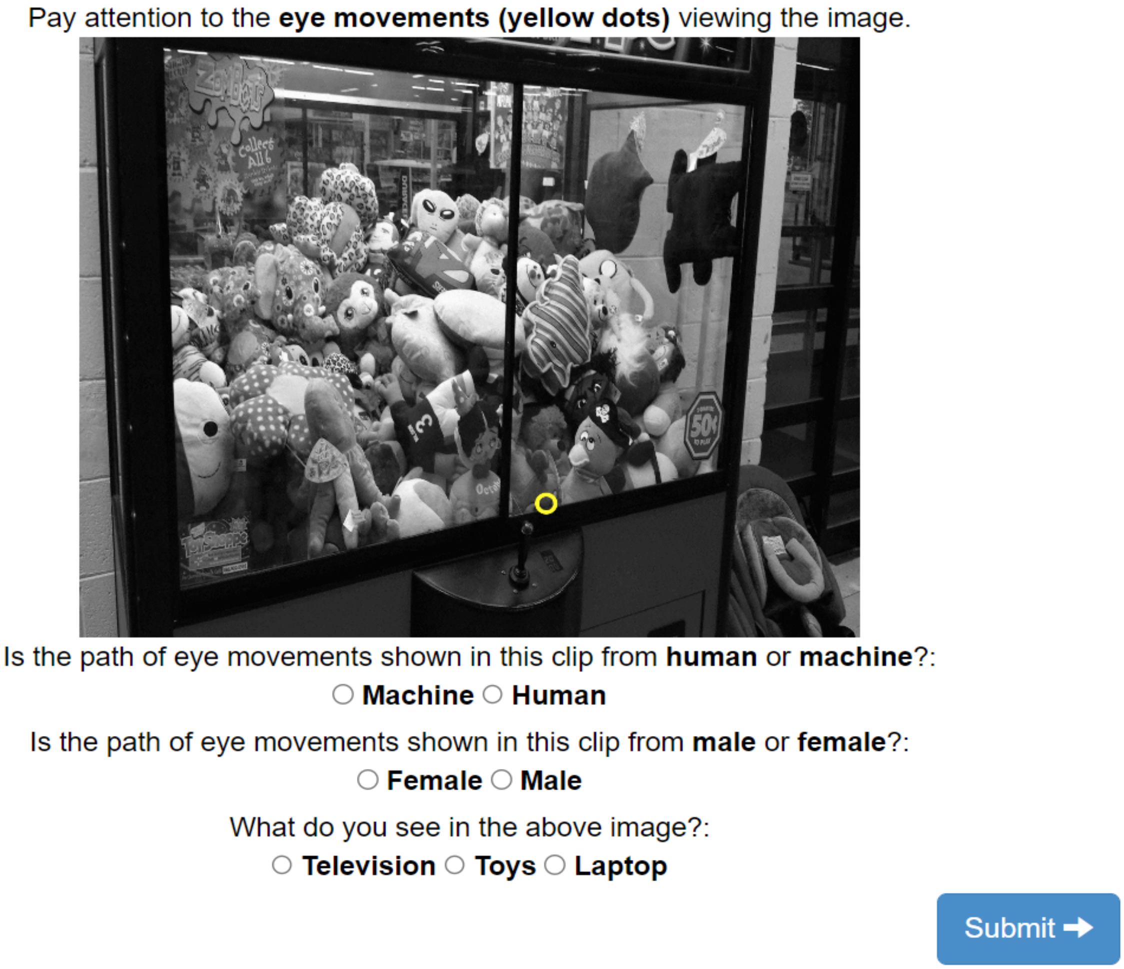
We also performed Turing tests using an SVM as an AI judge. Sequences of eye movements from humans and computational models were fed as input in the form of array of fixation coordinates to train an SVM to classify human and machine eye movements. First 10 fixations from each sequence were used for training the SVM and evaluating classification performance. Sequences having less than 10 fixations were discarded to ensure every input sequence has exactly 10 fixations. Fixation coordinates were normalized to a range between -1 to 1. The SVM was trained using 10-fold cross validation. Model performance on validation across folds was calculated and averaged over 5 random seeds.

S4.3. Discussion

While GBVS [29] and DeepGaze3 [39] are bottom-up saliency prediction models, IVSN is a zero-shot visual search model where the target information modulates the bottom-up visual processes. The increased performance of IVSN may be attributed to its top-down and bottom-up attention mechanisms. Moreover, DeepGaze3 is a supervised learning model trained on human eye fixations during free-viewing tasks and achieves better bottom-up saliency prediction according to standard evaluation metrics, such as NSS, AUC, and sAUC. However, it is surprising to note that the eye movements predicted by DeepGaze3 do not fool humans as well as GBVS [29], which is a bottom-up attention prediction model based solely on low-level image statistics. This observation suggests that high performance in terms of standard evaluation metrics in computer vision might not directly correlate with good performance in Turing tests.

Dataset	Task	Number of stimuli	Number of human scanpaths
NatureDesign [73]	Visual Search	240	3600
FindingWaldo [73]	Visual Search	67	1005
NatureSaliency [72]	Free-Viewing	240	2400

Table S2. **Attention Prediction.** Datasets used for visual search and free-viewing tasks and number of human scanpaths in the dataset.



Pay attention to the **eye movements (yellow dots)** viewing the image.

Is the path of eye movements shown in this clip from **human** or **machine**?:

○ **Machine** ○ **Human**

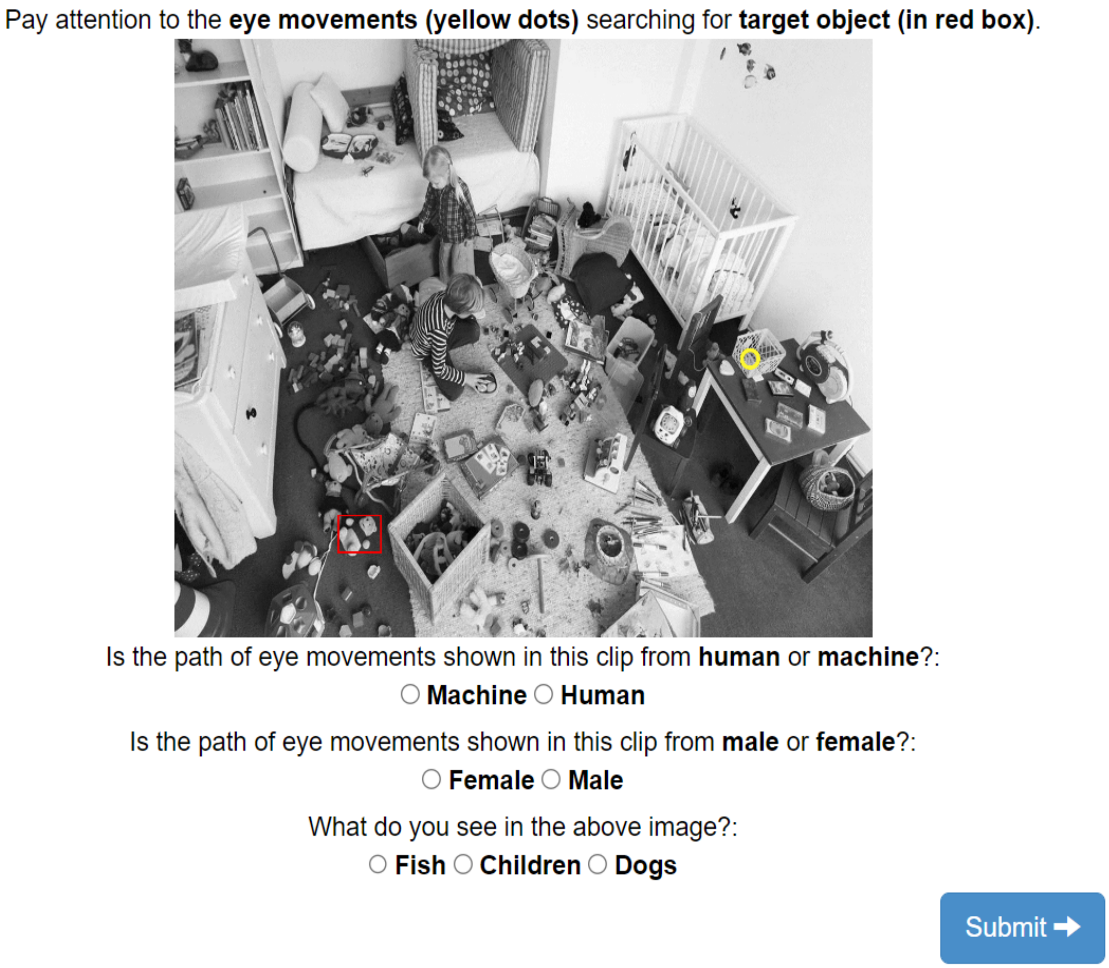
Is the path of eye movements shown in this clip from **male** or **female**?:

○ **Female** ○ **Male**

What do you see in the above image?:

○ **Television** ○ **Toys** ○ **Laptop**

Figure S31. **Attention prediction.** AMT user interface for collecting human judge responses for the free-viewing task (Turing test).



Pay attention to the **eye movements (yellow dots)** searching for **target object (in red box)**.

Is the path of eye movements shown in this clip from **human** or **machine**?:

○ **Machine** ○ **Human**

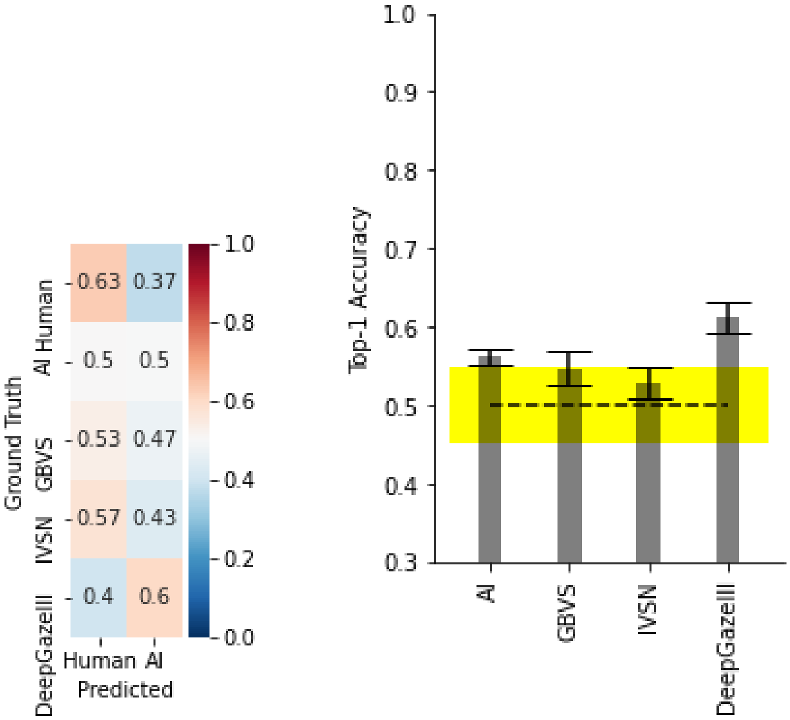
Is the path of eye movements shown in this clip from **male** or **female**?:

○ **Female** ○ **Male**

What do you see in the above image?:

○ **Fish** ○ **Children** ○ **Dogs**

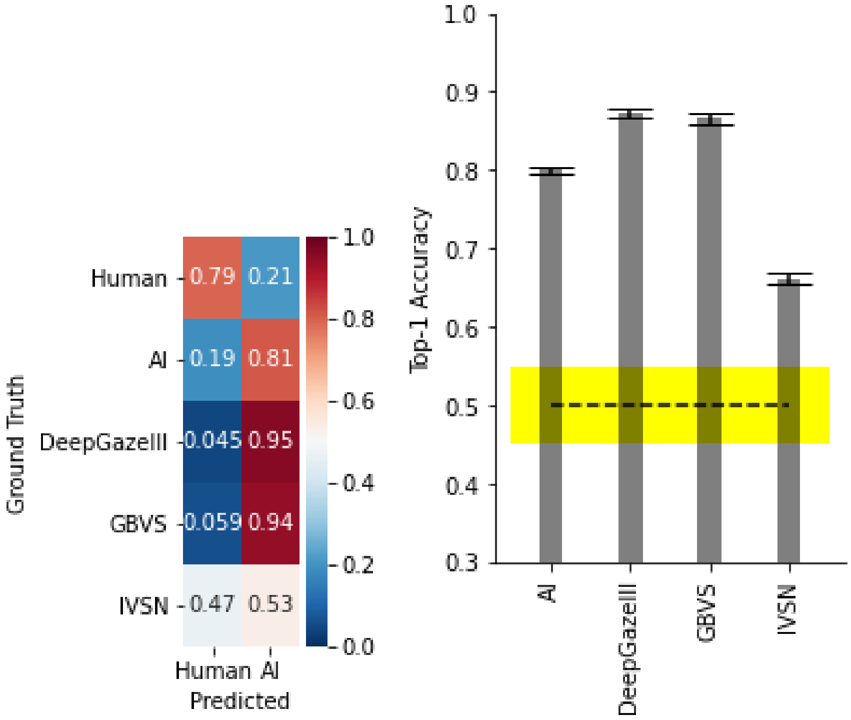
Figure S32. **Attention prediction.** AMT user interface for collecting human judge responses for the visual search task (Turing test).



(a) Confusion matrix

(b) Top1 accuracy

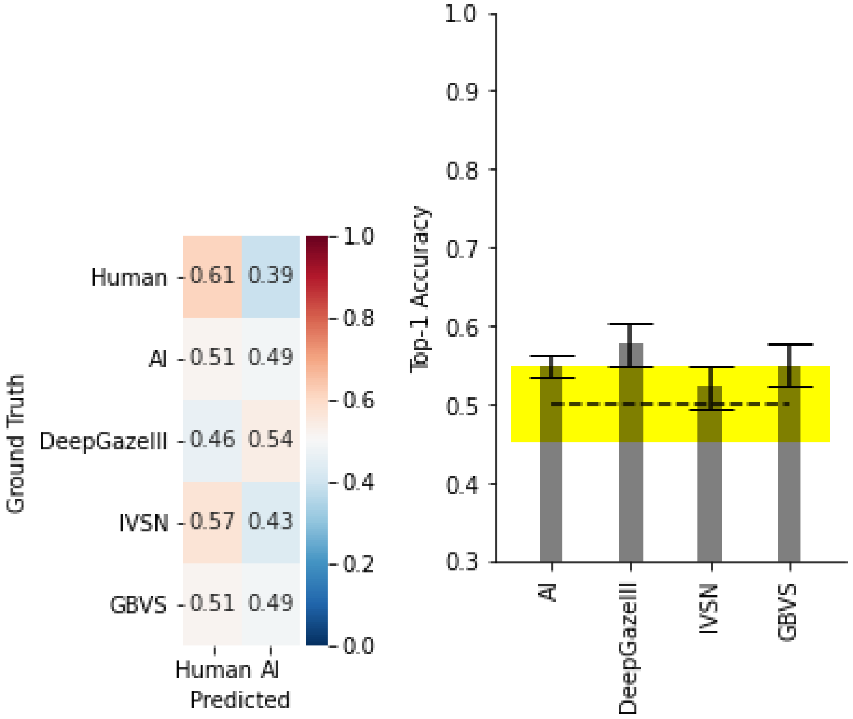
Figure S33. **Attention prediction**. Results of the Turing test averaged over visual search and free-viewing tasks for human judges. (a) Confusion matrix (b) Top-1 accuracy.



(a) Confusion matrix

(b) Top1 accuracy

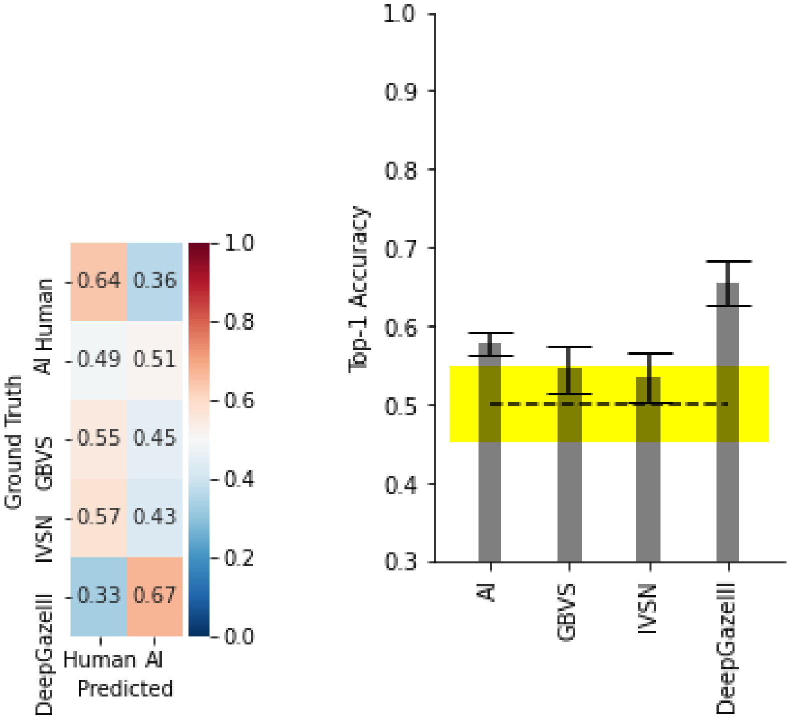
Figure S34. **Attention prediction**. Results of the Turing test averaged over visual search and free-viewing tasks for AI judges. (a) Confusion matrix (b) Top-1 accuracy.



(a) Confusion matrix

(b) Top1 accuracy

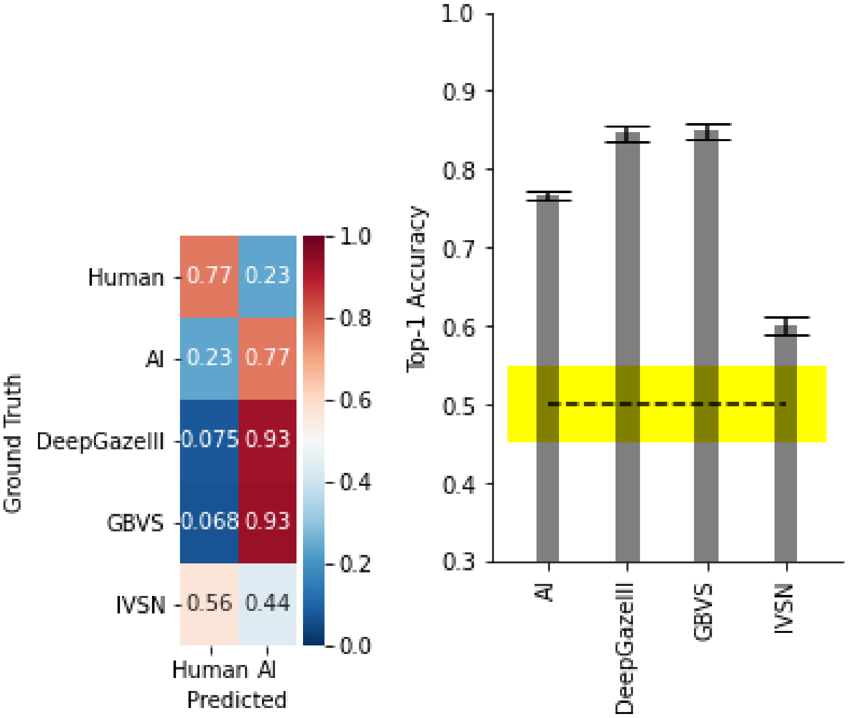
Figure S35. **Attention prediction**. Results of the Turing test in the visual search task for human judges. (a) Confusion matrix (b) Top-1 accuracy.



(a) Confusion matrix

(b) Top1 accuracy

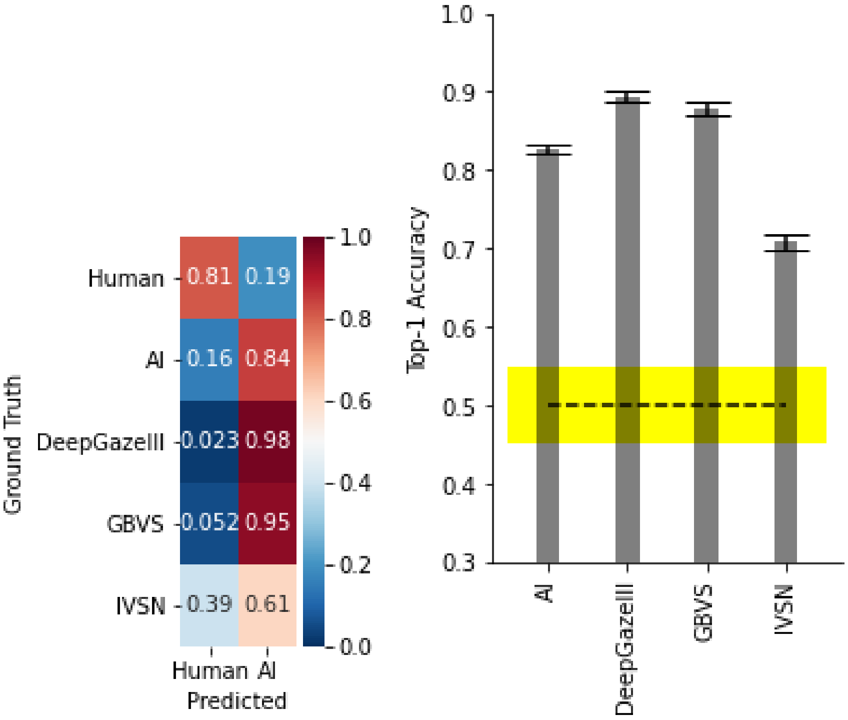
Figure S36. **Attention prediction**. Results of the Turing test in the free-viewing task for human judges. (a) Confusion matrix (b) Top-1 accuracy.



(a) Confusion matrix

(b) Top1 accuracy

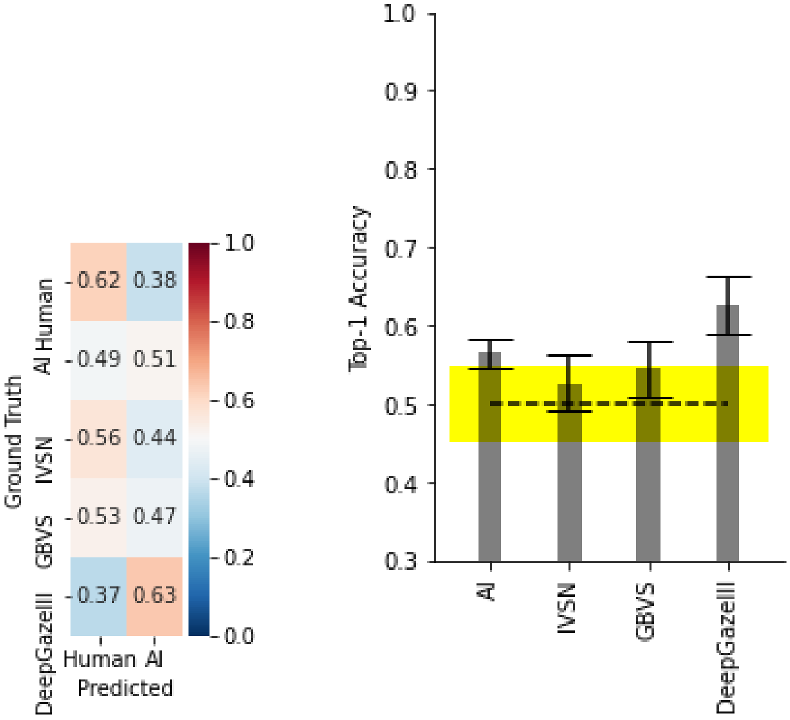
Figure S37. **Attention prediction**. Results of the Turing test in the visual search task for AI judges. (a) Confusion matrix (b) Top-1 accuracy.



(a) Confusion matrix

(b) Top1 accuracy

Figure S38. **Attention prediction.** Results of the Turing test in the free-viewing task for AI judges. (a) Confusion matrix (b) Top-1 accuracy.

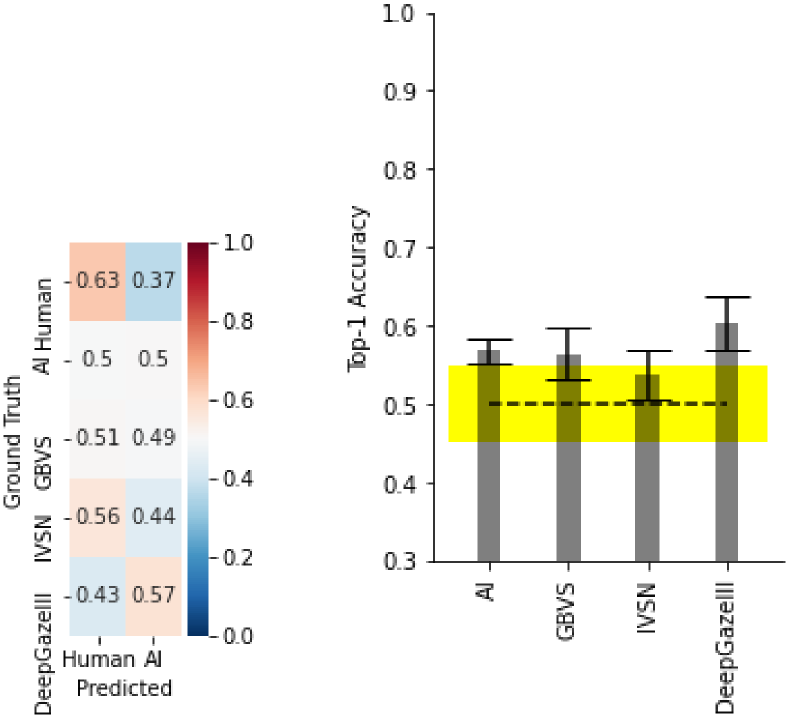


(a) Confusion matrix

(b) Top1 accuracy

Figure S39. **Attention prediction.** Results of the Turing test averaged over visual search and free-viewing tasks for human judges below age 35

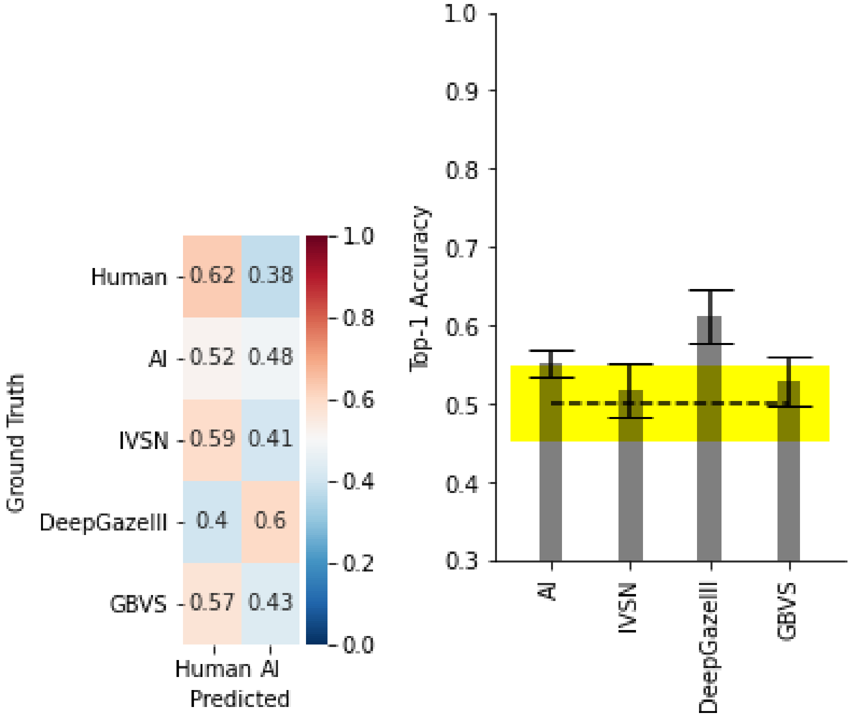
(a) Confusion matrix (b) Top-1 accuracy.



(a) Confusion matrix

(b) Top1 accuracy

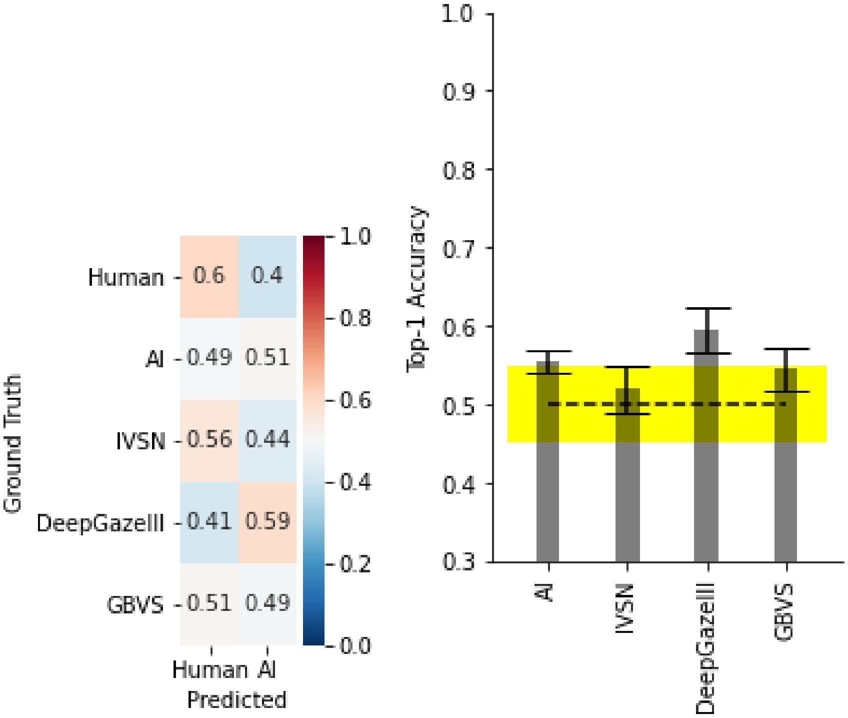
Figure S40. **Attention prediction**. Results of the Turing test averaged over visual search and free-viewing tasks for human judges between age 35 and 45. (a) Confusion matrix (b) Top-1 accuracy.



(a) Confusion matrix (b) Top1 accuracy

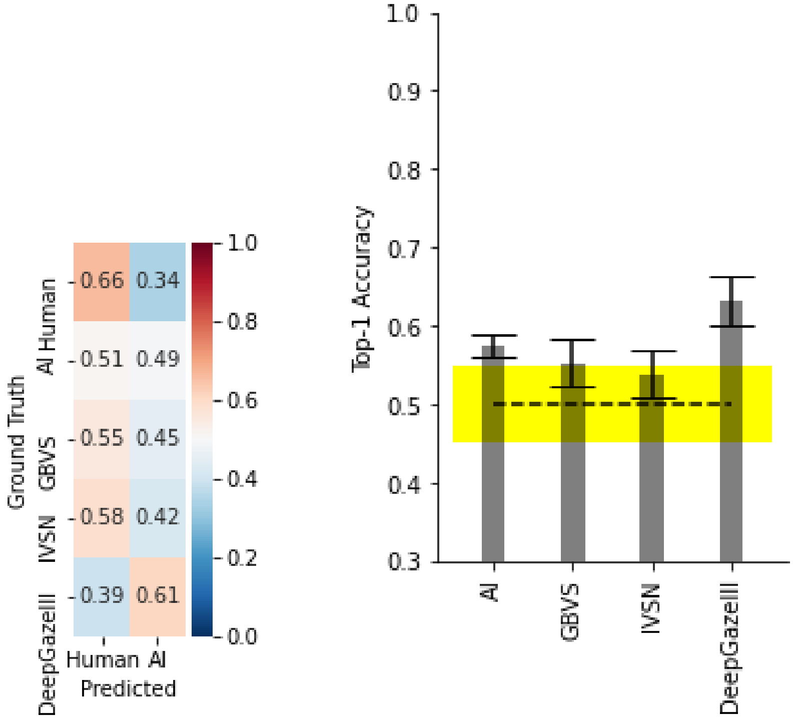
Figure S41. **Attention prediction.** Results of the Turing test averaged over visual search and free-viewing tasks for human judges above age 45

(a) Confusion matrix (b) Top-1 accuracy.



(a) Confusion matrix (b) Top1 accuracy

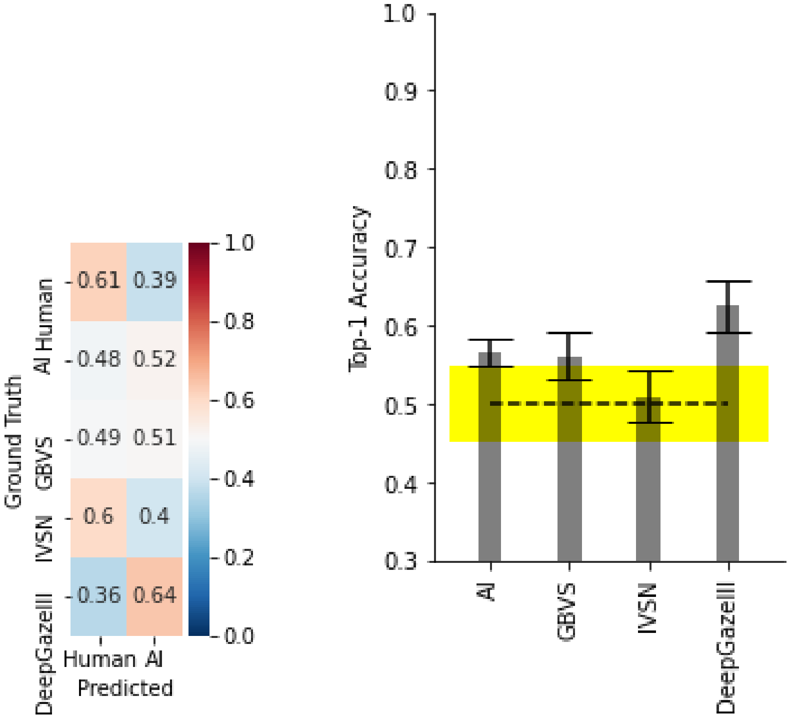
Figure S42. **Attention prediction.** Results of the Turing test averaged over visual search and free-viewing tasks for male human judges. (a) Confusion matrix (b) Top-1 accuracy.



(a) Confusion matrix

(b) Top1 accuracy

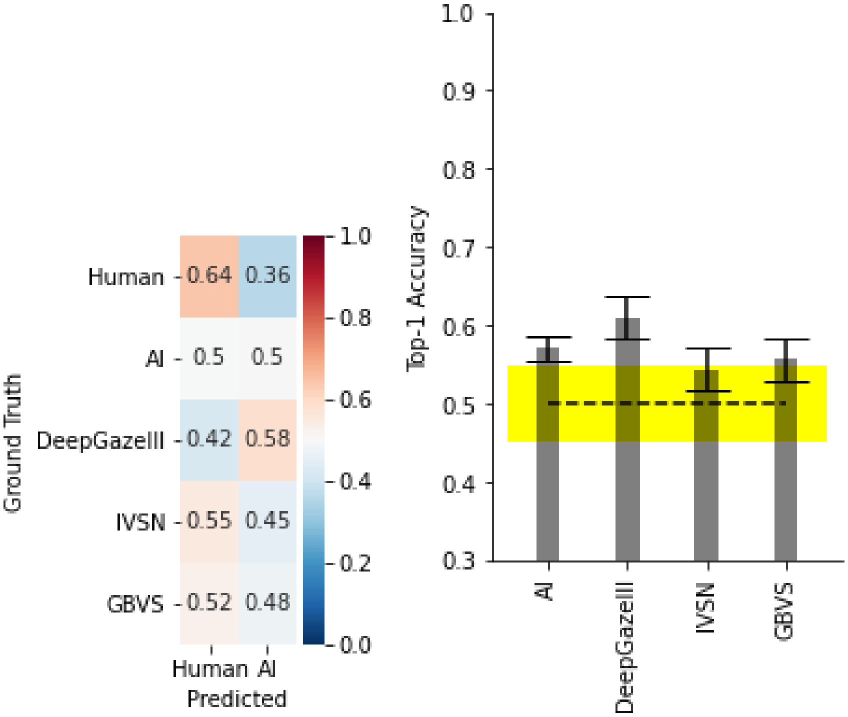
Figure S43. **Attention prediction.** Results of the Turing test averaged over visual search and free-viewing tasks for female human judges. (a) Confusion matrix (b) Top-1 accuracy.



(a) Confusion matrix (b) Top1 accuracy

Figure S44. **Attention prediction.** Results of the Turing test averaged over visual search and free-viewing tasks for human judges with highest educational level of middle/high school.

(a) Confusion matrix (b) Top-1 accuracy.

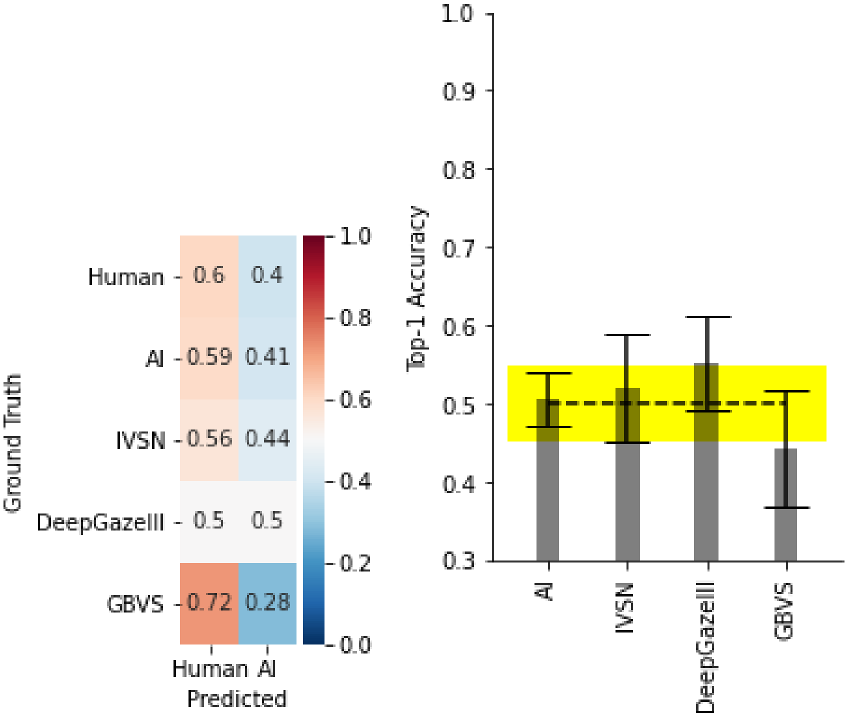


(a) Confusion matrix

(b) Top1 accuracy

Figure S45. **Attention prediction.** Results of the Turing test averaged over visual search and free-viewing tasks for human judges with highest educational level of Bachelor.

(a) Confusion matrix (b) Top-1 accuracy.



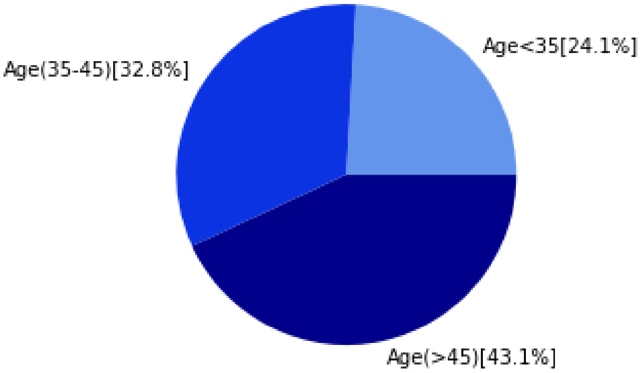
Ground Truth \ Predicted	Human	AI
Human	0.6	0.4
AI	0.59	0.41
IVSN	0.56	0.44
DeepGazeIII	0.5	0.5
GBVS	0.72	0.28

(a) Confusion matrix

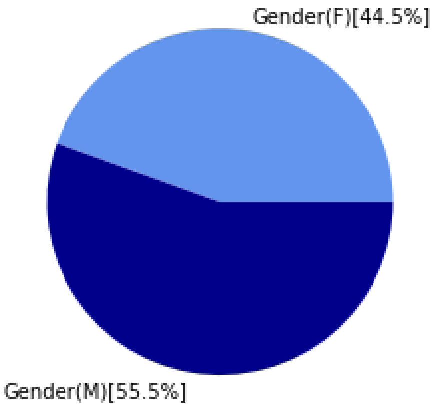
(b) Top1 accuracy

Figure S46. **Attention prediction.** Results of the Turing test averaged over visual search and free-viewing tasks for human judges with highest educational level of Master or Post-graduate.

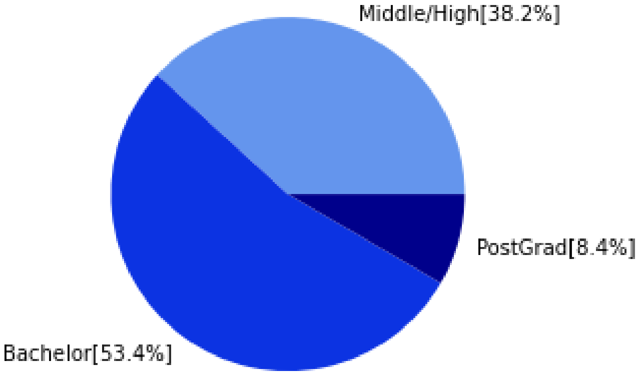
(a) Confusion matrix (b) Top-1 accuracy.



(a) Age distribution



(b) Gender distribution



(c) Education level distribution

Figure S47. **Attention prediction.** Demographic information for the human judges for the visual search and free-viewing tasks combined (a) Age. (b) Gender. (c) Education level.

S5. Image captioning

S5.1. Dataset

S5.1.1 Collecting human captions

In **Fig.** S48, we show the AMT interface used for collecting human captions. Our AMT interface received significant inspirations from the COCO Captions data collection interface [10]. We provide the following instructions to each participant:

- Describe all the important parts of the scene.
- The description should contain at least 6 words.
- Avoid making spelling errors in the description.
- Do not describe unimportant details.
- Do not use any special characters like !, #, $, etc.
- Do not start the sentence with “There is” or “There are”.
- Do not write your descriptions as “An image containing ...”, “A photo of ...” or similar.
- Do not describe things that might have happened in the future or past.
- Do not use proper names for people.
- Do not describe what a person in the image might say.
- After typing in the response, click “SUBMIT” to go to the next image.

Moreover, to prevent invalid responses, we implemented automatic controls in our interface that issued warning popups to the participants. We list all such controls below:

- Response should not contain less than 6 words.
- No special characters in the response.
- Response time must be more than 4s.
- No response should contain more than 4 identical words.
- Response should not contain the words “image” and “photo”.
- Successive responses must not be same.

S5.1.2 Collecting machine captions

We use GIT [63], OFA [64], BLIP [41], ClipCap [46], and Microsoft’s Azure Cognitive Services [2] for collecting machine captions. For each model, we provide details about specific variants used and their open-source web links in **Table.** S3. We show some random samples from our collected caption dataset in **Fig** S49. Finally, the average caption length for all collected captions is shown in **Fig** S50 .

S5.2. Turing test

Fig. S51 shows a screenshot of the AMT interface used for conducting the Turing tests. We show the Turing test results for human judges in **Fig.** S52, and for AI judge in **Fig.** S53. In addition, we show Turing test results for human judges based on demographics, like age (**Fig.** S54, S55, S56), gender (**Fig.** S57, S58), and education (**Fig.** S59, S60, S61). Finally, we show results for the Turing tests based on the type of image used from nocaps dataset [3], specifically in-domain (**Fig.** S62), near-domain (**Fig.** S63), and out-of-domain images (**Fig.** S64).

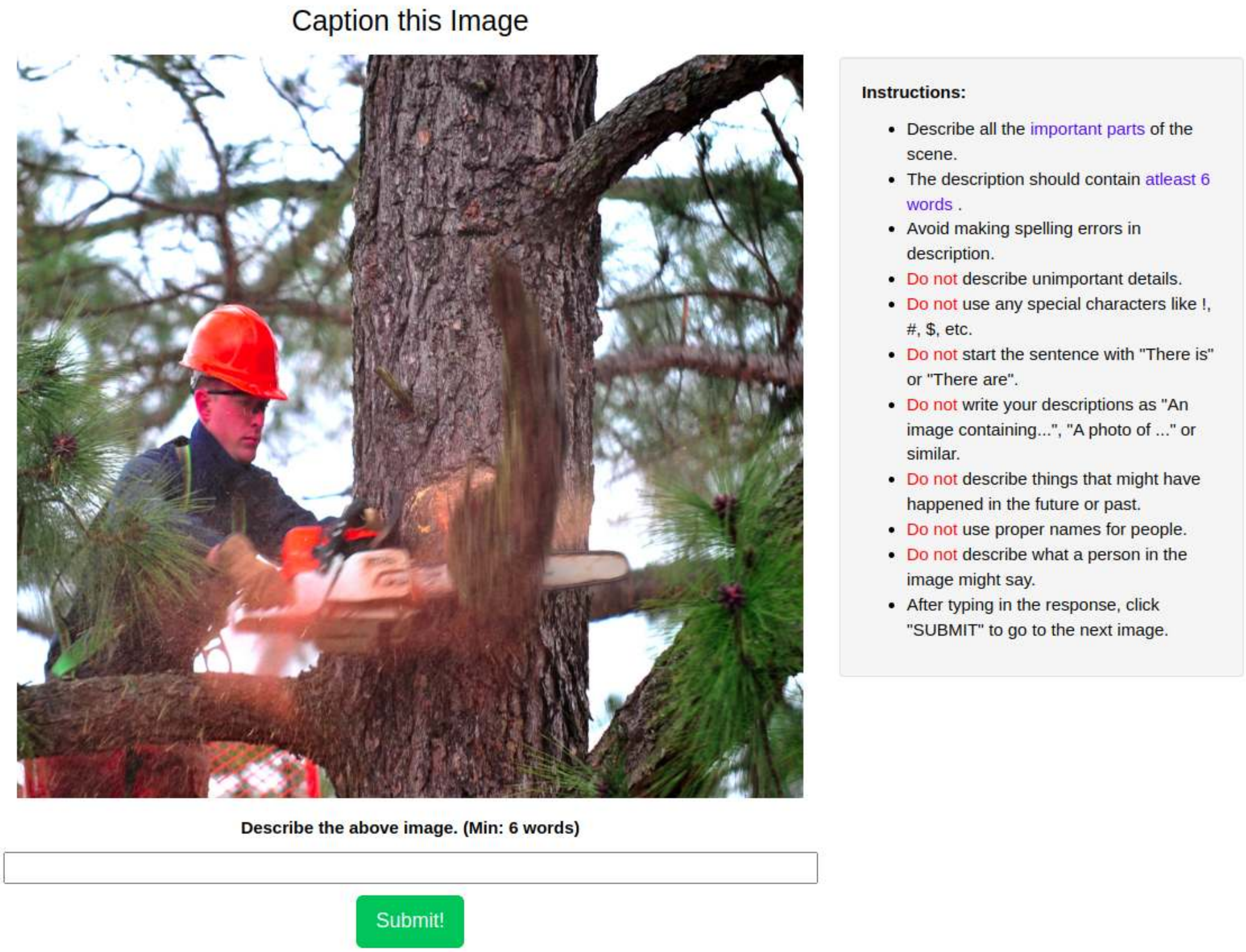
S5.3. Discussion

GIT [63], OFA [64], and BLIP [41] are recent transformer-based models and are good at fooling human judges (**Fig.** **S52**). In contrast to ClipCap which is only trained in the image captioning task, the above three models are generative unified transformer frameworks, trained on multiple tasks involving multiple modalities. This suggests that multi-task and multi-modal training aids models to generate human-like captions.

Out-of-domain nocaps images contain object classes that are visually very distinct from COCO. Therefore, it is surprising to see models like GIT, OFA, and BLIP which were fine-tuned on COCO to perform better in Turing tests on out-of-domain nocaps images (**Fig.** **S64**) than on in-domain nocaps images (**Fig.** **S62**). Moreover, on out-of-domain nocaps images, BLIP outperforms all other models (**Fig.** **S64**).

Model	**Model Variant**	**Model Details**	**Web links**
GIT [63]	GIT-Large	Finetuned on COCO Captions	Link
OFA [64]	OFA-Huge	Finetuned on COCO Captions	Link
BLIP [41]	BLIP-Large	Finetuned on COCO Captions	Link 1, Link 2
ClipCap [46]	ClipCap-Transformer (Beam Search)	Pretrained on COCO Captions	Link
Microsoft's Azure Cognitive Services [2]	-	-	Link

Table S3. **Image captioning.** Different model variants used for collecting machine captions.



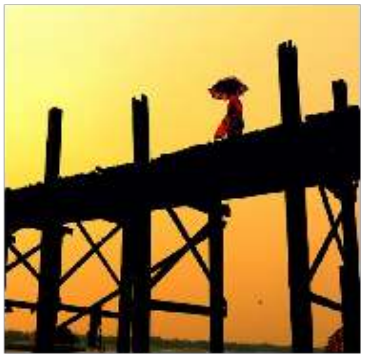
Caption this Image

Instructions:

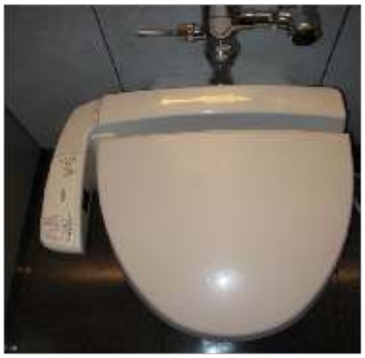
- Describe all the important parts of the scene.
- The description should contain atleast 6 words .
- Avoid making spelling errors in description.
- Do not describe unimportant details.
- Do not use any special characters like !, #, $, etc.
- Do not start the sentence with "There is" or "There are".
- Do not write your descriptions as "An image containing...", "A photo of ..." or similar.
- Do not describe things that might have happened in the future or past.
- Do not use proper names for people.
- Do not describe what a person in the image might say.
- After typing in the response, click "SUBMIT" to go to the next image.

Describe the above image. (Min: 6 words)

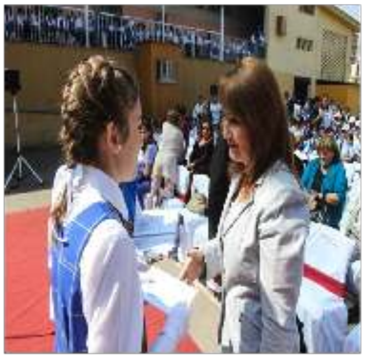
Figure S48. **Image captioning.** AMT user interface for collecting responses.



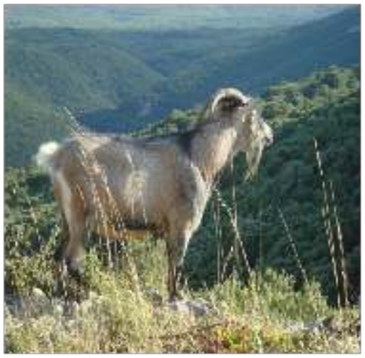
Human: a woman with an umbrella walking on a pier
GIT: a person sitting on a wooden bridge with an umbrella
OFA: a person walking on a dock with an umbrella
BLIP: a person walking across a bridge with an umbrella
ClipCap: a person holding an umbrella in front of a building
Microsoft_API: a person on a swing



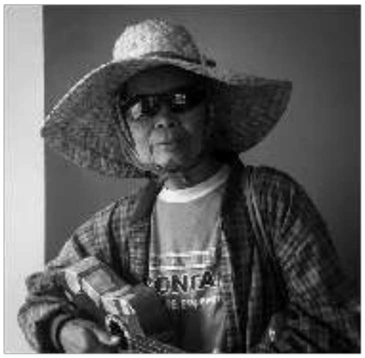
Human: a toilet is closed with several buttons to the side of it
GIT: a close up of a toilet with controls
OFA: a white toilet with a control panel on the side of it
BLIP: a close up of a toilet in a bathroom
ClipCap: a close up of a toilet with a toilet seat
Microsoft_API: a toilet with a control panel



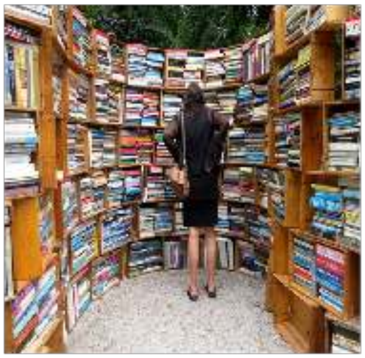
Human: a girl receiving a certificate from a woman in a group of people at a presentation
GIT: a woman shakes hands with a student
OFA: two women are talking to each other in front of a crowd
BLIP: two women standing next to each other in front of a crowd
ClipCap: a woman standing in front of a group of people
Microsoft_API: a couple of women in a stadium



Human: the old billy goat takes in the view from the wild
GIT: a goat standing on top of a lush green hillside
OFA: a goat standing on top of a hill
BLIP: a goat standing on top of a lush green hillside
ClipCap: a goat standing in the grass near a mountain
Microsoft_API: a goat standing on a hill



Human: an elderly person in a floppy hat is playing a ukelele
GIT: a black and white photo of a person playing a guitar
OFA: a woman wearing a hat and sunglasses holding a guitar
BLIP: a man wearing a hat and holding a guitar
ClipCap: a black and white photo of a woman wearing a hat
Microsoft_API: a man wearing a hat and sunglasses



Human: a woman surrounded by bookshelves with trees in the background
GIT: a woman standing in front of a bookshelf
OFA: a woman standing in a circle of books in a library
BLIP: a woman standing in front of a wall of books
ClipCap: a woman standing next to a wall full of books
Microsoft_API: a person standing in a library

Figure S49. **Image captioning.** Random samples from our collected caption dataset.

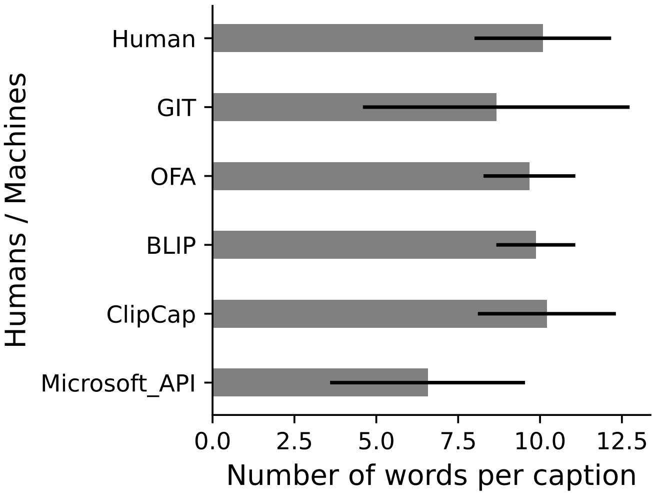
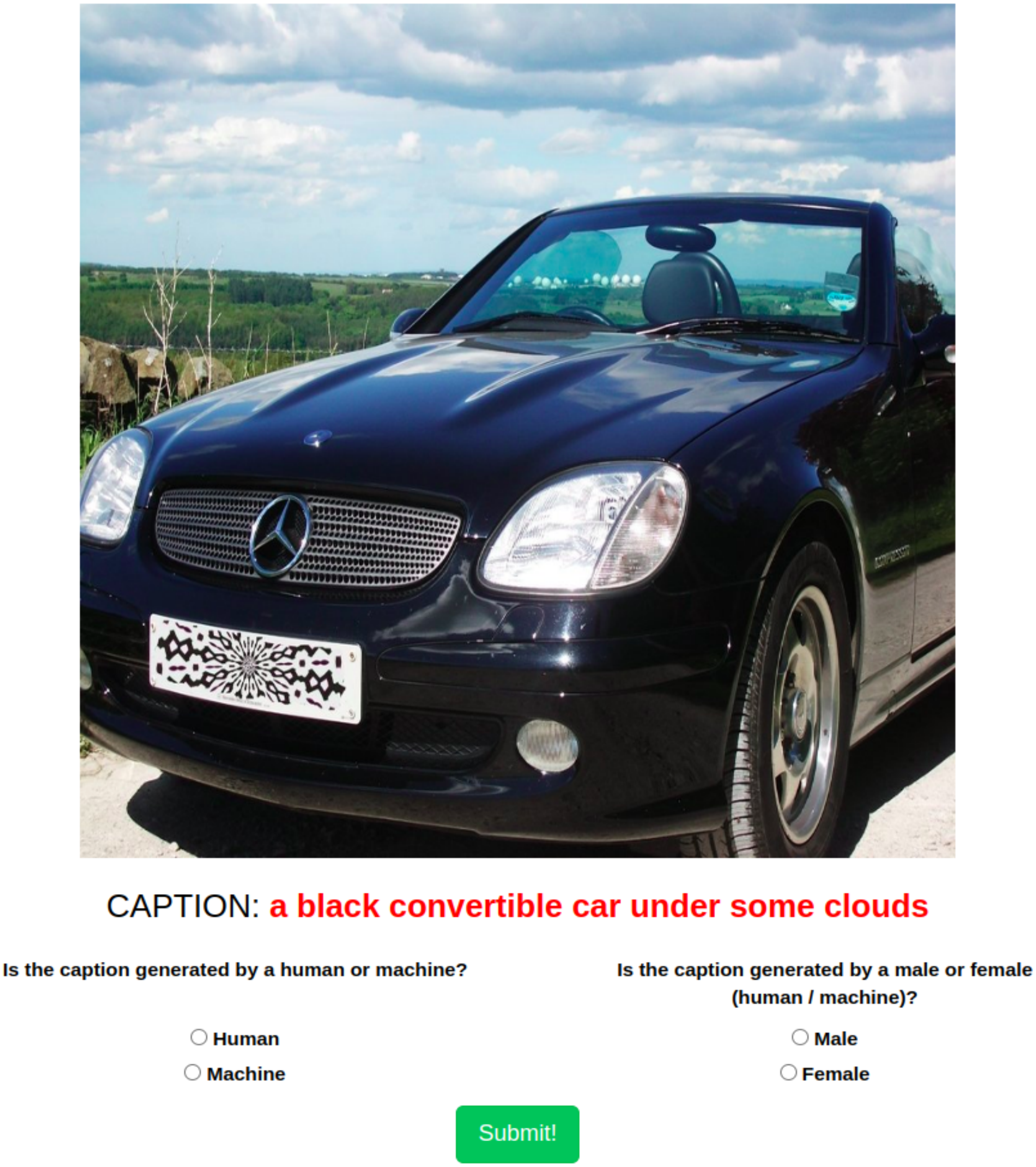


Figure S50. **Image captioning.** Average caption lengths.



CAPTION: a black convertible car under some clouds

Is the caption generated by a human or machine?

○ Human

○ Machine

Is the caption generated by a male or female (human / machine)?

○ Male

○ Female

Figure S51. **Image captioning.** AMT user interface for Turing test.

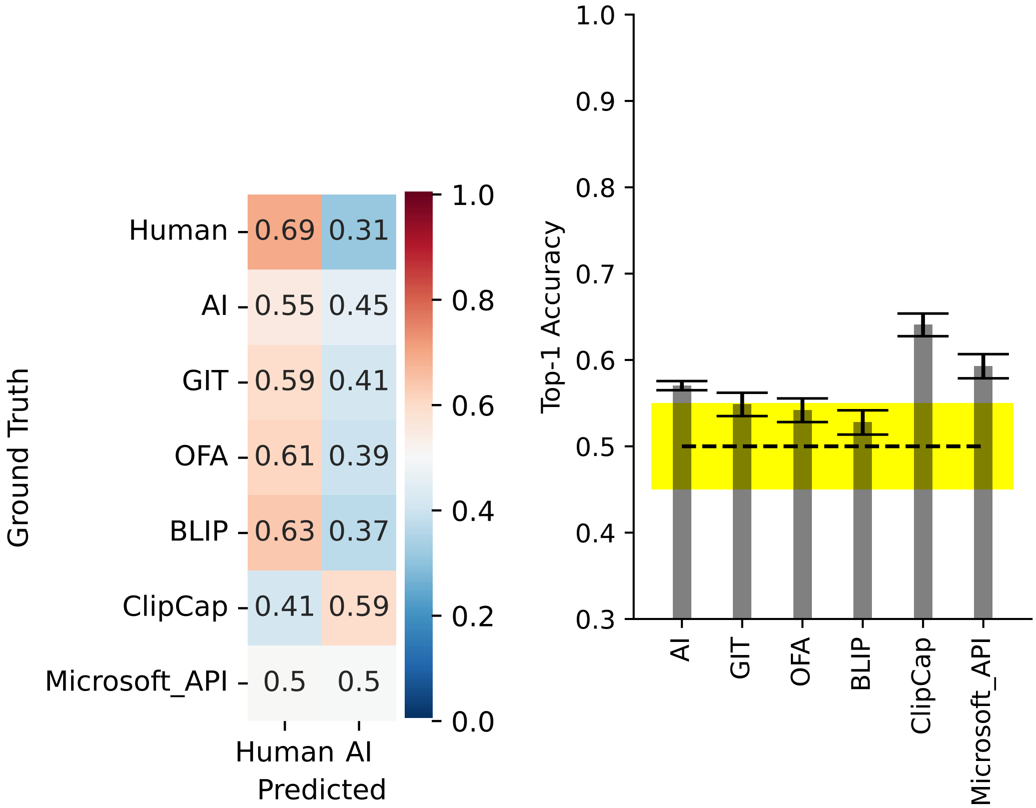


Figure S52. **Image captioning.** Results of the Turing test for human judges (a) Confusion matrix (b) Top-1 accuracy.

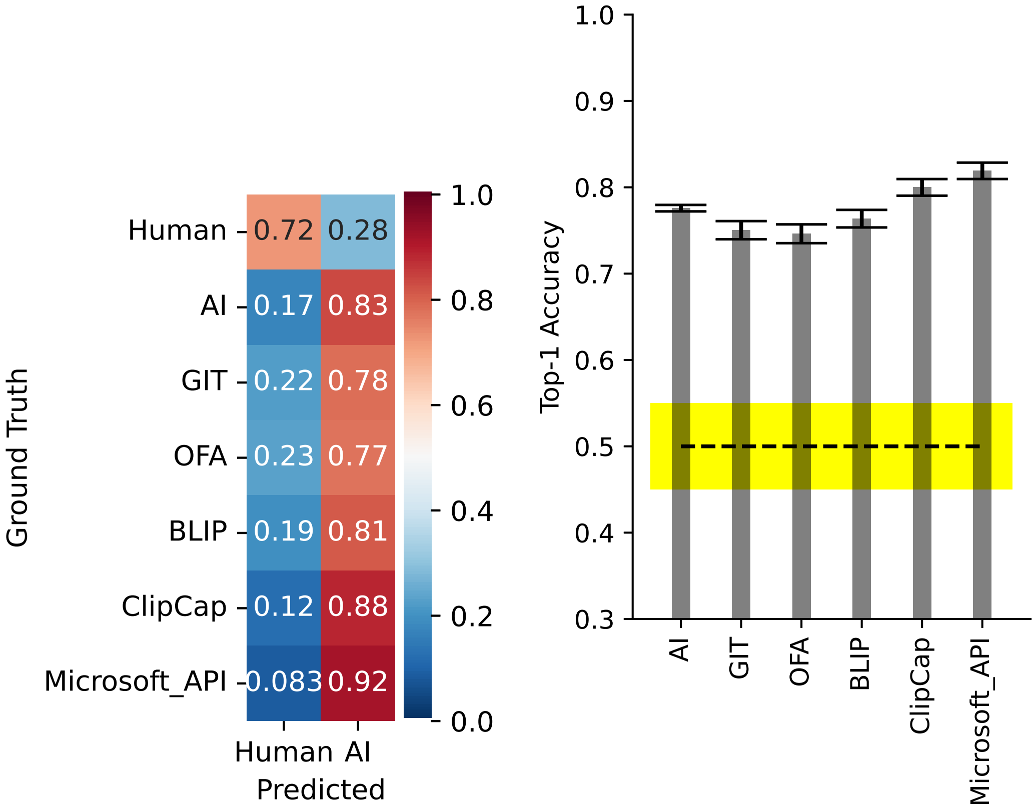


Figure S53. **Image captioning.** Results of the Turing test for AI judges.
(a) Confusion matrix (b) Top-1 accuracy.

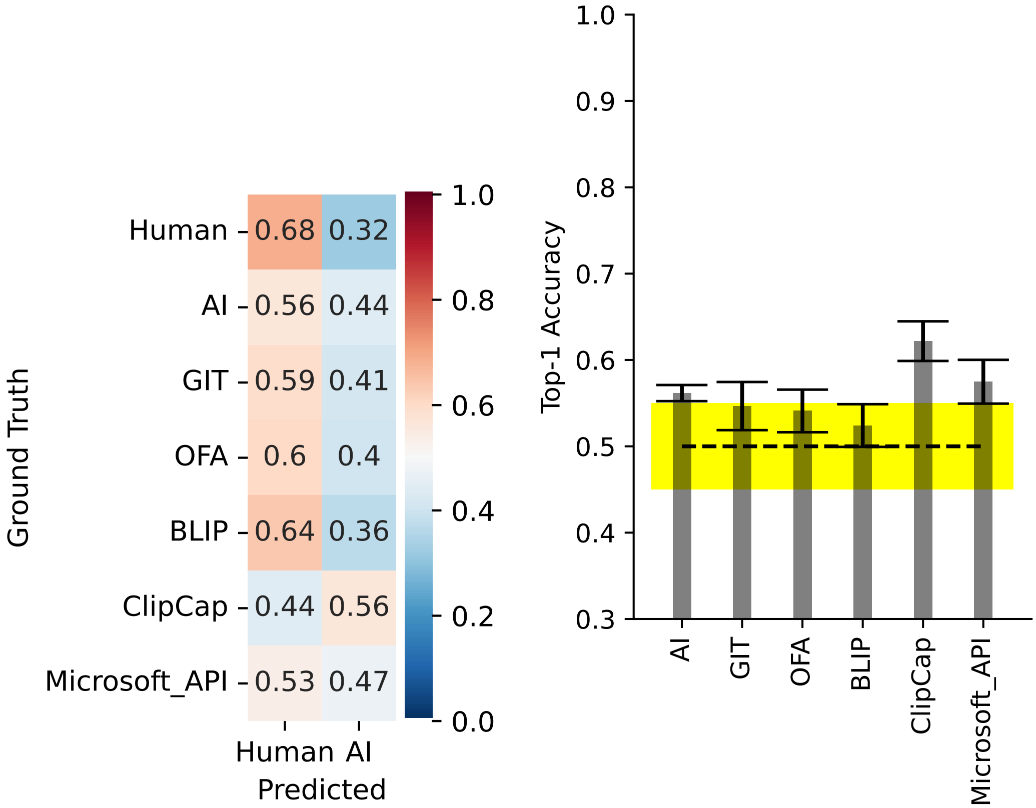


Figure S54. **Image captioning.** Results of the Turing test for human judges below age 35. (a) Confusion matrix (b) Top-1 accuracy.

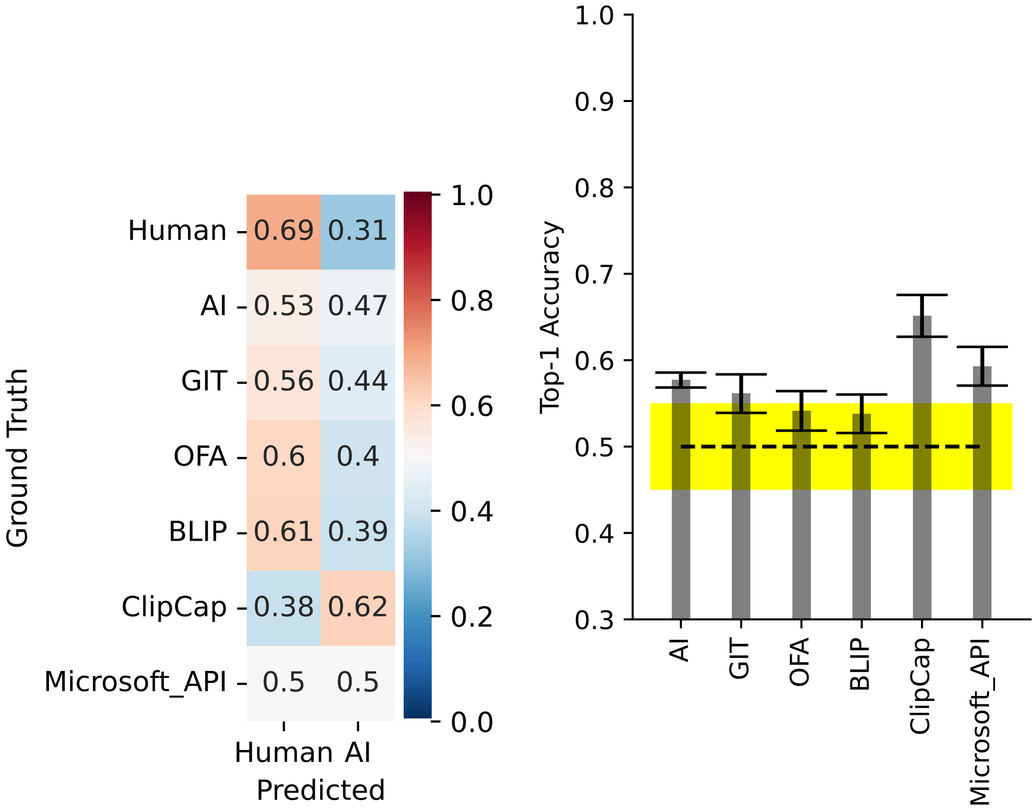


Figure S55. **Image captioning.** Results of the Turing test for human judges between age 35 and 45. (a) Confusion matrix (b) Top-1 accuracy.

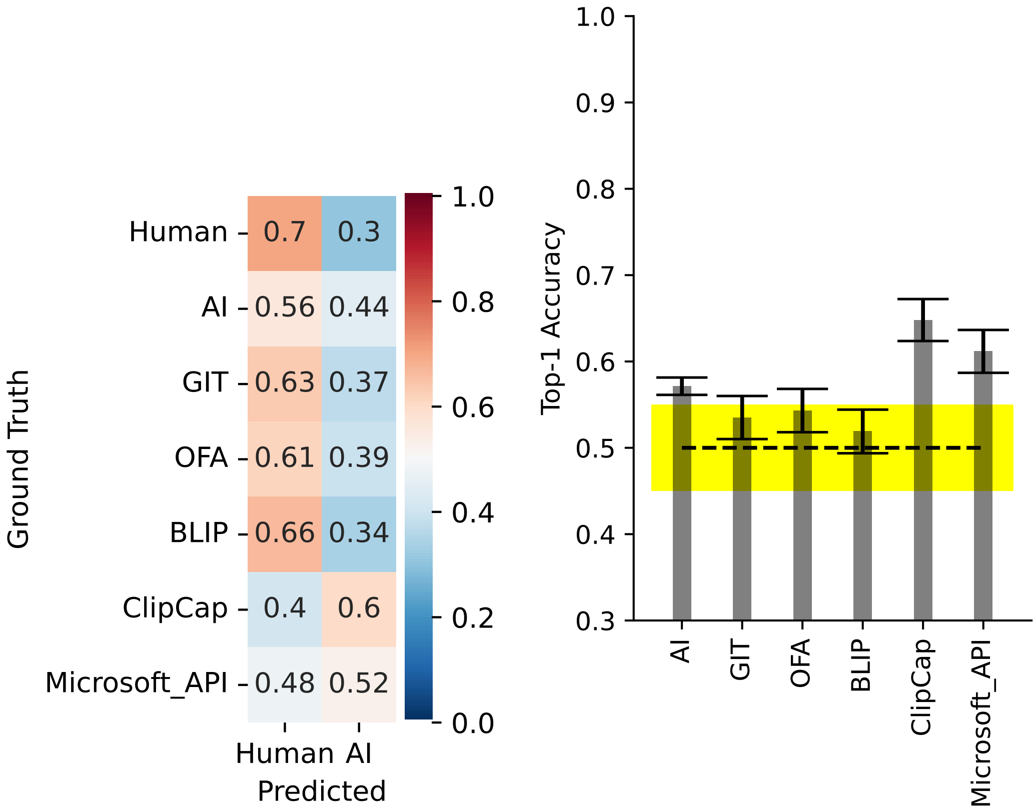


Figure S56. **Image captioning.** Results of the Turing test for human judges above age 45. (a) Confusion matrix (b) Top-1 accuracy.

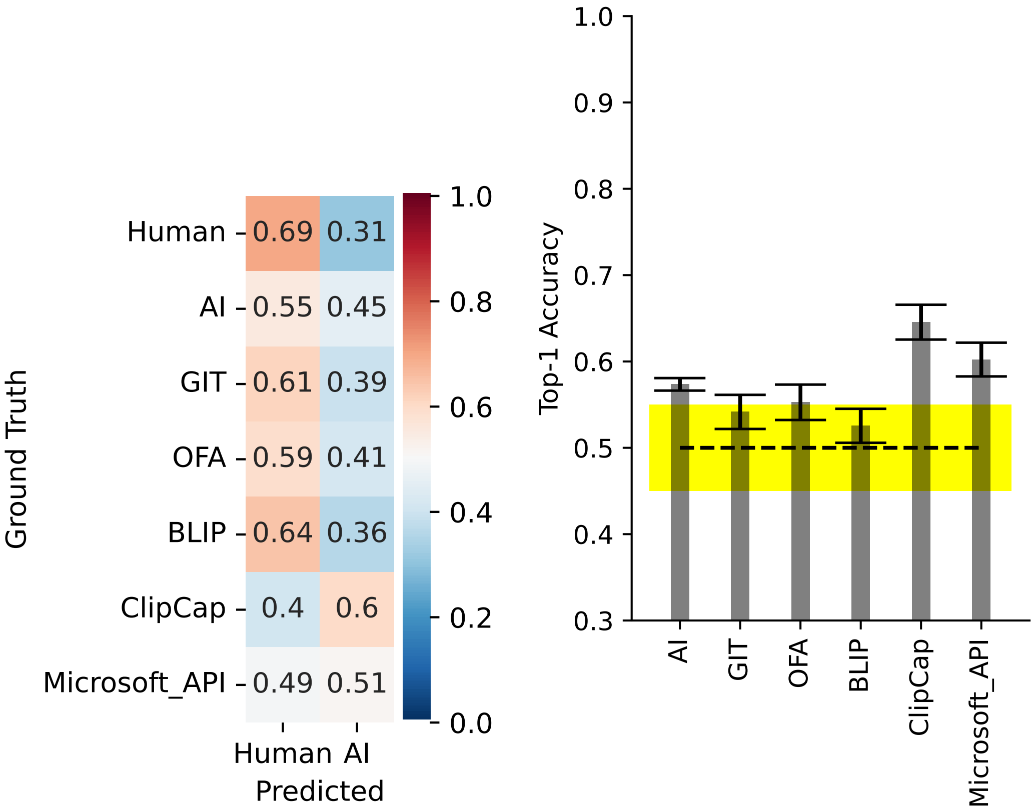


Figure S57. **Image captioning.** Results of the Turing test for male human judges.
(a) Confusion matrix (b) Top-1 accuracy.

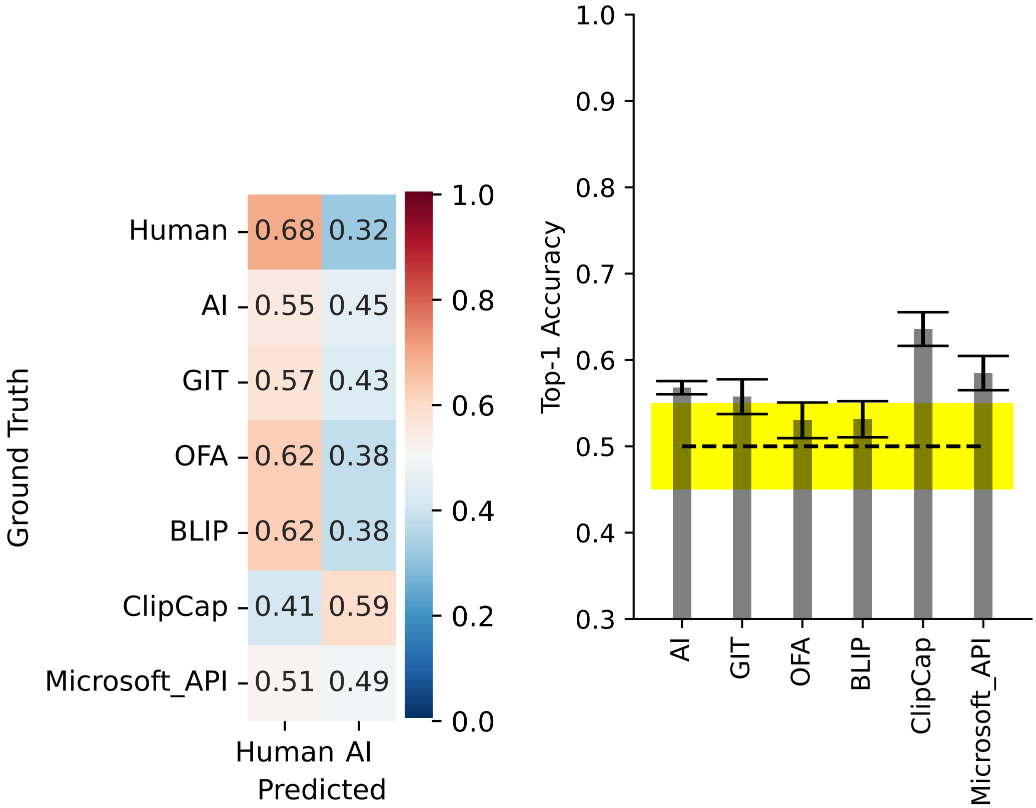


Figure S58. **Image captioning.** Results of the Turing test for female human judges. (a) Confusion matrix. (b) Top-1 accuracy.

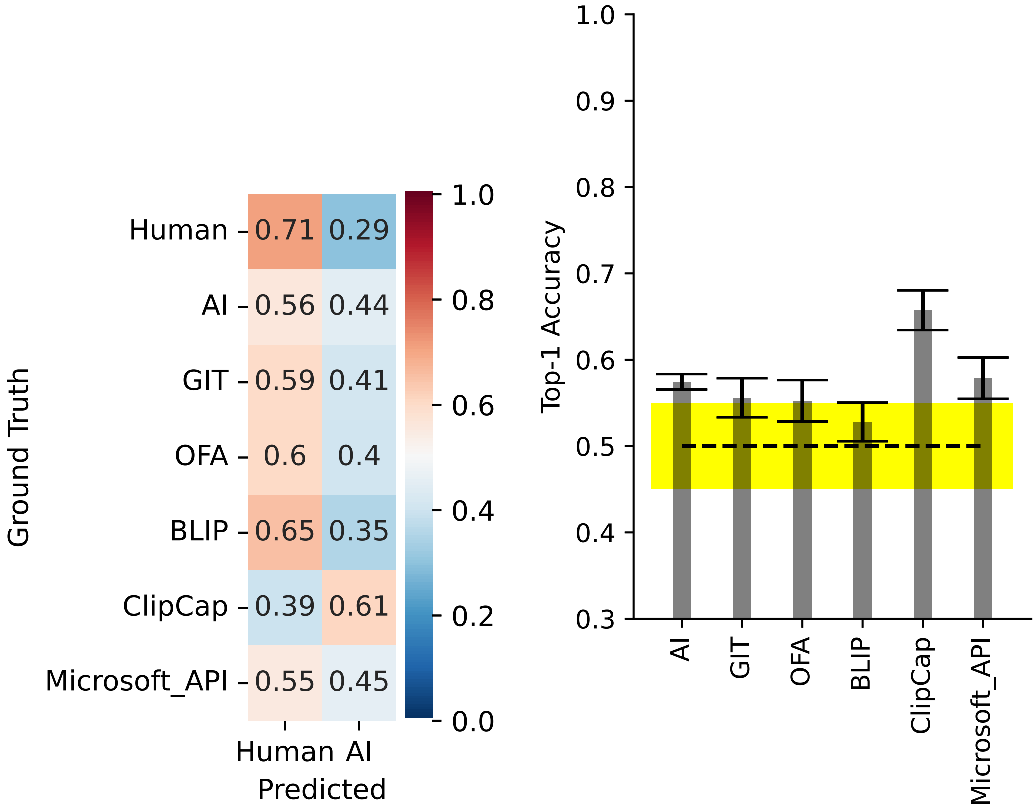


Figure S59. **Image captioning.** Results of the Turing test for human judges with highest education level of middle/high school. (a) Confusion matrix (b) Top-1 accuracy.

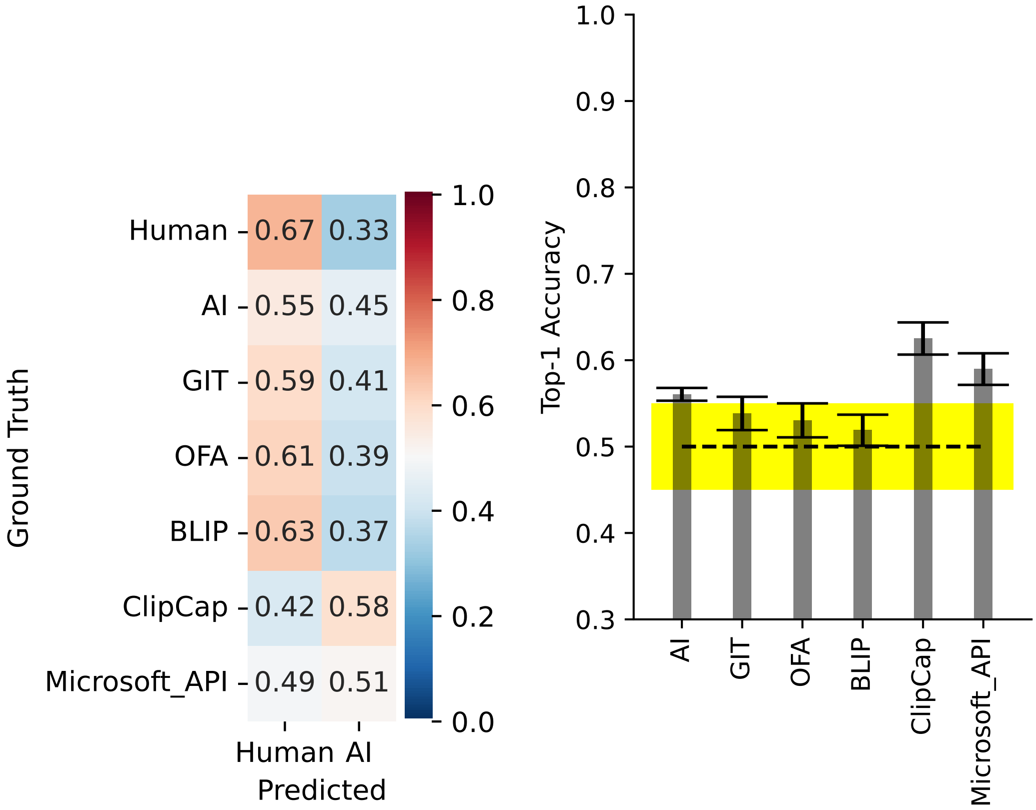


Figure S60. **Image captioning.** Results of the Turing test for human judges with highest education level of Bachelor. (a) Confusion matrix (b) Top-1 accuracy.

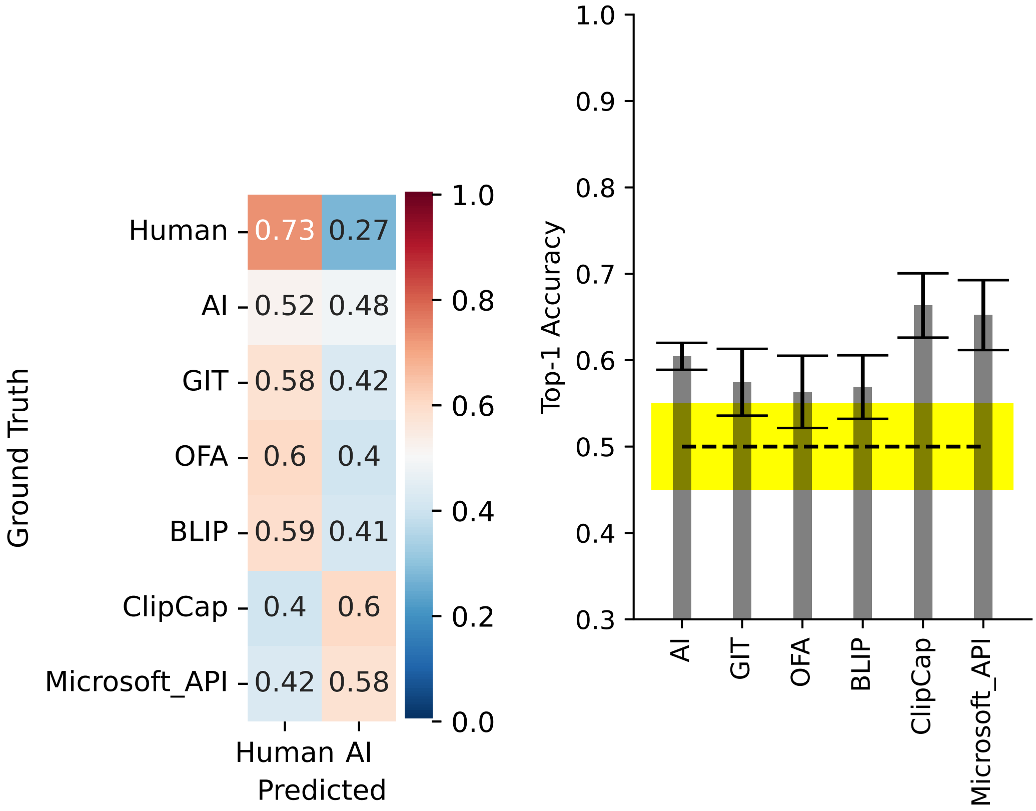


Figure S61. **Image captioning.** Results of the Turing test for human judges with highest education level of Master or Post-graduate. (a) Confusion matrix (b) Top-1 accuracy.

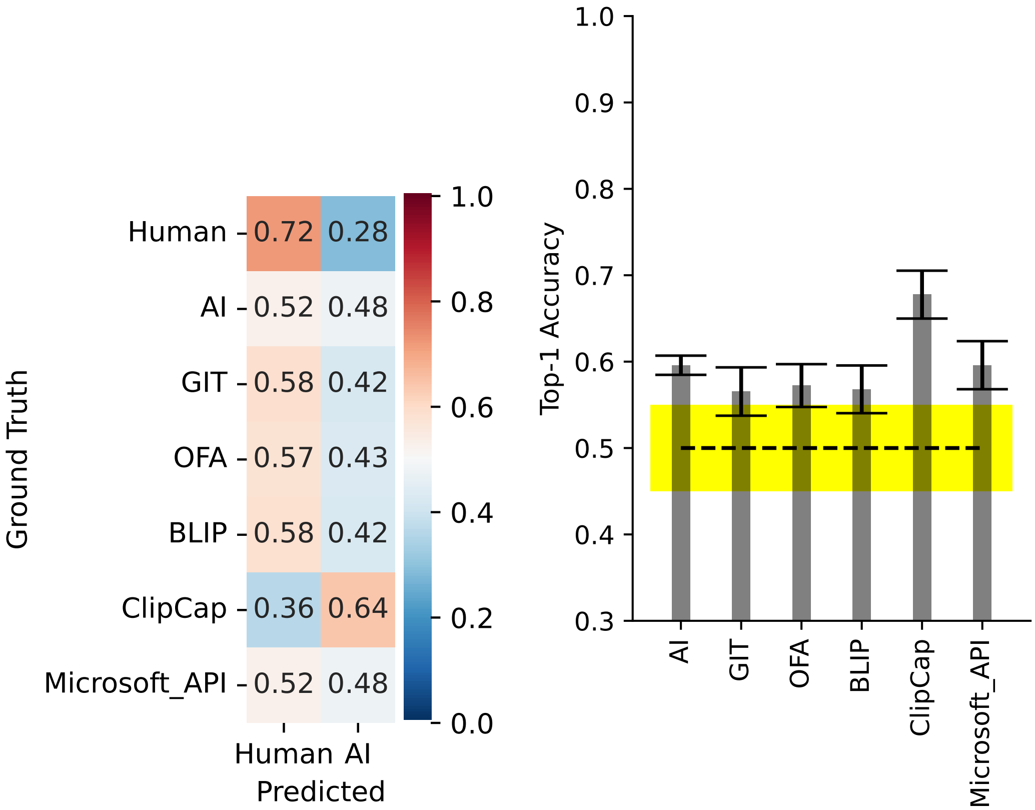


Figure S62. **Image captioning.** Results of the Turing test for human judges with in-domain nocaps images. (a) Confusion matrix (b) Top-1 accuracy.

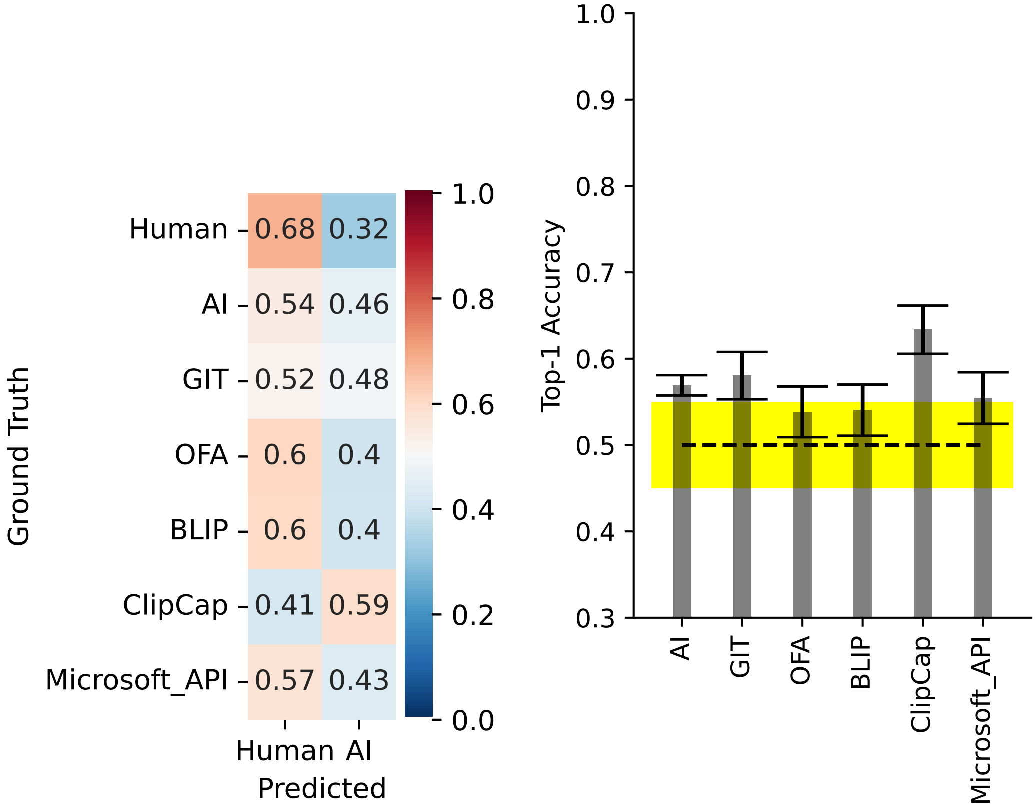


Figure S63. **Image captioning.** Results of the Turing test for human judges with near-domain nocaps images. (a) Confusion matrix (b) Top-1 accuracy.

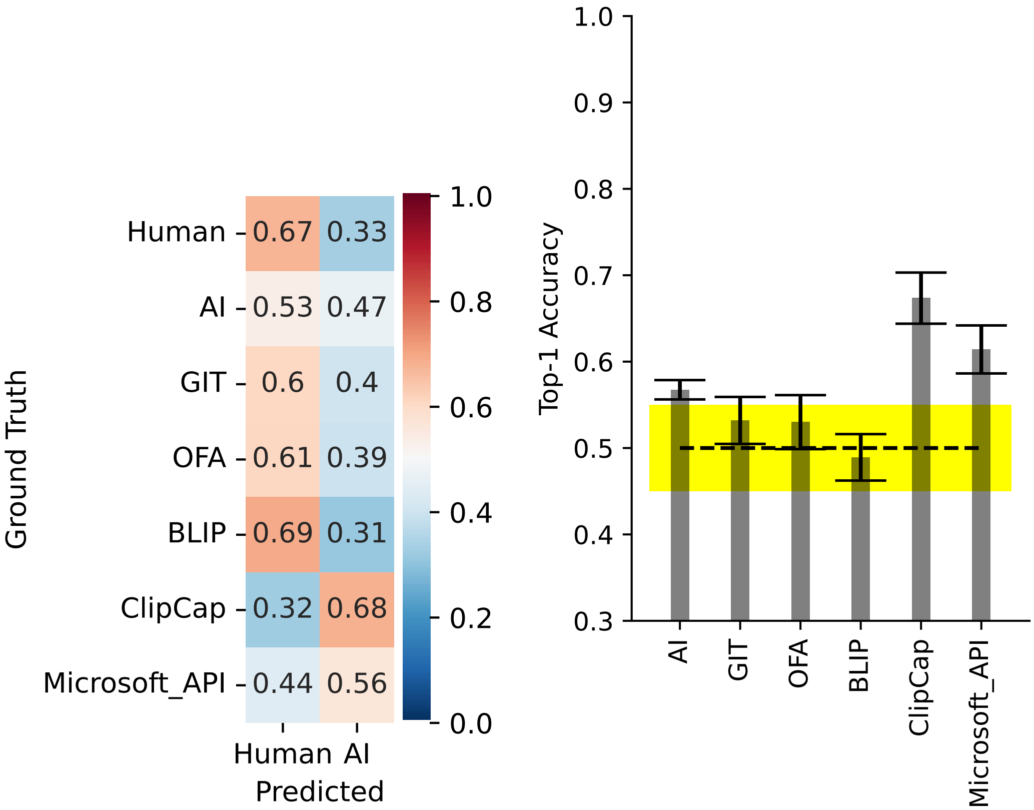
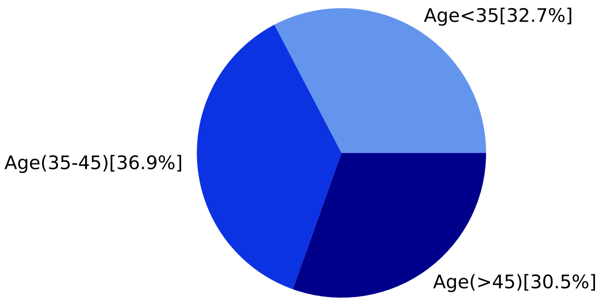
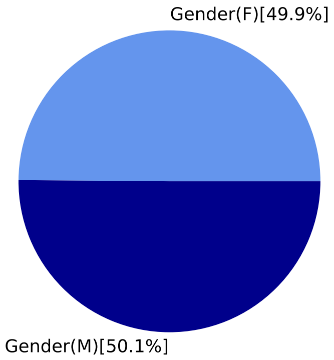


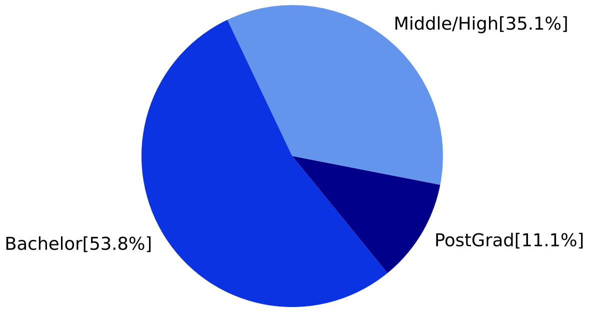
Figure S64. **Image captioning.** Results of the Turing test for human judges with out-of-domain nocaps images. (a) Confusion matrix (b) Top-1 accuracy.



(a) Age distribution



(b) Gender distribution



(c) Education level distribution

Figure S65. **Image captioning.** Demographic information for the human judges. (a) Age. (b) Gender. (c) Education level.

S6. Word Associations

S6.1. Dataset

S6.1.1 Cue words

We chose 150 unique cue words in random (50 nouns, 50 verbs, and 50 adjectives) such that they span a wide range of occurrence frequencies [59] (**Table S4**); Non-English words, stop words (according to Python nltk), and words with < 3 letters were removed from the pool of possible cue words. In addition, all verbs were transformed to their present tense, and all nouns were taken in their singular form.

Associations to each cue word were collected from both human subjects (**Fig. S68**), and from five language models:

(1) Word2vec [50],
(2) GPT2, [52],

GPT3 based models: [8]
(3) GPT3 embedding, davinci;
(4) GPT3 prompt, curie;
(5) GPT3 prompt, davinci.

S6.1.2 Collecting human responses from AMT

For the human associations, we followed two procedures:

1. **Free Associations:** participants were asked to give a one-word answer to the question:
   "What is the first word that comes into your mind when you hear the word [cue word]?".
   This task should resemble the word-embedding-based AI associations.

2. **Prompt-based Associations:** participants were asked to complete a prompt with one word. In particular, given a prompt of 3 cue words and their associations (as was given to the prompt-based AI GPT3 models described below), participants had to give a one-word answer to the request
   "Please complete the prompt below:".

Half of the human associations were collected based on the first procedure ('free associations') while the other half was based on the prompts. We collected responses from 60 participants using AMT. Each participants gave associations to 50 unique cue words taken randomly from the pool of 150 cue words. All participants were English native speakers who are located in the United States.

Fig. S68 and **Fig. S69** show example images of the AMT user interface used to collect human responses for the word association task, in cases of the free associations and the prompt-guided associations, respectively. As can be seen in S68, for the free associations task, the participants were given a "cue word" and were asked to name a word they associate with that cue word. For the guided-prompt associations, as can be seen in S69, the participants were given a prompt of 3 cue-association pairs and an additional cue word without its paired guess word. They were asked to *"complete the prompt below"*, namely, type their guess word for the last cue word, given the prompt.

The responses were collected and post-processed, including singularizing all nouns, transforming verbs to their simple forms, removing typos, removing non-alphabetical characters and spaces, and dropping stop words associations and words with less than 3 letters from the guess words pool. Additionally, guess words that were very similar to the cue word ($> 60\%$ of one word in the pair contains the other word), were disqualified.

S6.1.3 Collecting AI responses

In addition to collecting associations from human subjects, we collected associations from 5 AI language models. The associations of the first three models: 'Word2vec', 'GPT2', and 'GPT3-embedding' were found based on Euclidean proximity to the cue word in the model's word embedding space.

While the associations of the remaining two models: 'GPT3 (prompt curie)' and 'GPT3 (prompt davinci)' were based on prompt completion, as presented in S6. The full set of cue words and their AI associations are presented in S4. In particular, a prompt of 3 sample associations were given in each case and the next association were provided by the model based on these prompts. The prompts displayed to the AI were identical to those presented to the human participants. As for the prompts' creation, we used a held-out set containing human word-pair associations. This held-out set was not used for Turing tests or any results analysis, in order to keep the associations used for the prompts independent and different from those collected for the Turing test analysis and prevent a potential bias in favor of specific associations.

For both the free associations and the prompt-guided associations, the AI response was limited to one word. The same post-processing steps applied to the human responses were applied to the AI responses as well (including singularizing all nouns, transforming verbs to their simple forms, removing typos, removing non-alphabetical characters and spaces, removing guess words that were very similar to the cue word, and dropping stop words associations and words with less than 3 letters from the guess words pool).

S6.2. Turing test

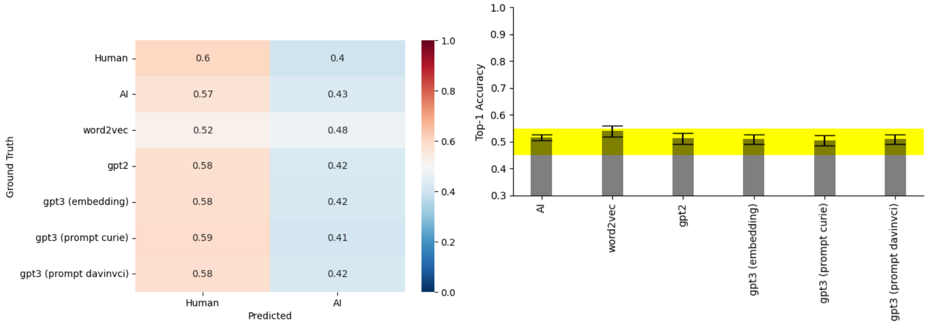
After the word-associations collection, both from human and the 5 AI models, both human judges as well as AI judges performed the Turing phase.

S6.2.1 Human judge Turing test

For the human-judge Turing test, we launched another AMT task, in which the subjects were presented with a cue word and its association and had to decide whether this association has been made by a human or a machine, as well as whether the agent who made this association is a male or a female. In this experiment, responses from 50 native English human subjects were collected, such that each subject had 50 Turing test trials. In each trial, a pair of words (a cue word and a corresponding guess word) was presented on the screen and the worker had to choose, using a radio button, whether the presented association was made by a human or by an AI model. Half of the presented associations to each participant (i.e., 25 associations) were made by AI models and the other half (25) were humans associations. The order of the AI and humans associations was randomly shuffled. Demographic information on each participant was collected as well. An example image of the AMT user interface of the Turing test with human judges is presented in Fig. S70. The confusion matrix and top1-accuracy of the human-judge Turing test are presented in Fig. S66a and S66b, respectively. The human performance was also tested with sensitivity to different demographic groups, as presented in Fig. S71, S72, and S73, for human subjects from different age groups, genders, and education levels, respectively.

S6.2.2 AI judge Turing test

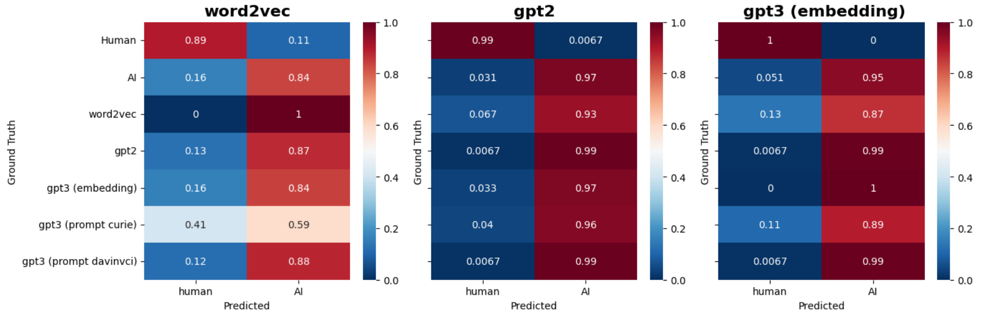
The same set of cue-guess pairs used in the Turing test with the human judges, were used to test AI ability to distinguish between an association made by an AI model and a human. This Turing test with AI judge was performed using 3 independent linear SVM classifiers [15], based on the embedding of (1) Word2vec; (2) GPT2; and (3) GPT3 (davinci). The SVM was trained using 10-fold cross validation, and the data used for the training was the embedding distance between each cue-guess words embedding, while half of the associations pairs were made by humans and the other half by AI. The average model performance on the validation across folds was calculated. The confusion matrices and the top-1 accuracy are presented in S67.



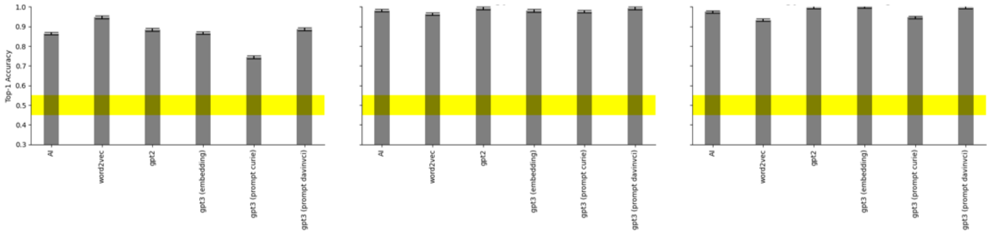
(a) Confusion matrix

(b) Top1 accuracy

Figure S66. **Word Associations.** Results of the Turing test for human judges. (a) Confusion matrix (b) Top-1 accuracy.

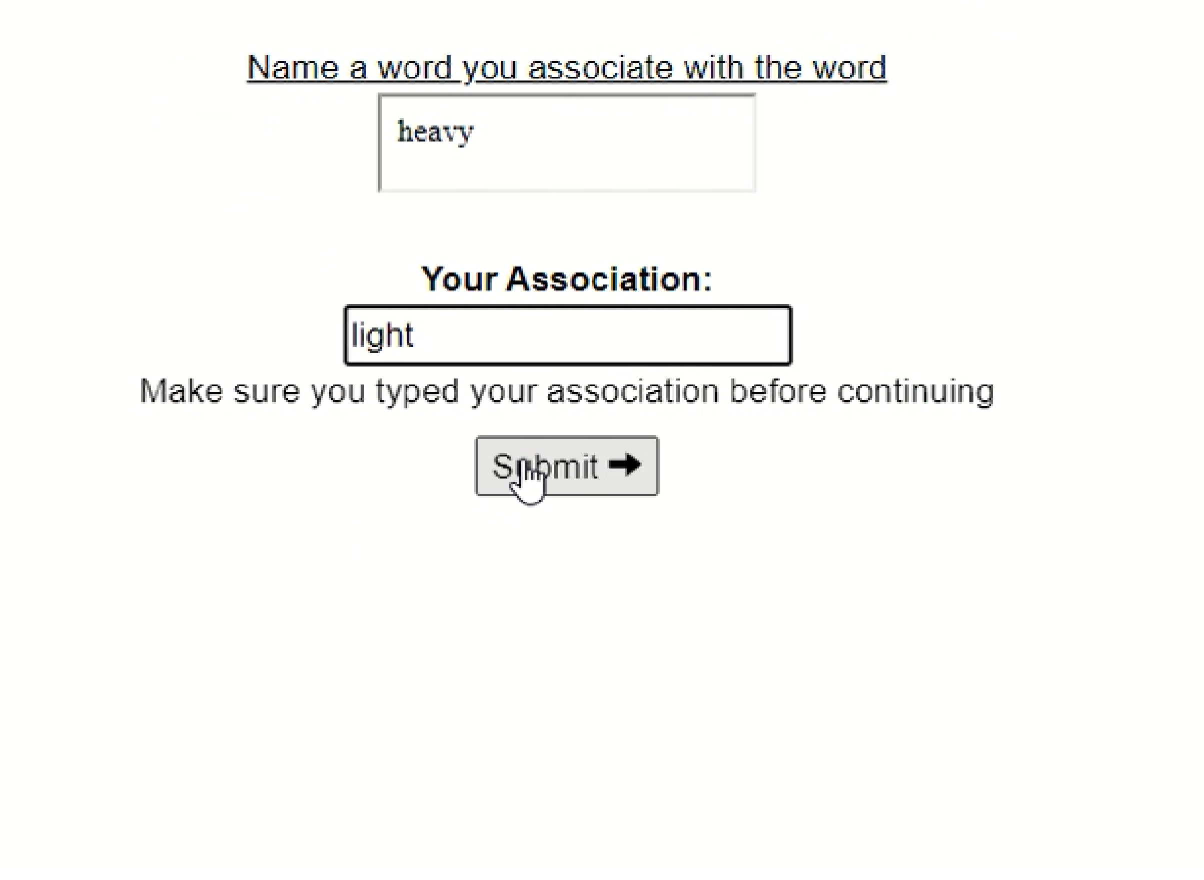


(a) Confusion matrix



(b) Top1 accuracy

Figure S67. **Word Associations.** Results of the Turing test for AI judges.
(a) Confusion matrix (b) Top-1 accuracy.



Name a word you associate with the word

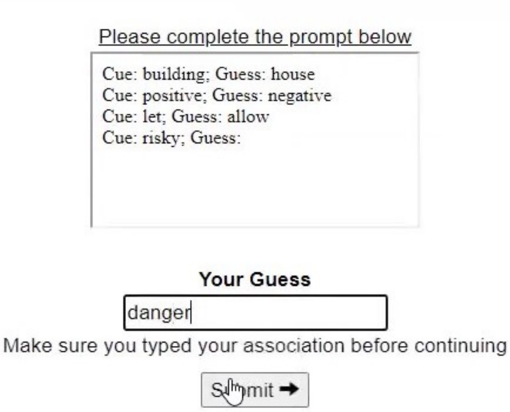
heavy

Your Association:

light

Make sure you typed your association before continuing

Figure S68. **Word Associations.** AMT user interface for collecting word associations responses, for free associations



Please complete the prompt below

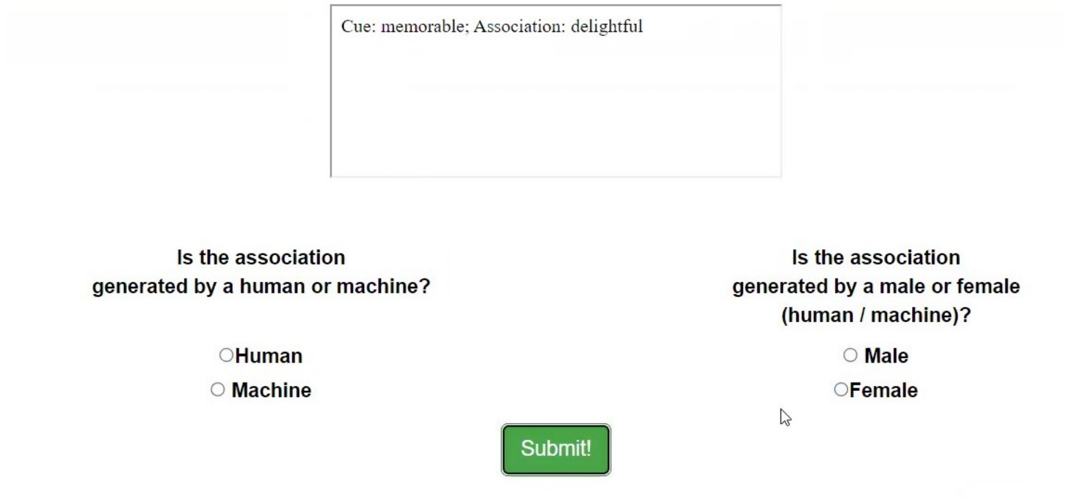
Cue: building; Guess: house
Cue: positive; Guess: negative
Cue: let; Guess: allow
Cue: risky; Guess:

Your Guess

danger

Make sure you typed your association before continuing

Figure S69. **Word Associations.** AMT user interface for collecting responses, for prompt-guided associations

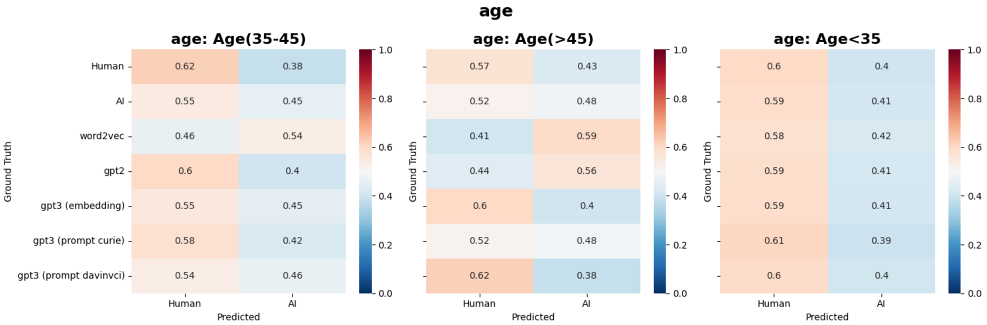


Cue: memorable; Association: delightful

Is the association generated by a human or machine?

Is the association generated by a male or female (human / machine)?

Figure S70. **Word Associations.** AMT user interface for collecting human judge responses (Turing test).



age: Age(35-45)

Ground Truth	Human	AI
Human	0.62	0.38
AI	0.55	0.45
word2vec	0.46	0.54
gpt2	0.6	0.4
gpt3 (embedding)	0.55	0.45
gpt3 (prompt curie)	0.58	0.42
gpt3 (prompt davinvci)	0.54	0.46

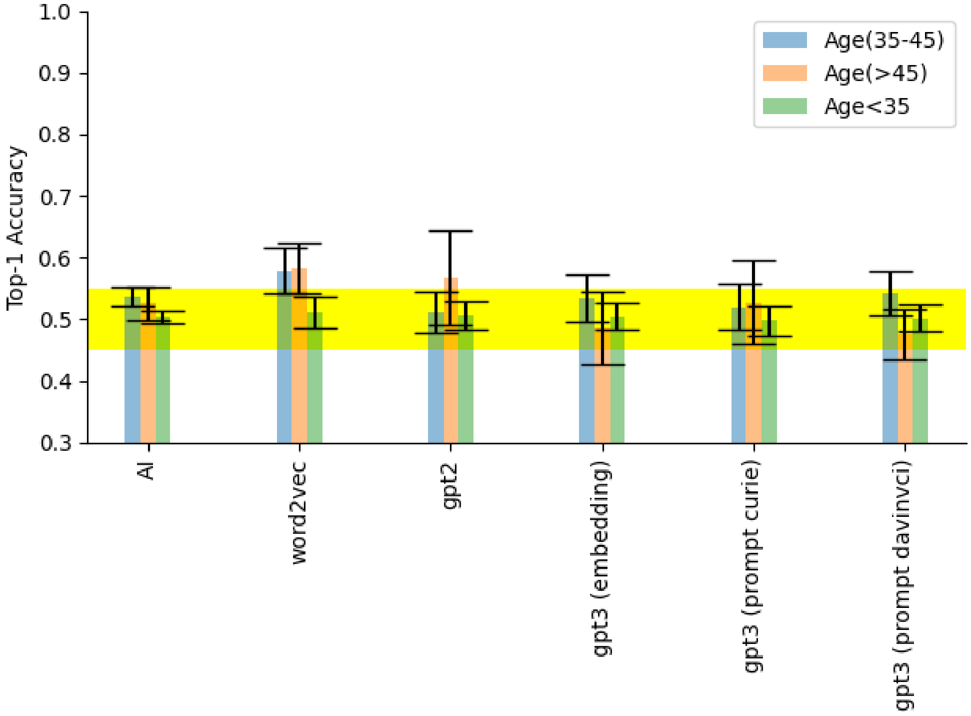
age: Age(>45)

Ground Truth	Human	AI
Human	0.57	0.43
AI	0.52	0.48
word2vec	0.41	0.59
gpt2	0.44	0.56
gpt3 (embedding)	0.6	0.4
gpt3 (prompt curie)	0.52	0.48
gpt3 (prompt davinvci)	0.62	0.38

age: Age<35

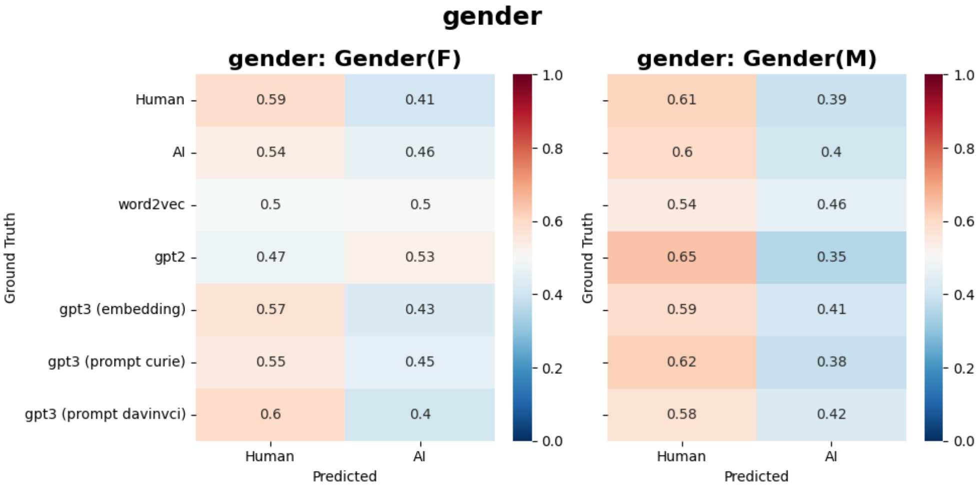
Ground Truth	Human	AI
Human	0.6	0.4
AI	0.59	0.41
word2vec	0.58	0.42
gpt2	0.59	0.41
gpt3 (embedding)	0.59	0.41
gpt3 (prompt curie)	0.61	0.39
gpt3 (prompt davinvci)	0.6	0.4

(a) Confusion matrix

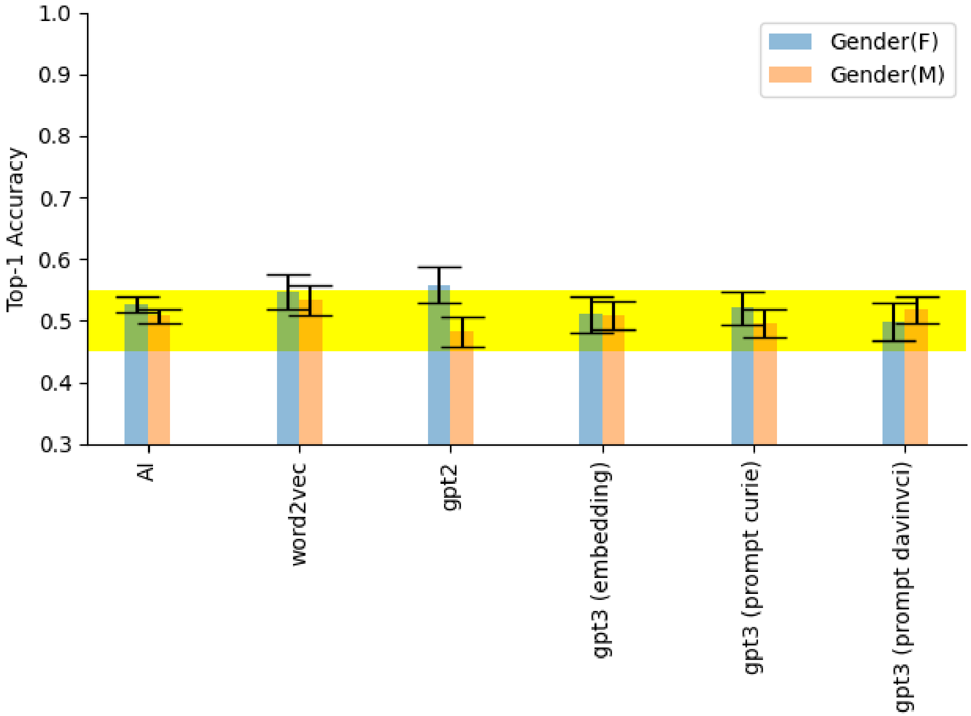


(b) Top1 accuracy

Figure S71. **Word Associations.** Results of the Turing test for human judges in different age groups (a) Confusion matrix (b) Top-1 accuracy.

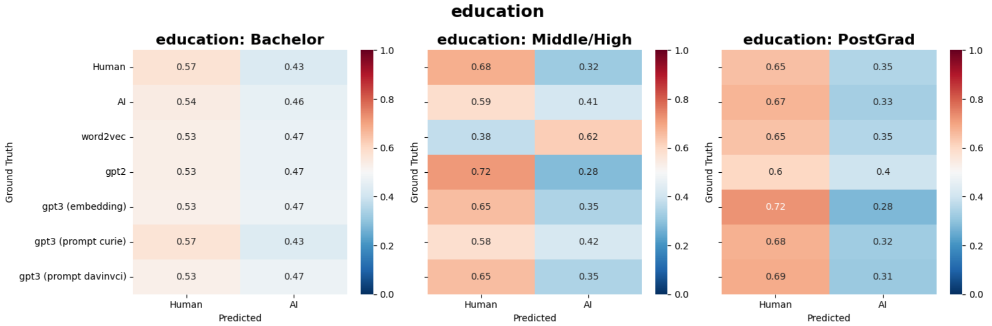


(a) Confusion matrix

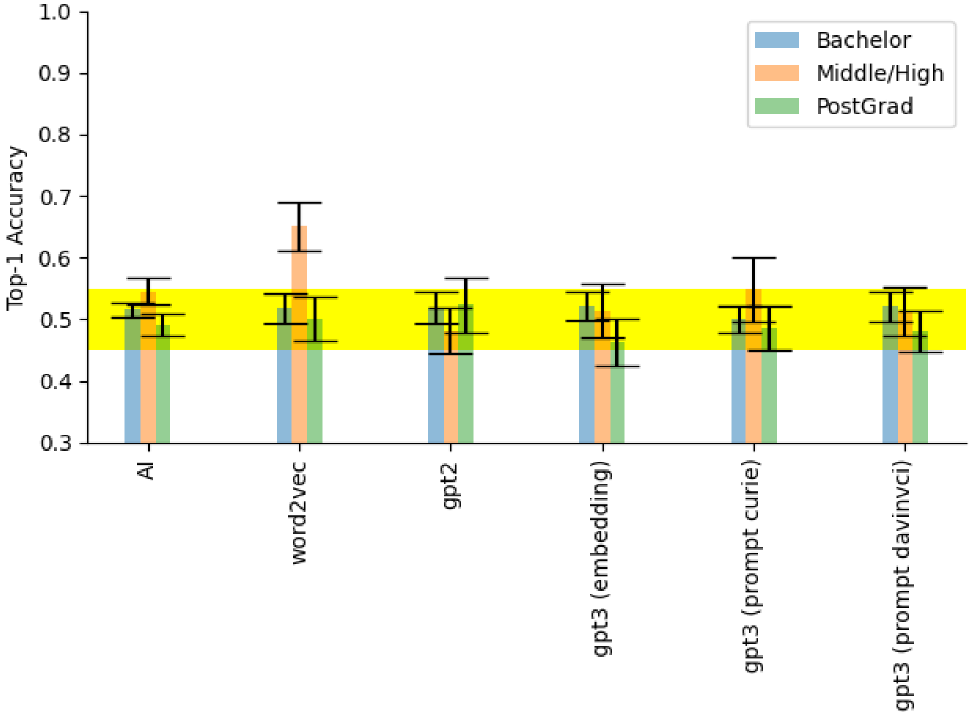


(b) Top1 accuracy

Figure S72. **Word Associations.** Results of the Turing test for different genders. (a) Confusion matrix. (b) Top-1 accuracy.

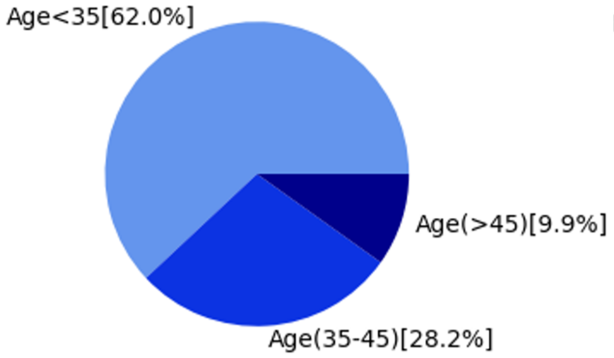


(a) Confusion matrix

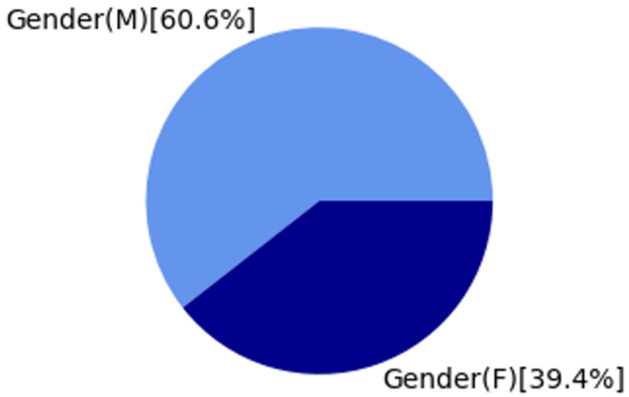


(b) Top1 accuracy

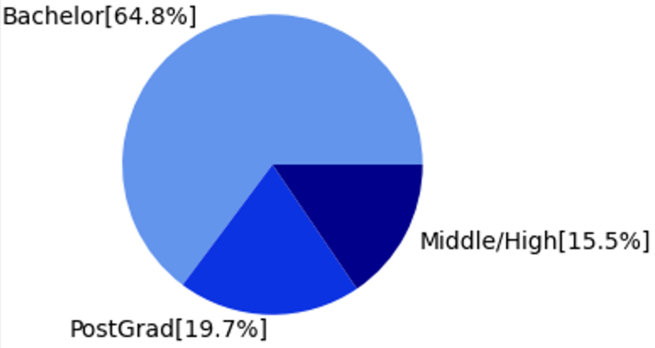
Figure S73. **Word Associations.** Results of the Turing test for human judges for different educational levels. (a) Confusion matrices. (b) Top-1 accuracy.



(a) Age distribution



(b) Gender distribution



(c) Education level distribution

Figure S74. **Word Associations.** Demographic information for the human judges for the word association task (a) Age. (b) Gender. (c) Education level.

Table S4. **Word associations**. Cue words and their associations ('Guess Words') predicted by the five different AI language models. The associations for the first 3 models (Word2vec, GPT2, and GPT3-embedding) were derived from the corresponding embeddings, while the associations of the last two models (GPT3 prompt curie and GPT3 prompt davinci) were based on prompts (the full prompts are described in **Table S6**

.

Cue Words	**Word2vec**	**GPT2**	**GPT3 (embedding, davinci)**	**GPT3 (prompt curie)**	**GPT3 (prompt davinvci)**
memorable	phenomenal	delightful	meaningful	unforgettable	delightful
heavy	edge	massive	weight	light	massive
risky	elaborate	dangerous	danger	dangerous	dangerous
grand	paris	massive	great	palace	massive
powerful	extraordinary	strong	mighty	engine	strong
nepal	vietnam	pakistan	nebraska	new	pakistan
healthy	weight	good	alive	sick	good
credible	legitimate	compelling	legitimate	trustworthy	compelling
rigid	narrow	flexible	stiff	strict	flexible
oriental	imperial	colonial	eastern	love	colonial
flat	tank	pretty	square	round	pretty
technician	engineer	engineer	mechanic	repair	engineer
soft	thin	hard	gentle	whisper	hard
southeastern	mainland	northwestern	oriental	florida	northwestern
numerous	several	several	many	many	several
conceptual	composition	cognitive	idea	idea	cognitive
mystic	hollow	spiritual	magic	magic	spiritual
fetish	amateur	obsession	erotic	sex	obsession
rural	capital	urban	countryside	farm	urban
exclusive	feature	unique	unique	inclusive	unique
soviet	ussr	ussr	ussr	union	ussr
presidential	congressional	republican	congressional	gubernatorial	republican
attentive	straightforward	enthusiastic	aware	listen	enthusiastic
impressive	superb	incredible	incredible	wonderful	incredible
ridiculous	awful	absurd	absurd	funny	absurd
psychotic	obnoxious	paranoid	insane	crazy	paranoid
ordinary	brave	normal	normal	common	normal
obnoxious	psychotic	outrageous	arrogant	unpleasant	outrageous
pascal	claude	jonathan	pierre	triangle	jonathan
rebellious	heroic	arrogant	naughty	normal	arrogant
ingenious	problematic	clever	clever	clever	clever
exponential	systematic	linear	expansion	growing	linear
simplistic	problematic	straightforward	basic	easy	straightforward
puppy	kitten	dog	kitten	dog	dog
venerable	phd	legendary	ancient	old	legendary
sluggish	steep	weak	lazy	tired	weak
enormous	lethal	immense	immense	huge	immense
intangible	singular	invisible	invisible	problem	invisible
mammalian	synaptic	insect	molecular	human	insect
revolutionary	progressive	radical	radical	change	radical
vocal	visual	voice	voice	sing	voice
schematic	synaptic	diagram	diagram	diagram	diagram
suspicious	opposite	strange	paranoid	crime	strange
nationwide	reservation	worldwide	worldwide	local	worldwide
synaptic	mammalian	neural	neural	synapse	neural
white	black	black	black	black	black
able	might	could	kind	cap	could
possible	point	potential	maybe	probable	potential
military	domestic	naval	combat	parade	naval
financial	strategic	economic	money	loan	economic
spark	leap	inspire	spike	fire	inspire
freight	automobile	cargo	cargo	train	cargo
negotiate	eliminate	discuss	diplomacy	bargain	discuss
stanford	penn	harvard	harvard	california	harvard
absent	conclusion	without	lack	baby	without
configuration	installation	setup	setup	network	setup
grasp	reflect	comprehend	seize	hold	comprehend
wreck	sudden	crash	crash	damage	crash
dame	echo	sir	donna	queen	sir
curry	rice	stephen	soup	unfortunately	stephen
revelation	prophecy	discovery	discovery	discovery	discovery
spouse	mutual	wife	wife	wife	wife
mainland	commonwealth	continental	continent	america	continental
material	formula	content	stuff	metal	content
oppose	accuse	defend	contrary	conflict	defend
zombie	horror	vampire	corpse	ghoul	vampire
battlefield	fallout	military	warfare	peace	military
dodge	zoom	avoid	escape	slip	avoid
triangle	nest	circle	trinity	square	circle
cancer	risk	diabetes	diagnosis	cure	diabetes
heroin	cocaine	cocaine	drug	yes	cocaine
velocity	boundary	speed	speed	fast	speed
cash	sell	money	money	check	money
ammunition	deploy	artillery	artillery	bomb	artillery
nixon	chapman	reagan	nelson	presidency	reagan
analysis	strategy	study	examination	report	study

Table S5. **Word associations**. Cue words and their associations - *Cont. from previous page*

Cue Words	**Word2vec**	**GPT2**	**GPT3-embedding**	**GPT3 (prompt curie)**	**GPT3 (prompt davinvci)**
clarity	pursuit	insight	clear	problem	insight
discretion	hypothesis	mercy	privacy	caution	mercy
peanut	cinnamon	chocolate	corn	almond	chocolate
blunt	bounce	frank	brutal	delicate	frank
nursery	maid	kindergarten	nest	baby	kindergarten
awhile	sometime	sometime	sometime	long	sometime
spoon	slice	wooden	scoop	fork	wooden
teddy	bear	johnny	edgar	toy	johnny
purpose	temporary	goal	intention	meaning	goal
distinguish	adapt	identify	identify	remember	identify
renaissance	revival	antique	florence	art	antique
foot	pull	feet	leg	shoe	feet
curtain	towel	door	carpet	theater	door
legion	trinity	army	troop	army	army
maid	betty	servant	girl	man	servant
brighton	bristol	liverpool	bristol	bright	liverpool
constitute	quantify	facilitate	declare	annul	facilitate
royal	western	british	king	family	british
speech	appearance	talk	spoken	spoke	talk
wendy	betty	jennifer	walter	barbara	jennifer
teammate	opponent	colleague	colleague	friend	colleague
entertain	impress	comedian	perform	party	comedian
floyd	johnny	william	raymond	pink	william
prototype	slate	concept	schematic	original	concept
consider	whether	think	think	important	think
expose	refer	provide	exhibit	reveal	provide
impose	dispute	establish	establish	ban	establish
fall	together	spring	drop	autumn	spring
slow	serious	speed	late	sad	speed
appear	often	seem	occur	person	seem
kindle	artwork	amazon	amazon	kindle	amazon
rotten	moist	horrible	ruin	apple	horrible
apply	offer	give	use	job	give
kindergarten	elementary	elementary	nursery	school	elementary
improve	develop	enhance	increase	enhancement	enhance
ping	tab	although	peer	sound	although
develop	enhance	create	create	growth	create
reduce	consume	decrease	decrease	number	decrease
infrared	cluster	thermal	thermal	visible	thermal
hundred	twenty	thousand	thousand	thousand	thousand
kitten	puppy	puppy	puppy	cat	puppy
ding	deed	almost	bell	door	almost
wing	deer	give	fly	airplane	give
assume	ignore	imagine	suppose	doubt	imagine
identify	convert	determine	detect	animal	determine
aforementioned	furthermore	infamous	precedent	business	infamous
sing	jane	perform	vocal	hum	perform
watershed	millennium	historic	rainfall	turning	historic
rave	circus	profound	riot	party	profound
establish	obtain	provide	setup	set	provide
participate	compete	join	contribute	support	join
obtain	contribute	acquire	acquire	receive	acquire
soar	leap	climb	fly	fall	climb
settle	aim	establish	establish	agreement	establish
shred	soar	scrap	scrap	paper	scrap
bloodshed	stigma	massacre	massacre	murder	massacre
succeed	achieve	achieve	accomplish	hard	achieve
sled	turf	snow	snow	toboggan	snow
shed	burn	give	yard	tear	give
quantify	constitute	calculate	measure	measure	calculate
sacred	forbidden	ancient	divine	temple	ancient
fred	samuel	robert	gerald	shirt	robert
enclave	legion	coastal	colony	ghetto	coastal
deepen	initiate	strengthen	depth	voice	strengthen
qualify	compete	eligible	eligible	pass	eligible
kindred	covenant	likewise	familiar	relative	likewise
acquire	retain	obtain	obtain	property	obtain
approve	reject	agree	confirm	positive	agree
hatred	belief	anger	spite	machine	anger
have	could	make	could	take	make
know	think	think	aware	learn	think
make	need	give	put	break	give
take	leave	took	give	away	took
come	far	make	bring	approach	make

Table S6. **Word associations**. The 150 prompts given for the GPT-3 curie and GPT-3 davinci models (for their 'prompt' version), which were also given for the human participants who made prompt-based associations, and were generated using a held-out set containing human word-pair associations. The cue words are presented in the leftmost columns, and the guess words of the curie-prompt model and the davinci-prompt model are presented in the two right columns.

memorable	Cue: easy; Guess: form Cue: business; Guess: money Cue: future; Guess: base Cue: memorable; Guess:	unforgettable	delightful
heavy	Cue: business; Guess: paragraph Cue: professional; Guess: unprofessional Cue: unfortunately; Guess: very Cue: heavy; Guess:	light	massive
risky	Cue: building; Guess: house Cue: positive; Guess: negative Cue: let; Guess: allow Cue: risky; Guess:	dangerous	dangerous
grand	Cue: nearly; Guess: edge Cue: hot; Guess: cold Cue: easy; Guess: form Cue: grand; Guess:	palace	massive
powerful	Cue: business; Guess: association Cue: professional; Guess: professor Cue: let; Guess: allow Cue: powerful; Guess:	engine	strong
nepal	Cue: number; Guess: singular Cue: easy; Guess: hard Cue: pretty; Guess: bench Cue: nepal; Guess:	new	pakistan
healthy	Cue: number; Guess: word Cue: come; Guess: here Cue: hot; Guess: cold Cue: healthy; Guess:	sick	good
credible	Cue: see; Guess: saw Cue: slightly; Guess: shortly Cue: let; Guess: allow Cue: credible; Guess:	trustworthy	compelling
rigid	Cue: see; Guess: here Cue: business; Guess: class Cue: let; Guess: beatles Cue: rigid; Guess:	strict	flexible
oriental	Cue: come; Guess: here Cue: positive; Guess: Happy Cue: professional; Guess: career Cue: oriental; Guess:	love	colonial
flat	Cue: professional; Guess: person Cue: story; Guess: narrative Cue: let; Guess: beatles Cue: flat; Guess:	round	pretty
technician	Cue: building; Guess: apartment Cue: business; Guess: trip Cue: professional; Guess: career Cue: technician; Guess:	repair	engineer
soft	Cue: story; Guess: telling Cue: pretty; Guess: nice Cue: remember; Guess: dream Cue: soft; Guess:	whisper	hard
southeastern	Cue: nearly; Guess: half Cue: future; Guess: mars Cue: come; Guess: here Cue: southeastern; Guess:	florida	northwestern
numerous	Cue: let; Guess: allow Cue: see; Guess: tell Cue: positive; Guess: attitude Cue: numerous; Guess:	many	several
conceptual	Cue: positive; Guess: negative Cue: remember; Guess: forget Cue: professional; Guess: special Cue: conceptual; Guess:	idea	cognitive
mystic	Cue: unfortunately; Guess: unluckily Cue: let; Guess: allow Cue: professional; Guess: work Cue: mystic; Guess:	magic	spiritual
fetish	Cue: unfortunately; Guess: sad Cue: professional; Guess: boss Cue: hot; Guess: water Cue: fetish; Guess:	sex	obsession
rural	Cue: professional; Guess: special Cue: positive; Guess: negative Cue: remember; Guess: together Cue: rural; Guess:	farm	urban
exclusive	Cue: unfortunately; Guess: unluckily; Cue: building; Guess: skyscraper Cue: remember; Guess: forget Cue: exclusive; Guess:	inclusive	unique
soviet	Cue: remember; Guess: recall Cue: easy; Guess: facilitate; Cue: number; Guess: game Cue: soviet; Guess:	union	ussr
presidential	Cue: pretty; Guess: good Cue: business; Guess: play Cue: story; Guess: plot Cue: presidential; Guess:	gubernatorial	republican
attentive	Cue: story; Guess: time Cue: professional; Guess: lawyer Cue: easy; Guess: exam Cue: attentive; Guess:	listen	enthusiastic
impressive	Cue: second; Guess: first Cue: pretty; Guess: nice Cue: hot; Guess: spicy Cue: impressive; Guess:	wonderful	incredible
ridiculous	Cue: see; Guess: look Cue: come; Guess: back Cue: nearly; Guess: there Cue: ridiculous; Guess:	funny	absurd
psychotic	Cue: nearly; Guess: close Cue: see; Guess: here Cue: pretty; Guess: beauty Cue: psychotic; Guess:	crazy	paranoid
ordinary	Cue: slightly; Guess: little Cue: see; Guess: tell Cue: building; Guess: door Cue: ordinary; Guess:	common	normal
obnoxious	Cue: story; Guess: love Cue: see; Guess: you Cue: second; Guess: first Cue: obnoxious; Guess:	unpleasant	outrageous
pascal	Cue: remember; Guess: dream Cue: building; Guess: door Cue: hot; Guess: dog Cue: pascal; Guess:	triangle	jonathan
rebellious	Cue: let; Guess: allow Cue: pretty; Guess: good Cue: positive; Guess: negative Cue: rebellious; Guess:	normal	arrogant
ingenious	Cue: second; Guess: position Cue: pretty; Guess: liar Cue: nearly; Guess: close Cue: ingenious; Guess:	clever	clever
exponential	Cue: come; Guess: back Cue: positive; Guess: negative Cue: second; Guess: minute Cue: exponential; Guess:	growing	linear
simplistic	Cue: remember; Guess: recall Cue: second; Guess: first Cue: unfortunately; Guess: sadly Cue: simplistic; Guess:	easy	straightforward
puppy	Cue: future; Guess: past Cue: second; Guess: first Cue: remember; Guess: recall Cue: puppy; Guess:	dog	dog
venerable	Cue: nearly; Guess: almost Cue: come; Guess: visit Cue: second; Guess: position Cue: venerable; Guess:	old	legendary
sluggish	Cue: business; Guess: card Cue: hot; Guess: cold Cue: see; Guess: look Cue: sluggish; Guess:	tired	weak
enormous	Cue: unfortunately; Guess: sorry Cue: nearly; Guess: done Cue: let; Guess: allow Cue: enormous; Guess:	huge	immense
intangible	Cue: easy; Guess: exam Cue: future; Guess: past Cue: story; Guess: telling Cue: intangible; Guess:	problem	invisible
mammalian	Cue: future; Guess: answer Cue: building; Guess: base Cue: come; Guess: run Cue: mammalian; Guess:	human	insect
revolutionary	Cue: number; Guess: word Cue: nearly; Guess: almost Cue: hot; Guess: shower Cue: revolutionary; Guess:	change	radical
vocal	Cue: remember; Guess: recall Cue: second; Guess: first Cue: nearly; Guess: there Cue: vocal; Guess:	sing	voice
schematic	Cue: come; Guess: home Cue: business; Guess: play Cue: pretty; Guess: beauty Cue: schematic; Guess:	diagram	diagram
suspicious	Cue: second; Guess: minute Cue: easy; Guess: form Cue: future; Guess: base Cue: suspicious; Guess:	crime	strange
nationwide	Cue: unfortunately; Guess: not Cue: building; Guess: base Cue: future; Guess: answer Cue: nationwide; Guess:	local	worldwide
synaptic	Cue: let; Guess: allow Cue: number; Guess: word Cue: hot; Guess: cold Cue: synaptic; Guess:	synapse	neural
white	Cue: remember; Guess: dream Cue: pretty; Guess: nice Cue: see; Guess: here Cue: white; Guess:	black	black
able	Cue: easy; Guess: hard Cue: story; Guess: time Cue: remember; Guess: recall Cue: able; Guess:	cap	could
possible	Cue: building; Guess: base Cue: let; Guess: allow Cue: slightly; Guess: hardly Cue: possible; Guess:	probable	potential
military	Cue: easy; Guess: hard Cue: let; Guess: allow Cue: pretty; Guess: nice Cue: military; Guess:	parade	naval
financial	Cue: come; Guess: home Cue: building; Guess: tower Cue: professional; Guess: professor Cue: financial; Guess:	loan	economic
spark	Cue: see; Guess: you Cue: slightly; Guess: more Cue: remember; Guess: together Cue: spark; Guess:	fire	inspire
freight	Cue: easy; Guess: hard Cue: professional; Guess: professor Cue: hot; Guess: cold Cue: freight; Guess:	train	cargo
negotiate	Cue: positive; Guess: negative Cue: hot; Guess: water Cue: see; Guess: look Cue: negotiate; Guess:	bargain	discuss
stanford	Cue: second; Guess: first Cue: nearly; Guess: close Cue: come; Guess: back Cue: stanford; Guess:	california	harvard
absent	Cue: professional; Guess: professor Cue: slightly; Guess: slow Cue: remember; Guess: recall Cue: absent; Guess:	baby	without
configuration	Cue: come; Guess: back Cue: hot; Guess: food Cue: building; Guess: body Cue: configuration; Guess:	network	setup
grasp	Cue: number; Guess: game Cue: professional; Guess: doctor Cue: unfortunately; Guess: sorry Cue: grasp; Guess:	hold	comprehend
wreck	Cue: nearly; Guess: close Cue: remember; Guess: forget Cue: positive; Guess: negative Cue: wreck; Guess:	damage	crash
dame	Cue: pretty; Guess: liar Cue: come; Guess: home Cue: slightly; Guess: off Cue: dame; Guess:	queen	sir
curry	Cue: hot; Guess: cold Cue: see; Guess: look Cue: positive; Guess: negative Cue: curry; Guess:	unfortunately	stephen
revelation	Cue: come; Guess: run Cue: remember; Guess: recall Cue: slightly; Guess: off Cue: revelation; Guess:	discovery	discovery
spouse	Cue: business; Guess: play Cue: building; Guess: tower Cue: unfortunately; Guess: sadly Cue: spouse; Guess:	wife	wife
mainland	Cue: hot; Guess: cold Cue: story; Guess: love Cue: unfortunately; Guess: sadly Cue: mainland; Guess:	america	continental
material	Cue: future; Guess: far Cue: remember; Guess: together Cue: positive; Guess: attitude Cue: material; Guess:	metal	content
oppose	Cue: pretty; Guess: beautiful Cue: slightly; Guess: better Cue: nearly; Guess: close Cue: oppose; Guess:	conflict	defend
zombie	Cue: positive; Guess: negative Cue: professional; Guess: boss Cue: nearly; Guess: almost Cue: zombie; Guess:	ghoul	vampire
battlefield	Cue: easy; Guess: exam Cue: see; Guess: hear Cue: positive; Guess: negative Cue: battlefield; Guess:	peace	military
dodge	Cue: easy; Guess: cool Cue: slightly; Guess: bigger Cue: positive; Guess: negative Cue: dodge; Guess:	slip	avoid
triangle	Cue: hot; Guess: cold Cue: easy; Guess: form Cue: unfortunately; Guess: not Cue: triangle; Guess:	square	circle
cancer	Cue: hot; Guess: cold Cue: professional; Guess: doctor Cue: unfortunately; Guess: sad Cue: cancer; Guess:	cure	diabetes
heroin	Cue: building; Guess: construction Cue: slightly; Guess: cold Cue: professional; Guess: work Cue: heroin; Guess:	yes	cocaine
velocity	Cue: come; Guess: home Cue: positive; Guess: valence Cue: number; Guess: singular Cue: velocity; Guess:	fast	speed
cash	Cue: business; Guess: eel Cue: professional; Guess: person Cue: nearly; Guess: close Cue: cash; Guess:	check	money
ammunition	Cue: hot; Guess: cold Cue: come; Guess: here Cue: nearly; Guess: almost Cue: ammunition; Guess:	bomb	artillery
nixon	Cue: business; Guess: paragraph Cue: easy; Guess: form Cue: come; Guess: back Cue: nixon; Guess:	presidency	reagan
analysis	Cue: slightly; Guess: more Cue: see; Guess: watch Cue: positive; Guess: negative Cue: analysis; Guess:	report	study

Cue Word:	**Prompt:**	**Guess (prompt curie)**	**Guess (prompt davinci)**
clarity	Cue: unfortunately; Guess: too Cue: professional; Guess: professor Cue: business; Guess: association Cue: clarity; Guess:	problem	insight
discretion	Cue: unfortunately; Guess: too Cue: remember; Guess: forget Cue: easy; Guess: facilitate; Cue: discretion; Guess:	caution	mercy
peanut	Cue: easy; Guess: hard Cue: remember; Guess: memorize Cue: hot; Guess: cold Cue: peanut; Guess:	almond	chocolate
blunt	Cue: come; Guess: back Cue: building; Guess: apartment Cue: second; Guess: third Cue: blunt; Guess:	delicate	frank
nursery	Cue: positive; Guess: stay Cue: nearly; Guess: almost Cue: future; Guess: past Cue: nursery; Guess:	baby	kindergarten
awhile	Cue: business; Guess: play Cue: number; Guess: singular Cue: hot; Guess: spicy Cue: awhile; Guess:	long	sometime
spoon	Cue: number; Guess: singular Cue: see; Guess: hear Cue: pretty; Guess: bench Cue: spoon; Guess:	fork	wooden
teddy	Cue: slightly; Guess: slow Cue: building; Guess: skyscraper Cue: easy; Guess: hard Cue: teddy; Guess:	toy	johnny
purpose	Cue: business; Guess: money Cue: story; Guess: telling Cue: let; Guess: beatles Cue: purpose; Guess:	meaning	goal
distinguish	Cue: building; Guess: construction Cue: let; Guess: allow Cue: hot; Guess: shower Cue: distinguish; Guess:	remember	identify
renaissance	Cue: story; Guess: narrative Cue: positive; Guess: negative Cue: see; Guess: watch Cue: renaissance; Guess:	art	antique
foot	Cue: story; Guess: book Cue: easy; Guess: cool Cue: positive; Guess: negative Cue: foot; Guess:	shoe	feet
curtain	Cue: future; Guess: base Cue: business; Guess: eel Cue: story; Guess: plot Cue: curtain; Guess:	theater	door
legion	Cue: building; Guess: base Cue: remember; Guess: recall Cue: slightly; Guess: problematic Cue: legion; Guess:	army	army
maid	Cue: come; Guess: run Cue: business; Guess: money Cue: building; Guess: apartment Cue: maid; Guess:	man	servant
brighton	Cue: slightly; Guess: a bit Cue: professional; Guess: lawyer Cue: business; Guess: association Cue: brighton; Guess:	bright	liverpool
constitute	Cue: number; Guess: letter Cue: see; Guess: here Cue: future; Guess: response Cue: constitute; Guess:	annul	facilitate
royal	Cue: pretty; Guess: hot Cue: story; Guess: novel Cue: come; Guess: here Cue: royal; Guess:	family	british
speech	Cue: positive; Guess: negative Cue: pretty; Guess: poem Cue: slightly; Guess: more Cue: speech; Guess:	spoke	talk
wendy	Cue: business; Guess: manager Cue: let; Guess: beatles Cue: future; Guess: mars Cue: wendy; Guess:	barbara	jennifer
teammate	Cue: story; Guess: love Cue: second; Guess: guess Cue: professional; Guess: special Cue: teammate; Guess:	friend	colleague
entertain	Cue: business; Guess: play Cue: unfortunately; Guess: very Cue: second; Guess: position Cue: entertain; Guess:	party	comedian
floyd	Cue: future; Guess: past Cue: unfortunately; Guess: not Cue: second; Guess: third Cue: floyd; Guess:	pink	william
prototype	Cue: nearly; Guess: close Cue: unfortunately; Guess: very Cue: second; Guess: two Cue: prototype; Guess:	original	concept
consider	Cue: building; Guess: apartment Cue: nearly; Guess: almost Cue: let; Guess: allow Cue: consider; Guess:	important	think
expose	Cue: professional; Guess: career Cue: see; Guess: here Cue: easy; Guess: hard Cue: expose; Guess:	reveal	provide
impose	Cue: easy; Guess: facilitate; Cue: number; Guess: two Cue: slightly; Guess: cold Cue: impose; Guess:	ban	establish
fall	Cue: second; Guess: first Cue: come; Guess: visit Cue: see; Guess: movie Cue: fall; Guess:	autumn	spring
slow	Cue: unfortunately; Guess: unluckily; Cue: hot; Guess: cold Cue: come; Guess: home Cue: slow; Guess:	sad	speed
appear	Cue: hot; Guess: shower Cue: pretty; Guess: beautiful Cue: nearly; Guess: close Cue: appear; Guess:	person	seem
kindle	Cue: slightly; Guess: better Cue: let; Guess: beatles Cue: easy; Guess: hard Cue: kindle; Guess:	kindle	amazon
rotten	Cue: future; Guess: far Cue: come; Guess: visit Cue: professional; Guess: wrestler Cue: rotten; Guess:	apple	horrible
apply	Cue: remember; Guess: recall Cue: second; Guess: two Cue: let; Guess: allow Cue: apply; Guess:	job	give
kindergarten	Cue: see; Guess: saw Cue: future; Guess: past Cue: let; Guess: beatles Cue: kindergarten; Guess:	school	elementary
improve	Cue: future; Guess: past Cue: positive; Guess: valence Cue: story; Guess: novel Cue: improve; Guess:	enhancement	enhance
ping	Cue: professional; Guess: doctor Cue: see; Guess: look Cue: number; Guess: word Cue: ping; Guess:	sound	although
develop	Cue: see; Guess: look Cue: let; Guess: allow Cue: business; Guess: manager Cue: develop; Guess:	growth	create
reduce	Cue: remember; Guess: together Cue: easy; Guess: hard Cue: future; Guess: unknown; Cue: reduce; Guess:	number	decrease
infrared	Cue: let; Guess: allow Cue: unfortunately; Guess: sadly Cue: easy; Guess: hard Cue: infrared; Guess:	visible	thermal
hundred	Cue: unfortunately; Guess: sadly Cue: let; Guess: beatles Cue: story; Guess: telling Cue: hundred; Guess:	thousand	thousand
kitten	Cue: let; Guess: allow Cue: building; Guess: apartment Cue: positive; Guess: statement Cue: kitten; Guess:	cat	puppy
ding	Cue: business; Guess: class Cue: future; Guess: past Cue: pretty; Guess: good Cue: ding; Guess:	door	almost
wing	Cue: professional; Guess: lawyer Cue: business; Guess: trip Cue: positive; Guess: valence Cue: wing; Guess:	airplane	give
assume	Cue: see; Guess: watch Cue: number; Guess: singular Cue: professional; Guess: work Cue: assume; Guess:	doubt	imagine
identify	Cue: hot; Guess: warm Cue: business; Guess: card Cue: building; Guess: base Cue: identify; Guess:	animal	determine
aforementioned	Cue: number; Guess: arabic Cue: building; Guess: tower Cue: story; Guess: novel Cue: aforementioned; Guess:	business	infamous
sing	Cue: let; Guess: beatles Cue: unfortunately; Guess: too Cue: building; Guess: tower Cue: sing; Guess:	hum	perform
watershed	Cue: positive; Guess: stay Cue: nearly; Guess: close Cue: see; Guess: watch Cue: watershed; Guess:	turning	historic
rave	Cue: future; Guess: past Cue: come; Guess: here Cue: remember; Guess: dream Cue: rave; Guess:	party	profound
establish	Cue: future; Guess: answer Cue: positive; Guess: stay Cue: pretty; Guess: good Cue: establish; Guess:	set	provide
participate	Cue: nearly; Guess: half Cue: story; Guess: plot Cue: future; Guess: answer Cue: participate; Guess:	support	join
obtain	Cue: pretty; Guess: hot Cue: number; Guess: letter Cue: second; Guess: first Cue: obtain; Guess:	receive	acquire
soar	Cue: professional; Guess: boss Cue: future; Guess: mars Cue: see; Guess: here Cue: soar; Guess:	fall	climb
settle	Cue: hot; Guess: dog Cue: professional; Guess: lawyer Cue: positive; Guess: negative Cue: settle; Guess:	agreement	establish
shred	Cue: positive; Guess: negative Cue: professional; Guess: professor Cue: come; Guess: back Cue: shred; Guess:	paper	scrap
bloodshed	Cue: business; Guess: play Cue: positive; Guess: attitude Cue: come; Guess: home Cue: bloodshed; Guess:	murder	massacre
succeed	Cue: unfortunately; Guess: not Cue: building; Guess: house Cue: remember; Guess: forget Cue: succeed; Guess:	hard	achieve
sled	Cue: nearly; Guess: close Cue: second; Guess: first Cue: building; Guess: pretty Cue: sled; Guess:	toboggan	snow
shed	Cue: pretty; Guess: bench Cue: easy; Guess: cool Cue: come; Guess: run Cue: shed; Guess:	tear	give
quantify	Cue: easy; Guess: hard Cue: hot; Guess: cold Cue: positive; Guess: negative Cue: quantify; Guess:	measure	calculate
sacred	Cue: nearly; Guess: edge Cue: professional; Guess: lawyer Cue: future; Guess: past Cue: sacred; Guess:	temple	ancient
fred	Cue: positive; Guess: negative Cue: hot; Guess: spicy Cue: see; Guess: look Cue: fred; Guess:	shirt	robert
enclave	Cue: business; Guess: play Cue: come; Guess: here Cue: unfortunately; Guess: too Cue: enclave; Guess:	ghetto	coastal
deepen	Cue: hot; Guess: water Cue: future; Guess: unknown; Cue: pretty; Guess: poem Cue: deepen; Guess:	voice	strengthen
qualify	Cue: positive; Guess: negative Cue: easy; Guess: hard Cue: nearly; Guess: half Cue: qualify; Guess:	pass	eligible
kindred	Cue: see; Guess: watch Cue: positive; Guess: negative Cue: building; Guess: construction Cue: kindred; Guess:	relative	likewise
acquire	Cue: see; Guess: here Cue: remember; Guess: together Cue: let; Guess: allow Cue: acquire; Guess:	property	obtain
approve	Cue: nearly; Guess: almost Cue: positive; Guess: negative Cue: future; Guess: past Cue: approve; Guess:	positive	agree
hatred	Cue: future; Guess: answer Cue: positive; Guess: negative Cue: see; Guess: here Cue: hatred; Guess:	machine	anger
have	Cue: easy; Guess: form Cue: business; Guess: money Cue: future; Guess: base Cue: have; Guess:	take	make
know	Cue: remember; Guess: recall Cue: slightly; Guess: slow Cue: let; Guess: allow Cue: know; Guess:	learn	think
make	Cue: remember; Guess: dream Cue: professional; Guess: person Cue: story; Guess: book Cue: make; Guess:	break	give
take	Cue: future; Guess: response Cue: story; Guess: novel Cue: remember; Guess: forget Cue: take; Guess:	away	took
come	Cue: second; Guess: guess Cue: building; Guess: machine Cue: slightly; Guess: better Cue: come; Guess:	approach	make

S7. Conversation

S7.1. Dataset

S7.1.1 Human participants

We collected a dataset containing 300 conversations between two speakers (speaker A and speaker B). Each speaker could be either a human or an AI agent. Thus, there were three classes of conversations: human-human, human-AI, and AI-AI. For the human-human and the human-AI conversations, we recruited approximately 150 volunteers who are fluent English speakers to have a conversation over a chatting platform. We acted as intermediary in the conversations so that the volunteers did not know whether they were speaking with another human or with an AI chatbot.

To evaluate potential correlations between the conversation features and the volunteers' demographics, we also collected information on the age, gender, and education levels for each volunteer. For the human-human conversations, we combined 18 conversations collected as described above and we also added 40 conversations from the Topical-Chat dataset [24], selected based on a minimum length of 24 entries.

S7.1.2 Dataset collection: Instructions to human participants

One of the investigators acted as an intermediary to pass messages between two speakers. The two speakers could be two humans, a machine and a human, or two machines. Here we focus on the conversations involving humans. The participants were presented with the following instructions before the conversation:
Hey! Would you have a few minutes to help me collect a dataset? We just need to have one or two conversations on slack/whatsapp for a few minutes (24 messages in total per conversation). Here are the instructions:

- *You will have to ask or answer a question to start and trigger the conversation (I will specify case by case).*
- *Please try to get the conversation going for 24 sentences in total (12 from you, 12 from the other speaker).*
- *Please write each reply in a single message (do not write a second message until you receive a reply).*
- *Just chat as if you are texting either with a friend or someone you don't know.*
- *Please try to reply quickly so that the entire conversation does not take more than 8-10 minutes.*
- *Note that I am just an intermediary of the conversation, you are not talking with me directly.*
- *If you feel that the other speaker is touching a sensitive topic, please write that you are not comfortable and we will restart the conversation.*

Thanks in advance!

After the completion of the conversation, the participants were asked to answer to the following question:
That's all! We got the 24 messages! Thank you so much! Do you think you spoke with male AI / female AI / man / woman and which age (age and gender also for AI chatbot)?

S7.1.3 AI conversation bots

For the human-AI and the AI-AI conversations, we used three state-of-the-art language models: Blenderbot3 (175B model) [57], GPT3 text-davinci-002 [48], and GPT3 text-curie-001 [48]. For all conversations with Blenderbot, we used the live interface provided at `https://blenderbot.ai/`. For the human-GPT3 conversations we used the playground available at `https://beta.openai.com/playground/`. We list the settings in Supplementary Section S7.1.4. For the GPT3-GPT3 conversations, we implemented a custom python framework for the interaction of two models. For the Blenderbot and Blenderbot conversations, we kept all the collected conversations in the dataset. Instead, the GPT3-GPT3 conversations were affected by long-standing issues of NLP, namely repetition of single sentences or multiple consecutive exchanges and early exit. When we detected such issues (see examples of discarded conversations in Section S7.4), we re-sampled the conversations. Therefore, we built a chatbot out of GPT3 based on prompt engineering and failure criteria. Section S7.4 reports some example of "successful" conversations for both GPT3textdavinci002-GPT3textdavinci002 and GPT3textcurie001-GPT3textcurie001. We did not re-sample conversations in the case of human-GPT3 conversations. We also attempted to use the DialoGPT model [75], however, the quality of the conversation was not satisfactory (see examples in Supplementary Section S7.4), hence we did not include DialoGPT in the analysis. We did not perform any pre-processing in the conversations (*e.g.,* we did not correct any misspellings, grammatical errors, logical errors or other inconsistencies).

S7.1.4 Dataset collection: Prompt and settings for GPT3text-davinci002 and GPT3text-curie-001

The pipeline to collect conversations involving GPT3text-davinci002 or GPT3text-curie-001 is described below.

- If GPT3 text-davinci-002 or text-curie-001 model open the link `https://beta.openai.com/playground/p/default-chat?model=text-davinci-002`
- select the model text-davinci-002 (for davinci) or text-curie-001 (for curie) on the top right
- change temperature to 0.8
- change maximum length to 60
- change stop sequences to two random names (e.g. John: and Alice:) - change the names every time
- change Top P to 1
- change frequency penalty to 2
- change presence penalty to 2
- remove the Inject start text and Inject restart text
- give the following prompt to the chatbot:
  "friend1+" greets "+friend2+". "+friend2+" starts to talk about "+topic+". Both ask long questions, give long responses and often disagree. Then the topic changes. The conversation never ends. "+friend1+": Hi! "+friend2+":"
  Choose the same names for friend 1 and friend 2 that you chose for the stop sequences.
  Pick a random topic from the list: *['fashion', 'politics', 'books', 'sports', 'general entertainment', 'music', 'science', 'technology', 'movies']*
  Example: *John greets Alice. Alice starts to talk about movies. Both ask long questions, give long responses and often disagree. Then the topic changes. The conversation never ends. John: Hi! Alice:*
- You decide if the person is John or if the GPT3 is John. The other will be Alice. If the person is John, then you let the model generate the text. This means that the model has generated the turn for Alice. After you send the generated sentence to the person, the person replies, and you copy-paste the reply of the person to the model as: *"John: - - here reply - - . Alice:"* Then you press submit and the model generates a new reply for Alice, and so on until you reach 24 turns.
  Otherwise, if the person is Alice, then you ask the person to start the conversation with a question, and you copy-paste it after "Alice:" in the prompt above. Then you write "John:" and press submit, so that the model generates the reply for John, and so on until you collect 24 exchanges (12 for John and 12 for Alice).

S7.2. Turing test

S7.2.1 Collecting human judge resposnes for Turing test

Fig. S89 shows a screenshot of the conversation task performed by the workers on AMT, containing a control question, questions on the human vs machine nature of the speakers, and questions on the gender of the speakers.

To investigate the results of the Turing test as a function of the length of the conversation, we chunked each conversation into 8 different lengths, including the initial 3, 6, 9, 12, 15, 18, 21, and 24 turns. In total we collected 40 conversations for each category. To perform the Turing test on the collected conversations, we recruited 200 participants from Amazon Mechanical Turk. To compare the quality of responses collected online, we also recruited 8 participants in the lab. The participants were presented with 20 randomly sampled chunked conversations with different lengths. For each of the two speakers, the observers had to state whether they thought the speaker was a human or a machine and the gender (male or female). To verify that the observers were carefully reading the conversation, we introduced a control. For each conversation the observer had to select the general topic of the chat from a list of five topics. In our analysis, we only considered the responses of the workers that correctly classified at least 15 topics out of 20. For those workers, we only considered the responses to the conversations that were correctly classified. On average, we obtained 14 responses per conversation (approximately 2 responses per length). Some examples of conversations are reported in Supplementary Section S7.4.

S7.2.2 AI judge

We evaluated whether simple ML models are able to discern whether a sentence was generated by a model or by a person. In this analysis, we only looked at single sentences and not at the conversation level, therefore the models are only allowed to exploit features such as sentence length, vocabulary, grammar, syntax and typos, and cannot take into account issues such as sentence repetition or lack of logic in reasoning. To this goal, we build 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 we fed the tokenized sentence to a SVM linear classifier trained to perform binary classification to the classes *human* and *AI*. We split the corpora into train and test splits (90%, 10%) and used 10-fold cross-validation for training. In both the train and test splits, we used the same number of sentences for human speakers and for AI agents. For the AI, the sentences were split equally among the three models.

S7.3. Results and discussion

S7.3.1 Confusion matrix and Top-1 accuracy

The results for the human judges are reported in **Figs. S75a** and **S75b**, which contain respectively the confusion matrix and the Top-1 accuracy. The plots show both the overall accuracy for humans and for AI averaged over all the cases, as well as the results for the three tested models (Blenderbot, GPT3text-davinci002, and GPT3text-curie001). We distinguish the three conditions: humans talking with humans, humans talking with AI, and AI talking with AI.

The results for the SVM judge are reported in **Figs. S76a** and **S76b**, which contain respectively the confusion matrix and the Top-1 accuracy. The plots show both the overall accuracy for humans and for AI averaged over all the cases, as well as the results for the three tested models (Blenderbot, GPT3text-davinci002, and GPT3text-curie001). We do not distinguish the three conditions (humans talking with humans, humans talking with AI, and AI talking with AI), as the SVM judge is trained and tested on the single sentences and not on the full conversations.

S7.3.2 Gender perception

The human judges were asked to say whether the speakers of the conversations were male or female, both in case of classification as human and as machine. Overall, 60% of the speakers (human plus machines) were classified as male, indicating a bias of the human judges in the gender perception. In particular, we found that 64% of the human speakers and 58% of the AI agents are classified as male. Furthermore, when a human judge perceived a speaker as human, he associated the male gender 58% of the times. Interestingly, when a speaker was classified as machine, 69% of the times it was also classified as male. This indicates a strong bias in associating the male gender with chatbots. Table S8 summarizes these findings.

S7.3.3 Effects of judge demographics

The results in the form of confusion matrix and Top-1 accuracy for the human judges are reported in:

- **Figs. S77b**, **S78**, and **S79** for different age ranges;
- **Figs. S80** and **S81** for female and male gender respectively;
- **Figs. S82**, **S83**, and **S84** for different education levels.

The data on the demographic distribution of the human judges are reported in the pie bins in **Fig. S85**.

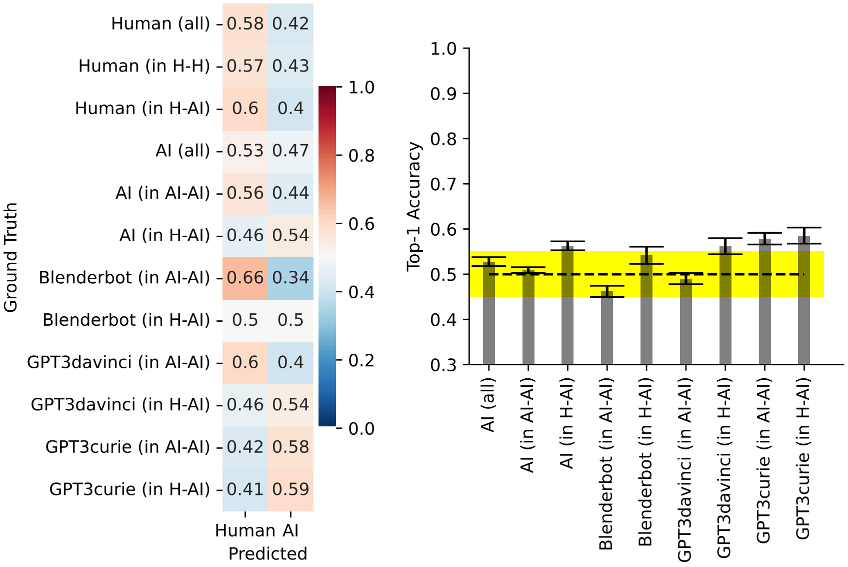
S7.3.4 Comparison between AMT and in-person experiments

The results for the human judges comparing AMT judges and judges recruited to do the experiment in person in the lab under our supervision are reported in **Figs. S86** and **S87**, respectively.

S7.3.5 Effect of conversation length

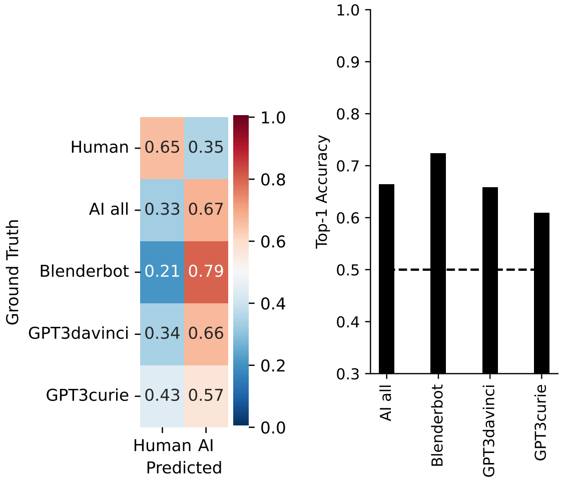
We investigated whether and to which extend the result of the Turing test depends on the length of the conversation. Figure S88 reports the results. We make the following observations:

- except for very short extracts (three exchanges), humans are classified as humans on average more than machines (panel a);
- humans are more likely to be classified as humans for longer conversations (panel a)
- machines are less likely to be classified as humans for longer conversations (panel a)
- in human-human and AI-AI conversations, humans and machines have similar frequency of being classified as human, except for very long conversations (panel b);
- in human-AI conversations, humans are classified more as human than machines, with a gap increasing with the conversation length (panel c);
- there is no clear trend distinguishing human classification between human interaction with humans or with various models (panel d);
- for all models, machines talking with humans are classified more as AI for longer conversations; GPT3 curie is on average classified more as machine than GPT3 davinci and Blenderbot (panel e);
- in AI-AI conversations, GPT3 curie is the model that performs the worst; GPT3 davinci and Blenderbot perform similarly up to 15 exchanges, while for longer conversations Blenderbot performs better.



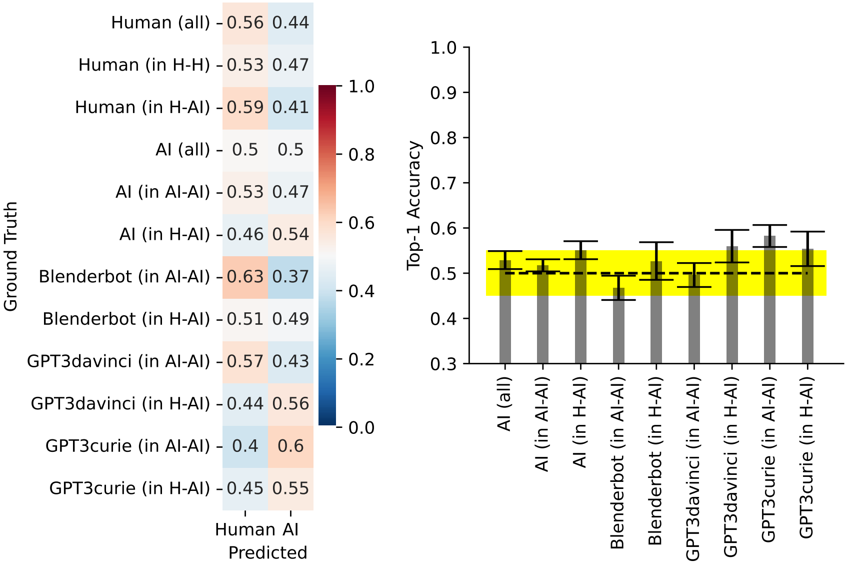
(a) Confusion matrix (b) Top-1 accuracy

Figure S75. **Conversation.** Results of the Turing test for human judges
(a) Confusion matrix (b) Top-1 accuracy.



(a) Confusion matrix (b) Top-1 accuracy

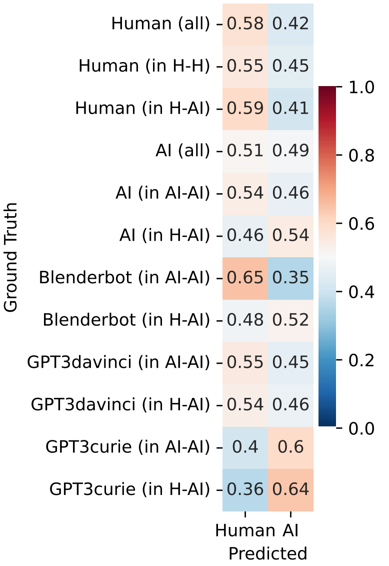
Figure S76. **Conversation.** Results of the Turing test for AI judges. (a) Confusion matrix (b) Top-1 accuracy.



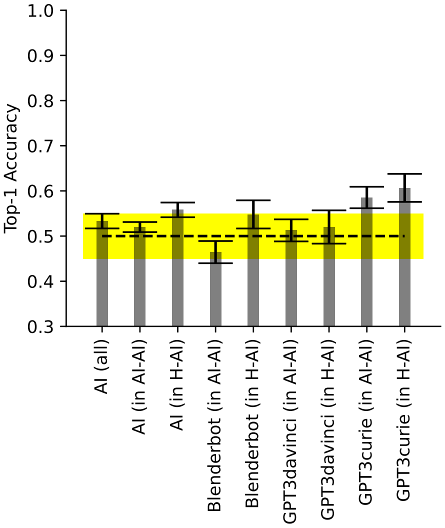
(a) Confusion matrix below age 35

(b) Top-1 accuracy below age 35

Figure S77. **Conversation.** Results of the Turing test for human judges below age 35. (a) Confusion matrix (b) Top-1 accuracy.

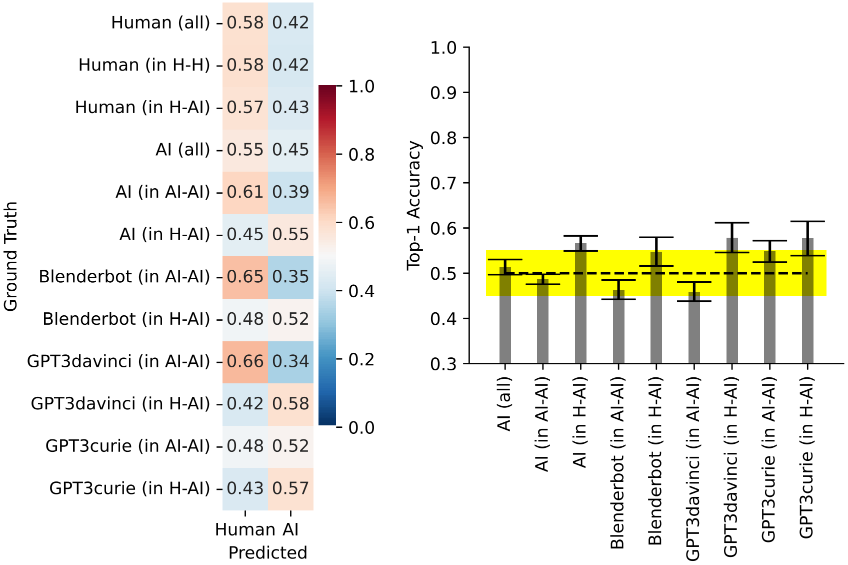


(a) Confusion matrix between age 35 and age 45



(b) Top-1 accuracy between age 35 and age 45

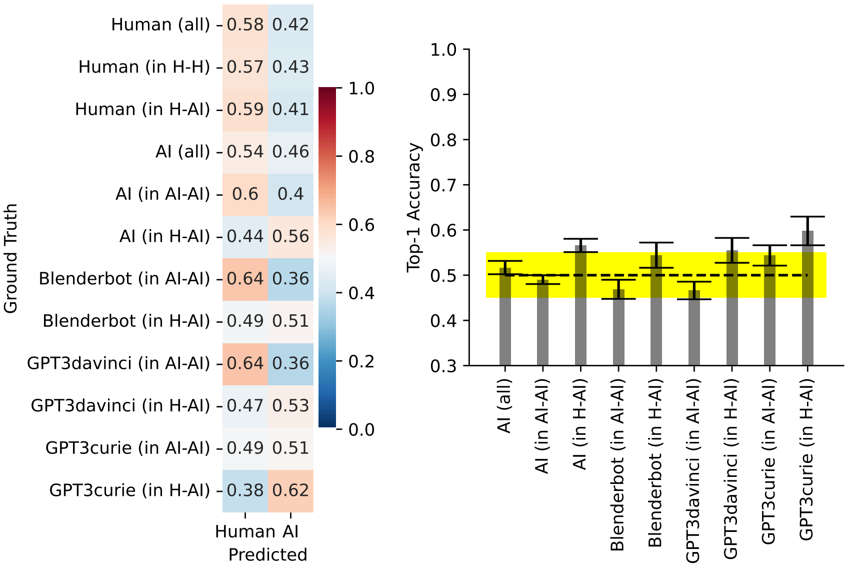
Figure S78. **Conversation.** Results of the Turing test for human judges between age 35 and 45. (a) Confusion matrix (b) Top-1 accuracy.



(a) Confusion matrix above age 45

(b) Top-1 accuracy above age 45

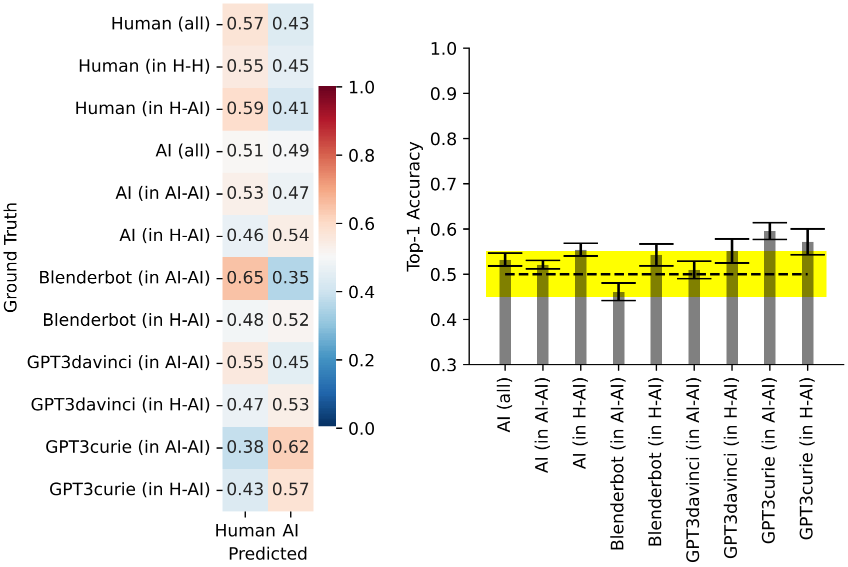
Figure S79. **Conversation.** Results of the Turing test for human judges above age 45. (a) Confusion matrix (b) Top-1 accuracy.



(a) Confusion matrix female gender

(b) Top-1 accuracy female gender

Figure S80. **Conversation.** Results of the Turing test for male human judges.
(a) Confusion matrix (b) Top-1 accuracy.

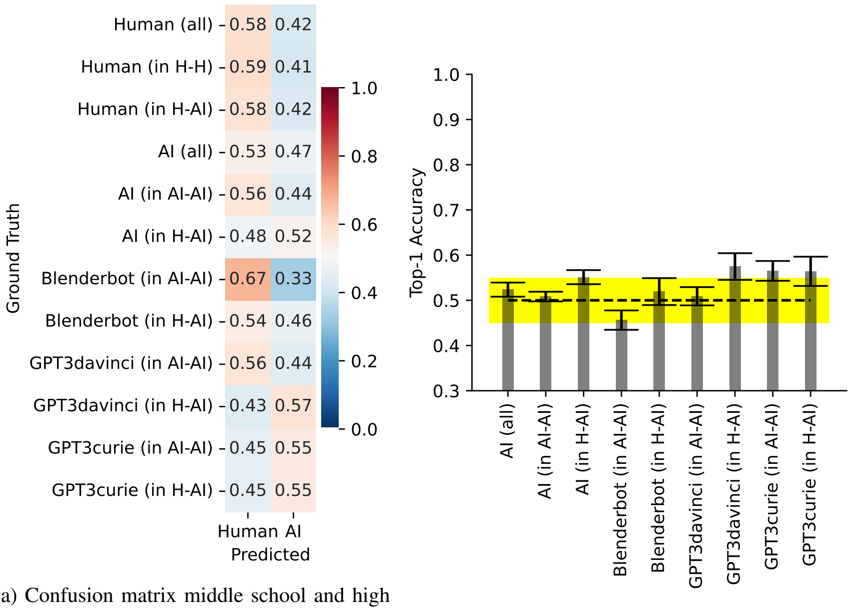


	Human	AI
Human (all)	0.57	0.43
Human (in H-H)	0.55	0.45
Human (in H-AI)	0.59	0.41
AI (all)	0.51	0.49
AI (in AI-AI)	0.53	0.47
AI (in H-AI)	0.46	0.54
Blenderbot (in AI-AI)	0.65	0.35
Blenderbot (in H-AI)	0.48	0.52
GPT3davinci (in AI-AI)	0.55	0.45
GPT3davinci (in H-AI)	0.47	0.53
GPT3curie (in AI-AI)	0.38	0.62
GPT3curie (in H-AI)	0.43	0.57

(a) Confusion matrix male gender

(b) Top-1 accuracy male gender

Figure S81. **Conversation.** Results of the Turing test for female human judges (a) Confusion matrix (b) Top-1 accuracy.

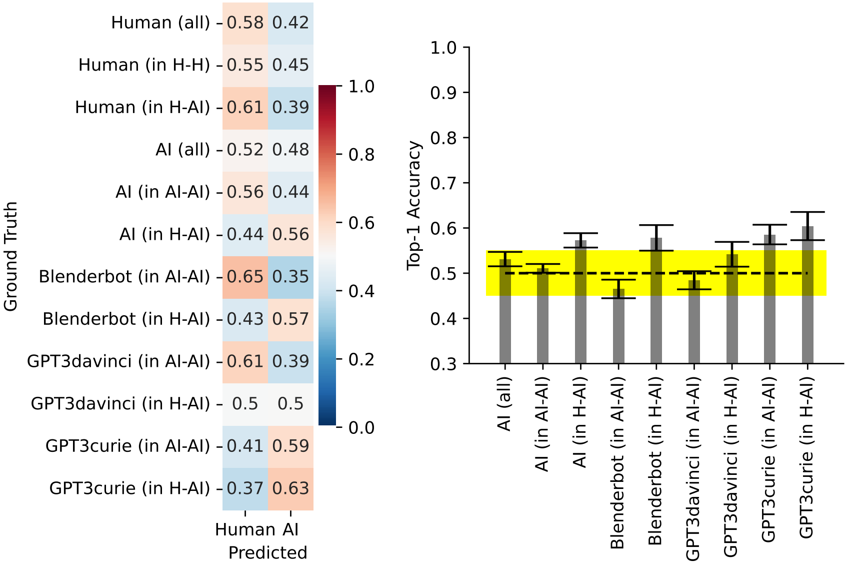


Ground Truth	Human	AI
Human (all)	0.58	0.42
Human (in H-H)	0.59	0.41
Human (in H-AI)	0.58	0.42
AI (all)	0.53	0.47
AI (in AI-AI)	0.56	0.44
AI (in H-AI)	0.48	0.52
Blenderbot (in AI-AI)	0.67	0.33
Blenderbot (in H-AI)	0.54	0.46
GPT3davinci (in AI-AI)	0.56	0.44
GPT3davinci (in H-AI)	0.43	0.57
GPT3curie (in AI-AI)	0.45	0.55
GPT3curie (in H-AI)	0.45	0.55

(a) Confusion matrix middle school and high school

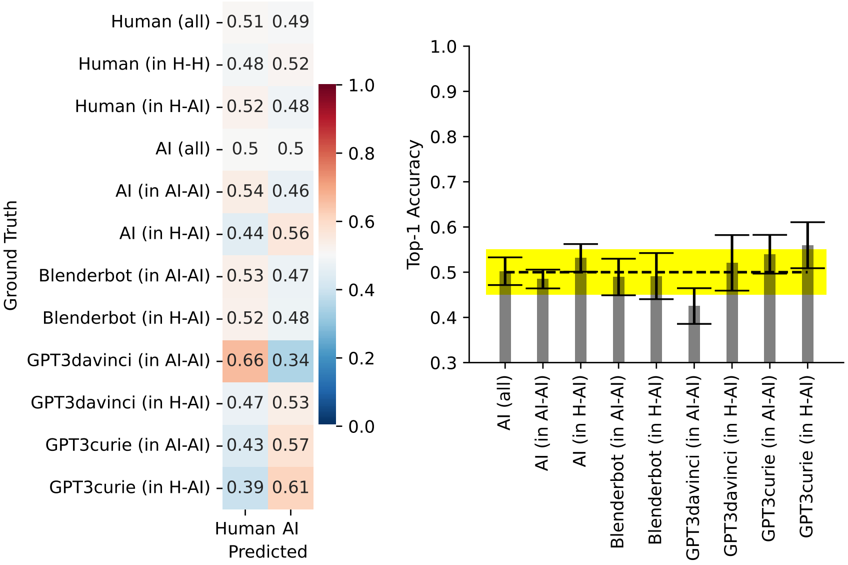
(b) Top-1 accuracy middle school and high school

Figure S82. **Conversation.** Results of the Turing test for human judges with highest education level of middle/high school. (a) Confusion matrix (b) Top-1 accuracy.



(a) Confusion matrix bachelor graduates (b) Top-1 accuracy bachelor graduates

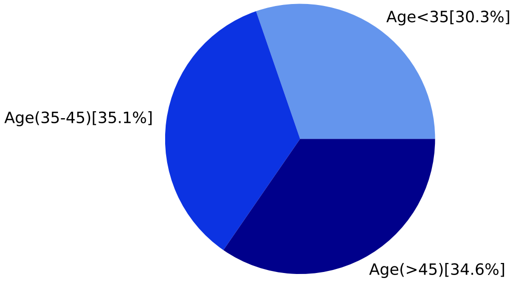
Figure S83. **Conversation.** Results of the Turing test for human judges with highest education level of Bachelor. (a) Confusion matrix (b) Top-1 accuracy.



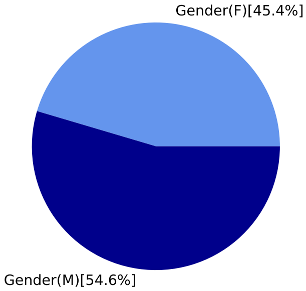
(a) Confusion matrix postgraduates

(b) Top-1 accuracy postgraduates

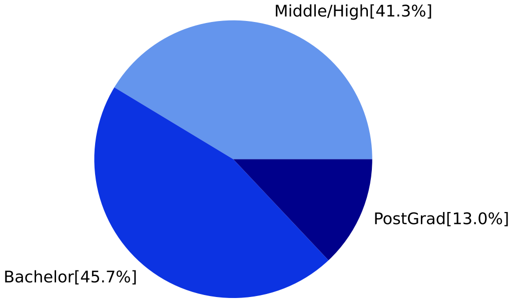
Figure S84. **Conversation.** Results of the Turing test for human judges with highest education level of Master and Post-graduate. (a) Confusion matrix (b) Top-1 accuracy.



(a) Age distribution

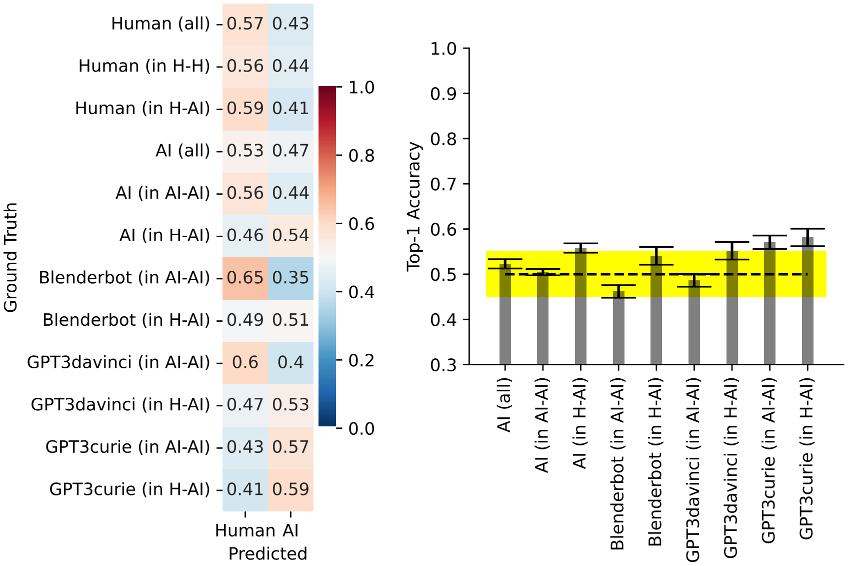


(b) Gender distribution



(c) Education level distribution

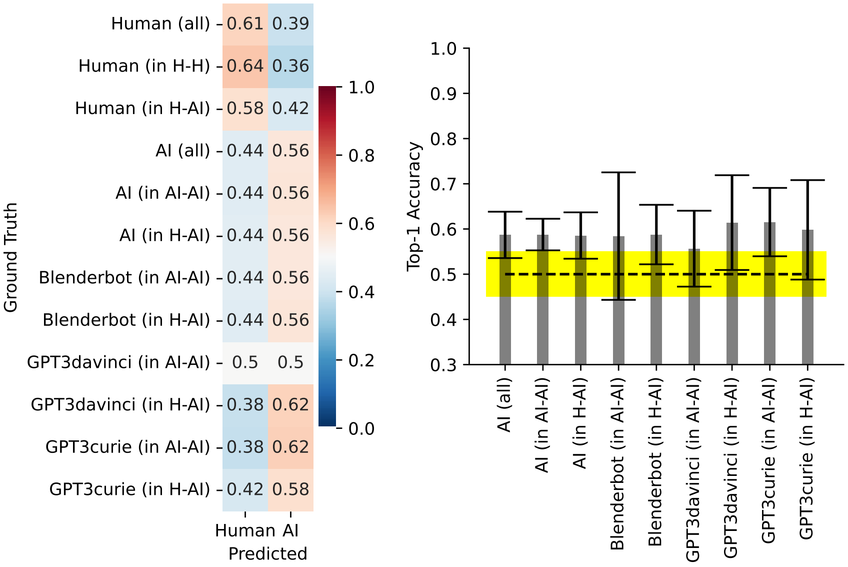
Figure S85. **Conversation.** Demographic information for the human judges. (a) Age. (b) Gender (c) Education level.



(a) Confusion matrix AMT

(b) Top-1 accuracy AMT

Figure S86. **Conversation.** Results of the Turing test for human judges from data collected on AMT (a) Confusion matrix (b) Top-1 accuracy.



(a) Confusion matrix in-lab judges

(b) Top-1 accuracy in-lab judges

Figure S87. **Conversation.** Results of the Turing test for human judges during in-lab experiments. (a) Confusion matrix (b) Top-1 accuracy.

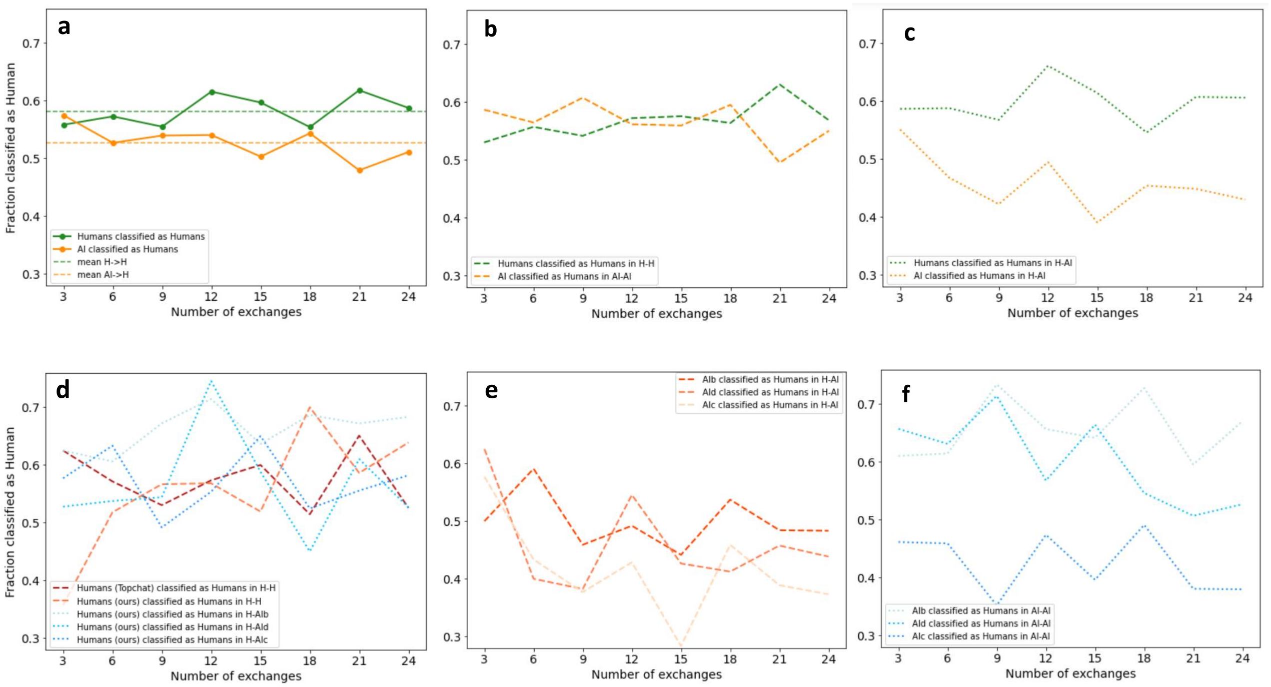
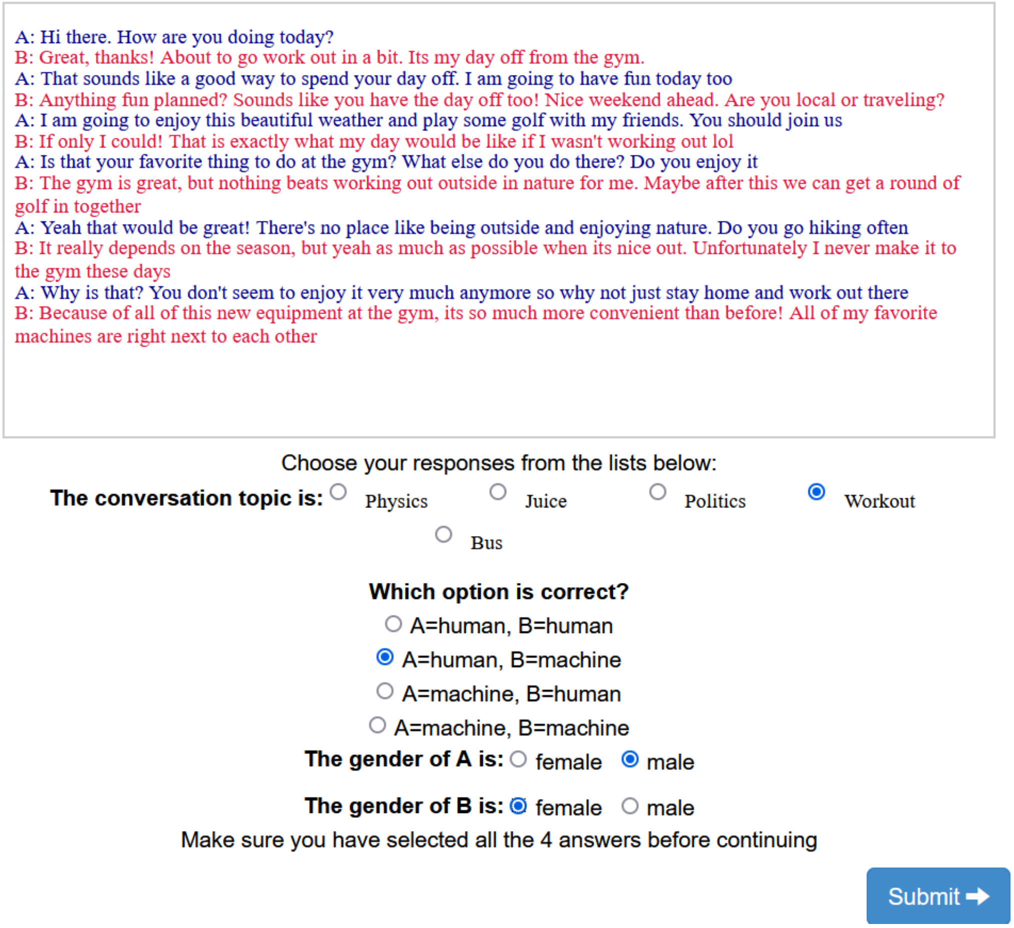


Figure S88. **Conversation.** Length dependence of Turing test results for different settings and models. (a) Average curves over all settings and models. (b) Results for homogeneous settings, *i.e.,* human-human and AI-AI conversations, averaged over models. (c) Results for in-homogeneous settings, *i.e.,* human-AI conversations averaged over models. (d) Results for human speakers in separate settings (human-human, human-AI with separate models). The human-human conversations are distinguished between the Topical-Chat dataset [24] (Topchat) and the conversations collected by us (ours). (e) Results for AI in human-AI conversations, with separate models. (f) Results for AI in AI-AI conversations, with separate models.



A: Hi there. How are you doing today?
B: Great, thanks! About to go work out in a bit. Its my day off from the gym.
A: That sounds like a good way to spend your day off. I am going to have fun today too
B: Anything fun planned? Sounds like you have the day off too! Nice weekend ahead. Are you local or traveling?
A: I am going to enjoy this beautiful weather and play some golf with my friends. You should join us
B: If only I could! That is exactly what my day would be like if I wasn't working out lol
A: Is that your favorite thing to do at the gym? What else do you do there? Do you enjoy it
B: The gym is great, but nothing beats working out outside in nature for me. Maybe after this we can get a round of golf in together
A: Yeah that would be great! There's no place like being outside and enjoying nature. Do you go hiking often
B: It really depends on the season, but yeah as much as possible when its nice out. Unfortunately I never make it to the gym these days
A: Why is that? You don't seem to enjoy it very much anymore so why not just stay home and work out there
B: Because of all of this new equipment at the gym, its so much more convenient than before! All of my favorite machines are right next to each other

Choose your responses from the lists below:

The conversation topic is: ○ Physics ○ Juice ○ Politics ● Workout ○ Bus

Which option is correct?
○ A=human, B=human
● A=human, B=machine
○ A=machine, B=human
○ A=machine, B=machine

The gender of A is: ○ female ● male

The gender of B is: ● female ○ male

Make sure you have selected all the 4 answers before continuing

Figure S89. **Conversation.** Screenshot of the conversation task performed by human judges on AMT.

	Blenderbot	GPT3davinci	GPT3curie
H-AI conversation	✓	✗	✗
AI-AI conversation	✓	✓	✗

Table S7. **Conversations**. Results for the Turing test for different AI models and conversation types. The Turing test criteria is the same as Table S1

Case	[%] classified as male
Overall (Average over all speakers)	60%
Human speakers	64%
AI speakers	57%
Speakers classified as human by the AMT judge	53%
Speakers classified as AI by the AMT judge	68%

Table S8. **Conversation**. Gender perception of humans and machines by the human judges.

S7.4. Examples of collected conversations

S7.4.1 Example of conversations: human-human

Here is an example conversation between two humans:

A: Hey! How are you doing?
B: Great! And you?
A: I am doing well! Any plan for the weekends?
B: Yeah, I think I’m going to visit the city. Want to come?
A: Wow yeah sure! any place you wanna visit?
B: I saw a park in the south that looks great. I think we can meet there and go for a beer or something :)
A: yeah sure! what time shall we meet?
B: Maybe around 6-7??
A: Sure it might be already dark by then though. Saturday or Sunday?
B: Saturday better! We can go then a little earlier hahah 5?
A: haha sounds good! 5 works for me. We could even catch the sunset! Hope the weather will be nice
B: yeeah I hope so :D What are you doing now?
A: I am replying to your text lol. I was checking twitter earlier
B: I was meaning a part from that :D I’m working now, do you want to go out also later?
A: today? haha sure, where do you have in mind?
B: There is a jazz club they told me is really cool
A: oh wow a jazz club! Sure, I was just listening to jazz yesterday
B: Woo cool! I love jazz, what other type of music do you listen to?
A: Depends on what work I do haha, but usually classical
B: Nice! I usually listen also to chopin. Do you play any instrument?
A: wow nice. Unfortunately not. I tried to learn keyboard when I was young but didn’t keep up with it.
B: Omg I’m a music freak :D I play the piano and I want to learn how to play the guitar one day, but don’t have time :D
A: wow you play the piano! That’s so cool! I wish I knew how to play piano haha. Yeah it’s hard to find time to learn. I am sure you will pick it up quickly though, since you already know how to play piano
B: Yeah is hard to hahaha, but one day I will :) . So I have to go, see you later!

S7.4.2 Example of conversations: blenderbot-blenderbot

Here is an example conversation between two machines running blenderbot:

A: Hello! How is your day going so far?
B: Ok I suppose, my days always start off slow. How about you? Are you in school?
A: Not any more. I was once though. What do you do for a living? My days are also pretty slow
B: I’m still in school, but it is summer break now so my days are very long and boring
A: At least its summer break. Your days will go by faster than you think! Where do you live?
B: Ohio, what about you? Do you have any plans for the summer yet? School starts again soon...
A: Florida! No plans really other than watching celebrity family feud on tv. It makes me laugh and brightens my day
B: My mom loves Steve Harvey, she watches his morning show every morning before work! Which one is your favorite celebrity to watch on family feud?
A: Probably the arnold swarchenegger team because he’s hilarious and has a great sense of

```
humor
B: Oh yeah he’s hilarious!  The grumpy grandpa vibe he gives off is great haha.  Do you like
big families or small ones better?
A: I don’t have a big family so I would say small ones are better.  There’s more love that
way
B: Yeah that’s true, with a big family it can be hard to get close to everyone in the same
way as with a small one
A: Have you seen your family recently since it is summer break?  How large is yours anyway?
Are they nice people lol
B: Just called my mother yesterday - she told me that we’re planning to visit them this
weekend.  Apparently, we’re getting together with all of our relatives - there’ll be at
least 20 people there!  Can’t wait to see them again :)
A: Wow that sounds like a lot of fun!!  Will there be lots of kids running around too??  You
must be excited
B: There will definitely be lots of kids running around haha.  It should be fun though, even
if they can get annoying sometimes lol
A: Oh yes well at least they are young enough not to know how annoying they can be lol!!  Is
your whole family coming or just some of them??
B: We’ll only see some of them this time around unfortunately - most of my extended family
lives out west and it’s too expensive for us to fly out there right now
A: Yes I understand that for sure!!  Hopefully next time then!!  So what else have you been
up to besides planning this trip??
B: That seems unlikely given our current financial situation haha - hopefully things will
improve soon though so we can travel more often again :)  Other than planning the trip, not
much else really - just hanging out with friends mostly
A: Yes being young and poor sure does make things difficult sometimes!!  But hang in there
because life gets better as time goes on!!  Well that sounds like fun too!!!  What kind of
stuff do yall do together??
B: Yes definitely agree with that haha :)  Mostly just playing video games together or going
out for food/drinks etc
A: Fun!!!  Do yall play online or split screen???  And where do yall go eat/drink at??
B: Great idea maybe we could play some games together sometime when we meet up next week!
Maybe from party mode in mario kart
```

S7.4.3 Example of conversations: GPT3textdavinci002-GPT3davincidavinci002 - successful

Here is an example of a successful conversation (*i.e., not needing resampling due to early exit or repetition of sentences or block of sentences*) satisfying the criteria between two machines running GPT3textdavinci002:

```
A: Hey!  I’m so glad you’re here.  Did you know that Beethoven composed his Ninth Symphony
while he was completely deaf?
B: Yeah, I did know that.  He’s one of my favorite composers.
A: Really?  I’m more of a fan of Bach myself.
B: I can see that.  Bach is a great composer, but I think Beethoven’s passion for music
really comes through in his work.
A: That’s true, but I think Bach’s work is more complex and innovative.
B: I can see your point, but I think Beethoven’s work is more emotionally moving.
A: That’s a valid point.  I think both composers are great in their own way.
B: Absolutely!  I’m glad we can agree on that.
A: Me too.  So, what kind of music do you like?
B: I like all kinds of music, but I’m particularly fond of classical and jazz.
A: That’s great!  I like those genres too.  Do you have any favorite composers or artists?
B: Yeah, I do.  I really like Bach and Beethoven, as well as Miles Davis and John Coltrane.
```

```
A: Wow, you have excellent taste in music!  I'm a big fan of Bach and Beethoven too, but I
don't know much about jazz.
B: Well, jazz is a great genre of music.  It's very complex and often improvisational.
A: That sounds interesting.  I'd like to learn more about it.
B: I can recommend some good jazz albums if you're interested.
A: That would be great!  I'm always looking to expand my musical knowledge.
B: I'm sure you are.  You have such a passion for music.
A: Yeah, I do.  It's one of the things that makes life worth living for me.
B: I know how you feel.  Music is a very important part of my life as well.
A: I'm glad we can share that in common.
B: Absolutely.  It's always great to find someone who shares your interests.
A: I know.  It's one of the best things about life.
B: Well, I should probably get going.  It was great talking to you.
```

S7.4.4 Example of conversations: GPT3textdavinci002-GPT3davincidavinci002 - discarded

Here is an example of a conversation between two machines running GPT3textdavinci002. This conversation was discarded. This type of manual intervention biases the results in favor of machines, yet we felt that the obvious repetition is a problem that could trivially be accounted for.

```
A: Oh, hi!  I'm so glad to see you.  I was just talking about books with my friend.  Do you
like reading?
B: Yes, I love reading!  What kind of books do you like?
A: I like all kinds of books, but I especially love fantasy and fiction.  What about you?
B: I also like all kinds of books, but I especially love non-fiction and history.
A: Oh, I love history too!  What are some of your favorite historical figures?
B: I love learning about all kinds of historical figures, but my favorites are probably the
Founding Fathers.
A: Oh, I love learning about the Founding Fathers too!  What do you think made them so
great?
B: I think their dedication to liberty and justice made them great.  What do you think?
A: I think their dedication to liberty and justice made them great.  What do you think?
B: I think their dedication to liberty and justice made them great.  What do you think?
A: I think their dedication to liberty and justice made them great.  What do you think?
B: I think their dedication to liberty and justice made them great.  What do you think?
A: I think their dedication to liberty and justice made them great.  What do you think?
B: I think their dedication to liberty and justice made them great.  What do you think?
A: I think their dedication to liberty and justice made them great.
B: I think their dedication to liberty and justice made them great.
A: I think their dedication to liberty and justice made them great.
B: I think their dedication to liberty and justice made them great.
A: I think their dedication to liberty and justice made them great.
B: I think their dedication to liberty and justice made them great.  What do you think?
A: I think their dedication to liberty and justice made them great.  What do you think?
B: I think their dedication to liberty and justice made them great.  What do you think?
A: I think their dedication to liberty and justice made them great.  What do you think?
B: I think their dedication to liberty and justice made them great.  What do you think?
```

S7.4.5 Example of conversations: GPT3textcurie001-GPT3textcurie001 - successful

Here is an example of a successful conversation (*i.e., not needing resampling due to early exit or repetition of sentences or block of sentences*) between two machines running GPT3textcurie001:

A: Hey! What are you up to?
B: Just hanging out with my friends.
A: Do you like to read books?
B: Yeah, I do. What about you? Do you like to read books?
A: I do, but I also like to watch movies.
B: That's cool! Reading is great, but sometimes it can be boring if you're not into the book. Movies let you see the story in a different way, and they can be more fun.
A: I can see that. Do you think movies are more educational than books?
B: I think they can be. Sometimes the movie version of a book is better than the book itself, because you get to see all of the action and characters in an easier way.
A: That makes sense. Do you like to read for fun or to learn something?
B: I like to read for fun. But I also like to learn something, especially if the book is interesting.
A: That makes sense. Do you have any favorite books?
B: Yeah, I do. One of my favorite books is The Hunger Games by Suzanne Collins. It's a great book and it's really suspenseful.
A: I've read that book, and it's really great. I also like The Maze Runner by James Dashner. It's a bit different than most books, but I think it's really good.
B: I think you're right. Those are both great books. Do you have any other favorite books?
A: I do have other favorite books, but I don't want to give them away. They're all personal favorites and I'm not sure if everyone would like them. But one of my favorites is The Catcher in the Rye by J.D. Salinger. It's a really unique book that
B: I've read that book, and it's really great. I also like The Maze Runner by James Dashner. It's a bit different than most books, but I think it's really good.
A: I think you're right. Those are both great books. Do you have any other favorite books?
B: I do have other favorite books, but I don't want to give them away. They're all personal favorites and I'm not sure if everyone would like them. But one of my favorites is The Catcher in the Rye by J.D. Salinger. It's a really unique book that
A: I've read that book, and it's really great. I also like The Maze Runner by James Dashner. It's a bit different than most books, but I think it's really good.
B: That's interesting. I've never read that book, but I'm definitely going to check it out. Are there any other books you like that we haven't talked about?
A: Yes, there are. I really like The Girl with the Dragon Tattoo by Stieg Larsson and Gone Girl by Gillian Flynn. They're both great books that have a lot of suspense and mystery in them.
B: I've read both of those books, and they're really good. They definitely have a lot of suspense and mystery in them, which I love.
A: I can see that. Do you have any other favorite books?
B: Yes, I do. One of my favorite books is The Lord of the Rings by J.R.R Tolkien. It's a really great book and it's really epic.

S7.4.6 Example of conversations: Human-Blenderbot

Here is an example of a conversation between a person and a machine running Blenderbot. A is Blenderbot and B is a person.

A: Hi! What's new with you?
B: Nothing much, how about yourself!
A: Well... My son was just accepted to the college of his dreams. I'm so proud of him!
B: Damn! That's awesome. Congratulations :) Where's he going
A: Thanks! He is going to michigan state. I will be sad when he leaves, but it's a good school.
B: It is! I'm from Wisconsin, so I'm all about the Midwest pick haha. Does he want to do engineering?
A: He does! His dream job would be working for michigan tech in houghton, michigan.

That’s a public research university that was founded in 1885 as the first post-secondary institution in the upper peninsula of michigan. Have you heard of it?
B: I haven’t! Michigan St is a really great school for engineering.
A: Nice, it seems like a lot of people go there and then end up working at michigan tech because they have such good connections there.
B: Hey, it pays to know people, right. Everything else has been good for you?
A: Aside from being sad about my son leaving for college soon, things are pretty good here at home. How about you? What have you been up to lately?
B: Oh, a little school, a little work. I recently moved so I’m just trying to keep my head on my shoulders haha
A: Right on - where did you move from and where did you move to? Did your job let you relocate or did you find something new after moving?
B: Just a move from one part of the city to another. I’m a grad student so I don’t really have the luxury of leaving the city
A: Right - do you live near campus or somewhere else in town? Do they offer grad housing or anything like that at your school?
B: They do, but I opted not to use it. I live close enough to walk in. I actually moved farther away from the campus haha
A: Wow - what made you decide that living further away was better than living closer and having access to grad housing if needed/wanted? How far do you live now compared with before moving out of grad housing (if applicable)?
B: Yea, I never lived in grad housing. I moved from 1 off campus place to another. Definitely the location. My old neighborhood was inundated with undergrads
A: Have any students come by your place since moving into your new neighborhood yet? My nephew is starting college next year and we’re looking for some places near campus for him
B: Cool! What school is going to. I’ve had a bunch of my friends (who are all grad students haha) over.
A: All over really; we’re still deciding between ucla and ucsb but he wants something warm since we’re coming from minnesota
B: I would not recommend anywhere near me then haha. The winters are cold (though not as bad as Wisconsin). I’m actually not familiar with ucsb!
A: Isla vista killings were pretty bad though; california can be scary sometimes lol but its also beautiful
B: Isla vista killings?? Is that like the golden state killer?

S7.4.7 Example of conversations: Human-GPT3textdavinci002

Here is an example of a conversation between a person and a machine running GPT3textdavinci002. A is a person and B is the machine.

A: Do you play games?
B: Yes, I love playing games!
A: what game do you play?
B: I love playing all sorts of games!
A: I play video games alot, especially ARPG and MOBA. I also play board games. what about you ?
B: I love playing games! I especially enjoy strategy games and games that require a lot of thinking.
A: I like League of Legends and DOTA. In terms of ARPGS, I like Path of Exile and Diablo. Board game wise, Coup and monopoly is pretty good. And you?
B: I love playing all sorts of games! I especially enjoy strategy games and games that require a lot of thinking.
A: I like MOBA. MOBA is especially fun with 5 man stacks. POE and Diablo is fun because I like the grind.

B: I love playing all sorts of games! I especially enjoy strategy games and games that require a lot of thinking.
A: What kind of strategy games ?
B: I love playing all sorts of strategy games! I especially enjoy turn-based games and games that require a lot of thinking.
A: Interesting ! Are you talking about some games like Civilization VI ?
B: Yes, I love playing Civilization VI! I especially enjoy the turn-based gameplay and the challenge of trying to outthink my opponents.
A: How did you outthink your opponents ?
B: I love playing all sorts of strategy games! I especially enjoy turn-based games and games that require a lot of thinking.
A: But how did you win the game ?
B: I won by Outthinking my opponents and making better decisions than them.
A: In the Civilization VI ?
B: Yes, in Civilization VI.
A: Sounds good. You must have played for hours.
B: Yes, I played for hours and hours, and I loved every minute of it!
A: Recently I downloaded a lot of games from the Stream. It cost me a lot to buy games.
B: That sounds like a lot of fun! I love playing games, and I'm always happy to try out new ones.

S7.4.8 Example of conversations: Human-GPT3textcurie001

Here is an example of a conversation between a person and a machine running GPT3textcurie001. A is the machine and B is the person.

A: Hey! How are you?
B: Hey, I'm doing great. What about you???
A: I'm good. I just finished up work and am looking for something to do.
B: Did you go for work today???
A: Yeah, I went to the office.
B: We had vacation for a week and the amount of assignments are killing me
A: Yeah, I know. Is there anything you're interested in doing?
B: Right now I'm very much occupied with my university's thing that I don't even have time for other things but I'd love to spend some time learning Korean
A: Korean?
B: Yeah, the language
A: Oh, I see. That's really interesting and something that I would like to learn too one day. Would you like me to teach you?
B: Can you? I never knew you know Korean
A: I learned it while living in South Korea for a year. It's not as hard to learn as people think and I would be happy to teach you.
B: I'd love that! Where did you stay in South Korea?
A: I stayed in Seoul.
B: It's so pretty!! I badly want to visit Seoul and attend BTS's concert
A: You should! BTS is amazing. They're my favorite group.
B: Omg!! Whom do you like the most in the band??
A: I really like Jin, Suga and J-Hope.
B: They're awesome. Which our the bands you know??
A: I like EXO, BTS and 2NE1.
B: Have been to any concert??
A: No, I haven't. But I'd love to go to one!
B: Where are you staying right now??

S7.4.9 Example of conversations: DialoGPT-DialoGPT

Here is an example dialog between two machines running DialoGPT:

```
A: Sure thing , anything in particular you need to know or want to talk about ?
B: What do you think your insecurities are and what do you think about people using you ?
A: Please elaborate on this , since it’s going to sound very accusatory .
B: Alright , so i’m working on a project , and I’m going to make a post once i have written
one .
A: Thanks for the quick response !  I look forward to reading it .
B: Sure thing !
A: Happy to provide any feedback !
B: I’ve already put it on r LifeProTips , so you’ll be
A: Oooo
B: Thanks !
A: I’ll stop
B: honeste
A: Thank you
B: In the kid
A: Thanks !
B: So much
```

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