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 Conversations, 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.

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.
### Table 1. Specifications of six Turing tasks

<table>
<thead>
<tr>
<th>Task</th>
<th>Num. Stimulus</th>
<th>Num. Turing Tests</th>
<th>Sources of Datasets</th>
<th>AI models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Object detection</td>
<td>808</td>
<td>1,975</td>
<td>self-collect, MSCOCO [42]</td>
<td>Google Vision API, Microsoft Azure Cognitive Services, Amazon Rekognition, Detectron2 [69]</td>
</tr>
<tr>
<td>Attention Prediction</td>
<td>547</td>
<td>2,160</td>
<td>NatureDesign [73], FindingWaldo [73], NatureSaliency [72]</td>
<td>IVSN [27,72,73], DeepGaze3 [39], GBVS [29]</td>
</tr>
<tr>
<td>Image Captioning</td>
<td>1,000</td>
<td>8,140</td>
<td>self-collect, MSCOCO [42], nocaps [3]</td>
<td>GIT [63], OPA [64], BLIP [65], ClipCap [16], Microsoft’s Azure Cognitive Services [2]</td>
</tr>
<tr>
<td>Word Association</td>
<td>1,500</td>
<td>3,550</td>
<td>self-collect</td>
<td>Word2Vec [59], GPT2 [52], GPT3-embedding (davinci) [8], GPT3-prompt (text-curie-001) [8], GPT3-prompt (text-Davinci-002) [8]</td>
</tr>
<tr>
<td>Conversation</td>
<td>300×8 lengths</td>
<td>4,120</td>
<td>self-collect</td>
<td>GPT3-text-davinci-002 [48], GPT3-text-curie-001 [48], Blenderbot [57], DialogPT [24]</td>
</tr>
</tbody>
</table>

#### 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 [5]. In total, 3,140 image-response pairs were collected (785 from humans and 2,355 from AI models, Table 1).

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

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

#### 3.2. Object detection

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

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

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

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

#### 3.3. Attention prediction

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

**Turing test (human judge).** Separate Turing tests were launched for eye movements from free-viewing tasks (80...
judges) and visual search tasks (100 judges) (Fig. S31, S47 and Fig. S32, S47). We presented infinitely repeating animated clips of eye movements from humans or model predictions with a maximum of 15 fixations to human judges on AMT. Each judge had to identify if the eye movements were from a human or a computational model. As a control, judges were also asked to answer “What do you see in the presented clip?” with one correct answer among 3 options. Responses from judges with a score < 7 out of 12 were not considered in the analyses.

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

### 3.4. Image captioning

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

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

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

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

### 3.5. Word associations

**Dataset.** We chose 150 unique cue words (50 nouns, 50 verbs, and 50 adjectives), spanning a wide range of occurrence frequencies [59] (Table S4; see Section S6 for multiple additional controls). Associations to each cue word were collected from human subjects (Fig. S68, S69, S74), and from the following language models: Word2vec [59], 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-davinci-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 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.
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 and 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 images 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 comparisons 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
S6.1 Dataset 90
S6.1.1 Cue words 90
S6.1.2 Collecting human responses from AMT 90
S6.1.3 Collecting AI responses 90
S6.2 Turing test 91
S6.2.1 Human judge Turing test 91
S6.2.2 AI judge Turing test 91
S7 Conversation

S7.1 Dataset
S7.1.1 Human participants
S7.1.2 Dataset collection: Instructions to human participants
S7.1.3 AI conversation bots
S7.1.4 Dataset collection: Prompt and settings for GPT3text-davinci002 and GPT3text-curie-001

S7.2 Turing test
S7.2.1 Collecting human judge responses for Turing test
S7.2.2 AI judge

S7.3 Results and discussion
S7.3.1 Confusion matrix and Top-1 accuracy
S7.3.2 Gender perception
S7.3.3 Effects of judge demographics
S7.3.4 Comparison between AMT and in-person experiments
S7.3.5 Effect of conversation length

S7.4 Examples of collected conversations
S7.4.1 Example of conversations: human-human
S7.4.2 Example of conversations: blenderbot-blenderbot
S7.4.3 Example of conversations: GPT3textdavinci002-GPT3davincidavinci002 - successful
S7.4.4 Example of conversations: GPT3textdavinci002-GPT3davincidavinci002 - discarded
S7.4.5 Example of conversations: GPT3textcurie001-GPT3textcurie001 - successful
S7.4.6 Example of conversations: Human-Blenderbot
S7.4.7 Example of conversations: Human-GPT3textdavinci002
S7.4.8 Example of conversations: Human-GPT3textcurie001
S7.4.9 Example of conversations: DialoGPT-DialoGPT

List of Supplementary Figures

S1 Color estimation. AMT user interface for collecting responses.
S2 Color estimation. Random samples from our collected color estimation dataset.
S3 Color estimation. AMT user interface for collecting human judge responses (Turing test).
S4 Color estimation. Random samples from our collected Turing tests.
S5 Color estimation. Results of the Turing test for human judges.
S6 Color estimation. Results of the Turing test for human judges with age level below 35.
S7 Color estimation. Results of the Turing test for human judges with age level 35-45.
S8 Color estimation. Results of the Turing test for human judges with age above 45.
S9 Color estimation. Results of the Turing test for human judges with education level below Bachelors.
S10 Color estimation. Results of the Turing test for human judges with education level of Bachelors.
S11 Color estimation. Results of the Turing test for human judges with education level above Bachelors.
S12 Color estimation. Results of the Turing test for human judges with Male gender.
S13 Color estimation. Results of the Turing test for human judges with Female gender.
S14 Color Estimation. Demographic information for the human judges for the color estimation task
S15 Color estimation. Results of the Turing test for AI judges
S16 Object detection. AMT user interface for collecting responses.
S17 Object detection. Random samples from our collected object detection dataset.
S18 Object detection. AMT user interface for collecting human judge responses (Turing test).
S19 Object detection. Random samples from our collected Turing tests.
S20 Object detection. Results of the Turing test for human judges.
S21 Object detection. Results of the Turing test for human judges with age below 35.
S22 Object detection. Results of the Turing test for human judges with age 35-45.
S23 Object detection. Results of the Turing test for human judges with age above 45.
S24 Object detection. Results of the Turing test for human judges with education level below Bachelors.
S25 Object detection. Results of the Turing test for human judges with education level of Bachelors.
S26 Object detection. Results of the Turing test for human judges with education level above Bachelors.
**Image captioning.** Demographic information for the human judges. (a) Age. (b) Gender. (c) Education level. 89

**Word Associations.** Results of the Turing test for human judges. 92

**Word Associations.** Results of the Turing test for AI judges. 93

**Word Associations.** AMT user interface for collecting word associations responses, for free associations. 94

**Word Associations.** AMT user interface for collecting responses, for prompt-guided associations. 95

**Word Associations.** AMT user interface for collecting human judge responses (Turing test). 96

**Word Associations.** Results of the Turing test for human judges in different age groups. 97

**Word Associations.** Results of the Turing test for different genders. 98

**Word Associations.** Results of the Turing test for human judges for different educational levels. 99

**Word Associations.** Demographic information for the human judges for the word association task. 100

**Conversation.** Results of the Turing test for human judges. 109

**Conversation.** Results of the Turing test for AI judges. 110

**Conversation.** Results of the Turing test for human judges below age 35. 111

**Conversation.** Results of the Turing test for human judges between age 35 and 45. 112

**Conversation.** Results of the Turing test for human judges above age 45. 113

**Conversation.** Results of the Turing test for male human judges. 114

**Conversation.** Results of the Turing test for female human judges. 115

**Conversation.** Results of the Turing test for human judges with highest education level of middle/high school. 116

**Conversation.** Results of the Turing test for human judges with highest education level of Bachelor. 117

**Conversation.** Results of the Turing test for human judges with highest education level of Master and Post-graduate. 118

**Conversation.** Demographic information for the human judges. 119

**Conversation.** Results of the Turing test for human judges from data collected on AMT. 120

**Conversation.** Results of the Turing test for human judges during in-lab experiments. 121

**Conversation.** Length dependence of Turing test results for different settings and models. 122

**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 et al. [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.
Table S1. **Summary of Turing test results for human judges.**

A tick (✓) indicates that the AI models pass the Turing test (0.45 ≤ 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 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. Similarly, Fig. S9, Fig. S10 and Fig. S11 show the performance of human judges across different education levels, 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

Submit ➡

Figure S1. Color estimation. AMT user interface for collecting responses.
Figure S2. **Color estimation.** Random samples from our collected color estimation dataset.
Please answer below.

Figure S3. **Color estimation**. AMT user interface for collecting human judge responses (Turing test).
Figure S4. **Color estimation.** Random samples from our collected Turing tests.
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.
Figure S6. **Color estimation.** Results of the Turing test for human judges with age level below 35.

(a) Confusion matrix (b) Top-1 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.
Figure S8. **Color estimation.** Results of the Turing test for human judges with age above 45.
(a) Confusion matrix (b) Top-1 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.
Figure S10. **Color estimation.** Results of the Turing test for human judges with education level of Bachelors.

(a) Confusion matrix (b) Top-1 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.
Figure S12. **Color estimation.** Results of the Turing test for human judges with Male gender.  
(a) Confusion matrix (b) Top-1 accuracy.
Figure S13. **Color estimation.** Results of the Turing test for human judges with Female gender.
(a) Confusion matrix (b) Top-1 accuracy.
Figure S14. **Color Estimation.** Demographic information for the human judges for the color estimation task
(a) Age. (b) Gender. (c) Education level.
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 unknown. 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.
Figure S17. **Object detection.** Random samples from our collected object detection dataset.
Figure S18. **Object detection.** AMT user interface for collecting human judge responses (Turing test).
Figure S19. **Object detection.** Random samples from our collected Turing tests.
Figure S20. **Object detection.** Results of the Turing test for human judges. (a) Confusion matrix (b) Top-1 accuracy.
Figure S21. **Object detection.** Results of the Turing test for human judges with age below 35. (a) Confusion matrix (b) Top-1 accuracy.
Figure S22. **Object detection.** Results of the Turing test for human judges with age 35-45. (a) Confusion matrix (b) Top-1 accuracy.
Figure S23. **Object detection.** Results of the Turing test for human judges with age above 45.

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

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

(a) Confusion matrix (b) Top-1 accuracy.
Figure S28. **Object detection.** Results of the Turing test for human judges with Female gender. (a) Confusion matrix (b) Top-1 accuracy.
Figure S29. **Object Detection.** Demographic information for the human judges for the object detection task
(a) Age. (b) Gender. (c) Education level.
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/10SYSrXuWIEvTQTMMH2Lpsq3ejfM3M0l9g_/view?usp=sharing and here https://drive.google.com/file/d/1JxIdbrHNEEElq6HGS1Yf7NC3-00v5ow0/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.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Task</th>
<th>Number of stimuli</th>
<th>Number of human scanpaths</th>
</tr>
</thead>
<tbody>
<tr>
<td>NatureDesign [73]</td>
<td>Visual Search</td>
<td>240</td>
<td>3600</td>
</tr>
<tr>
<td>FindingWaldo [73]</td>
<td>Visual Search</td>
<td>67</td>
<td>1005</td>
</tr>
</tbody>
</table>

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).
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.
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.
Figure S35. **Attention prediction.** Results of the Turing test in the visual search task for human judges.

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

(a) Confusion matrix (b) Top-1 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.
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.
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.
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.
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.
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.
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.
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.
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.
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).
<table>
<thead>
<tr>
<th>Model</th>
<th>Model Variant</th>
<th>Model Details</th>
<th>Web links</th>
</tr>
</thead>
<tbody>
<tr>
<td>GIT [63]</td>
<td>GIT-Large</td>
<td>Finetuned on COCO Captions</td>
<td>Link</td>
</tr>
<tr>
<td>OFA [64]</td>
<td>OFA-Huge</td>
<td>Finetuned on COCO Captions</td>
<td>Link</td>
</tr>
<tr>
<td>BLIP [41]</td>
<td>BLIP-Large</td>
<td>Finetuned on COCO Captions</td>
<td>Link 1, Link 2</td>
</tr>
<tr>
<td>ClipCap [46]</td>
<td>ClipCap-Transformer (Beam Search)</td>
<td>Pretrained on COCO Captions</td>
<td>Link</td>
</tr>
<tr>
<td>Microsoft’s Azure Cognitive Services [2]</td>
<td>-</td>
<td>-</td>
<td>Link</td>
</tr>
</tbody>
</table>

Table S3. **Image captioning.** Different model variants used for collecting machine captions.
Figure S48. **Image captioning.** AMT user interface for collecting responses.
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

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

Submit ➡

Figure S68. **Word Associations.** AMT user interface for collecting word associations responses, for free associations
Figure S69. Word Associations. AMT user interface for collecting responses, for prompt-guided associations.
Figure S70. **Word Associations.** AMT user interface for collecting human judge responses (Turing test).
Figure S71. **Word Associations.** Results of the Turing test for human judges in different age groups
(a) Confusion matrix (b) Top-1 accuracy.
Figure S72. **Word Associations.** Results of the Turing test for different genders.
(a) Confusion matrix. (b) Top-1 accuracy.
Figure S73. **Word Associations.** Results of the Turing test for human judges for different educational levels.

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

<table>
<thead>
<tr>
<th>Cue Words</th>
<th>Word2vec</th>
<th>GPT2</th>
<th>GPT3 (embedding, davinci)</th>
<th>GPT3 (prompt curie)</th>
<th>GPT3 (prompt davinci)</th>
</tr>
</thead>
<tbody>
<tr>
<td>heavy</td>
<td>edge</td>
<td>massive</td>
<td>weight</td>
<td>light</td>
<td>massive</td>
</tr>
<tr>
<td>risky</td>
<td>elaborate</td>
<td>dangerous</td>
<td>danger</td>
<td>dangerous</td>
<td></td>
</tr>
<tr>
<td>grand</td>
<td>paris</td>
<td>massive</td>
<td>great</td>
<td>palace</td>
<td>massive</td>
</tr>
<tr>
<td>powerful</td>
<td>extraordinary</td>
<td>strong</td>
<td>mighty</td>
<td>engine</td>
<td>strong</td>
</tr>
<tr>
<td>nepal</td>
<td>vietnam</td>
<td>pakistan</td>
<td>nebraska</td>
<td>new</td>
<td>pakistan</td>
</tr>
<tr>
<td>healthy</td>
<td>weight</td>
<td>good</td>
<td>alive</td>
<td>sick</td>
<td>good</td>
</tr>
<tr>
<td>credible</td>
<td>legitimate</td>
<td>compelling</td>
<td>legitimate</td>
<td>trustworthy</td>
<td>compelling</td>
</tr>
<tr>
<td>rigid</td>
<td>narrow</td>
<td>flexible</td>
<td>stiff</td>
<td>strict</td>
<td>flexible</td>
</tr>
<tr>
<td>oriental</td>
<td>imperial</td>
<td>colonial</td>
<td>eastern</td>
<td>love</td>
<td>colonial</td>
</tr>
<tr>
<td>soft</td>
<td>thin</td>
<td>hard</td>
<td>gentle</td>
<td>whisper</td>
<td>hard</td>
</tr>
<tr>
<td>southeastern</td>
<td>mainland</td>
<td>northwestern</td>
<td>oriental</td>
<td>florida</td>
<td>northwestern</td>
</tr>
<tr>
<td>numerous</td>
<td>several</td>
<td>several</td>
<td>many</td>
<td>many</td>
<td>several</td>
</tr>
<tr>
<td>conceptual</td>
<td>composition</td>
<td>cognitive</td>
<td>idea</td>
<td>idea</td>
<td>cognitive</td>
</tr>
<tr>
<td>inquisitive</td>
<td>hollow</td>
<td>spiritual</td>
<td>magic</td>
<td>magic</td>
<td>spiritual</td>
</tr>
<tr>
<td>fetish</td>
<td>amateur</td>
<td>obsession</td>
<td>erotic</td>
<td>sex</td>
<td>obsession</td>
</tr>
<tr>
<td>rural</td>
<td>capital</td>
<td>urban</td>
<td>countryside</td>
<td>farm</td>
<td>urban</td>
</tr>
<tr>
<td>exclusive</td>
<td>feature</td>
<td>unique</td>
<td>unique</td>
<td>inclusive</td>
<td>unique</td>
</tr>
<tr>
<td>soviet</td>
<td>ussr</td>
<td>ussr</td>
<td>ussr</td>
<td>union</td>
<td>ussr</td>
</tr>
<tr>
<td>presidential</td>
<td>congressional</td>
<td>republican</td>
<td>republican</td>
<td>gubernatorial</td>
<td>republican</td>
</tr>
<tr>
<td>attractive</td>
<td>enthusiastic</td>
<td>awesome</td>
<td>listen</td>
<td>enthusiastic</td>
<td></td>
</tr>
<tr>
<td>impressive</td>
<td>superb</td>
<td>incredible</td>
<td>incredible</td>
<td>wonderful</td>
<td>incredible</td>
</tr>
<tr>
<td>ridiculous</td>
<td>awful</td>
<td>absurd</td>
<td>absurd</td>
<td>funny</td>
<td>absurd</td>
</tr>
<tr>
<td>psychotic</td>
<td>noxious</td>
<td>paranoid</td>
<td>insane</td>
<td>crazy</td>
<td>paranoid</td>
</tr>
<tr>
<td>ordinary</td>
<td>brave</td>
<td>normal</td>
<td>normal</td>
<td>normal</td>
<td>normal</td>
</tr>
<tr>
<td>enormous</td>
<td>lethal</td>
<td>immense</td>
<td>immense</td>
<td>huge</td>
<td>immense</td>
</tr>
<tr>
<td>intangible</td>
<td>singular</td>
<td>invisible</td>
<td>problem</td>
<td>invisible</td>
<td></td>
</tr>
<tr>
<td>manmade</td>
<td>synthetic</td>
<td>insect</td>
<td>molecular</td>
<td>human</td>
<td>insect</td>
</tr>
<tr>
<td>revolutionary</td>
<td>progressive</td>
<td>radical</td>
<td>radical</td>
<td>change</td>
<td>radical</td>
</tr>
<tr>
<td>vocal</td>
<td>visual</td>
<td>voice</td>
<td>voice</td>
<td>sing</td>
<td>voice</td>
</tr>
<tr>
<td>schematic</td>
<td>synaptic</td>
<td>diagram</td>
<td>diagram</td>
<td>diagram</td>
<td></td>
</tr>
<tr>
<td>suspicious</td>
<td>opposite</td>
<td>strange</td>
<td>crime</td>
<td>strange</td>
<td></td>
</tr>
<tr>
<td>nationwide</td>
<td>reservation</td>
<td>worldwide</td>
<td>worldwide</td>
<td>local</td>
<td>worldwide</td>
</tr>
<tr>
<td>synaptic</td>
<td>mammalian</td>
<td>neural</td>
<td>neural</td>
<td>synaptic</td>
<td>neural</td>
</tr>
<tr>
<td>white</td>
<td>black</td>
<td>black</td>
<td>black</td>
<td>black</td>
<td>black</td>
</tr>
<tr>
<td>able</td>
<td>might</td>
<td>could</td>
<td>kind</td>
<td>cap</td>
<td>could</td>
</tr>
<tr>
<td>possible</td>
<td>potential</td>
<td>maybe</td>
<td>probable</td>
<td>probable</td>
<td>potential</td>
</tr>
<tr>
<td>military</td>
<td>domestic</td>
<td>naval</td>
<td>combat</td>
<td>parade</td>
<td>naval</td>
</tr>
<tr>
<td>financial</td>
<td>strategic</td>
<td>economic</td>
<td>money</td>
<td>loan</td>
<td>economic</td>
</tr>
<tr>
<td>spark</td>
<td>leap</td>
<td>inspire</td>
<td>spike</td>
<td>fire</td>
<td>inspire</td>
</tr>
<tr>
<td>freight</td>
<td>automobile</td>
<td>cargo</td>
<td>cargo</td>
<td>train</td>
<td>cargo</td>
</tr>
<tr>
<td>negotiate</td>
<td>eliminate</td>
<td>discuss</td>
<td>diplomacy</td>
<td>bargain</td>
<td>discuss</td>
</tr>
<tr>
<td>stanford</td>
<td>penn</td>
<td>harvard</td>
<td>harvard</td>
<td>california</td>
<td>harvard</td>
</tr>
<tr>
<td>absent</td>
<td>conclusion</td>
<td>without</td>
<td>baby</td>
<td>without</td>
<td></td>
</tr>
<tr>
<td>configuration</td>
<td>installation</td>
<td>setup</td>
<td>setup</td>
<td>network</td>
<td>setup</td>
</tr>
<tr>
<td>grasp</td>
<td>reflect</td>
<td>comprehend</td>
<td>seize</td>
<td>hold</td>
<td>comprehend</td>
</tr>
<tr>
<td>wreck</td>
<td>sudden</td>
<td>crash</td>
<td>crash</td>
<td>damage</td>
<td>crash</td>
</tr>
<tr>
<td>dame</td>
<td>echo</td>
<td>sin</td>
<td>donna</td>
<td>queen</td>
<td>sir</td>
</tr>
<tr>
<td>curry</td>
<td>rice</td>
<td>stephen</td>
<td>soup</td>
<td>unfortunately</td>
<td>stephen</td>
</tr>
<tr>
<td>revelation</td>
<td>prophecies</td>
<td>discovery</td>
<td>discovery</td>
<td>discovery</td>
<td></td>
</tr>
<tr>
<td>spouse</td>
<td>mutual</td>
<td>wife</td>
<td>wife</td>
<td>wife</td>
<td></td>
</tr>
<tr>
<td>main</td>
<td>land</td>
<td>main</td>
<td>continent</td>
<td>america</td>
<td>continental</td>
</tr>
<tr>
<td>material</td>
<td>formula</td>
<td>content</td>
<td>metal</td>
<td>content</td>
<td></td>
</tr>
<tr>
<td>oppose</td>
<td>accuse</td>
<td>defend</td>
<td>contrary</td>
<td>conflict</td>
<td>defend</td>
</tr>
<tr>
<td>zombie</td>
<td>horror</td>
<td>vampire</td>
<td>corpse</td>
<td>ghoul</td>
<td>vampire</td>
</tr>
<tr>
<td>battlefield</td>
<td>defeat</td>
<td>military</td>
<td>warfare</td>
<td>peace</td>
<td>military</td>
</tr>
<tr>
<td>dodge</td>
<td>zoom</td>
<td>avoid</td>
<td>escape</td>
<td>slip</td>
<td>avoid</td>
</tr>
<tr>
<td>triangle</td>
<td>next</td>
<td>circle</td>
<td>trinity</td>
<td>square</td>
<td>circle</td>
</tr>
<tr>
<td>cancer</td>
<td>risk</td>
<td>diabetes</td>
<td>diagnosis</td>
<td>cure</td>
<td>diabetes</td>
</tr>
<tr>
<td>heroin</td>
<td>cocaine</td>
<td>drug</td>
<td>yes</td>
<td>cocaine</td>
<td>yes</td>
</tr>
<tr>
<td>velocity</td>
<td>boundary</td>
<td>speed</td>
<td>speed</td>
<td>fast</td>
<td>speed</td>
</tr>
<tr>
<td>cash</td>
<td>sell</td>
<td>money</td>
<td>money</td>
<td>check</td>
<td>money</td>
</tr>
<tr>
<td>ammunition</td>
<td>deploy</td>
<td>artillery</td>
<td>artillery</td>
<td>bomb</td>
<td>artillery</td>
</tr>
<tr>
<td>nixon</td>
<td>chapman</td>
<td>reagan</td>
<td>nelson</td>
<td>presidency</td>
<td>reagan</td>
</tr>
<tr>
<td>analysis</td>
<td>strategy</td>
<td>study</td>
<td>examination</td>
<td>report</td>
<td>study</td>
</tr>
<tr>
<td>Cue Words</td>
<td>Word2vec</td>
<td>GPT2</td>
<td>GPT3-embedding</td>
<td>GPT3 (prompt curie)</td>
<td>GPT3 (prompt davinci)</td>
</tr>
<tr>
<td>-----------</td>
<td>----------</td>
<td>------</td>
<td>----------------</td>
<td>---------------------</td>
<td>-----------------------</td>
</tr>
<tr>
<td>clarity</td>
<td>pursuit</td>
<td>insight</td>
<td>clear</td>
<td>problem</td>
<td>insight</td>
</tr>
<tr>
<td>discretion</td>
<td>hypothesis</td>
<td>privacy</td>
<td>caution</td>
<td>mercy</td>
<td></td>
</tr>
<tr>
<td>peanut</td>
<td>cinnamon</td>
<td>chocolate</td>
<td>corn</td>
<td>almond</td>
<td>chocolate</td>
</tr>
<tr>
<td>blunt</td>
<td>bounce</td>
<td>frank</td>
<td>brutal</td>
<td>delicate</td>
<td>frank</td>
</tr>
<tr>
<td>nursery</td>
<td>maid</td>
<td>kindergarten</td>
<td>nest</td>
<td>baby</td>
<td>kindergarten</td>
</tr>
<tr>
<td>awhile</td>
<td>sometime</td>
<td>sometime</td>
<td>sometime</td>
<td>sometime</td>
<td>sometime</td>
</tr>
<tr>
<td>spoon</td>
<td>slice</td>
<td>wooden</td>
<td>scoop</td>
<td>fork</td>
<td>wooden</td>
</tr>
<tr>
<td>teddy</td>
<td>bear</td>
<td>johnny</td>
<td>Edgar</td>
<td>toy</td>
<td>johnny</td>
</tr>
<tr>
<td>purpose</td>
<td>temporary</td>
<td>goal</td>
<td>intention</td>
<td>Meaning</td>
<td>goal</td>
</tr>
<tr>
<td>distinguish</td>
<td>adapt</td>
<td>identify</td>
<td>identify</td>
<td>identify</td>
<td>identify</td>
</tr>
<tr>
<td>renaissance</td>
<td>revival</td>
<td>antique</td>
<td>Florence</td>
<td>art</td>
<td>antique</td>
</tr>
<tr>
<td>infrared</td>
<td>tool</td>
<td>fall</td>
<td>flag</td>
<td>shoe</td>
<td>fire</td>
</tr>
<tr>
<td>curtain</td>
<td>towel</td>
<td>door</td>
<td>carpet</td>
<td>theater</td>
<td>door</td>
</tr>
<tr>
<td>legion</td>
<td>trinity</td>
<td>army</td>
<td>troop</td>
<td>army</td>
<td>army</td>
</tr>
<tr>
<td>mind</td>
<td>Betty</td>
<td>servant</td>
<td>girl</td>
<td>man</td>
<td>servant</td>
</tr>
<tr>
<td>brighton</td>
<td>bristol</td>
<td>Liverpool</td>
<td>Bry</td>
<td>Bristol</td>
<td>Liverpool</td>
</tr>
<tr>
<td>constitute</td>
<td>quantity</td>
<td>facilitate</td>
<td>declare</td>
<td>Annul</td>
<td>Facilitate</td>
</tr>
<tr>
<td>royal</td>
<td>western</td>
<td>British</td>
<td>King</td>
<td>Family</td>
<td>British</td>
</tr>
<tr>
<td>speech</td>
<td>appearance</td>
<td>Talanta</td>
<td>Spoken</td>
<td>Spoke</td>
<td></td>
</tr>
<tr>
<td>wendy</td>
<td>Betty</td>
<td>Jennifer</td>
<td>Walter</td>
<td>Barbara</td>
<td>Jennifer</td>
</tr>
<tr>
<td>teammate</td>
<td>opponent</td>
<td>colleague</td>
<td>colleague</td>
<td>Friend</td>
<td>Colleague</td>
</tr>
<tr>
<td>whatever</td>
<td>impress</td>
<td>Cornelius</td>
<td>Perkin</td>
<td>Party</td>
<td>Cornelius</td>
</tr>
<tr>
<td>lloyd</td>
<td>Johnny</td>
<td>Raymond</td>
<td>Pink</td>
<td>William</td>
<td></td>
</tr>
<tr>
<td>prototype</td>
<td>slate</td>
<td>Concept</td>
<td>Schematic</td>
<td>Original</td>
<td>Concept</td>
</tr>
<tr>
<td>consider</td>
<td>Whether</td>
<td>Think</td>
<td>Think</td>
<td>Important</td>
<td>Think</td>
</tr>
<tr>
<td>expose</td>
<td>refer</td>
<td>Provide</td>
<td>Exhibit</td>
<td>Reveal</td>
<td>Provide</td>
</tr>
<tr>
<td>impose</td>
<td>dispute</td>
<td>Establish</td>
<td>Establish</td>
<td>Ban</td>
<td>Establish</td>
</tr>
<tr>
<td>fall</td>
<td>Together</td>
<td>Spring</td>
<td>Drop</td>
<td>Autumn</td>
<td>Spring</td>
</tr>
<tr>
<td>slow</td>
<td>Serious</td>
<td>Speed</td>
<td>Late</td>
<td>Sad</td>
<td>Speed</td>
</tr>
<tr>
<td>appear</td>
<td>Often</td>
<td>Seem</td>
<td>Occur</td>
<td>Person</td>
<td>Seem</td>
</tr>
<tr>
<td>kindle</td>
<td>Artwork</td>
<td>Amazon</td>
<td>Kindle</td>
<td>Amazon</td>
<td></td>
</tr>
<tr>
<td>rotten</td>
<td>Moth</td>
<td>Horrible</td>
<td>Ruin</td>
<td>Apple</td>
<td>Horrible</td>
</tr>
<tr>
<td>apply</td>
<td>Offer</td>
<td>Give</td>
<td>Use</td>
<td>Job</td>
<td></td>
</tr>
<tr>
<td>kindergarten</td>
<td>Elementary</td>
<td>Elementary</td>
<td>Nursery</td>
<td>School</td>
<td>Elementary</td>
</tr>
<tr>
<td>improve</td>
<td>Develop</td>
<td>Enhance</td>
<td>Increase</td>
<td>Enhancement</td>
<td>Enhance</td>
</tr>
<tr>
<td>jing</td>
<td>Tab</td>
<td>Peer</td>
<td>Sound</td>
<td>Although</td>
<td></td>
</tr>
<tr>
<td>develop</td>
<td>Enhance</td>
<td>Create</td>
<td>Create</td>
<td>Growth</td>
<td>Create</td>
</tr>
<tr>
<td>reduce</td>
<td>Consume</td>
<td>Decrease</td>
<td>Decrease</td>
<td>Number</td>
<td>Decrease</td>
</tr>
<tr>
<td>hundred</td>
<td>Thousand</td>
<td>Thousand</td>
<td>Thousand</td>
<td>Thousand</td>
<td>Thousand</td>
</tr>
<tr>
<td>kitchen</td>
<td>Puppy</td>
<td>Puppy</td>
<td>Puppy</td>
<td>Cat</td>
<td>Puppy</td>
</tr>
<tr>
<td>ding</td>
<td>Deed</td>
<td>Almost</td>
<td>Bell</td>
<td>Door</td>
<td>Almost</td>
</tr>
<tr>
<td>bring</td>
<td>Give</td>
<td>Almost</td>
<td>Give</td>
<td>Almost</td>
<td></td>
</tr>
<tr>
<td>assume</td>
<td>Ignore</td>
<td>Imagine</td>
<td>Suppose</td>
<td>Doubt</td>
<td>Imagine</td>
</tr>
<tr>
<td>identify</td>
<td>Convert</td>
<td>Determine</td>
<td>Detect</td>
<td>Animal</td>
<td>Determine</td>
</tr>
<tr>
<td>thousand</td>
<td>Hundred</td>
<td>Twenty</td>
<td>Thousand</td>
<td>Thousand</td>
<td>Thousand</td>
</tr>
<tr>
<td>watershed</td>
<td>Millennium</td>
<td>Historic</td>
<td>Historical</td>
<td>Historical</td>
<td>Historcal</td>
</tr>
<tr>
<td>river</td>
<td>Circuit</td>
<td>Profound</td>
<td>Riot</td>
<td>Riot</td>
<td>Profound</td>
</tr>
<tr>
<td>establish</td>
<td>Obtain</td>
<td>Provide</td>
<td>Setup</td>
<td>Set</td>
<td>Provide</td>
</tr>
<tr>
<td>participate</td>
<td>Comlete</td>
<td>Join</td>
<td>Contribute</td>
<td>Support</td>
<td>Join</td>
</tr>
<tr>
<td>obtain</td>
<td>Contribute</td>
<td>Acquire</td>
<td>Acquire</td>
<td>Receive</td>
<td>Acquire</td>
</tr>
<tr>
<td>seel</td>
<td>Aim</td>
<td>Establish</td>
<td>Establishment</td>
<td>Agreement</td>
<td>Establish</td>
</tr>
<tr>
<td>settle</td>
<td>Aim</td>
<td>Establish</td>
<td>Establish</td>
<td>Agreement</td>
<td>Establish</td>
</tr>
<tr>
<td>shred</td>
<td>Soar</td>
<td>Scrap</td>
<td>Scrap</td>
<td>Scrap</td>
<td>Scrap</td>
</tr>
<tr>
<td>bloodstream</td>
<td>Stigma</td>
<td>Massacre</td>
<td>Massacre</td>
<td>Murder</td>
<td>Massacre</td>
</tr>
<tr>
<td>succeed</td>
<td>Achieve</td>
<td>Achieve</td>
<td>Accomplish</td>
<td>Hard</td>
<td>Achieve</td>
</tr>
<tr>
<td>sled</td>
<td>Surf</td>
<td>Snow</td>
<td>Snow</td>
<td>Toboggan</td>
<td>Snow</td>
</tr>
<tr>
<td>shed</td>
<td>Burn</td>
<td>Give</td>
<td>Yard</td>
<td>Tear</td>
<td>Give</td>
</tr>
<tr>
<td>quantify</td>
<td>Constitute</td>
<td>Calculate</td>
<td>Measure</td>
<td>Measure</td>
<td>Calculate</td>
</tr>
<tr>
<td>sacred</td>
<td>Forbidden</td>
<td>Ancient</td>
<td>Divine</td>
<td>Temple</td>
<td>Ancient</td>
</tr>
<tr>
<td>fred</td>
<td>Samson</td>
<td>Robert</td>
<td>Gerald</td>
<td>Shirt</td>
<td>Robert</td>
</tr>
<tr>
<td>slave</td>
<td>Colonize</td>
<td>Colony</td>
<td>Colony</td>
<td>Colony</td>
<td>Colony</td>
</tr>
<tr>
<td>deepen</td>
<td>Initiate</td>
<td>Strengthen</td>
<td>Depth</td>
<td>Voice</td>
<td>Strengthen</td>
</tr>
<tr>
<td>quality</td>
<td>Compete</td>
<td>Eligible</td>
<td>Eligible</td>
<td>Pass</td>
<td>Eligible</td>
</tr>
<tr>
<td>kindness</td>
<td>Covenan</td>
<td>Likewise</td>
<td>Familiar</td>
<td>Relative</td>
<td>Likewise</td>
</tr>
<tr>
<td>acquire</td>
<td>Retain</td>
<td>Obtain</td>
<td>Obtain</td>
<td>Property</td>
<td>Obtain</td>
</tr>
<tr>
<td>approve</td>
<td>Reject</td>
<td>Agree</td>
<td>Confirm</td>
<td>Positive</td>
<td>Agree</td>
</tr>
<tr>
<td>hatred</td>
<td>Better</td>
<td>Anger</td>
<td>Spite</td>
<td>Machine</td>
<td>Anger</td>
</tr>
<tr>
<td>have</td>
<td>Could</td>
<td>Make</td>
<td>Could</td>
<td>Take</td>
<td>Make</td>
</tr>
<tr>
<td>talk</td>
<td>Think</td>
<td>Aware</td>
<td>Learn</td>
<td>Think</td>
<td></td>
</tr>
<tr>
<td>take</td>
<td>Need</td>
<td>Give</td>
<td>Put</td>
<td>Break</td>
<td>Give</td>
</tr>
<tr>
<td>make</td>
<td>Leave</td>
<td>Take</td>
<td>Away</td>
<td>Look</td>
<td></td>
</tr>
<tr>
<td>come</td>
<td>Far</td>
<td>Make</td>
<td>Bring</td>
<td>Approach</td>
<td>Make</td>
</tr>
</tbody>
</table>

102
| 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: unemployed Cue: uncertainly; Guess: very; heavy; Guess: light | massive |
| risky | Cue: building; Guess: house Cue: let; Guess: allow Cue: risky; Guess: dangerous | dangerous |
| grand | Cue: nearly; Guess: edge Cue: hot; Guess: cold; easy; Guess: form Cue: grand; Guess: dangerous | 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; pretty; Guess: bench Cue: nepal; Guess: Nepal | new |
| healthy | Cue: number; Guess: word Cue: come; Guess: here Cue: hot; Guess: cold; healthy; Guess: straight | sick |
| credible | Cue: see; Guess: here Cue: business; Guess: class Cue: let; Guess: allow Cue: credible; Guess: trustworthy | good |
| oriental | Cue: come; Guess: here Cue: positive; Guess: Happy Cue: professional; Guess: career Cue: oriental; Guess: love | compelling |
| flat | Cue: professional; Guess: person Cue: story; Guess: narrative Cue: let; Guess: beats 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: soft | whisper |
| southeastern | Cue: nearly; Guess: half Cue: future; Guess: mars Cue: come; Guess: here Cue: southeastern; Guess: many | 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 |
| mystic | Cue: unfortunately; Guess: unlockly Cue: let; Guess: allow Cue: professional; Guess: work Cue: mystic; Guess: sex | cognitive |
| fellish | Cue: professional; Guess: special Cue: positive; Guess: negative Cue: remember; Guess: together Cue: rural; Guess: farm | spiritual |
| exclusive | Cue: unfortunately; Guess: unlockly Cue: building; Guess: skyscraper Cue: remember; Guess: forget Cue: exclusive; Guess: inclusive | unique |
| soviet | Cue: second; Guess: recall Cue: easy; Guess: game Cue: soviet; Guess: union | republican |
| presidential | Cue: story; Guess: time Cue: professional; Guess: lawyer Cue: easy; Guess: exam Cue: attentive; Guess: energetic |
| impressive | Cue: second; Guess: first; pretty; Guess: secretive; Guess: impressive; Guess: wonderful | enthusiastic |
| ridiculous | Cue: see; Guess: look Cue: come; Guess: back; Guess: nearly; Guess: there Cue: ridiculous; Guess: funny | absurd |
| ordinary | Cue: slightly; Guess: little Cue: see; Guess: tell Cue: building; Guess: door Cue: ordinary; Guess: common |
| obnoxious | Cue: story; Guess: love Cue: see; Guess: first Cue: obnoxious; Guess: Jonathan | normal |
| pascal | Cue: remember; Guess: dream Cue: building; Guess: door Cue: hot; Guess: dog Cue: pascal; Guess: Jonathan | arrogant |
| rebellious | Cue: let; Guess: allow Cue: pretty; Guess: good Cue: positive; Guess: negative Cue: rebellious; Guess: clever |
| ingenuous | Cue: second; Guess: position Cue: pretty; Guess: nearly; Guess: close; Guess: ingenious; Guess: clever |
| exponential | Cue: come; Guess: back; positive; Guess: negative; second; Guess: minute; exponential; Guess: exponential |
| simplistic | Cue: remember; Guess: recall Cue: second; Guess; first; unfortunately; Guess: sadly; simplistic; Guess: linear |
| puppy | Cue; future; Guess; past; second; Guess; first; reminder; Guess; recall Cue; puppy; able; Guess | straightforward |
| venerable | Cue: nearly; Guess: almost; come; Guess: visit; second; Guess; position Cue: venerable; Guess | dog |
| sluggish | Cue: business; Guess: card Cue: hot; Guess; cold; see; Guess; look Cue; sluggish; Guess | legendary |
| intangible | Cue: easy; Guess: exam Cue: future; Guess: past; story; Guess: telling; Guess: intangible; Guess | problem |
| mammalian | Cue: future; Guess: answer Cue: building; Guess: base Cue: come; Guess: run; mammalian; Guess | insect |
| revolutionary | Cue: number; Guess; word Cue: nearly; Guess; almost; hot; Guess; shower; revolution; Guess | radical |
| vocal | Cue: come; Guess: home Cue: business; Guess: play Cue: pretty; Guess; beauty Cue: schematic; Guess | sing |
| schematic | Cue; come; Guess; home Cue; business; Guess; play; pretty; Guess; beauty; schematic; Guess | diagram |
| suspicious | Cue; second; Guess; minute; easy; Guess; form Cue; future; Guess; base Cue; suspicious; Guess | crime |
| nationwide | Cue; unfortunately; Guess; not Cue; building; Guess; base; future; Guess; answer; nationwide; Guess | local |
| synthetic | Cue; let; Guess; allow; number; Guess; word Cue; hot; Guess; cold; synthetic; Guess | neural |
| white | Cue; remember; Guess; dream Cue; pretty; Guess; nice; see; Guess; here; white; Guess | black |
| able | easy; Guess; hard; story; Guess; time; remember; Guess; recall; able; Guess | cap |
| possible | Cue; building; Guess; base; let; Guess; allow; slightly; Guess; hardly; possible; Guess | probable |
| military | easy; Guess; hard; let; Guess; allow; pretty; Guess; nice; Cue; military; Guess | potential |
| financial | Cue; unfortunately; Guess; not; building; Guess; base; future; Guess; answer; nation; Guess | loan |
| apt | easy; Guess; hard; hot; Guess; cold; see; Guess; look; easy; Guess; cold; Guess; easy; Guess; cold; Guess | fire |
| freight | easy; Guess; hard; professional; Guess; professor Cue; hot; Guess; cold; freight; Guess | train |
| negotiate | easy; Guess; negative Cue; hot; Guess; water Cue; see; Guess; look; negotiate; Guess | bargain |
| stamford | Cue; second; Guess; first Cue; nearly; Guess; close; Guess; come; Guess; back; Guess; stamford; Guess | california |
| absent | Cue; professional; Guess; professor Cue; slightly; Guess; slow; Cue; remember; Guess; recall Cue; absent; Guess | baby |
| configuration | Cue; come; Guess; back; Guess; hot; Guess; food; Cue; building; Guess; body; Cue; configuration; Guess | without |
| grasp | Cue; number; Guess; game Cue; professional; Guess; doctor; Guess; unfortunately; Guess; sorry; Cue; grasp; Guess | hold |
| wreck | nearly; Guess; close; remember; Guess; forget; Cue; positive; Guess; negative; Cue; wreck; Guess | damage |
| dame | pretty; Guess; liar Cue; come; Guess; home Cue; slightly; Guess; off; Cue; dame; Guess | queen |
| curry | hot; Guess; cold; see; Guess; look; positive; Guess; negative; curry; Guess | sir |
| revelation | Cue; come; Guess; run Cue; remember; Guess; recall; slightly; Guess; off; Cue; revelation; Guess | fortunately |
| spouse | Cue; business; Guess; play Cue; building; Guess; tower Cue; unfortunately; Guess; sadly; spouse; Guess | discovery |
| mainland | Cue; story; Guess; story; Guess; nearly; Guess; sad; mainland; Guess | discovery |
| material | Cue; future; Guess; far Cue; remember; Guess; together; positive; Guess; attitude Cue; material; Guess | world |
| oppose | Cue; pretty; Guess; beautiful Cue; slightly; Guess; better Cue; nearly; Guess; close Cue; oppose; Guess | conflict |
| zombie | Cue; positive; Guess; negative Cue; professional; Guess; boss Cue; nearly; Guess; almost Cue; zombie; Guess | ghoul |
| battlefield | Cue; easy; Guess; exam Cue; see; Guess; hear Cue; positive; Guess; negative Cue; battlefield; Guess | peace |
| dodgy | Cue; easy; Guess; cool Cue; slightly; Guess; bigger Cue; positive; Guess; negative; dodgy; Guess | slip |
| triangle | Cue; hot; Guess; cold; easy; Guess; form Cue; unfortunately; Guess; not; triangle; Guess | avoid |
| cancer | Cue; hot; Guess; cold; professional; Guess; doctor; Cue; professional; Guess; cancer; Guess | square |
| heroin | Cue; building; Guess; construction Cue; slightly; Guess; cold; professional; Guess; work; heroin; Guess | circle |
| velocity | Cue; come; Guess; home Cue; positive; Guess; valence Cue; number; Guess; singular Cue; velocity; Guess; check |
| cash | Cue; business; Guess; role Cue; professional; Guess; person Cue; nearly; Guess; close; cash; Guess | bomb |
| ammunition | Cue; hot; Guess; cold; come; Cue; here; nearly; Guess; almost Cue; ammunition; Guess; bomb |
| nixon | Cue; business; Guess; paragraph Cue; easy; Guess; form Cue; come; Guess; back; Guess; nixon; Guess | artillery |
| analysis | Cue; slightly; Guess; more Cue; see; Guess; watch Cue; positive; Guess; negative; analysis; Guess | report |

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 the cues were generated by the prompt-davicin model and the davinci-promt model are presented in the two right columns.
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.
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 responses 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.
Figure S75. Conversation. Results of the Turing test for human judges
(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.
Figure S77. **Conversation.** Results of the Turing test for human judges below age 35.

(a) Confusion matrix (b) Top-1 accuracy.
Figure S78. **Conversation.** Results of the Turing test for human judges between age 35 and 45. (a) Confusion matrix (b) Top-1 accuracy.
Figure S79. **Conversation.** Results of the Turing test for human judges above age 45. (a) Confusion matrix (b) Top-1 accuracy.
Figure S80. **Conversation.** Results of the Turing test for male human judges.

(a) Confusion matrix (b) Top-1 accuracy.
Figure S81. **Conversation.** Results of the Turing test for female human judges
(a) Confusion matrix (b) Top-1 accuracy.
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.
Figure S83. Conversation. Results of the Turing test for human judges with highest education level of Bachelor. 
(a) Confusion matrix (b) Top-1 accuracy.
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.
Figure S85. **Conversation.** Demographic information for the human judges.
(a) Age. (b) Gender (c) Education level.
Figure S86. Conversation. Results of the Turing test for human judges from data collected on AMT
(a) Confusion matrix (b) Top-1 accuracy.
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

Submit ➤

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

<table>
<thead>
<tr>
<th>Case</th>
<th>Blenderbot</th>
<th>GPT3 davinci</th>
<th>GPT3 curie</th>
</tr>
</thead>
<tbody>
<tr>
<td>H-AI conversation</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>AI-AI conversation</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Case</th>
<th>[%] classified as male</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall (Average over all speakers)</td>
<td>60%</td>
</tr>
<tr>
<td>Human speakers</td>
<td>64%</td>
</tr>
<tr>
<td>AI speakers</td>
<td>57%</td>
</tr>
<tr>
<td>Speakers classified as human by the AMT judge</td>
<td>53%</td>
</tr>
<tr>
<td>Speakers classified as AI by the AMT judge</td>
<td>68%</td>
</tr>
</tbody>
</table>
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 hahahaha, 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
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. What do you think?
B: I think their dedication to liberty and justice made them great. What do you think?
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