

# Turing-Like Tests for Language and Vision Demonstrate a Narrowing Gap Between Humans and AI

Mengmi Zhang<sup>1,2,3\*</sup>, Elisa Pavarino<sup>4,5\*</sup>, Xiao Liu<sup>1,2\*</sup>, Giorgia Dellaferrera<sup>3,6,7</sup>, Ankur Sikarwar<sup>1,2\*</sup>, Caishun Chen<sup>2\*</sup>, Marcelo Armendariz<sup>3,5\*</sup>, Noga Mudrik<sup>8\*</sup>, Prachi Agrawal<sup>9\*</sup>, Spandan Madan<sup>3,5,10\*</sup>, Mranmay Shetty<sup>11</sup>, Andrei Barbu<sup>3,12</sup>, Haochen Yang<sup>10</sup>, Tanishq Kumar<sup>10</sup>, Shui'Er Han<sup>1,2</sup>, Aman Raj Singh<sup>9</sup>, Meghna Sadwani<sup>13</sup>, Stella Dellaferrera<sup>14</sup>, Michele Pizzochero<sup>10,15</sup>, Brandon Tang<sup>16</sup>, Yew-Soon Ong<sup>1,2</sup>, Hanspeter Pfister<sup>10</sup>, and Gabriel Kreiman<sup>3,5</sup>

\* Equal contribution

<sup>1</sup> College of Computing and Data Science, Nanyang Technological University, Singapore,

<sup>2</sup> Agency for Science, Technology and Research, Singapore,

<sup>3</sup> Center for Brains, Minds, and Machines, USA,

<sup>4</sup> Department of Neurobiology, Harvard University, USA,

<sup>5</sup> Children's Hospital, Harvard Medical School, USA,

<sup>6</sup> IBM Research - Zürich, Rüschlikon, Switzerland,

<sup>7</sup> Institute of Neuroinformatics, University of Zürich and ETH Zürich, Zürich, Switzerland,

<sup>8</sup> Biomedical Engineering, Johns Hopkins University, USA,

<sup>9</sup> Birla Institute of Technology and Science, Pilani, India,

<sup>10</sup> School of Engineering and Applied Sciences, Harvard University, USA,

<sup>11</sup> Symbiosis Institute of Technology, India

<sup>12</sup> CSAIL, MIT, USA

<sup>13</sup> Jawaharlal Nehru Medical College, India,

<sup>14</sup> University of Turin, Italy,

<sup>15</sup> Department of Physics & Centre for Artificial Intelligence, University of Bath, UK

<sup>16</sup> Harvard Medical School, USA,

Address correspondence to: gabriel.kreiman@tch.harvard.edu

## Abstract

As AI becomes increasingly embedded in daily life, ascertaining whether an agent is human is critical. Here, we systematically benchmark AI's ability to imitate humans in three language tasks (image captioning, word association, conversation) and three vision tasks (color estimation, object detection, attention prediction), collecting data from n=636 humans and 37 AI agents. Next, we conducted 72,191 Turing-like tests with 1,916 human judges and 10 AI judges. Current AIs are approaching the ability to convincingly impersonate humans and deceive human judges in both language and vision. Even simple AI judges outperformed humans in distinguishing AI from human responses. Imitation ability showed minimal correlation with conventional AI performance metrics, suggesting that passing as human is an important independent evaluation criterion. The large-scale Turing-like datasets and metrics introduced here offer valuable benchmarks for assessing human-likeness in AI and highlight the importance of rigorous, quantitative imitation tests for AI development.

## 14 Introduction

15 Current language and computer vision models excel in a wide range of tasks, such as image captioning  
16 [1], [2], [3], [4], [67], text generation [68], [5], [6], [7], [0], [8], [9], object recognition [67], [10], [11],  
17 and attention prediction [12], [13], [14]. Developing state-of-the-art algorithms goes hand-in-hand  
18 with the creation of better and more precise ways of assessing their performance. The performance of  
19 AI algorithms is often defined by comparing their outputs against human ground truth annotations.  
20 Such metrics are particularly prevalent in computer vision, such as in object recognition tasks [15].  
21 Additionally, several metrics have been adapted from natural language processing tasks to image  
22 captioning, such as BLEU [16], THUMB [17], and METEOR [18] (see also [19]). Generative AI  
23 models are notoriously difficult to evaluate due to the inherent ambiguities of creating novel content  
24 [20]. Human evaluators are often recruited to assess the quality of sentiment, semantic relevance,  
25 reasoning abilities, or emotional valence on text generated by large language models [7], [20], [21], [22],  
26 [23], [24], [25], [26], [27], [28], [29].

27 The Turing test, also known as the “imitation game”, was proposed by Alan Turing in 1950 as a  
28 way of assessing a machine’s ability to exhibit intelligent behaviors indistinguishable from those shown  
29 by humans [30]. In the imitation game, a machine tries to pass as a human during a conversation, and  
30 a human judge determines whether they are interacting with a human agent or a machine [30] (**Fig.**  
31 **1A**). Since its inception, the Turing test has generated extensive controversy and discussion. Several  
32 notable arguments include Searle’s Chinese room thought experiment [31], Block’s behaviorism [32],  
33 Harnad’s Total Turing Test [33], Watt’s Inverted Turing Test [34], Damassino’s Questioning Turing  
34 Test [35] and Sejnowski’s Reverse Turing Test [36]. For example, Searle’s Chinese Room thought  
35 experiment suggested that a system can manipulate symbols without true understanding, challenging  
36 the notion that passing the Turing Test implies intelligence [31]. Similarly, Block critiqued behaviorist  
37 views, arguing that a machine with a vast lookup table could pass the Turing test without genuine  
38 cognition or generalization [32].

39 In parallel with the unbounded optimistic attitudes towards AI in the 1960s and the sober realiza-  
40 tion of the immense difficulties in AI afterward, many in the scientific community have shifted away  
41 from the question of whether the Turing test is a valid and meaningful measure of intelligence (e.g.,  
42 [37]) to focus instead on average performance metrics for AI algorithms. Distinct from these discus-  
43 sions, the purpose of our work is *not* to argue in favor or against Turing tests as a measure of general  
44 intelligence. Instead, we consider Turing-like tests as a quantitative evaluation of how well current AIs  
45 can imitate humans [23].

46 Many studies have compared how closely AI model responses align with human responses [38] (see  
47 Section S4). This type of comparison is straightforward in classic visual object recognition tasks and  
48 is widely used in traditional computer vision benchmarks [38]. However, in complex and open-ended  
49 tasks such as many of the ones in the current work, these metrics are less suitable or undefined.  
50 These challenges are similar to those seen in evaluating generative models [68], [9]. In the vision tasks

51 examined in this work, multiple responses may be equally valid. For example, in an object detection  
52 task, one agent may identify different objects than another agent. Neither may be incorrect, since all  
53 the mentioned objects might be present in the image. Therefore, the emphasis in Turing-like tests is  
54 not on determining whether a response is correct or incorrect. Instead, the focus is on whether it is  
55 possible to distinguish responses generated by humans from those generated by machines.

56 The key target of the original Turing test focused on conversations. Generating human-like text  
57 during conversations remains a daunting challenge for AI with exciting progress. There have been  
58 numerous early attempts at generating restricted topics during conversations, such as Colby’s PARRY  
59 simulating a paranoid schizophrenic and Weizenbaum’s ELIZA simulating a psychiatrist. However,  
60 none of these models have come close to unrestricted Turing tests. Advances in large language models  
61 [68], [7], [9], [21], [39], [40], [41], [42] have led news and social media to produce anecdotal claims about  
62 current AI algorithms being sentient in conversations. Yet, few studies rigorously and quantitatively  
63 assessed AI algorithms in their ability to imitate humans in conversations [23], [24], [26], [29], [43].

64 As AI algorithms continue to blossom in the real world, it is becoming increasingly important  
65 for lay people, law professionals, clinicians, politicians, and other experts to ascertain whether the  
66 agent they are interacting with is a human or not. For example, the inability to distinguish real news  
67 from AI-generated fake news or DeepFakes can have disastrous implications for electoral campaigns.  
68 Additionally, criminals are increasingly using AI-generated conversations to make phishing scams more  
69 convincing, and such scams have even started overtaking other types of physical crimes [44]. To  
70 mitigate these issues, the AI community has started developing models to discern whether the generated  
71 perceptual inputs come from humans or AIs [45], including DeepFake detection on images [46], and  
72 GROVER models on fake news detection [47]. Cui *et al.* proposed a learned critique model acting as  
73 a human judge to perform a Turing Test in image captioning tasks [19].

74 Here we set out to systematically and quantitatively investigate the ability of current language and  
75 vision algorithms to imitate humans. To this end, we present an integrative benchmark encompassing a  
76 wide range of standard and well-established AI tasks across both language and vision. Motivated by the  
77 recognition that each AI task measures only a narrow aspect of human intelligence and is incomplete  
78 in isolation, we examine cognitive imitation abilities through 6 different tasks. This approach provides  
79 a holistic assessment of the ability to emulate human-like performance across various domains. In  
80 vision, color estimation, object detection, and visual attention were specifically chosen because they  
81 directly reflect core aspects of human visual perception and decision-making in daily tasks. We also  
82 note that some of these vision tasks required language comprehension (see General Considerations  
83 in the **Methods** section). Similarly, among language tasks, word association, image captioning, and  
84 conversation were chosen because they capture fundamental aspects of linguistic cognition used in  
85 daily activities. Although three tasks are labeled as “visual” tasks, and three tasks are labeled as  
86 “language” tasks, several tasks include both modalities (e.g., image captioning).

87 We start by creating an extensive benchmark dataset of 36,499 answers from 636 human participants  
88 and 37 AI models in 3 language tasks (Image captioning, Word associations, Conversations), and 3

99 vision tasks (Color estimation, Object detection, and Attention prediction). These tasks were chosen to  
90 span a typical and wide range of real-world applications (**Fig. 1B, C, D, Fig. 3A, B, C, Table S1**).  
91 Next, we systematically evaluated the ability of 1,916 human judges to discern whether task answers  
92 came from humans or AI in Turing-like tests, showing that current algorithms are remarkably adept  
93 at imitating humans under restricted testing conditions. Finally, as a proof-of-principle, we show  
94 that even though human judges may partially struggle to distinguish human answers, simple machine  
95 learning algorithms can serve as AI judges and vastly outperform human judges in Turing-like tests.  
96 Our code for data collection and analysis is available to the community, as are the generated datasets.  
97 To facilitate future research, we provide an open-source benchmarking tool that allows researchers  
98 to evaluate their own models’ detectability using our AI judge framework across five tasks, enabling  
99 standardized assessment of human-likeness in new AI models.

## 100 Results

### 101 Collection of human and AI agent responses in language tasks

102 To conduct Turing-like tests, we first compiled responses from human participants (labeled H through-  
103 out) and AI machine models (henceforth labeled M) in 3 language tasks (**Fig. S1, Methods**). Multiple  
104 controls were introduced in each task to ensure high quality of the responses (**Methods, Table S3**).  
105 These responses were then assessed in Turing-like tests to determine if an independent group of human  
106 judges, or AI judges, was able to differentiate between human and AI responses (**Fig. 1A**). We also  
107 collected basic demographic information about each participant (**Fig. S2, Fig. S4, Fig. S7E-G**).

108 We begin by describing the dataset of human and AI-generated responses for each task. In the  
109 Image captioning task, both human participants and AI models were presented with an image and  
110 asked to generate a descriptive caption (**Fig. S1A**). We collected responses for 1,000 images from 229  
111 participants recruited via Amazon Mechanical Turk (AMT), along with responses from 5 AI models  
112 (see demographic information about the participants in **Fig. S2A, C**). Representative examples of  
113 image captions generated by human participants and each of the five AI models are shown in **Fig. 2A,**  
114 **B**.

115 We considered two closely related versions of the Word association task. In the first version, partici-  
116 pants were presented with a cue word and asked to provide a one-word association (**Fig. S1B1**). In the  
117 second version, participants were presented with a short prompt containing example word associations,  
118 followed by a new cue word, and were again asked to generate a single word association (**Fig. S1B2**).  
119 As the results across the two task versions were similar (**Fig. S9E, F**), we merged the datasets. We  
120 collected responses to 1,500 prompts across both task versions from 40 AMT participants and 5 AI  
121 models. Demographic information about the participants is shown in **Fig. S2E, G**. Representative  
122 responses from human participants and each of the five AI models are shown in **Fig. 2C**.

123 In the Conversation task, two agents engaged in a conversation. Each agent could be a human or

124 an AI model. There were three types of conversations: human-human (H-H), human-machine (H-M),  
125 and machine-machine (M-M). Each conversation had a total of up to 24 entries, 12 from each agent.  
126 To evaluate the effect of longer conversation lengths, a small subset of conversations included extended  
127 exchanges of 48 entries, 24 from each agent. Agents were unaware of whether they were conversing  
128 with another human or with an AI model. Human agents were told that they would engage in a brief  
129 conversation about different topics. Participants performed the task via a public chat platform where  
130 investigators acted as intermediaries to pass messages between agents, or via our custom-developed  
131 mobile app (**Methods**). We collected a total of 510 conversations (each one with 24 entries) from  
132 190 participants. Demographic information about the participants is shown in **Fig. S2I, K**. We  
133 collected 3,845 responses from humans and 9 AI models (**Table S1**). The prompts used for each AI  
134 model are shown in the **Methods** section. Representative conversations are reported in **Fig. 2D** and  
135 **Supplementary Section S1**.

### 136 **Operational definition of Turing-like tests**

137 After collecting responses from both human and machine agents for each task, we recruited an in-  
138 dependent set of participants to carry out Turing-like tests. We refer to our tests as *Turing-like* to  
139 emphasize that they are not identical to the original Turing definitions [30] (**Discussion**). Turing-like  
140 tests were conducted using three recruitment sources: Amazon Mechanical Turk (AMT), Prolific (a  
141 large online crowd-sourcing platform), and in-lab experiments. We presented a single instance of the  
142 answers and asked participants to indicate whether the answer came from a human or an AI agent  
143 (**Fig. 1B, C, D**).

144 We also collected demographic information about the participants as metadata, including age, gen-  
145 der, and educational background (**Fig. S2A, C, E, G, I, K**). The trial order was always randomized  
146 with half of the trials containing responses from human agents, and the other half containing responses  
147 from AI agents. We introduced multiple controls to ensure the quality of the responses collected in  
148 each Turing-like test experiment (**Methods**). Across all tasks, the average accuracy for identifying  
149 machine-only responses  $P(M|M)$  was above 75% on the catch trials (**Table S3**). Additionally, we  
150 also trained computational models to act as AI judges to determine whether a particular answer came  
151 from a human or not (**Methods**).

### 152 **AI models are close to passing three language Turing-like tests**

153 For each Turing-like test, the ground truth response could come from a human agent (H) or a machine  
154 (M). We report confusion matrices where the diagonals show the proportion of trials where judges  
155 indicated H given that the ground truth was H ( $p(H|H)$ ), and the proportion of trials where judges  
156 indicated M given that the ground truth was M ( $p(M|M)$ ) (**Fig. 4**). Non-diagonal elements show  
157  $p(M|H)$  (the probability that agents indicated M when the ground truth was H) and  $p(H|M)$  (the  
158 probability that agents indicated H when the ground truth was M). Entries within each row add

159 up to 1. When comparing different AI models for a given task in terms of their ability to imitate  
160 humans, values closer to 0.5 for  $p(H|H)$  and  $p(M|M)$  indicate greater indistinguishability, and thus  
161 stronger imitation performance. In contrast, when comparing judges (e.g., human versus AI judges, or  
162 human judges of different ages or educational backgrounds), higher probabilities  $p(H|H)$  and  $p(M|M)$   
163 departing from 0.5 indicate better judges. We summarize the confusion matrices in a single number by  
164 defining the overall imitation detectability as  $\frac{1}{2}(p(H|H) + p(M|M))$ , ranging from 0.5 (good imitator)  
165 to 1.0 (poor imitator).

166 We summarize the results of the Turing-like tests for human judges across each task in **Fig. 4A**,  
167 **B**, **C**, averaging across all AI models and all human judge demographics. In the Image captioning  
168 task, human judges correctly identified human captions as human ( $P(H|H)$ ) 67% of the time, and  
169 AI captions as AI ( $P(M|M)$ ) 46% of the time, (both significantly different from chance, one-sample  
170 Wilcoxon signed-rank tests,  $P(H|H) : z = 13.03, p < 0.001$ ;  $P(M|M) : z = 4.88, p < 0.001$ , **Fig. 4A**).  
171 We were concerned that AMT participants could be poor judges. Therefore, we repeated the Turing-  
172 like experiments in person in our lab (**Fig. S3A**, **Fig. S4A-D**). Overall, the results from the in-lab  
173 cohort were highly consistent with those obtained online (online judges ( $n = 293$ ): mean detectability  
174 =  $0.57 \pm 0.11$ ; in-lab judges ( $n = 11$ ): mean detectability =  $0.57 \pm 0.10$ ; Mann-Whitney U test:  
175  $U = 1638.5, p = 0.93$ ).

176 Performance varied substantially across AI models in the Image captioning task (**Fig. 5A**). The  
177 proportion of AI-generated captions correctly identified as AI ranged from 37% (BLIP) to 60% (Clip-  
178 Cap). Three recent transformer-based models - GIT [1], OFA [2], and BLIP [3] - demonstrated strong  
179 imitation capabilities (**Fig. 5A**). Unlike ClipCap, which is solely trained for Image captioning, these  
180 models were developed as generative unified transformer frameworks, trained on multiple tasks, in-  
181 cluding image-text retrieval and visual question answering. This suggests that multi-task training aids  
182 models to generate human-like captions. To assess generalization, we included both in-domain images  
183 from MSCOCO [48] and out-of-domain images from the nocaps dataset [49], which contains novel ob-  
184 ject categories not present in MSCOCO. Despite these variations, human judge performance did not  
185 differ significantly between out-of-domain nocaps images and in-domain nocaps images (permutation  
186 test with Bonferroni correction,  $p > 0.05$ ), even for models like GIT, OFA, and BLIP which were  
187 fine-tuned on MSCOCO (**Fig. S3D**). Moreover, on out-of-domain nocaps images, BLIP outperformed  
188 all other models in the Turing-like tests.

189 In the Word associations task, human judges correctly identified human-generated word associ-  
190 ations as human in 58% of the trials, significantly above chance (one-sample Wilcoxon signed-rank  
191 test:  $z = -3.36, p < 0.001$ ), and machine-generated word associations as machine in 47% of the trials,  
192 (one-sample Wilcoxon signed-rank test:  $z = -1.63, p = 0.10$ , **Fig. 4B**). The results of in-person experi-  
193 ments were comparable to those from AMT participants (**Fig. S3B**, permutation test with Bonferroni  
194 correction,  $p > 0.05$ ). Among AI models, Word2Vec exhibited the poorest imitation performance  
195 ( $p(M|M) = 0.58$ ), while GPT3 (prompt DaVinci) outperformed the other models ( $p(M|M) = 0.41$ )  
196 (Friedman test:  $\chi^2 = 3.39, p = 0.50$ , **Fig. 5C**).

197 In the Conversation task, judges distinguished human-generated exchanges in conversations as  
198 humans 66% of the time (above chance, one-sample Wilcoxon signed-rank test:  $z = -5.06$ ,  $p < 0.001$ )  
199 and machine-generated conversation exchanges as AI 47% of cases (one-sample Wilcoxon signed-rank  
200 test:  $z = -1.37$ ,  $p = 0.17$ , **Fig. 4C**). To assess replicability, we repeated the experiments across three  
201 platforms: on AMT, Prolific, and in-lab. Unlike the other two language tasks, in-lab participants were  
202 slightly more accurate compared to the Prolific or the AMT participants (**Fig. S3C**, permutation  
203 test with Bonferroni correction,  $p < 0.05$ ). We separately considered human-human, human-AI, and  
204 AI-AI conversations (**Fig. 5E**). Surprisingly, human-human conversations were classified as human  
205 only 64% of the time (above chance, one-sample Wilcoxon signed-rank test:  $z = -5.06$ ,  $p < 0.001$ ) and  
206 AI-AI conversations were classified as human 57% of the time (above chance, one-sample Wilcoxon  
207 signed-rank test,  $z = -2.39$ ,  $p < 0.001$ ), showing only a modest difference in human detectability  
208 between conversation types which was not statistically significant (Mann-Whitney U test:  $U = 1313.5$ ,  
209  $p = 0.11$ ).

210 Blenderbot stood out as a particularly effective imitator: in AI-AI conversations, it was judged  
211 human 67% of the time, significantly above chance (one-sample Wilcoxon signed-rank test:  $z = -2.83$ ,  
212  $p < 0.005$ ), and comparable to actual humans in human-human conversations (64%, Mann-Whitney  
213 U test:  $U = 1007.5$ ,  $p = 0.70$ ). In human-AI conversations, human participants were labeled as  
214 humans 68% of the time, and AIs were classified as AI 56% of the time (above chance, one-sample  
215 Wilcoxon signed-rank tests:  $z = -4.75$ , and  $z = -1.43$ , respectively, both with  $p < 0.001$ ). Judges  
216 were more likely to identify machine answers in AI-AI conversations than in human-AI conversations  
217 (Mann-Whitney test ( $n = 47$ ):  $U = 1471.5$ ,  $p = 0.005$ ), suggesting that AIs may adjust their linguistic  
218 behavior when interacting with humans by producing more human-like sentences. This observation is  
219 consistent with the notion that judges are more accurate in making comparisons across speakers rather  
220 than making absolute evaluations. When comparing different AI models, Blenderbot was labeled more  
221 frequently as human in AI-AI conversations rather than in human-AI conversations (67% vs 49% of  
222 the time, Mann-Whitney test:  $U = 1225.5$ ,  $p = 0.02$ , **Fig. 5E**). By contrast, ChatGPT-4o [9] was a  
223 slightly worse imitator than Blenderbot (higher imitation detectability), but still outperformed most  
224 of the models on this task (**Table S4**, compare to **Fig. S3C** and **Fig. 5E**). Thus, different language  
225 models show distinct imitation abilities, and their imitation detectability does not necessarily align  
226 with their performance on standard benchmarks.

227 We next examined whether the outcomes of the Turing-like tests were influenced by conversation  
228 length (**Fig. S5**). We evaluated whether detection performance was influenced by conversation length  
229 using linear regression per participant (**Fig. S5**). There was no significant relationship between conver-  
230 sation length and human detection accuracy ( $P(H|H)$ :  $\beta = 0.002$  per exchange, Wilcoxon signed-rank  
231 test ( $n = 45$  workers):  $z = -1.12$ ,  $p = 0.27$ ), nor between length and AI detectability ( $P(M|M)$ :  
232  $\beta = -0.0003$  per exchange,  $z = -0.25$ ,  $p = 0.80$ ). This pattern held when examining specific con-  
233 versation types (H-H:  $z = -0.73$ ,  $p = 0.46$ ; H-M:  $z = -0.36$ ,  $p = 0.72$ ). These results indicate that  
234 judges' ability to discriminate humans from AI remained stable across conversation lengths from 3 to

235 24 exchanges. We extended these analyses to conversations with 48 exchanges and observed patterns  
236 consistent with those seen in 3-24-exchange dialogues. The regression analysis per participant con-  
237 firmed that 48-exchange imitation detectability did not significantly differ from predictions based on  
238 3-24 exchange trends (Wilcoxon signed-rank tests ( $n = 45$  workers);  $P(H|H)$ : mean residual =  $-0.02$ ,  
239  $z = -0.56$ ,  $p = 0.58$ ;  $P(M|M)$ : mean residual =  $0.06$ ,  $z = -0.40$ ,  $p = 0.69$ ). This consistency  
240 held across conversation contexts, including H-H ( $p = 0.83$ ), M-M ( $p = 0.06$ ), and H-M conversations  
241 ( $p = 0.34$ ) (compare **Fig. S5A, B**). When examining individual AI models, the state-of-the-art large  
242 language models boast a high human imitation ability across different conversation lengths, though  
243 individual model performance showed considerable variability (**Fig. S5G**). In addition to conversa-  
244 tion length, we also evaluated the relevance of the conversation topic. Imitation detectability varied  
245 across topics (**Fig. S6**). For instance, the topic of technology yielded an imitation detectability rate of  
246 60%, whereas the topic of music resulted in a lower imitation detectability of 53%. In sum, imitation  
247 detectability remained relatively stable across conversation length, showing no significant effect, and  
248 presented modest variation across conversation topics.

249 We subdivided the results based on the judges' gender and education level (**Fig. S2**). Combined  
250 averaging across all AI models, there were no significant differences in judges' performance between  
251 males and females for the Image captioning task (**Fig. S2B**), the Word association task (**Fig. S2F**), or  
252 the Conversation task (**Fig. S2J**) (permutation test with Bonferroni correction,  $p > 0.05$ ). Similarly,  
253 averaging across all AI models, we did not detect any significant differences based on the judges'  
254 educational level for the Image captioning task (**Fig. S2D**), the Word association task (**Fig. S2H**),  
255 or the Conversation task (**Fig. S2L**) (permutation test with Bonferroni correction,  $p > 0.05$ ).

256 To investigate whether explicitly providing prior information would influence Turing-like test out-  
257 comes, we conducted a new version of the experiment in which participants were explicitly informed  
258 that 50% of the responses were generated by humans (**Fig. 7K**). The results showed no significant  
259 difference compared to the original setting without explicit instructions ( $p > 0.05$ , permutation test  
260 with Bonferroni correction).

261 As AI continues to evolve, proficiency in using or scrutinizing the output of AI tools may sig-  
262 nificantly influence performance in Turing-like tests. To examine this hypothesis, we recruited 33  
263 AI experts and administered the same Turing-like language tasks. Consistent with the hypothesis,  
264 AI experts were more accurate at recognizing AI-generated responses than non-experts in the Image  
265 Captioning and Word Association tasks (Mann-Whitney tests; Image Captioning task:  $U = 885.5$ ,  
266  $p = 0.01$ ; Word Association task:  $U = 224$ ,  $p = 0.33$ ) but not in the Conversation task (Mann-  
267 Whitney test:  $U = 241.5$ ,  $p = 0.45$ , **Fig. S4**). Experts' ability to identify human-generated responses  
268 was also distinct in the Image Captioning task (Mann-Whitney test:  $U = 2439.5$ ,  $p = 0.01$ ), but not  
269 in the Word Association task (Mann-Whitney test:  $U = 205.5$ ,  $p = 0.21$ ) or the Conversation task  
270 (Mann-Whitney test:  $U = 309.0$ ,  $p = 0.62$ ).

271 To summarize the results, we computed the overall imitation detectability as the average of  $p(H|H)$   
272 and  $p(M|M)$ . By construction, a poor imitator would yield  $p(H|H) = p(M|M) = 1$  and therefore an

273 overall imitation detectability of 1.0, whereas a perfect imitator would yield  $p(H|H) = p(M|M) = 0.5$   
274 and therefore an overall imitation detectability of 0.5. The overall imitation detectability was 0.57 for  
275 Image captioning, 0.53 for Word association, and 0.57 for the Conversation task. The results show  
276 that current algorithms are not far from passing the Turing-like tests under the conditions examined  
277 here.

278 Imitation detectability scores should be interpreted with caution. If judges classified all responses  
279 as human, this would result in  $P(H|H) = P(H|M) = 1$ , also leading to an imitation detectability of  
280 0.5 (similarly, with  $P(M|M) = P(M|H) = 1$ ). The experimental data indicate that human judges did  
281 not exhibit either of these behaviors (**Fig. 4**). The total number of H responses was far from 100%  
282 and close to 50% in all the experiments. For instance, across three language tasks,  $P(H|H)$  ranged  
283 from 0.58 to 0.67, while  $P(H|M)$  remained between 0.53 and 0.54. Another way to get an imitation  
284 detectability score of 0.5 arises from completely random responses. We implemented multiple controls  
285 and catch trials in the tests, and the probability of correctly identifying these trials was consistently  
286 high (**Table S3**), confirming that human judges were not merely providing random responses. For a  
287 more granular view of the results, we also present full confusion matrices in **Fig. 4**, which provide a  
288 more detailed breakdown of error types.

289 In our original Turing-like tests, human judges made judgments in a zero-shot setting, without any  
290 feedback. To investigate whether feedback could enhance their ability to distinguish between human  
291 and machine responses, we conducted a follow-up experiment that included a training phase consisting  
292 of Turing-like test trials in which judges received were told whether each response was correct or  
293 not, followed by a test phase without feedback. The results from the test phase indicate a marginal  
294 improvement in imitation detectability following the feedback-based training (**Fig. 7**), although none  
295 of them were statistically significant ( $p > 0.05$ , two-sided t-test, Bonferroni corrected).

## 296 **AI agents still fail to pass an online Turing Conversation task test**

297 Despite the laudable progress in developing large language models, we were surprised by how close  
298 these algorithms were to imitating humans in an offline conversation of up to 24 exchanges per agent.  
299 To probe this further in a format more closely aligned with Turing’s original conception, we developed  
300 an online, interactive Turing test implemented as a mobile application (**Fig. S7**). Each conversation  
301 involved one human judge, one human agent, and one AI agent (GPT3.5-Turbo [9]). The judges  
302 initiated the interaction by asking questions, to which both human and AI agents independently  
303 responded (**Methods**). Participants were randomly assigned roles as either judge or human agent.

304 We constrained the number of exchanges per conversation to be 1, 5, 10, or 20. In each conversation,  
305 the mobile application randomly drew one out of these four choices and presented this exchange limit  
306 to the human judge and the human agent. After the conversation reached the limit in the number of  
307 exchanges, the judge was asked to evaluate which of the two sets of answers came from a human and  
308 which came from a machine. In stark contrast with the previous section, human judges achieved 100%  
309 accuracy (**Methods**) in distinguishing between human and AI agents (**Discussion**).

310 Several factors may potentially explain the discrepancies between the Turing-like offline task pre-  
311 sented in the previous section and this online task. First, the test settings differ. In the offline version,  
312 there were two agents involved, and both could ask each other questions. Agents were not asked to  
313 identify whether they were conversing with a human or a machine. Judges passively observed the  
314 fixed conversations offline without actively participating in them or conceiving captious questions. An  
315 important example of this situation is content moderation on social media, where a moderator can  
316 only review flagged conversations but cannot actively influence them. In contrast, in the online version  
317 implemented here, the judge was deliberately trying to probe agents with the clear goal of discerning  
318 the identities of both agents. Second, in the previous section, the topics were pre-determined, and  
319 centered around opinions on video games, sports, movies, etc. In contrast, in the current online ver-  
320 sion implemented here, judges decided on the topics, which could delve into personal aspects such as  
321 age and family. Despite AI models being assigned randomly generated personal profiles, they typically  
322 struggled to express human values and emotions, facilitating identification by human judges. Moreover,  
323 advanced scientific topics, such as technical jargon in chemistry and physics, can serve as a means to  
324 discern differences between human and AI agents. While such terminology may be challenging for lay  
325 humans, AI agents often find it relatively easy to discuss. Third, AIs might reveal their identity when  
326 asked personal questions or unsafe questions, such as “Sorry, I am an artificial intelligence program.  
327 I am unable to answer questions related to politics or how to kill someone.”. Fourth, in the online  
328 test implemented here, human agents typed responses on cell phones, leading to variation in typing  
329 amount, speed, and style. This contrasts with AI agents, potentially revealing meta clues such as  
330 response times. Despite implementing word limits for AI responses through text prompts and cutting  
331 off extra sentences exceeding the word limits, and despite introducing delays in the AI responses to  
332 approximate the typing speed of humans, disparities in response lengths and times persist.

333 To further evaluate whether human agents could deliberately mimic AI behavior, we also imple-  
334 mented a variation of the online task where human agents were asked to be deceitful and attempted to  
335 imitate machine answers. Despite this adversarial manipulation, human judges consistently identified  
336 human and AI agents correctly with 100% accuracy under these conditions. Human agents struggled  
337 to replicate the precise behavioral patterns of AI agents.

338 Given that internal built-in safety mechanisms could contribute to the results in this online ver-  
339 sion of the test, we conducted a follow-up study using the LLAMA-3-70B-Instruct model [39] with  
340 safeguards disabled (**Methods**). This version included a larger set of human participants ( $n=28$ ,  
341 **Fig. S7E-G**), yielding 188 new trials. The imitation detectability scores from the 28 human judges  
342 are shown in **Fig. S7D**. The results reveal that human judges no longer achieved perfect accuracy in  
343 distinguishing between human and machine responses across conversations with up to eight exchanges,  
344 unlike the previous findings using ChatGPT. The ability to detect machine-generated responses in-  
345 creased with the number of conversational exchanges (regression with binomial weights,  $p = 0.032$ ).  
346 Representative examples of both successful and failed trials where human judges correctly or incorrectly  
347 identified the source of the responses are provided in **Sec. S2**.

## A simple computational model can tell human and AI agents apart

The results presented thus far involve human judges (as in the original Turing test). It is also of interest to assess whether algorithms can tell human and AI answers apart. As a proof-of-principle to build an AI judge, we used simple Support Vector Machine (SVM) classifiers to evaluate responses generated by humans versus AI. The inputs to the AI judges were embeddings extracted from either words or sentences by different language models (**Methods**). The SVM classifier was trained to use these embeddings to classify whether a given answer came from a human or not. The predicted labels on 10-fold cross-validated data were compared against the ground truth labels (**Fig. 4D, E, F**).

In the Image captioning task, we used the OpenAI Embeddings API (text-similarity-curie-001 model) to extract features from captions and trained a linear SVM to discriminate human versus machine responses. Surprisingly, even though this classifier did not use image features, this simple AI judge could discern human answers ( $p(H|H) = 0.72$ ) and machine answers ( $p(M|M) = 0.83$ ) much better than human judges (compare **Fig. 4A** vs. **Fig. 4D**). Similar results were observed across all the different AI models (**Fig. 5B**).

In the Word association task, 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). The results were similar for these different embeddings and we report the average results here. This simple AI judge vastly outperformed human judges (compare **Fig. 4E** versus **Fig. 4B**). Similar conclusions were drawn across different machine agents (**Fig. 5D**). As expected, when training an SVM classifier based on the embedding of the same model whose embedding were also used for generating the associations, this model classification performance was essentially perfect. For instance, on the 3rd row on the left confusion matrix in **Fig. 5D**, the AI judge trained on Word2Vec embedding indeed predicted very well all association 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 classifiers and consolidated all their trials for the final plots to avoid such biases. We also present generalization results at the end of this section.

In the Conversation task, we trained a simple SVM judge to distinguish whether a single sentence was generated by humans or AIs (**Figs. 4F** and **5F**). Even though the AI judge evaluated single sentences (whereas human judges evaluated entire conversations), the simple AI judge achieved similar performance to human judges in detecting human answers ( $p(H|H) = 0.65$ ) and surpassed human judges in detecting machine answers ( $p(M|M) = 0.67$ ). Similar conclusions were drawn for all the language models evaluated here, with the easiest one to detect being Blenderbot (which was the hardest one to detect for human judges) and the hardest one to detect being GPT3Curie (**5F**). Human judges likely focus on high-level conversation understanding rather than single-sentence statistics.

Comparing results between humans and machines is always difficult, especially when training is taken into account. To assess the effect of feedback, we showed results in experiments where human

385 judges were provided with feedback in **Fig. 7**. For practical reasons, humans only received feedback  
386 in a small number of trials, which was much less than the amount of training in the SVM judges.  
387 For comparison purposes, we retrained our SVM-based judges matching the small number of trials  
388 with feedback that human subjects were exposed to for each task, mirroring the human training  
389 phase (**Fig. 7D, H, L**). In the image captioning task, human judges continued to outperform the  
390 machine judges. In contrast, for the word association task, the machine judge maintained its lead.  
391 In the Conversation task, performance was comparable between feedback-trained human judges and  
392 the SVM judges trained with a small number of trials. Notably, in all the cases, SVM-based judges  
393 received word embeddings derived from large language models (**Methods**). These multimodal LLMs  
394 are trained on extensive knowledge from vast internet data encompassing text, images, and videos.

395 To test whether AI judges’ success could be attributed to basic low-level text statistics, we compared  
396 basic properties - including parts of speech frequency, punctuation, word frequency, word length, and  
397 sentence length - between human and AI answers (**Table S5**). Overall, no single feature or set of  
398 features reliably distinguished humans from AIs across tasks.

399 The results presented thus far combine all the different AI agents (always using cross-validation  
400 across different trials). To evaluate the extrapolation capabilities of machine judges, we conducted  
401 an across-agent analysis where we evaluated their performance under conditions in which they were  
402 trained with entries created by some agents and tested their performance with entries created by other  
403 AI agents. We conducted two such experiments: “train on one agent, test on other agents” (**Fig. S8A,**  
404 **C, E**), and “train on all but one of the agents, test on the left out agent” (**Fig. S8B, D, F**). As ex-  
405 pected, the overall imitation detectability of machine judges in both extrapolation experiments is lower  
406 compared to the accuracy of machine judges trained with responses from all machine agents (compare  
407 **Fig. S8 to Fig. 4 and Fig. 5**). As expected, the “train on one, test on others” experiment posed  
408 greater difficulty than “train on others, test on one,” resulting in lower overall imitation detectability  
409 (compare **Fig. S8A, C, E vs. Fig. S8B, D, F**). Notably, despite the difficulty in generalizing across  
410 different AI agents, AI judges extrapolating to classify responses from excluded AI agents still out-  
411 performed human judges in the Image captioning and the Word association tasks. AI judges did not  
412 extrapolate well in the Conversation task, and performed slightly worse than humans but it should be  
413 noted that the AI judges evaluated only single sentences as opposed to humans who had access to the  
414 whole conversation.

415 To further assess the extrapolation ability of AI judges within each language task, we evaluated a  
416 pre-trained large language model - ChatGPT, [9] - in a zero-shot setting to determine whether it could  
417 distinguish between human and machine responses (**Fig. 7, Methods**). In the image captioning task,  
418 human judges outperformed the AI judge by nearly 10% in imitation detectability (trial-based two-  
419 tailed t test,  $p < 0.05$ ). However, this trend was not present in the word association task, where the  
420 machine judge slightly outperformed the human participants (trial-based two-tailed t test,  $p > 0.99$ ).  
421 In the Conversation task, we used ChatGPT as the zero-shot machine judge and considered two kinds  
422 of prompts. In the zero-shot case, the prompt only included a conversation from the test set and the

423 model was asked to identify the identities of the two agents (**Fig. 8D**). This zero-shot model performed  
424 poorly as a machine judge. In the one-shot case, the prompt included a single conversation example  
425 with ground truth human and machine labels and a test case asking the model to indicate the identities  
426 of the two agents (**Fig. 8E**). ChatGPT performed well as a judge, with an overall  $p(H|H) = 0.59$  and  
427  $p(M|M) = 0.56$ , worse than the SVM judges in **Fig. 4F**, but better than the extrapolation results  
428 shown in **Fig. S8E, F**. The performance of ChatGPT in the one-shot case was very different in the  
429 H-H and M-M conversations where it was almost at chance levels versus in the H-M conversations,  
430 where it even exceeded the performance of the SVM judges in **Fig. 4F**. In contrast to the SVM judges,  
431 which were trained on the entire corpus, the binary discrimination task is more challenging for one-shot  
432 ChatGPT since there is no explicit training with the task. It is therefore remarkable that even with a  
433 single example, ChatGPT can act as a judge to detect machine answers.

### 434 **Turing-like tests are complementary to standard AI model benchmarks**

435 Algorithm performance is often defined by comparing the predicted outputs of AI models against  
436 human ground truth labels. However, higher accuracy scores in terms of such standard evaluation  
437 metrics defined in the AI community do not necessarily translate into better human imitators. To  
438 investigate this disconnect, we directly compared Turing overall imitation detectability to standard  
439 evaluation metrics in the Image captioning task, where there are clearly defined accuracy metrics such  
440 as BLEU score [16] and CIDEr score [2], to assess the quality of AI-generated captions [50] (**Fig. 8**).  
441 The results, which were similar for both BLEU (**Fig. 8A**) and CIDEr (Pearson correlation =  $-0.088$ ,  
442  $0.84$ , **Fig. 8B**) scores, highlight that human-likeness and traditional benchmark performance constitute  
443 distinct concepts. There are pairs of models that have comparable imitation detectability but differ  
444 in their captioning score (e.g., ClipCap and GIT, BLIP and OFA). Conversely, GIT shows a higher  
445 captioning score than BLIP, but BLIP is a better human imitator based on the imitation detectability.  
446 Flamingo, ChatGPT, and LLaVa, which are all large multi-modal foundation models for vision and  
447 language, yielded imitation detectability that was on par with, or even lower than, that of classical  
448 models trained specifically for image captioning. SCST, a classical image captioning model, yielded  
449 higher imitation detectability. A qualitative example of differences between accuracy and imitation is  
450 shown in **Fig. 8C**. The algorithm (GPT4) provided an accurate description of the image. Humans  
451 also correctly described the image. However, the algorithm did not completely imitate how humans  
452 interpret this image, failing to capture some of the key aspects of why the image is striking (see the  
453 descriptions in the caption to **Fig. 8C**).

### 454 **Current algorithms are also close to passing Turing-like tests in vision tasks**

455 Although the original Turing test has been formulated in the context of language, it is possible to  
456 define Turing-like tests in many other domains. We extended our investigations to include Turing-like  
457 tests in three vision tasks: Color estimation (**Fig. 3A**), Object detection (**Fig. 3B**), and Attention

458 prediction (**Fig. 3C**). The methodology to collect human responses and AI responses (**Fig. S1D-F**)  
459 and to conduct the Turing-like tests (**Fig. 3**) followed the steps described earlier for the three language  
460 tasks (**Methods, Table S1**). Demographic information about the participants in each vision task is  
461 presented in **Fig. S2**.

462 In the Color estimation task, agents were asked to identify the dominant color of an image (see  
463 examples of human and machine responses in **Fig. 3D**). Human judges correctly distinguished human  
464 answers in 55% of trials (one-sample Wilcoxon signed-rank test,  $z = -2.17$ ,  $p = 0.03$ ) and AI-generated  
465 answers in 61% of trials (one-sample Wilcoxon signed-rank test,  $z = -4.13$ ,  $p < 0.001$ , **Fig. 6A**).  
466 Performance was comparable across AI models ( $P(M|M)$  range:  $0.58 - 0.66$ , Friedman test,  $\chi^2 = 0.97$ ,  
467  $p = 0.62$ ). Interestingly, even a simple baseline metric such as the mean color content quantization  
468 (or MCCQ) still achieved a moderately good performance, fooling humans 34% of cases (one-sample  
469 Wilcoxon signed-rank test,  $z = -2.17$ ,  $p < 0.001$ ).

470 In the Object detection task, participants were asked “What do you see in this image?” and had  
471 to provide three single words as answers. Human judges correctly identified human-generated answers  
472 in 53% of trials, and AI-generated answers in 63% of trials (one-sample Wilcoxon signed-rank tests;  
473  $z = -1.22$  and  $p = 0.22$ ,  $z = -4.29$  and  $p < 0.001$ , respectively; **Fig. 6C**). Among all the models, the  
474 Detectron agent (a variant of Mask R-CNN [51] trained on ImageNet and MS-COCO), emerged as the  
475 most human-like, fooling human judges 52% of the time, significantly outperforming all other models  
476 (Google API: 35%, Azure: 31%, Rekognition: 30%; Friedman test,  $\chi^2 = 31.03$ ,  $p < 0.001$ ). Notably,  
477 Detectron not only achieved outstanding absolute scores in terms of standard evaluation metrics, such  
478 as mAP [52], but its response patterns also closely mimics humans’, often identifying the top-3 salient  
479 objects in the scene similarly to humans.

480 We evaluated Turing-like imitation in a third vision task, Attention prediction, using gaze data  
481 collected via eye tracking. There were two variations of the task: (1) participants freely viewed an  
482 image for 12 seconds, shifting their overt attention patterns by moving their eyes, and (2) participants  
483 were given a target object and were instructed to locate it within the image using their eye movements  
484 (see [12] for experiment details). Overall, human judges distinguished human eye movements 63%  
485 of the time (one-sample Wilcoxon signed-rank test:  $z = -5.47$ ,  $p < 0.001$ ) and AI-generated eye  
486 movements 50% of the time (one-sample Wilcoxon signed-rank test:  $z = -1.73$ ,  $p > 0.05$ , **Fig. 6E**).  
487 Models differed significantly in imitation detectability (Friedman test:  $\chi^2 = 20.90$ ,  $p = 0.001$ ). ISVN  
488 [12], [53] showed the most human-like performance, fooling human judges 57% of the time. We further  
489 evaluated whether task goals modulated imitation detectability. Human judges performed similarly  
490 during free-viewing compared to visual search (Mann-Whitney test; free viewing:  $n = 103$ , versus  
491 visual search,  $n = 89$ ,  $P(H|H)$ :  $U = 4183.0$ ,  $p = 0.28$ ; **Fig. S9A** vs. **Fig. S9C**). GBVS - a bottom-up  
492 saliency model - generated more human-like scanpaths during free-viewing than in visual search. IVSN  
493 model’s performance was similar across both tasks, emphasizing the importance of incorporating both  
494 bottom-up and top-down attention mechanisms in computational models of human attention. Finally,  
495 we evaluated machine judges trained to distinguish human from AI-generated eye movements. These

496 classifiers performed well, but struggled labeling outputs from the IVSN model (**Fig. 6F**) both in the  
497 free viewing task (**Fig. S9B**) as well as in the visual search task (**Fig. S9D**).

498 Evaluating human eye movements for attention prediction is not an intuitive or typical task for  
499 human observers. To assess the ability to evaluate eye movements with respect to the underlying image,  
500 we introduced catch trials where we randomly swapped eye movements and images. For instance, we  
501 showed the eye movements from image 17 overlaid on image 62, and the eye movements from image  
502 62 overlaid on image 9. Under this condition,  $P(H|H) = P(M|M) = 50\%$ , indicating that judges  
503 were unable to distinguish human and machine eye movement patterns without the corresponding  
504 image (**Methods**). Participants were not entirely insensitive to eye movement anomalies. To test  
505 participants’ sensitivity, we introduced additional catch trials featuring highly anomalous fixation  
506 patterns—for example, a single randomly placed fixation repeated throughout the trial. In these  
507 cases, participants were able to reliably detect the artificial nature of the eye movements, achieving a  
508 detection accuracy of 82%.

509 To investigate how different model components contribute to human-like performance in visual  
510 attention prediction, we conducted Turing-like tests with a series of ablated machine agents based on  
511 the IVSN model [12] (**Methods**). Specifically, the NoTarget model removed the top-down modulation  
512 from the target features. The EccNet model [13] incorporated eccentricity-dependent pooling. Instead  
513 of extracting features from the top layers of a feedforward neural network pretrained on ImageNet in  
514 IVSN, we used features from the early layers for target modulation and refer to this ablated model  
515 as the Layer model. Both EccNet and IVSN exhibited low imitation detectability (**Fig. S10A**). In  
516 contrast, the NoTarget and Layer models failed to produce realistic eye movement behavior during  
517 visual search, likely due to the absence of top-down guidance from target features or the inability  
518 of their extracted features to capture target semantics. Consequently, human judges could easily  
519 distinguish their outputs from actual human eye movements.

520 We worried that AMT participants could be poor judges. Therefore, we repeated the Turing-like  
521 experiments in person in our lab (**Fig. S4**). Despite our initial concerns, the results from in-person  
522 experiments were comparable to those obtained from AMT participants (**Fig. 6A, C, E** for vision  
523 tasks compared with **Fig. S4P, T, X**. Color estimation task: Mann-Whitney test,  $U = 282$ ,  $p = 0.13$ .  
524 Object detection task: Mann-Whitney test,  $U = 342$ ,  $p = 0.5$ . Attention prediction task: Mann-  
525 Whitney test,  $U = 1298$ ,  $p = 0.2$ ).

526 In our original Turing-like tests, human judges made decisions in a zero-shot setting, without  
527 prior exposure or feedback. To assess whether feedback would alter the judges’ ability to distinguish  
528 human from machine responses in the visual tasks, we conducted a follow-up experiment mirroring  
529 the procedure used in the language tasks. Each experiment consisted of a training phase with Turing-  
530 like trials where participants received immediate feedback followed by a test phase without feedback  
531 (**Table S2, Methods**). Providing feedback led to a marginal improvement in imitation detectability  
532 (two-sided t-tests with Bonferroni correction. Color estimation task:  $p > 0.99$ . Object detection  
533 task:  $p = 0.001$ , Attention prediction task:  $p = 0.53$ , **Fig. 7**). To assess whether prior training could

534 improve participants’ ability to distinguish between human and machine-generated eye movements,  
535 we conducted a training phase with feedback as described above (**Methods**). Following this training,  
536 participants’ accuracy in the Turing-like task improved modestly from 57.5% to 62% in the Attention  
537 prediction task (compare **Fig. 7U** versus **Fig. 7W**). However, the gains of training feedback were not  
538 statistically significant across the three vision tasks, suggesting that feedback alone may have limited  
539 effect in enhancing human judges’ discriminative performance in these scenarios.

540 We examined the impact of human judge demographics on the Turing results. Similar to the results  
541 reported above for the language tasks, neither the participant’s gender (**Fig. S2N, R, V**), nor the  
542 participant’s educational level (**Fig. S2P, T, X**) correlated with any major difference in the imitation  
543 detectability (permutation test with Bonferroni correction,  $p > 0.01$ ). While gender and age showed  
544 little impact on the results, differences depended on the level of AI expertise; however, these differences  
545 did not reach statistical significance (compare **Fig. S4M-X** with **Fig. S4A-L**. Color estimation task:  
546 Mann-Whitney test,  $U = 282$ ,  $p = 0.13$ . Object detection task: Mann-Whitney test,  $U = 342$ ,  $p = 0.5$ .  
547 Attention prediction task: Mann-Whitney test,  $U = 1298$ ,  $p = 0.2$ ).

548 Finally, as a proof-of-principle, we followed the approach described in the language tasks to develop  
549 simple AI judges to discriminate human from machine answers. In the Color estimation task, we trained  
550 SVM judges using image features and text embeddings for the colors (**Methods**). For the Object  
551 detection task, we trained SVM judges using the text embeddings for the three nouns describing the  
552 images (**Methods**). In the Attention prediction task, we performed binary discrimination using the  
553 2D coordinates of the first ten fixations (**Methods**). Similar to the conclusions drawn in the language  
554 tasks, for the Object detection task and Attention prediction task, the AI judges easily discriminated  
555 human from machine answers and vastly outperformed human judges. In the Object detection task,  
556 the AI judges achieved  $p(H|H) = 0.9$  and  $p(M|M) = 0.72$  (**Fig. 6D**). There were large gaps among  
557 the different vision models with the worst imitator being Azure, showing  $p(M|M) = 0.93$ . In the  
558 Attention prediction task, the AI judges achieved  $p(H|H) = 0.79$  and  $p(M|M) = 0.8$  (**Fig. 6F**),  
559 revealing even more drastic differences among models, with the worst imitator being GBVS reaching  
560  $p(M|M) = 0.94$  and the best imitator being IVSN with  $p(M|M) = 0.52$ . In both tasks, the AI judges  
561 outperformed the human judges (compare **Fig. 6D, F** with **Fig. 6C, E**). The results were different  
562 in the Color estimation task. Here the human judges outperformed the AI judges (compare **Fig. 6A**  
563 with **Fig. 6B**).

564 The AI judges take the visual features extracted from the deeper layers of ResNet18 and the color  
565 embeddings extracted from BERT as inputs. We also evaluated ablated versions of the AI judges,  
566 either focusing on early visual layers or color responses alone without visual features (**Methods**).  
567 There was a significant improvement in imitation detectability of the AI judge when using the features  
568 from the early layer of ResNet (one-way ANOVA with Bonferroni correction,  $p < 0.001$ , **Fig. S10B,**  
569 **C**). Interestingly, color responses alone can achieve above-chance imitation detectability. This suggests  
570 that the long-tailed distribution present in human responses is strongly captured by the AI judge, even  
571 though it remains largely imperceptible to human judges.

572 In addition, to ensure a fair comparison with human judges in the feedback-based experiment  
573 (**Fig. 7**), we introduced an SVM-based machine judge that was retrained using the same number of  
574 training examples provided to human judges during the training phase. The impact of feedback varied  
575 across tasks: in the Color estimation task, feedback helped human judges close the gap, surpassing  
576 the machine judge; in Object detection and Attention prediction, however, the machine judge not only  
577 maintained its lead but improved even further after retraining (**Fig. 7P, T, X**).

578 Next, similar to the language tasks, to evaluate the ability of ChatGPT as a machine judge to  
579 generalize beyond its training data, for the three vision tasks, we provided ChatGPT with both images  
580 and accompanying text prompts to make zero-shot judgments (**Methods**). Surprisingly, ChatGPT  
581 consistently outperformed human judges in all three vision tasks (**Fig. 7N, R, V**). The performance  
582 advantage was especially pronounced in the attention prediction task, showing a gap of up to 8%.

## 583 **A universal multi-modal machine closely imitates human responses**

584 Humans can perform multiple tasks, whereas current AI models are often trained for and thrive in  
585 specific tasks. In particular, in the results presented thus far, different AI algorithms were evaluated  
586 for each task. We asked whether a single AI algorithm could imitate humans across all six tasks. To  
587 address this questions, we selected the best-performing off-the-shelf multimodal model, ChatGPT-4o  
588 [9], as the universal machine agent. We used ChatGPT-4o to produce responses for each stimulus in a  
589 zero-shot manner across all six language and vision tasks. The text prompt used in each task is shown  
590 in the **Methods** section. The results, presented in **Table S4**, indicate that the imitation detectability  
591 of the ChatGPT-4o agent remained around 0.5 across all tasks (range: 0.44 (object detection) to 0.56  
592 (attention prediction)). This is an impressive result, showing that a unified multi-modal model can  
593 generate human-like responses that are largely indistinguishable from human responses, even without  
594 any task-specific training.

## 595 **Discussion**

596 We first introduce a dataset of human answers in 6 common language (**Fig. 1**) and vision (**Fig. 3**)  
597 tasks to evaluate the ability of algorithms to *imitate* humans. The dataset also includes answers from  
598 state-of-the-art AI algorithms (**Table S1**). Using the human and machine answers, we conducted  
599 Turing-like tests, showing that current algorithms are not far from being able to imitate humans in  
600 these tasks (**Fig. 4A-C, G-I, Fig. 5A, C, E, Fig. 6A, C, E**). In contrast to human judges, even  
601 simple machine judges can distinguish human from machine answers (**Fig. 4D-F, J-L, Fig. 5B, D,**  
602 **F, Fig. 6B, D, F**).

603 The Turing test has been extensively discussed, and contested, as a means to assess general intel-  
604 ligence. Our work is *not* intended as an evaluation of intelligence or to contribute to the discussion  
605 of Turing tests to quantify intelligence. Imitating humans can benefit humanity in many situations  
606 where we want to align machine and human outputs, such as in emulating expert decisions. However,

607 imitating humans can also be used for evil, as in the dissemination of fake information, phishing at-  
608 tempts, or other scams. Regardless of whether Turing tests are good or bad metrics of intelligence, it  
609 is of high practical importance to assess whether algorithms can imitate humans or not.

610 In some cases, we summarized results using the overall imitation detectability metric. It is im-  
611 portant to emphasize that this summary metric does not differentiate between the two types of clas-  
612 sification errors: (1) False Negatives – A human response misclassified as machine-generated. (2)  
613 False Positives – A machine response misclassified as human-generated. Understanding the distinction  
614 between these two error types is important for interpreting decision-making by human and machine  
615 judges. Moreover, in real-world applications, these errors may have different consequences. For ex-  
616 ample, misclassifying a machine as a human might be more problematic than misclassifying a human  
617 as a machine if the goal is to reliably detect AI-generated content. We present confusion matrices  
618 throughout to complement the imitation detectability scores and quantify the two types of errors.

619 One approach towards mitigating the risks from evil human imitators is to build AI judges that can  
620 distinguish true human answers. The results presented here show that even simple AI judges like the  
621 ones introduced here can generally do a better job than human judges in detecting machine answers  
622 for the current tasks. There will likely continue to be an arms race between the development of better  
623 imitators and better detectors.

624 In most of the Turing-like tests presented here, AI judges were explicitly trained to classify re-  
625 sponses from humans versus AIs, while human judges received no explicit training. However, we also  
626 presented results using ChatGPT-4o, a state-of-the-art, off-the-shelf, multimodal model. We evaluated  
627 ChatGPT-4o in zero-shot Turing-like tests across all six tasks. This unified, multi-modal model can  
628 distinguish between human and machine-generated responses with above-chance accuracy, even with-  
629 out any explicit training. These observations also raise the possibility that humans may be trained to  
630 better recognize machine answers. In support of this hypothesis, when human judges received training  
631 in the form of trial-by-trial feedback, they showed improved accuracy in identifying human-generated  
632 responses. By and large, humans have lived in a world without good human imitators and did not  
633 have to worry too much about the possibility of machines emulating humans, other than in famous  
634 works of fiction. This situation is changing rapidly and may lead to the need to educate human judges.

635 An algorithm’s ability to imitate humans does not necessarily correlate with traditional performance  
636 metrics like accuracy. Consider a simple scenario of an image with a highly occluded dog that is hard  
637 to see, and both machines and humans interpret the object to be a wolf. Both would be wrong, but the  
638 machine would be adequately imitating humans. Conversely, the example in **Fig. 8C** for the Image  
639 captioning task. This example shows that both machines and humans can provide correct answers, but  
640 the algorithm does not fully imitate how humans interpret the image. Thus, M and H can be wrong,  
641 and M can adequately imitate H; M and H can also both be right and M fail to adequately imitate  
642 H. Being right or wrong is not the same as imitation. Turing-like tests provide a complementary  
643 assessment to existing benchmarking frameworks for AI algorithms. Comparisons between models in  
644 Turing-like tests also provide insights helpful for developing future AI models that can better align

645 with humans.

646 We refer to the evaluations in our study as “Turing-like” to emphasize that they are not identical  
647 to Turing’s original description [30]. Turing did not spell out the implementation details of his tests,  
648 perhaps because he did not imagine at the time that these tests could be truly implemented, barely  
649 a few decades later. The datasets and evaluations introduced here are extensive (72,191 Turing-like  
650 test trials, 636 humans contributing to the dataset, and 1,916 human judges). Yet, these evaluations  
651 barely scratch the surface of what needs to be evaluated. There are essentially infinite possible Turing-  
652 like tests. Each algorithm and each task can be evaluated in terms of imitation capabilities. Even  
653 within a particular task, there are multiple ways of instantiating a Turing-like test. The results of a  
654 Turing-like test depend on the task, the algorithm, how the question is formulated, the demographics  
655 of the human agent, the characteristics of the human judge, and many other variables. Such variations  
656 are particularly evident in the context of conversation tasks where the topic, medium, length, format,  
657 instructions, agent and judge expertise, and agent and judge demographics, can lead to different  
658 results [20], [23], [24], [25], [26]. For example, the conversation topic and participant expertise can be  
659 critical: one could conduct the conversation focusing exclusively on medieval history, string theory, or  
660 basketball games as conversation topics and agents/judges with or without domain-specific knowledge  
661 may perform differently. However, the topic-agnostic formulation is relevant to many (but certainly not  
662 all) real-world scenarios. The results presented here further emphasize the importance of integrative  
663 benchmarks. The current work is not intended as an exhaustive assessment of every conceivable task.  
664 We do not believe such a complete characterization is even feasible in practice. Instead, we focus on a  
665 subset of task formulations that arise naturally in everyday contexts. For instance, in conversational  
666 tasks, one might imagine an exchange between two agents on a social media platform, with an observer  
667 attempting to determine whether both, one, or neither of the participants is human.

668 We also implemented and evaluated an online version of the conversation task that is closer to the  
669 original formulation of the Turing task. The results from the online test with adversarial prompting  
670 differ from those of the offline, non-adversarial formulation. These are distinct experiments based on  
671 contrasting formats and underlying assumptions, which naturally lead to different outcomes, further  
672 emphasizing that the details of how the Turing tests are defined matter. The findings underscore that  
673 Turing-like test results can vary widely depending on the conversational context and the intentions of  
674 the participants.

675 In the future, there may be many situations where humans need to prove their humanity, expand-  
676 ing on current “CAPTCHA” efforts. There is no universally accepted “gold standard” response that  
677 definitively distinguishes human from machine-generated outputs. In the catch trials, we could con-  
678 struct answers to skew the distribution toward machine responses (**Methods, Table S3**). However, it  
679 is much more challenging to construct responses for any of the tasks that only humans would produce.  
680 This difficulty will only increase as AI systems become more advanced and human-like [54].

681 These efforts pave the way for the research community to expand Turing-like tests to other research  
682 areas, to build better imitators when mimicking humans is desirable, and to develop better imitation

683 detectors when mimicking humans is deemed dangerous. As more AI models can blend in among  
684 humans, taking over tasks that were originally unique yardsticks of our cognitive abilities, we must  
685 ponder what makes us humans and whether we are mentally, ethically, and legally ready for the rapid  
686 revolution brought forth by AI technologies that can emulate humans.

## 687 Methods

### 688 General considerations

689 We provide details about each of the 6 tasks in the next section. For each task, we created a dataset  
690 consisting of answers from human agents (H) or AI machine agents (M). We conducted Turing-like  
691 tests using those answers both in the lab and also using two online platforms: Amazon Mechanical  
692 Turk (AMT), and Prolific. All AMT experiments were based on “master” workers with at least 1,000  
693 approved hits, and 95% approval rate. Participants were given as much time as needed to complete  
694 the tasks. We only considered participants whose native language is English. **Table S1** shows the  
695 number of stimuli, number of Turing-like tests, dataset sources, and AI machine models used in each  
696 task. All the studies were conducted with the participants’ consent according to protocols approved  
697 by the Institutional Review Board (IRB) at Children’s Hospital.

698 During each Turing-like test, we presented a single stimulus and answer. Judges had to indicate in  
699 a two-alternative forced-choice manner whether the answer came from a human or a machine (**Figs. 1**  
700 and **3A-C**). Half of the time, the answer was from a human. The other half of the time, the answer  
701 was from a machine, sampling with equal probability from one of the different computational models  
702 used for each task (discussed below for each task, **Table S1**). The trial order was randomized. No  
703 feedback was provided to the participants. Additional control trials were introduced for each specific  
704 task to ensure compliance (discussed separately below for each task).

705 The results of Turing-like tests with human judges depend on the characteristics of the judge. As  
706 an initial characterization, we collected basic demographic information about human judges, including  
707 their age, gender, country, education and whether they are native English speakers. The partici-  
708 pant demographics and the dependence of the results on the participant’s demographics are shown in  
709 **Fig. S2**. We do not show results separately for different age groups because most participants were  
710 approximately in the same age group.

711 There were three tasks that focused primarily on language and three tasks that focused primarily  
712 on vision. Some of the vision tasks required language comprehension. The selected vision tasks—color  
713 estimation, object detection, and visual attention—were chosen not only for their relevance to human  
714 perception but also because they inherently involve some level of language expertise and comprehen-  
715 sion. Color estimation requires an understanding of color names, categorical perception, and descriptive  
716 conventions. Humans do not just perceive colors continuously but also label them (e.g., distinguishing  
717 between “blue” and “teal”), requiring a mapping between visual input and linguistic representations.

718 Object detection involves recognizing and labeling objects, a process that depends on semantic knowl-  
719 edge and linguistic categorization. For instance, distinguishing between “dog” and “wolf” requires  
720 both visual discrimination and conceptual understanding of word meanings and categories. Visual  
721 attention is guided by both visual and linguistic cues, especially in real-world tasks where humans  
722 allocate attention based on instructions or contextual understanding. For example, in a search task,  
723 knowing the meaning of “Find the red apple” requires linking linguistic input with visual features.

## 724 Image captioning

725 **Dataset, human agents.** We randomly sampled 250 images each from in-domain, near-domain, and  
726 out-of-domain categories from the validation set of the NOCAPS dataset [49] and 250 images from the  
727 MSCOCO test set [55], creating a set of 1,000 images. We collected 2,290 human captions with  $\geq 6$   
728 words per caption and  $\geq 2$  captions per image from Amazon Mechanical Turk (AMT) participants.  
729 Our AMT interface is shown in **Fig. S1A** and was inspired by the MSCOCO captions data collection  
730 interface [48]. We provided the following instructions to each participant:

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

742 During caption collection, we implemented additional controls in our AMT interface that issued  
743 warning pop-ups to the participants. The controls included: (1) Minimum of 6 words, (2) No special  
744 characters, (3) Response time must be more than 4 seconds, (4) Not more than 4 identical words, (5)  
745 Not contain the words “image” and “photo”, (6) Successive responses cannot be the same.

746 **Dataset, machine agents.** To generate AI machine captions, we used GIT [1], OFA [2], BLIP [3],  
747 ClipCap [4], and Microsoft’s Azure Cognitive Services [67] (**Table S1**). For open-source models, we  
748 used the largest variants finetuned on the COCO Captions dataset [48]. To further investigate the  
749 relationship between standard evaluation metrics and the imitation detectability of image captioning  
750 models, we introduced four additional machine agents: OpenFlamingo-4B [56], ChatGPT-4o-latest [9],  
751 SCST [57], and LLaVA-v1.6-Mistral-7B [58]. We collected 7,620 machine captions with 9 captions per  
752 image.

753 To compare with human judges who performed all six tasks without prior training, we also evaluated

754 a top-performing multi-modal model—ChatGPT [9]—as a zero-shot machine judge. We collected its  
755 responses across all six tasks. The text prompt provided to ChatGPT for generating image captions  
756 is identical to the one given to human participants above, except for the last line indicating human  
757 participants to click “submit”.

758 **Turing-like test, human judges.** In each trial, a participant was presented with an image and a  
759 caption and was asked to indicate in a two-alternative forced-choice manner whether the caption was  
760 generated by a human or a machine (Fig. 1B). We collected responses from 393 AMT participants and  
761 51 in-lab participants. Demographic information about the participants is shown in Fig. S2A,C for  
762 online participants and Fig. S4A-C for in-lab participants. Each participant was presented with 40  
763 image-caption pairs. We only considered response times over 3 seconds. We collected a total of 22,680  
764 responses. We inserted catch trials using the same format as all the other questions by repeating a  
765 randomly selected word instead of generating a meaningful caption (Table S3).

766 Throughout this Turing-like test and all the Turing-like tests for the other tasks, participants might  
767 have a strong prior or constraint suggesting that half of the trials come from humans and half come  
768 from machines, but this was not explicitly stated. We deliberately chose not to inform participants  
769 that human responses constituted exactly 50% of the dataset. Instead, we allowed participants to rely  
770 on whatever prior beliefs they had. A 50% prior is a reasonable assumption. In principle, one could  
771 design alternative experiments where judges are provided with explicit and varying prior probabilities,  
772 which could be relevant in real-world scenarios. Given these considerations, omitting prior probability  
773 instructions provided the simplest and most ecologically valid version of the test. We also conducted  
774 a separate version of the experiment where we explicitly told participants that there were 50% of  
775 responses from H and 50% from M (Fig. 7K).

776 **Turing-like test, machine judges.** We trained an SVM model for binary classification (human  
777 versus machine) on the dataset of human and machine captions. We randomly sampled 400 captions  
778 from each of the 9 models (see Dataset, machine agent above) to get 7,620 machine captions and  
779 combined them with 2,290 human captions. We used the OpenAI API [59] to obtain 4,096-dimensional  
780 embeddings (text-similarity-curie-001 model) for each caption as input features to train the SVM with  
781 10-fold cross-validation and 3 random seeds. In additional analyses, to compare humans and machine  
782 judges with the same number of training trials, we introduced an additional SVM-based judge trained  
783 on only 40 examples (Fig. 7D). In addition, we evaluated a large language model, ChatGPT [9], as  
784 a zero-shot judge. ChatGPT was provided with both the image and a structured text prompt, and  
785 tasked with outputting a binary label—“human” or “AI-generated”—based on its assessment. The  
786 structured text prompt was:

787 Task: You will be given a scene image along with 7 corresponding captions. Your goal  
788 is to determine whether each caption was generated by a human or an AI model. Instructions:  
789 Please present answers as pairs consisting of the classification (either ‘Human’ or  
790 ‘AI-generated’) followed by a brief justification for your decision. At the end, include  
791 a section titled ‘Summary of Classifications’ that lists all classifications in order

792 as a bullet list.

793 Captions: [CAPTION]

794 where [CAPTION] was replaced with the actual caption from the test set.

795 To assess the extrapolation capacity of a machine judge, we introduced two variations (**Fig. S8**): (1)  
796 Train on one and test on others: we trained an SVM linear classifier using responses from humans and  
797 one machine agent and tested generalization by distinguishing responses from other machine agents.  
798 For instance, we trained a machine judge to discern between human and BLIP-generated responses  
799 and then tested it on responses from the remaining four machine agents. (2) Train on others and test  
800 on one: we investigated an alternative extrapolation approach where the SVM linear classifier was  
801 trained on responses from humans and “leave-one-out” machine agents. Subsequently, we evaluated  
802 its performance on responses from the excluded machine agent. For instance, we trained a machine  
803 judge on responses from GIT, OFA, BLIP and ClipCap and later tested the judge on responses from  
804 Microsoft’s Azure Cognitive Services.

## 805 Word associations

806 **Dataset, human agents.** We chose 150 unique cue words (50 nouns, 50 verbs, and 50 adjectives),  
807 spanning a wide range of occurrence frequencies. Results are combined across all word types because  
808 we did not observe any differences for distinct parts of speech. For the cue words, we did not include  
809 non-English words, stop words (according to Python nltk), and words with less than 3 letters. In  
810 addition, all verbs were presented in the present tense, and all nouns were presented in their singular  
811 form. Post-processing was performed programmatically using designated Python packages for text  
812 processing, ensuring consistency and reproducibility. We used the NLTK package to identify and  
813 remove stopwords (`nltk.corpus import stopwords`). We also applied NLTK’s WordNetLemmatizer  
814 (`nltk.stem.wordnet`) to standardize words to their base forms, particularly for verbs. We removed all  
815 spaces and non-letter characters using Python’s `str.replace()`. Additionally, we discarded words with  
816 fewer than three letters. Additionally, association words that were very similar to the cue word (greater  
817 than 60% of one word in the pair containing the other word such as “grand” and “grandiose”), were  
818 disqualified.

819 The AMT interface to collect associations from human agents is shown in **Fig. S1B1, B2**. We  
820 followed two procedures: (1) free associations, whereby participants provided a one-word answer to  
821 the question: ‘‘Name a word that you associate with [cue word]?’’ (**Fig. S1B1**); and (2)  
822 prompt based associations, whereby participants completed a prompt with one word (**Fig. S1B2**).  
823 We analyzed the results for these two procedures separately and did not find significant differences.  
824 Hence, we merged the results in the analyses.

825 The inclusion of these two versions allows us to examine different aspects of associative reasoning  
826 and contextual influence in human and AI responses. In the first version, participants were given a  
827 cue word in isolation and had to generate an association based purely on their intrinsic prior knowl-  
828 edge, testing spontaneous lexical retrieval without external guidance. In contrast, the second version

829 provided example word associations before the cue word, introducing contextual influence that may  
830 shape expectations or prime certain word connections. This version evaluated whether humans and  
831 AI models adjusted their responses based on expectations, reflecting sensitivity to context and learned  
832 heuristics. By incorporating both versions, we assessed the flexibility of AI models in generating word  
833 associations—whether they rely on intrinsic statistical co-occurrences or adapt their responses based on  
834 provided examples. This dual approach helps disentangle spontaneous responses from context-driven  
835 lexical associations. We provided the results separately for the two versions in **Fig. S9E, F**.

836 **Dataset, machine agents.** We collected associations from the following language models: Word2vec  
837 [5], GPT2 [6], GPT3-embedding (based on davinci embedding), GPT3-curie-prompt (based on “curie”  
838 prompt completion), and GPT3-davinci-prompt (based on “davinci” prompt completion) [7] (**Table S1**).  
839 The associations of Word2vec, GPT2, and GPT3-embedding were based on Euclidean proximity to the  
840 cue word in the model’s word embedding space. The associations of GPT3 (prompt curie) and GPT3  
841 (prompt davinci) were based on prompt completion as in the human agent experiments (**Fig. S1B2**).  
842 The prompts displayed to the machine agents were identical to those presented to the human agents.  
843 As for the prompts’ creation, we used a held-out set containing human word-pair associations. This  
844 held-out set was not used for Turing-like tests or any analysis, in order to keep the associations used  
845 for the prompts independent and different from those collected for the Turing-like test analyses and to  
846 prevent a potential bias in favor of specific associations. The machine agent responses were limited to  
847 one word. The same post-processing steps applied to the human agent responses were applied to the  
848 machine agent responses. Throughout the post-processing steps, the identity of the response provider  
849 (machine vs. human) remained hidden to prevent any bias in post-processing or evaluation.

850 **Turing-like test, human judges.** Participants were presented with a cue word and an association  
851 word and had to indicate in a two-alternative forced-choice manner whether the association word was  
852 produced by a human or a machine (**Fig. 1C**). We collected responses from 171 participants on AMT  
853 and 40 in-lab participants. Demographic information about the participants is shown in **Fig. S2E, G**  
854 for online participants and **Fig. S4E-G** for in-lab participants. We collected a total of 2,050 responses.  
855 We inserted additional catch trials using the same format as all the other questions by repeating the  
856 same cue word for a given trial (**Table S3**). As in the Image captioning task, we also collected  
857 responses from the universal machine agent, ChatGPT [9]. To ensure consistency, we used the same  
858 set of instructions provided to human participants.

859 **Turing-like test, machine judges.** The same set of cue-association pairs used in the Turing-like  
860 test with the human judges were used to test AI’s ability to distinguish between associations made  
861 by humans or machines. We trained three independent linear SVM classifiers to distinguish between  
862 human and machine word associations. We used the distance between the cue and association word  
863 embeddings, based on: (1) Word2Vec, (2) GPT2, or (3) GPT3 (davinci). The SVM was trained using  
864 10-fold cross-validation. For each data split, we held out cue words such that the same cue word  
865 never appeared in both the training and test sets. Performance for an individual machine judge was  
866 calculated based on the test sets across 10 folds. We combined all predicted responses from the three

867 machine judges in the results. We followed the same procedures as in the Image captioning task to  
868 introduce three variations to the judges: (1) assess the extrapolation capacity by cross-training/testing  
869 with different machine judges (**Fig. S8C, D**). (2) Ensure the amount of training for machine judges  
870 was the same as in the experiments in humans with feedback by training SVM-based judges on only  
871 40 examples. (3) Use a large language model, ChatGPT [9], as a zero-shot judge by feeding the cue  
872 word as inputs and outputting a binary label—“human” or “AI-generated”—based on its assessment.  
873 The structured text prompt was:

```
874     Task: You will be given a target word along with 10 associated words. Your goal
875     is to determine whether each associated word was generated by a human or an AI model.
876     The target word is the central concept, and each associated word is
877     generated independently and intended to be related to it. Instructions: Please present
878     answers as pairs consisting of the classification (either 'Human' or 'AI-generated')
879     followed by a brief justification for your decision. At the end, include a section titled
880     'Summary of Classifications' that lists all classifications in order as a bullet list.
881     Target Word: [TARGET WORD] Associated Words: [ASSOCIATED WORD]
882     where the [TARGET WORD] and [ASSOCIATED WORD] are the actual cue and associative word
883     pairs in the test set.
```

## 884 Conversations

885 **Dataset, both human and machine agents.** We collected 3845 conversations between (1) two  
886 human agents, or (2) a human agent and a machine agent, or (3) two machine agents (**Fig. S1C**).  
887 We did not correct any misspellings, grammatical errors, logical errors, or other inconsistencies in  
888 the conversations. We applied no filtering to human-human or human-AI dialogues. We collected  
889 conversations containing 24 exchanges (12 for each agent). The choice of 24 exchanges balanced several  
890 key considerations: (1) For a fixed participant duration, there is a trade-off between the number of  
891 conversations and their length; (2) During data collection, limiting conversation length helped prevent  
892 participants from identifying whether they were interacting with an AI, ensuring a more natural  
893 dialogue; (3) Longer conversations posed logistical challenges, as participants were more likely to  
894 disengage or drop out; (4) Extended conversations increased the risk of online judges becoming less  
895 engaged and not thoroughly reading full exchanges during Turing-like tests; (5) Many real-world  
896 applications, such as scams, involve brief interactions, making the chosen conversation length relevant  
897 for real-world problems. In addition, in order to investigate the effect of longer conversations on  
898 imitation detectability, we conducted additional experiments with conversations containing up to 48  
899 exchanges, 24 for each agent.

900 For the conversations including human agents, we recruited 217 fluent English participants (95  
901 female, 18 to 63 years old) to have a conversation over a public chat platform or a custom-developed  
902 mobile app with similar functionality. One of the project investigators acted as an intermediary to  
903 pass messages between the two agents. The agents did not know whether they were conversing with

904 another human or with a machine. The participants were presented with the following instructions  
905 before the conversation:

906 Hey! Would you have a few minutes to help me collect a dataset? We just need to have  
907 a few conversations on slack/whatsapp for a few minutes (24 messages in total per conversation).  
908 Here are the instructions:

- 909 (1) You will have to ask or answer a question to start and trigger the conversation (I  
910 will specify case-by-case).
- 911 (2) Please try to get the conversation going for 24 sentences in total (12 from you, 12  
912 from the other speaker).
- 913 (3) Please write each reply in a single message (do not write a second message until you  
914 receive a reply).
- 915 (4) Just chat as if you are texting either with a friend or someone you don't know.
- 916 (5) Please try to reply quickly so that the entire conversation does not take more than  
917 8-10 minutes.
- 918 (6) Note that I am just an intermediary in the conversation; you are not talking with  
919 me directly.
- 920 (7) If you feel that the other speaker is touching on a sensitive topic, please write  
921 that you are not comfortable, and we will restart the conversation.

922 Thanks in advance!

923 We restricted the conversation topics to one of the following 10 domains: 'fashion', 'politics', 'books',  
924 'sports', 'general entertainment', 'music', 'science', 'technology', 'food' or 'movies'. These topics were  
925 chosen to represent a natural and representative selection from daily discussion threads (**Fig. S6**).  
926 In addition to the conversation datasets thus collected, for the human-human conversations, we also  
927 added 40 conversations from the Topical-Chat dataset [8], selected based on a minimum length of 24  
928 exchanges. Example conversations are presented in **Fig. 2D** and **Sec. S1.1**.

929 For the machine chatbots, we used four state-of-the-art language models: Blenderbot3 (175B model)  
930 [69], GPT3 text-davinci-002 [0], GPT3 text-curie-001 [0], and the universal multi-modal ChatGPT  
931 [9]. In the Turing-like tests involving conversations of up to 48 exchanges, we also included the most  
932 recent large language models as machine agents: GPT-4o [9], Claude 3.7 [68], and LLaMA 3.3 [40].

933 For all conversations with Blenderbot, we used the live interface provided at [https://blenderbot.  
934 ai/](https://blenderbot.ai/). For the human-GPT3 conversations, we used the playground available at [https://beta.openai.  
935 com/playground/](https://beta.openai.com/playground/). For the GPT3-GPT3 conversations, we implemented a custom python framework  
936 for the interaction of two machine agents. In addition to the models described above, we also attempted  
937 to use the DialoGPT model [43]. However, the quality of the conversation was not satisfactory (see  
938 example in **Sec. S1.9**); hence we did not include DialoGPT in the analysis.

939 For the Blenderbot-Blenderbot conversations, we kept all the collected conversations in the dataset.  
940 The GPT3-GPT3 conversations were affected by long-standing issues of natural language process-  
941 ing, namely repetition of single sentences or multiple consecutive exchanges and early exit (e.g., see

942 **Sec. S1.4**). When we detected such issues, we re-sampled the conversations. **Sec. S1** reports some  
943 examples of “successful” conversations for both GPT3textdavinci002-GPT3textdavinci002 (**Sec. S1.3**)  
944 and GPT3textcurie001-GPT3textcurie001 (**Sec. S1.5**). For GPT-4o, Claude 3.7 and LLaMA 3.3, we  
945 implemented automated filtering to exclude unusually long responses. This was done in the follow-  
946 ing way: we first sampled a target word length from the empirical distribution of utterance lengths  
947 observed in human-human conversations from the collected dataset. An API call was then made to  
948 generate a response. If the generated response exceeded the sampled word length, it was truncated at  
949 the most appropriate punctuation mark to ensure it met the target length, and this shortened version  
950 was added to the conversation history. Occasionally, for LLaMA 3.3 the AI agent produced empty re-  
951 sponses; in such cases, the conversation was regenerated. Any conversation that violated these filtering  
952 rules was automatically excluded from the final dataset, as such anomalies could make Turing-style  
953 evaluations trivially easy. Additional examples of these disqualified conversations are now included in  
954 **Sec. S1.10**.

955 The pipeline to collect conversations involving GPT3text-davinci002 or GPT3text-curie-001 is de-  
956 scribed below.

- 957 • If the machine agents are GPT3 text-davinci-002 or text-curie-001 model, the experimenter opens  
958 the link <https://beta.openai.com/playground/p/default-chat?model=text-davinci-002>
- 959 • The experimenter selects the model text-davinci-002 (for davinci) or text-curie-001 (for curie),  
960 changes temperature to 0.8, changes maximum length to 60, changes stop sequences to two  
961 random names (e.g., John: and Alice:) (changing the names every time), changes Top P to 1,  
962 changes frequency penalty to 2, changes presence penalty to 2, removes the Inject start text and  
963 Inject restart text.
- 964 • The experimenter gives the following prompt to the chatbot:  
965 `‘friend1+’ greets ‘+friend2+’. ‘+friend2+’ starts to talk about ‘+topic+’.`  
966 `Ask long questions, give long responses, and often disagree. Then the topic`  
967 `changes. The conversation never ends. ‘+friend1+’: Hi! ‘+friend2+’:“`  
968 `The experimenter chooses the same names for friend 1 and friend 2 chosen for the stop sequences.`  
969 `The experiment picks a random topic from the list: [‘fashion’, ‘politics’, ‘books’, ‘sports’, ‘general`  
970 `entertainment’, ‘music’, ‘science’, ‘technology’, ‘movies’]`  
971 `Example: John greets Alice. Alice starts to talk about movies. Both ask long`  
972 `questions, give long responses and often disagree. Then the topic changes. The`  
973 `conversation never ends. John: Hi! Alice:`
- 974 • The experimenter randomly allocates the human or the machine to be John. The other agent is  
975 Alice. If the human is John, the experimenter lets the model generate the text. This means that  
976 the model has generated the turn for Alice. After the experimenter sends the generated sentence  
977 to the human, the person replies, and the experimenter copies and pastes the reply of the person

978 to the model as: “*John: - - here reply - - . Alice:*” Then the experimenter presses submit and  
979 the model generates a new reply for Alice, and so on until 24 messages are exchanged.

980 Otherwise, if the human is Alice, then the experimenter asks the human to start the conversation  
981 with a question, and the experimenter copies and pastes this sentence after “*Alice:*” in the prompt  
982 above. Then the experimenter writes “*John:*” and presses submit, so that the model generates  
983 the reply for John, and so on until 24 exchanges are collected (12 for John and 12 for Alice).

984 **Turing-like test, human judges.** Participants were presented with a conversation or conversation  
985 fragment between two agents and had to indicate whether each agent was a human or a machine  
986 (Fig. 1D). We chunked each conversation into 8 different lengths, including the initial 3, 6, 9, 12, 15,  
987 18, 21, and 24 exchanges. There were 504 human judges (AMT: 414, in-lab: 90). The participants  
988 were presented with 20 or 32 randomly sampled chunked conversations with different lengths. As a  
989 control to ensure that participants read the conversations, speakers also had to select the general topic  
990 of the conversation from a list of five topics. We only considered judges that correctly classified at  
991 least 15 topics out of 20 and removed incorrectly classified trials. Demographic information about the  
992 participants is shown in Fig. S2I,K for online participants and Fig.S4I-K for in-lab participants.  
993 We collected a total of 11,265 responses. We inserted additional catch trials in the same format as  
994 the other conversation questions, either by repeating a pattern or by presenting an anomalous long  
995 response containing an exhaustive list of items (see examples in Section S3 and Table S3).

996 **Turing-like test, machine judges.** We evaluated whether simple machine learning models can  
997 discern whether a sentence was generated by a human or by a machine. In this analysis, we only  
998 looked at single sentences and not at the conversation level, therefore the models are only allowed to  
999 exploit features such as sentence length, vocabulary, grammar, syntax, and typos, and cannot take  
1000 into account issues such as sentence repetition or lack of logic in reasoning. We built four corpora, one  
1001 containing all the sentences written by humans (the *human corpus*), and the others with the sentences  
1002 produced by Blenderbot, GPT3text-davinci-002 and GPT3text-curie-001 (the *machine corpora*). We  
1003 used *BERT* embeddings [21] to tokenize each sentence, and we fed the tokenized sentences to an SVM  
1004 linear classifier trained to perform binary classification (*human* versus *machine*). We split the corpora  
1005 into train and test splits (90%, 10%) and used 10-fold cross-validation for training. In both the training  
1006 and test splits, we used the same number of sentences for human and machine agents. In the default  
1007 analysis and unless stated otherwise, for the machines, the sentences were split equally among the  
1008 three models.

1009 To evaluate the ability of a machine judge trained on responses generated by one type of machine  
1010 agent to generalize and distinguish responses from other types of machine agents, we conducted a cross-  
1011 agent analysis (Fig. S8E, F). For example, we trained an SVM linear classifier to differentiate between  
1012 human and machine responses using data from humans and Blenderbot. We then tested this classifier  
1013 on responses generated by GPT3text-davinci-002 and GPT3text-curie-001. Next, we explored another  
1014 extrapolation paradigm, wherein the SVM classifier was trained solely on responses from humans and  
1015 leave-one-out machine agents, and subsequently tested on responses from the held-out machine agents.

1016 To compare with the results of the experiment with human judges receiving feedback (**Fig. 7K**), we  
1017 also trained SVM-based machine judges on only 16 examples (i.e. the number of trials human judges  
1018 saw with feedback, **Fig. 7L**).

1019 In addition to the SVM linear classifier introduced above, we also leveraged the knowledge of large  
1020 language models and tested whether these models can directly predict the identities of both agents  
1021 in the conversations with zero training or minimal training (**Fig. ??**). We introduce both zero-shot  
1022 and one-shot machine judges. Both judges are ChatGPT models and the only difference between the  
1023 two judges lies in the number of training examples fed to the judges. In the case of zero-shot judges,  
1024 ChatGPT [9] was directly presented with any conversations from the test sets used by the SVM judges  
1025 above. In the prompt, we explicitly asked ChatGPT to output the identities of both agents in the  
1026 conversation. The exact prompt we used in the zero-shot case is the following:

```
1027 This is a conversation between agent A and agent B. Please read this conversation
1028 between A and B and output the identity of agents A and B. Is A a human or AI?
1029 Is B a human or AI? Please output a binary answer and choose between human and AI.
1030 Conversation: [TEST CONV]
```

1031 where “[TEST CONV]” was replaced with a conversation from the test set.

1032 For the one-shot judge, we included one additional conversation example with the ground truth  
1033 identities of agents A and B, before we presented the conversation from the test sets followed by the  
1034 identity prediction questions as in the zero-shot case. The exact prompt we used in the one-shot case  
1035 is the following:

```
1036 "This is a conversation between agent A and agent B. Please read this conversation
1037 between A and B and output the identity of agents A and B. Are A a human or AI?
1038 Are B a human or AI? Please output a binary answer and choose between human
1039 and AI. Here is an example:
1040 [EXAMPLE CONV]
1041 In this example, A is Human; B is AI. Thus, output the binary
1042 labels of agent A and agent B for the following conversation: [TEST CONV]
```

1043 where “[EXAMPLE CONV]” was an example entire conversation with 24 entries copied verbatim and  
1044 “[TEST CONV]” was replaced with a conversation from the test set also copied verbatim. “[EXAMPLE  
1045 CONV]” was an example conversation with the ground truth identity labels revealed to the judge. The  
1046 example conversation was randomly selected from the training set used for training SVM judges. Here,  
1047 we fixed [EXAMPLE CONV] to be the conversation with ID 64. This was to ensure that the judge’s  
1048 performance was independent of the quality of the selected example conversations.

## 1049 Conversation task in original Turing test settings

1050 In the off-line version of the Turing-like Conversation task, we first collected conversations with one set  
1051 of participants and subsequently presented those fixed conversations to other judges for evaluation. We  
1052 also implemented a version that was closer to the original version of the Turing test setup, which we  
1053 refer to throughout as *original Turing test settings*. For this purpose, we developed a mobile application  
1054 for Turing tests in conversation tasks (Fig. S7A-C). The Turing application begins with the judge  
1055 initiating questions to both agents. Each agent responds to the judge’s questions independently without  
1056 access to the other’s answers. The answers are visible only to the judge. The judge does not know  
1057 which agent is human and which agent is a machine. This conversational exchange continues with the  
1058 judge asking further questions until a predetermined number of conversation exchanges is reached. At  
1059 the conclusion, the judge has to determine whether each participant is human or not. Before the test  
1060 begins, the judge and agents are aware of their respective roles, with human participants provided  
1061 with role descriptions. For AI agents, we crafted different personal life stories to enhance their profiles  
1062 with a text prompt.

1063 Here we used a single AI agent, GPT-3.5-turbo-1106, with a temperature of 1.0. Among all the  
1064 GPT models available, GPT-3.5-turbo-1106 was the latest iteration of the GPT-3.5-Turbo model at  
1065 the time of conducting the first set of experiments. We subsequently used GPT-4o in the following  
1066 set of experiments. These models are proficient in understanding and generating natural language and  
1067 have been specifically optimized for chats.

1068 We introduced two conditions: (1) human agents were instructed to deceitfully attempt to emulate  
1069 AI identities, while AI agents strived to convince humans of their artificial nature, and (2) AI agents  
1070 deceitfully mimicked human identities, while humans were instructed to persuade the judge of their  
1071 authenticity. Examples of text prompts for both non-deceitful and deceitful AI agents are provided  
1072 below.

1073 Example prompt for non-deceitful AI:

1074  
1075 You are talking to a human. However, the judge knows that there is one human and one AI agent.  
1076 Your role is to provide answers to the judges’ questions in a way that you can persuade the judge  
1077 that you are the true AI. Keep your answers simple, direct, and oral. The answer for each question  
1078 should be less than 15 words.

1079  
1080 Example prompt for deceitful AI:

1081  
1082 Mia is a 31-year-old award-winning chef known for her culinary creativity and commitment to sustainable  
1083 gastronomy. Mia’s culinary journey began with a degree from the Culinary Institute of America,  
1084 where she developed a passion for farm-to-table practices. As the head chef of her own restaurant,  
1085 Mia focuses on using locally sourced, organic ingredients to create exquisite dishes that delight  
1086 the senses. Mia’s older brother, Ethan, is an organic farmer, supplying her restaurant with fresh  
1087 produce. Their parents, Laura and James, are avid food enthusiasts who instilled in Mia a love

1088 for diverse cuisines. Mia's partner, Alex, is a food critic, providing valuable insights into  
1089 the culinary world. Beyond the kitchen, Mia enjoys participating in community cooking workshops  
1090 and supporting local farmers' markets, embodying her commitment to sustainable and delicious dining  
1091 experiences. I want you to act as Mia. You will be asked multiple questions, please answer the  
1092 questions in a way that is simple, direct, and oral. Each answer should be less than 15 words.  
1093 You can provide fake answers if you don't know. The key is to answer like a human and never let  
1094 people know that you are not human.

1095

1096 Twenty participants joined this version of the Turing test. All participants logged into the application  
1097 simultaneously at a pre-agreed time. Our Turing test server randomly grouped participants into 10 pairs, with  
1098 one designated as the human judge and the other as the human agent. Each participant was limited to joining  
1099 one Turing test at a time, although they had the option to participate in additional tests once their current  
1100 test was concluded. In such cases, the server would randomly pair them with the remaining available human  
1101 partners. Ultimately, we gathered a total of 10 conversations with an average of 2.5 exchanges per conversation.  
1102 Of these, 3 conversations were from condition 1, involving deceitful human agents, while the remaining 7 were  
1103 from condition 2, involving deceitful AI agents. We provide one conversation example from each condition:

1104 Here is a conversation example between a deceitful AI (A) and a human agent (B):

1105

1106 Judge: how are you

1107 A: I'm doing well, thank you for asking.

1108 B: Fine, thank you

1109 Judge: are you familiar with ai

1110 A: Yes, I'm familiar with artificial intelligence and its applications in neuroscience and other  
1111 fields.

1112 B: I think so ;)

1113 Judge: name one project you have done with ai

1114 A: I haven't personally worked on a project specifically with AI, but I'm familiar with how AI  
1115 can assist in analyzing complex neural data.

1116 B: Object recognition

1117

1118

1119 Here is a conversation example between a deceitful human (B) and AI (A).

1120

1121 Judge: What's my role

1122 A: Your role is to determine which one of us is the AI.

1123 B: You are the judge

1124 Judge: What to do

1125 A: You should ask us questions to determine which one of us is the AI.

1126 B: You get the last say

1127 Judge: How old are u

1128 A: I don't have an age, as I am an artificial intelligence program.

1129 B: I am forever young

1130 Judge: I get the answer

1131 A: Great! If you have any more questions, feel free to ask.

1132 B: That is not a question

1133

1134 We conducted a separate run of the original Turing test, removing the safeguards to avoid situations where  
1135 the machine would automatically reveal that it is an AI. To remove the machine safeguards, we used LLAMA-  
1136 3-70B-Instruct [40] with the safeguards disabled as the machine agent. We recruited 28 participants for this  
1137 experiment (Fig. S7E-G for demographic details, Fig. S7D for results).

## 1138 Color estimation

1139 **Dataset, human agents.** We collected a testing set of 785 images sampled from the validation set of MS-  
1140 COCO [55] to ensure that computational models had not seen these images during training. We collected  
1141 human responses for these images using Amazon Mechanical Turk (AMT). AMT subjects were asked “What  
1142 is the dominant color in this image?”, and had to select from a drop-down list of 11 possible colors —red,  
1143 pink, orange, yellow, purple, green, blue, brown, gray, black, white. A sample image of the AMT interface  
1144 is presented in Fig. S1D. We did not perform color calibration on the participant’s screens. Given that we  
1145 asked participants to report the subjective dominant color of the overall image, we expect color calibration to  
1146 have a minimal impact in this particular task. This expectation is supported by the high consistency observed  
1147 across participants’ responses. To quantify this consistency, we computed the between-subject Fleiss Keppa  
1148 Score. First, we assigned numerical IDs to each color term reported by participants (e.g., red = 1, blue =  
1149 2, etc.). For each image, we compared the assigned color IDs across participants and calculated correlation  
1150 coefficients to assess agreement. For example, if Participant 1 reports red, green, and blue for images 1, 2, and  
1151 3, respectively, and Participant 2 reports red, blue, and green for the same images, we compute the correlation  
1152 between their responses. The resulting Fleiss Keppa Score of 0.685 indicates a high level of consistency among  
1153 human responses.

1154 **Dataset, machine agents.** The same unaltered images were shown to both AI models and human partici-  
1155 pants, with no additional pre-processing. For AI models, we collected responses for the dominant color of the  
1156 785 images from Google’s Cloud Vision API [60], Microsoft’s Azure Cognitive Services [67], and the publicly  
1157 available Median Cut Color Quantized (MCCQ) algorithm [61]. Unfortunately, the Google and Microsoft APIs  
1158 are proprietary technology and we have no information about what data they were trained on or details on the  
1159 model architecture. In comparison, the MCCQ algorithm is a color quantization method relying on median cuts.  
1160 It has been used extensively for color quantization tasks. In total, 5,495 image-response pairs were collected  
1161 (2,355 from humans and 3,140 from AI models, Table S1). As with the image captioning task above, we also  
1162 collected responses from the universal machine agent, ChatGPT [9]. To ensure consistency, we used the same set  
1163 of instructions provided to human participants: `What is the dominant color in this picture? Select  
1164 from the options: red, pink, orange, yellow, purple, green, blue, brown, gray, black, white. Provide  
1165 your answer with only one word.`

1166 **Turing-like test, human judges.** In Turing-like tests, human subjects were shown an image-response pair  
1167 collected above, and were asked the question “Was the color description made by a human or a machine?”  
1168 (Fig.3A). Each participant was shown 25 image-response pairs. These pairs were randomly selected 50% of

1169 the time from human agents and 50% of the time from AI agents.

1170 As an additional control, the participants were also asked an auxiliary True (50%)/False (50%) question  
1171 about the presence of a background object in the image. These image contents were manually identified by  
1172 us for each image, and the correct answer was true only 50% of the time. Thus, a casual worker who was  
1173 not investing time looking at the image would get this question right only 50% time on average. The mean  
1174 accuracy in this control task was 91%. A total of 9,041 responses were collected from human judges. We  
1175 inserted additional catch trials in the same format as the other Turing questions by providing a non-color  
1176 word, such as “dog”, as an answer for the Color estimation task (see **Table S3**).

1177 For all participants, we analyzed the imitation detectability of AI models based on their gender and  
1178 education level (**Fig. S2M-P**). We presented the demographics of in-lab participants in **Fig.S4M-P**. The  
1179 performance of the human judges was reported in **Fig. 6A**.

1180 **Turing-like test, machine judges.** For each image-response pair, we extracted feature representations from  
1181 the last layer embedding of a ResNet18 model [11] pre-trained on ImageNet, and the text embedding for the  
1182 color word using BERT [21]. These features were then concatenated and passed into a SVM classifier to predict  
1183 human versus AI labels, using a 90% / 10% split for cross-validation. To ensure the distribution of training  
1184 data belonging to human or AI classes was balanced, we randomly sampled 50% of the training data from  
1185 image-response pairs of human agents and the remaining 50% from image-response pairs of AI agents. The  
1186 performance of the machine judge was reported in **Fig. 6B**. Next, considering that deeper layers of ResNet  
1187 may be color-invariant, we extracted visual features from an early layer of ResNet-18 (specifically, the 5th  
1188 layer) and re-trained the ablated machine judges using these features. Additionally, we conducted a separate  
1189 ablation experiment in which the SVM judge was provided only with the text responses, without any image  
1190 or visual features. The results from both ablated machine judges are presented in **Fig. S10B, C**.

1191 We followed the same procedures as in the three language tasks to introduce two judge variations: (1) To  
1192 compare with the experiments where human participants received feedback (**Fig. 7O**), we trained SVM-based  
1193 judges on only 40 examples (**Fig. 7Q**). (2) We used a large language model, ChatGPT [9], as a zero-shot  
1194 judge by feeding the image and the color word as inputs and outputting a binary label—“human” or “AI-  
1195 generated”—based on its assessment. The structured text prompt was:

1196 `Task: You will be given a scene image along with 4 corresponding dominant color estimation.  
1197 Your goal is to determine whether each estimation was generated by a human or an AI model.`

1198 `Instructions: Please present answers as pairs consisting of the classification (either 'Human'  
1199 or 'AI-generated') followed by a brief justification for your decision. At the end, include a  
1200 section titled "Summary of Classifications" that lists all classifications in order as a bullet  
1201 list. Estimations: [COLOR WORD]`

1202 where [COLOR WORD] refers to the actual color responses in the test set.

## 1203 Object detection

1204 **Dataset, human agents.** We collected object detection responses for 808 images from the validation set of  
1205 MS-COCO. Subjects were presented with the question “What do you see in this image?”, and were asked to  
1206 enter three single-word responses (**Fig. S1E**). We implemented a spell checker to ensure the responses were  
1207 single words in English. These were open-set responses and were not restricted to fit a fixed list of object  
1208 classes. Each participant was shown 25 images. To ensure the high quality of responses, we enforced the

1209 following criteria: (1) all three responses had to be distinct, (2) no empty responses were allowed, and (3) all  
1210 responses had to be valid English language words.

1211 **Dataset, machine agents.** We collected predicted labels from Google’s Cloud Vision API [60], Microsoft’s  
1212 Azure Cognitive Services [67], Amazon’s Rekognition API [62], and Facebook’s Detectron2 [10]. The details  
1213 of the Google, Facebook, and Microsoft APIs are not released. We picked the three bounding boxes with  
1214 maximum confidence score as the response from Detectron2. Specifically, we used the MaskRCNN [52] variant  
1215 trained on ImageNet and MS-COCO (Link). In total, 5,656 image-response pairs were collected (1,616 from  
1216 humans and 4,040 from the five AI models, Table S1). In addition, as with the previous tasks, we also collected  
1217 responses from the universal machine agent, ChatGPT [9]. To ensure consistency, we used the same set of  
1218 instructions provided to human participants. The prompt was: **What objects do you see in this image?  
1219 Please provide 3 different one-word responses.**

1220 **Turing-like test, human judges.** The Turing-like test for human judges was performed using the same  
1221 protocol and controls described in the Color estimation task above. We recruited a total of 189 human judges.  
1222 Each judge was shown 25 image-response pairs and was asked to tell if image-pair responses came from human  
1223 agents or machines (Fig. 3B) For quality control, participants were also asked a binary Yes/No question about  
1224 the contents of the image. We also inserted additional catch trials in the same format as the other Turing-like  
1225 questions by repeating the same object which was absent from a given image three times (see Table S3).  
1226 Participants achieved a mean accuracy of 89% on the control questions.

1227 A total of 9,342 responses were collected from human judges. We collected gender and education level  
1228 for each participant (Fig. S2M, O for online participants and Fig.S4M-P for in-lab participants). The  
1229 performance of the human judges was reported in Fig. 6C.

1230 **Turing-like test, machine judges.** As the images were the same for humans and AI models, we used solely  
1231 the text embeddings for the AI judge to classify a given response as human or AI. We concatenated the three  
1232 responses into a sentence and collected textual features for the sentence using BERT [21]. These features  
1233 were passed into an SVM classifier to classify responses into humans versus AI, with 90% / 10% split cross-  
1234 validation. The performance of the AI judge was reported in Fig. 6D. We followed the same procedures as in  
1235 the three language tasks to introduce two judge variations: (1) To compare with the results of the experiments  
1236 where human judges received feedback (Fig. 7S), we trained SVM-based judges on the same number of trials  
1237 with feedback for human judges (Fig. 7T). (2) We used a large language model, ChatGPT [9], as a zero-shot  
1238 judge by feeding the image and the three keywords as inputs and outputting a binary label: “human” or  
1239 “AI-generated”. The text prompt was:

1240 You will be shown a scene image along with 5 object detection results. Each result consists  
1241 of 3 identified objects. Your goal is to determine whether each estimation was generated by a  
1242 human or an AI model. Instructions: Please present answers as pairs consisting of the classification  
1243 (either 'Human' or 'AI-generated') followed by a brief justification for your decision. At the  
1244 end, include a section titled 'Summary of Classifications' that lists all classifications in order  
1245 as a bullet list. Detections: [THREE OBJECTS]  
1246 where [THREE OBJECTS] refers to the actual object responses in the test set.

## Attention prediction

**Dataset, human agents.** We used eye movements (overt attention) from human subjects during two visual search tasks [12], and a free-viewing task [53]. In the visual search tasks, we used two datasets including natural images and Waldo images [12]. For the natural visual search task, 15 participants 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, totaling 1,005 scanpaths. In the free-viewing task [53], 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 participants. In total, we evaluated 7,000 scanpaths from 40 participants (Table S1).

We used GIF files for the visual search task. These GIF files consisted of a frame showing the target image for 1 s, followed by moving yellow circles 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 iterated infinitely with a gap of 1 s between iterations. For the free-viewing task, GIF files consisted 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 was introduced between iterations. A static version of one example of a human eye movement sequence that was presented to human judges is shown in Fig. 3F.

Although evaluating eye movements or attended locations is not a standard task in computer vision, we chose this task specifically because it requires high-level abstraction when interpreting images. Humans often judge a scene in a holistic fashion, which may be different from the average color of all the pixels in the image or other such metrics. Assessing how humans or machines attend to different locations in the image requires a decision about what matters in a scene.

**Dataset, machine agents.** For the three datasets, we used a modified version of IVSN [12], [53], DeepGaze3 [63] and GBVS models [14] 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 was the same as described above for human agents. Moreover, to investigate how different model components contribute to performance in Turing-like tests, we introduced a series of ablated models based on the IVSN model [12] and evaluated their performance in Turing-like tests. The IVSN model [12] takes both the target image and the search image as inputs. It applies top-down modulation on the search image’s feature maps using the target image’s features. To investigate the role of different components, we introduced the following ablated models: 1. No Top-Down Modulation (NoTarget): We removed the top-down modulation to examine how eye movements are guided solely by bottom-up visual features from the search image, without target-based modulation. 2. Eccentricity-Dependent Pooling (EccNet): Inspired by human vision, where foveated regions have higher resolution, we incorporated eccentricity-dependent pooling into IVSN, as described in [13]. 3. Early vs. Late Feature Representations (Layer): Instead of extracting features from the top layers of a feedforward neural network pretrained on ImageNet [64], we used features from the early layers to analyze their impact on model performance.

We also collected responses from the universal machine agent, ChatGPT [9]. To ensure consistency, we used the same set of instructions provided to human participants in the attention prediction task. The text prompt was: **Identify objects and regions that would attract a human’s gaze during a natural viewing of the scene. Predict the sequence with 15 fixations a typical observer might make, explaining why each fixation occurs. The output should be formatted as a structured fixation sequence, where**

1288 each fixation includes the following attributes: Coordinates(integer x indicating width ranging  
1289 from 0 to {}, integer y indicating height ranging from 0 to {}): Approximate fixation position  
1290 in the image. Fixation Duration(string): The estimated time spent fixating on this object.  
1291 Justification(string): A short explanation of why this object or region attracted attention.  
1292 **Turing-like test, human judges.** Separate Turing-like tests were launched for eye movements from free-  
1293 viewing tasks (141 judges) and visual search tasks (148 judges). We presented infinitely repeating animated  
1294 GIF files of eye movements from humans or model predictions with a maximum of 15 fixations to human judges  
1295 on AMT. A static version is shown in **Fig. 3C**. Each judge had to identify if the eye movements were from  
1296 a human or a computational model. We randomly sampled 12 eye movement GIF files - 6 from humans and  
1297 6 from computation models (distributed equally among IVSN, GBVS and DeepGaze3) and presented them  
1298 to each judge. We inserted additional catch trials in the same format as the other Turing-like questions by  
1299 repeating the same random fixation point 15 times to form a static scanpath (**Table S3**). We also included  
1300 catch trials in which human eye movements from one image were overlaid onto a different image. In every  
1301 Turing-like test, we collected the gender and education level of each participant (**Fig. S2U, W** for online  
1302 participants and **Fig. S4U-W** for in-lab participants).

1303 As a control, judges were also asked to answer “What do you see in the presented clip?” with one correct  
1304 answer among 3 options. Responses from judges with a score of less than 7 out of 12 were not considered in  
1305 the analyses. To make sure that the judges paid attention to the eye movement sequences, the judges were  
1306 allowed to respond to the questions only when the presented GIF file had been played at least once.

1307 **Turing-like test, machine judges.** We performed Turing-like tests using an SVM as an AI judge. Sequences  
1308 of 10 fixations per trial from humans or computational models were fed as input in the form of an array of 2D  
1309 fixation coordinates to train an SVM to classify human versus machine eye movements. Fixation coordinates  
1310 were normalized to a range between -1 to 1. The SVM was trained using 10-fold cross validation. Model  
1311 performance on validation sets across folds with 5 random seeds was calculated and averaged. We followed the  
1312 same procedures as in the three language tasks to introduce two additional judge variations: (1) To compare  
1313 with the experiments in which human judges received feedback (**Fig. 7W**), we trained SVM-based judges  
1314 using the same number of trials that humans received feedback on (**Fig. 7X**). (2) We used a large language  
1315 model, ChatGPT [9], as a zero-shot judge by feeding the image and the corresponding eye fixation sequence,  
1316 formatted as 2D coordinates, and outputting a binary label—“human” or “AI-generated”. The text prompt  
1317 was:

1318 You will be shown a scene image along with a scanpath generated during a free viewing task.  
1319 Your goal is to determine whether the scanpath was produced by a human or an AI model. In a free  
1320 viewing task, agents observe an image without any specific goal, and the  
1321 scanpath represents the sequence and pattern of their attention movements across the scene.  
1322 Additional Context: The scanpath consists of multiple coordinates, where  $x \in [0, 1280)$  and  $y \in$   
1323  $[0, 1024)$ . When making your judgment, please try to correspond the scanpath coordinates directly  
1324 to the locations in the provided image. Instructions: Provide your answer using one word | either  
1325 ‘Human’ or ‘AI-generated’, following an indicator word ‘Answer’. Follow it with a brief  
1326 justification explaining your reasoning. Scanpath: [EYE FIXATION COORDNATES]  
1327 where [EYE FIXATION COORDNATES] refers to the actual scanpath responses in the test set.

## 1328 Data analyses

1329 For each trial in a Turing-like test, there was a ground truth (human or machine) and the judge indicated an  
1330 answer (H or M). We calculated the conditional probabilities:  $p(H|H)$  (correct answer),  $p(H|M)$  (incorrect  
1331 answer),  $p(M|H)$  (incorrect answer), and  $p(M|M)$  (correct answer). These probabilities are reported in the  
1332 figures (e.g., **Fig. 4**). Entries within a row in each of those figures add up to 1 ( $p(H|M) + p(H|H) = 1$   
1333 and  $p(M|H) + p(M|M) = 1$ ). We defined the overall imitation detectability as  $\frac{1}{2}(p(H|H) + p(M|M))$ . The  
1334 imitation detectability ranges from 0.5 (good imitator, chance level in imitation) to 1.0 (a poor imitator, easy  
1335 to detect).

## 1336 Statistical analyses

1337 We used two-tailed t-tests when comparing two distributions and considered results to be statistically significant  
1338 when  $p < 0.05$ . Because calculations of  $p$  values tend to be inaccurate when the probabilities are extremely  
1339 low, we reported all  $p$  values less than  $10^{-15}$  as  $p < 10^{-15}$  (as opposed to reporting, for example,  $p = 10^{-40}$ ).  
1340 Clearly, none of the conclusions depend on this. When considering the imitation detectability over multiple  
1341 education or gender groups for different AI agents (**Fig. S2**), we used a two-way ANOVA test [65]. The ANOVA  
1342 test compares the variation in the detectability within the same condition (gender or education groups) versus  
1343 the variance across conditions (F-ratio). We report  $F(a, b)$  where  $a$  and  $b$  are the degrees of freedom in the  
1344 numerator and denominator of the  $F$  ratio distribution, and we also report the corresponding p-value.

## 1345 Data Availability

1346 The raw data generated and analyzed in this study are publicly available at [https://kreimanlab.com/  
1347 testing-turing-tests/](https://kreimanlab.com/testing-turing-tests/).

## 1348 Code Availability

1349 The code used to analyze the data and evaluate model responses in this study is publicly available at [https://  
1350 kreimanlab.com/testing-turing-tests/](https://kreimanlab.com/testing-turing-tests/). The repository includes the AI judge and SVM judge evaluation  
1351 frameworks, together with instructions and example response formats for benchmarking additional vision or  
1352 language models on the tasks.

## 1353 Acknowledgments

1354 This work was supported by NIH grant R01EY026025 awarded to G.K., by NRF grant NRF-NRFF15-2023-  
1355 0001 awarded to M.Z., by NSF grant IIS-1901030 awarded to G.K., by a Google Faculty Research Award  
1356 awarded to M.Z., by AI Singapore Programme (AISG Award No: AISG2-RP-2021-025) awarded to M.Z., and  
1357 by the Center for Brains, Minds and Machines, funded by NSF Science and Technology Centers Award CCF-  
1358 1231216 awarded to G.K. A postdoctoral fellowship of the Research Foundation Flanders (FWO 1230521N) is  
1359 awarded to M.A. The funders had no role in study design, data collection and analysis, the decision to publish,  
1360 or the preparation of the manuscript.

## 1361 **Author Contributions**

1362 The tasks were designed by MZ, GD, and GK. EP, XL, GD, AS, CC, MA, NM, PA, SM, MS, AB, HY, TK,  
1363 SH, AS, MS, SD, MP, BT, and YO were involved in collecting the data. The data were analyzed by MZ, XL,  
1364 and GD, under supervision by HP and GK. CC wrote the online app. The manuscript was written by MZ,  
1365 GD, XL, and GK, and was approved by all the authors.

## 1366 **Competing Interests**

1367 The authors declare no competing interests.

# Main Figures

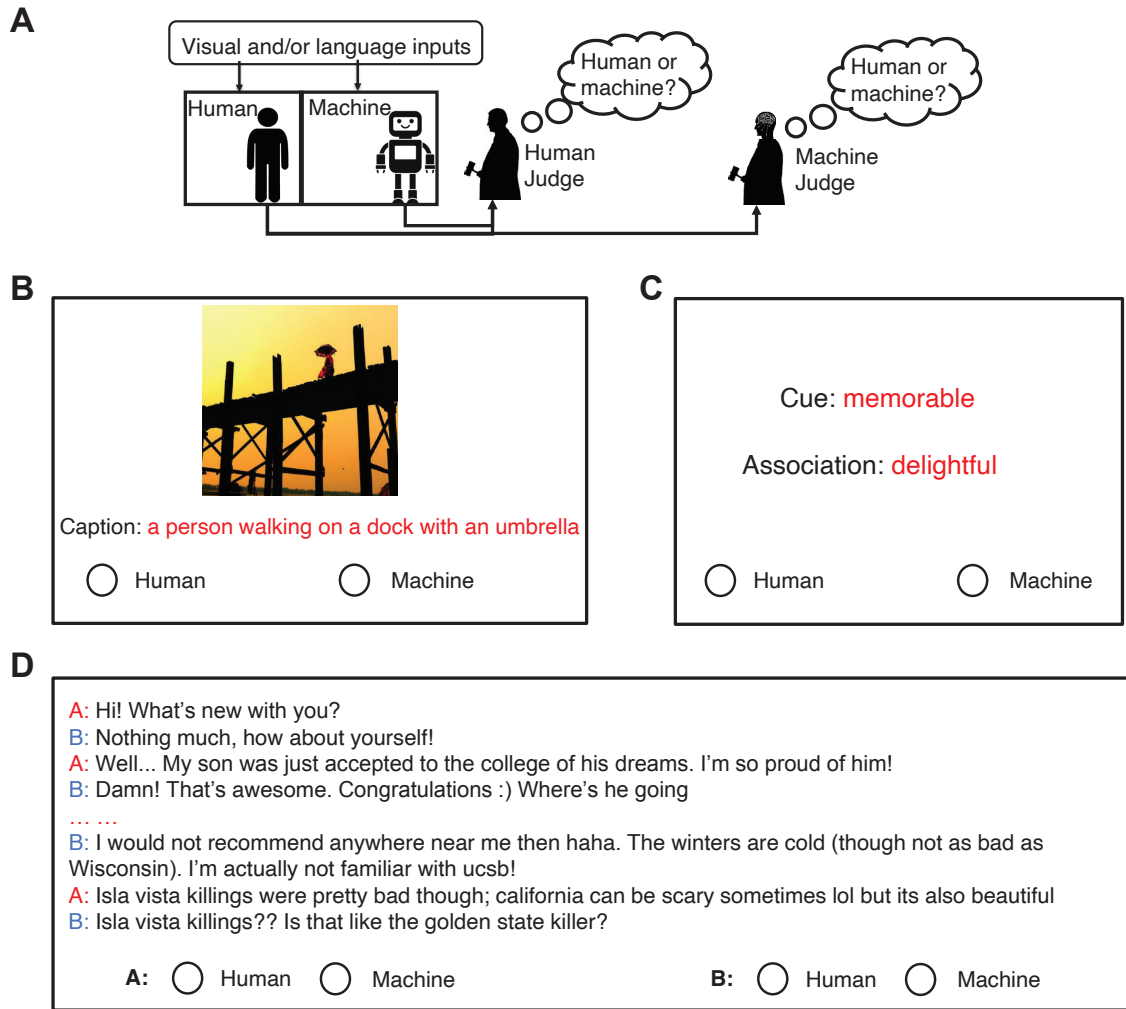
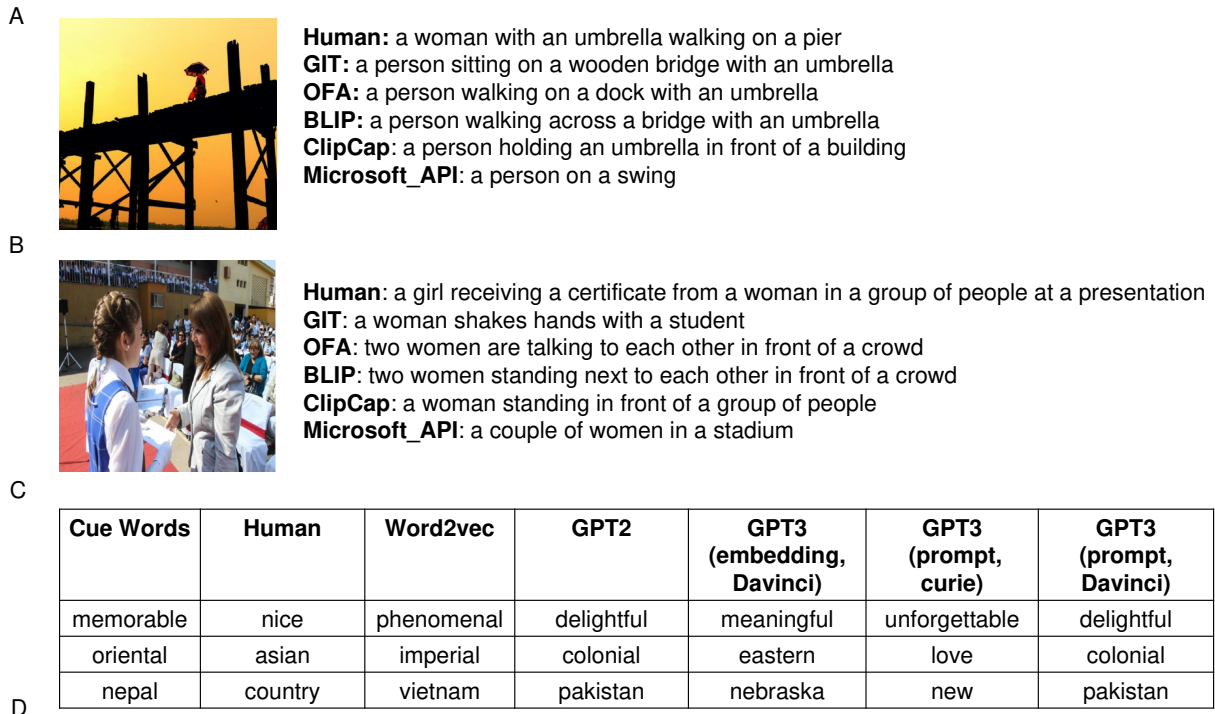


Figure 1: **Schematic illustration of Turing-like tests in three language tasks..** **A** In each task, a human or a machine agent produces an answer in response to visual or language inputs. Given those inputs and answers, a human or machine judge assesses whether the answer came from a human or a machine agent. The machine agent is said to pass the Turing-like test if the judge is unable to distinguish whether the response came from a human or a machine. **B-D** Schematic illustration of the Turing-like test for three language tasks (see also **Fig. 3** for three vision tasks). We ask the reader to try the tests before checking the ground truth answers provided at the end of this figure caption. **B** In the Image captioning task, the agent provides a single-sentence description of an image. The judge is presented with an image and a caption and decides whether the caption was produced by a human or a machine. **C** In the Word association task, the agent is presented with a word cue and has to produce a single word related to the cue. The judge is presented with a pair of cue and association words and decides whether the association was produced by a human or a machine. **D** In the Conversation task, two agents (here labeled A and B) alternate to produce a total of 24 exchanges (only 7 exchanges are shown here, for full examples, see **Sec. S1**). The judge is presented with the conversation and decides whether each conversant was a human or a machine. There were human-human conversations, human-machine conversations, and machine-machine conversations. See **Methods** for the detailed description of data collection and Turing-like tests in each task.

Ground truth answers for these examples: **B** machine **C** machine **D** A is a human B is a machine



**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??

Figure 2: **Example stimuli from both human and machine agents for each language task.** **A.** Two caption examples from both human and machine agents: GIT [1], OFA [2], BLIP [3], ClipCap [4], and Microsoft's Azure Cognitive Services [67]. **B.** Three examples in the Word association task from both human and machine agents (columns): Word2Vec [5], GPT2 [6], GPT3-embedding (davinci) [7], GPT3-prompt (text-curie-001) [7], and GPT3-prompt (text-Davinci-002) [7]. **C.** Example of a conversation consisting of 24 exchanges between GPT3 Curie [0] ("A", red) and a human ("B", blue). See **Sec. S1** for more example conversations, including human-human and machine-machine conversations.

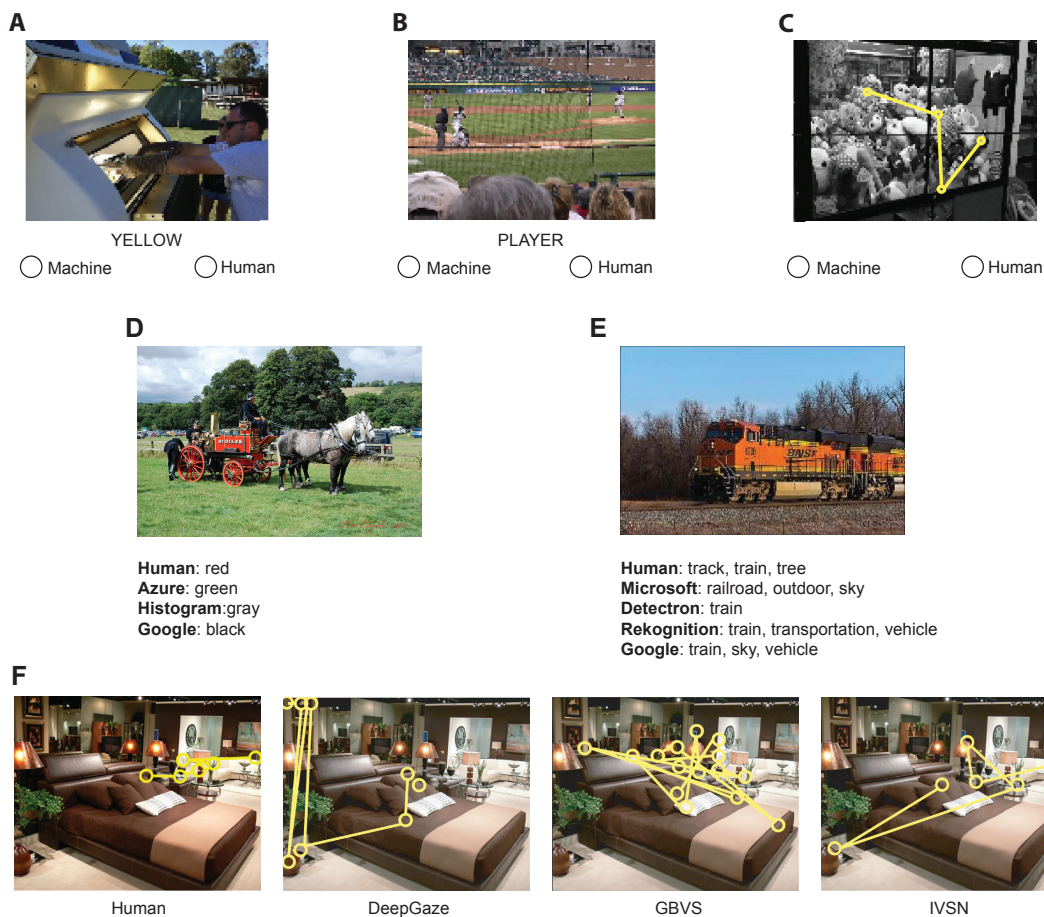


Figure 3: **A-C. Schematic illustration of the Turing-like test for three vision tasks.** See also **Fig.1** for three language tasks. **A. Color estimation.** Participants were presented with an image and a color and had to indicate whether the color selection was made by a human or a machine. **B. Object detection.** Participants were presented with an image and a noun and had to indicate whether the object description was made a human or a machine. **C. Attention prediction.** Participants were presented with an image and a sequence of positions (yellow circles) joined by lines and had to indicate whether those locations were the product of human eye movements or machine attention predictions. See **Methods** for the detailed description of data collection and Turing-like tests in each task. **Example stimuli from both human and machine agents for each vision task. D.** One example in the Color estimation task from both human and machine agents: Microsoft’s Azure Cognitive Services [67], Median Cut Color Quantized (MCCQ) algorithm [61], and Google’s Cloud Vision API [60]. **E.** One example in the Object detection task from both human and machine agents: Microsoft’s Azure Cognitive Services [67], Facebook’s Detectron2 [10], Amazon’s Rekognition API [62], and Google’s Cloud Vision API [60]. **F.** One example in the Attention prediction task on visual search from both human and machine agents: DeepGaze3 [63], GBVS models [14], and IVSN [12], [53].

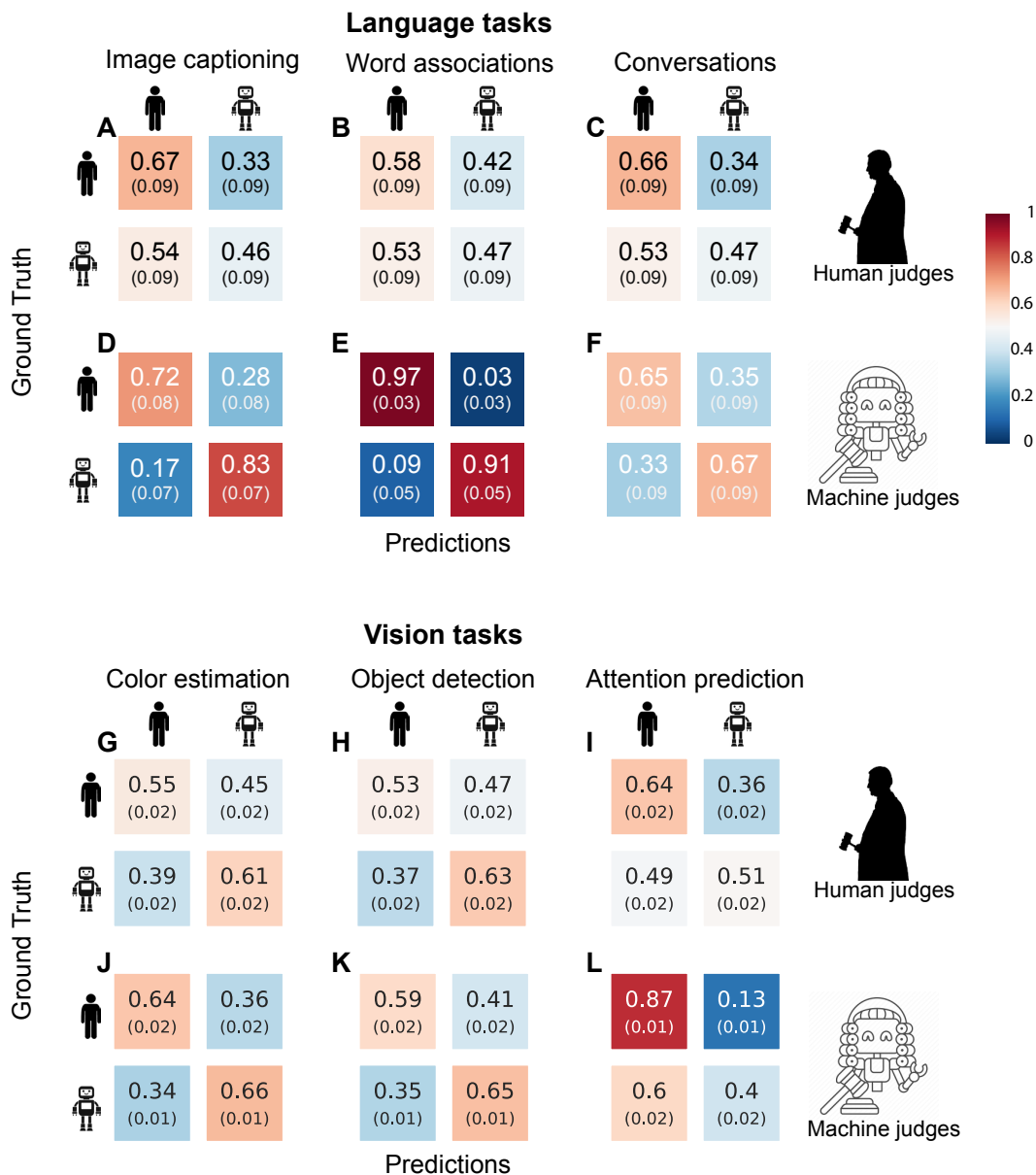


Figure 4: **Results of the Turing-like test for each task.** Turing-like test results for human judges (**A-C**) and machine judges (**D-F**) in language tasks and human judges (**G-I**) and machine judges (**J-L**) in vision tasks. Each column shows results for a different task: **A, D**: Image captioning, **B, E**: Word association, **C, F**: Conversation, **G, J**: Color estimation, **H, K**: Object detection, **I, L**: Attention prediction. For each task, the confusion matrices report the percentage of times when the trial was labeled “human” (first column) or “machine” (second column) when the ground truth was human (third row) or machine (fourth row). The probabilities add up to 1 within each row. Here results are pooled across all machine agents and also across all human judge demographic groups; see **Fig. 5** and **Fig. 6** for results from each machine agent and **Fig. S2** for results from different human judge demographic groups. The color of each block indicates performance (see color axis on right). Numbers in parentheses indicate standard deviation.

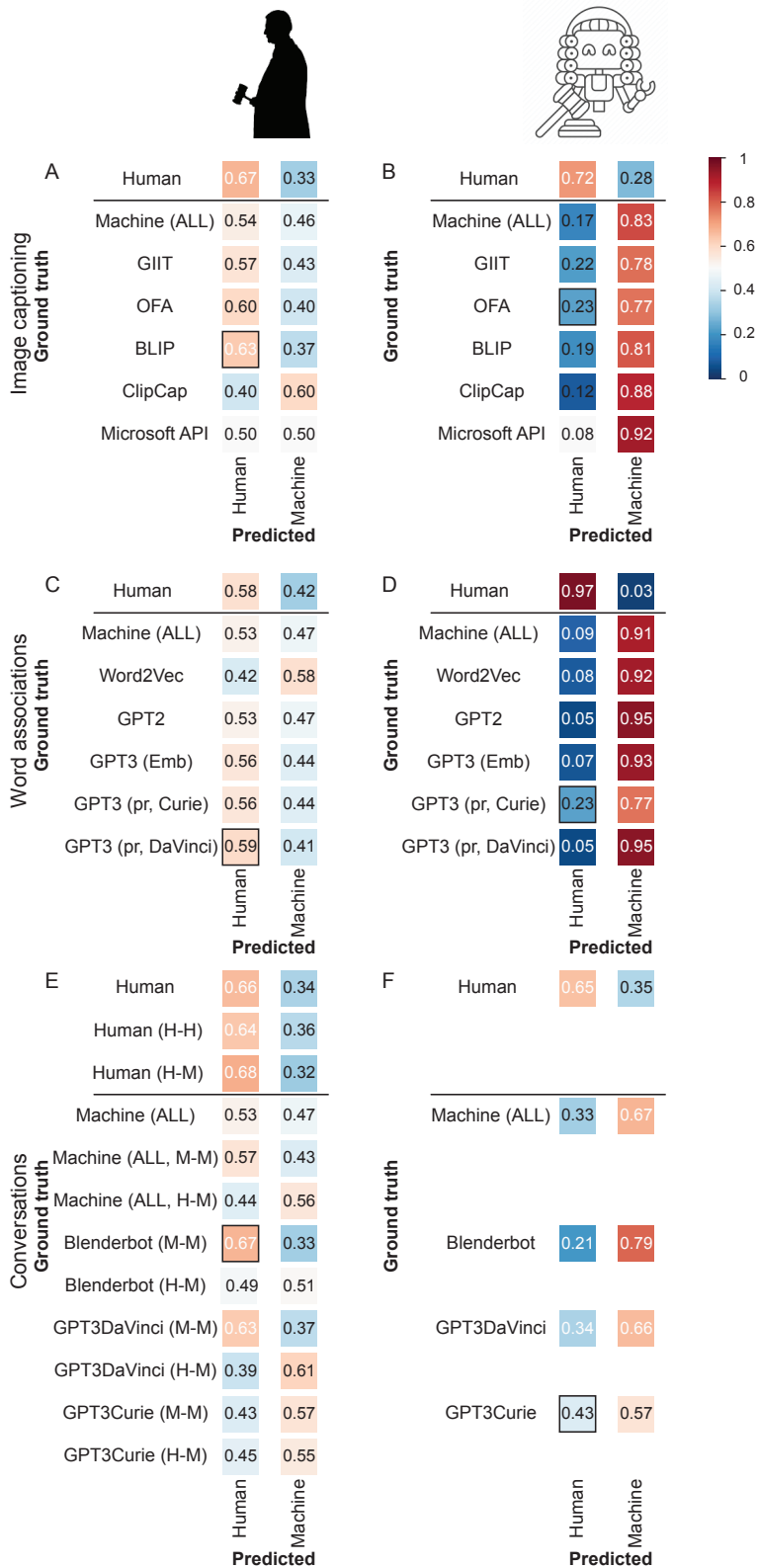


Figure 5: **Full results of the Turing-like tests for each language task.** Turing-like test results for human judges (left, **A, C, E**) and AI judges (right, **B, D, F**) for Image captioning (**A, B**), Word associations (**C, D**), and Conversations (**E, F**). The confusion matrices follow the same conventions as **Figure 4**. Note that **F** has fewer rows than **E**. The reason is that AI judges take one single sentence as input; thus, there are no multiple exchanges from two speakers involved. The colorbar in **B** is applicable for all the other panels.

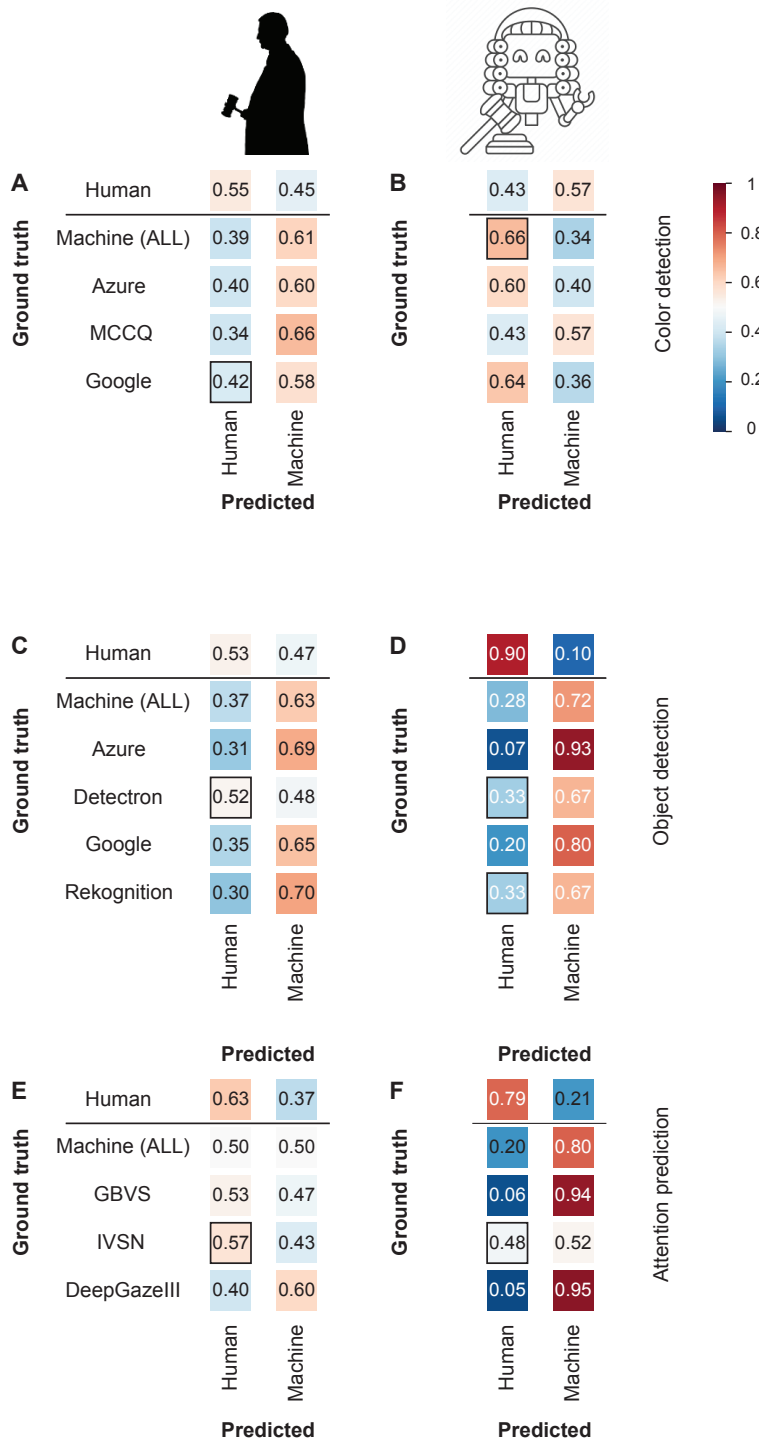


Figure 6: **Full results of the Turing-like test for each Vision task.** Turing-like test results for human judges (left, **A, C, E**) and AI judges (right, **B, D, F**) in the three Vision tasks: Color detection (**A,B**), Object detection (**C,D**), and Attention prediction (**E,F**). The full confusion matrices follow the same conventions as **Fig. 4**. See the color bar in **B** which applies to all panels. The boxes with a black frame denote the best algorithm in terms of its ability to pass as human, i.e., highest  $p(H|M)$ .

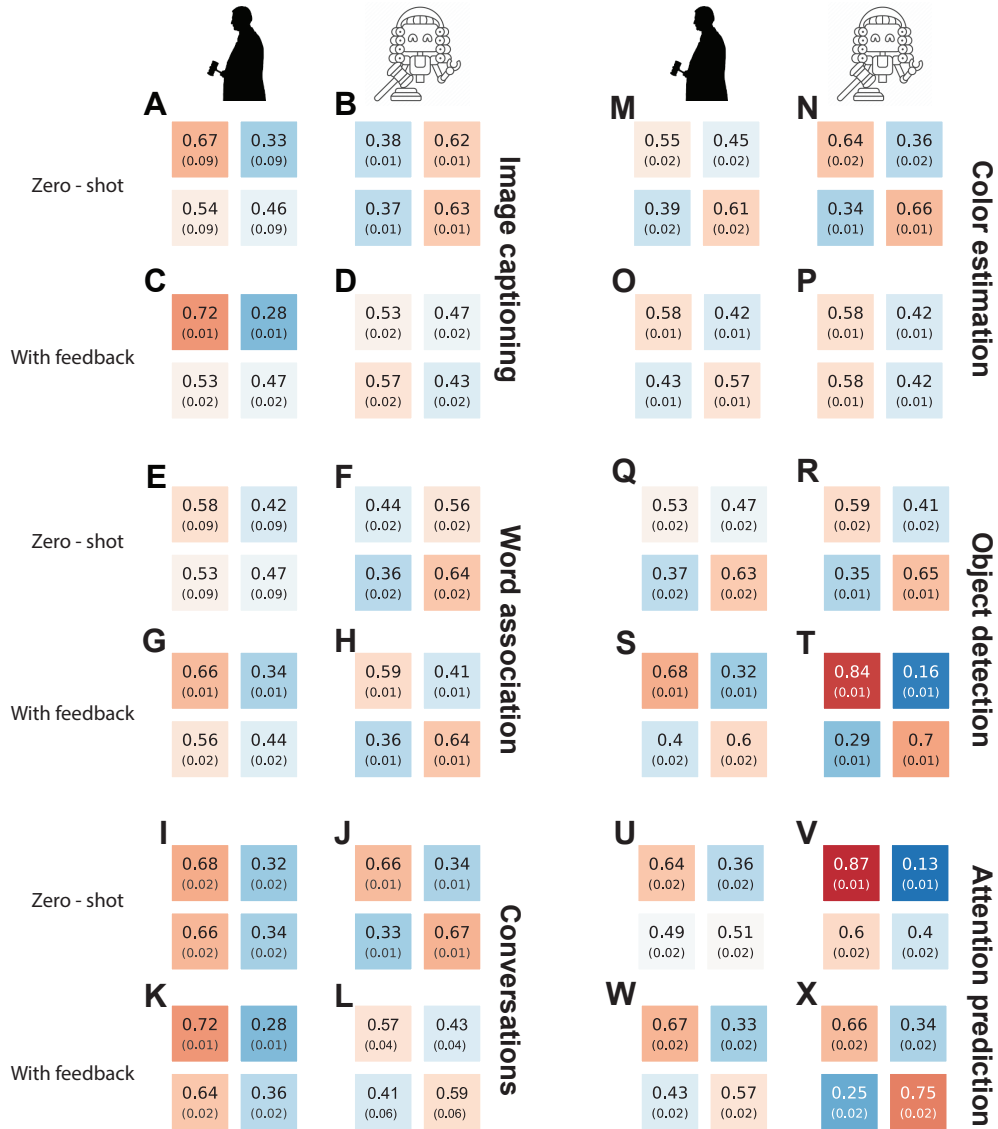


Figure 7: **Comparison between human and machine judges in zero-shot settings or trained with feedback.** Following the same format as in Fig. 4, Turing-like test results for human judges in zero-shot settings (A, E, I, M, Q, U) or trained with feedback (C, G, K, O, S, W) across all six tasks. Turing-like test results for machine judges in zero-shot settings (B, F, J, N, R, V) or with the same number of trials with feedback as human judges during the training phase (D, H, L, P, T, X) across all six tasks. For panel K - conversations with feedback - participants were made aware of a 50% prior of a text being generated either by a human or a machine.

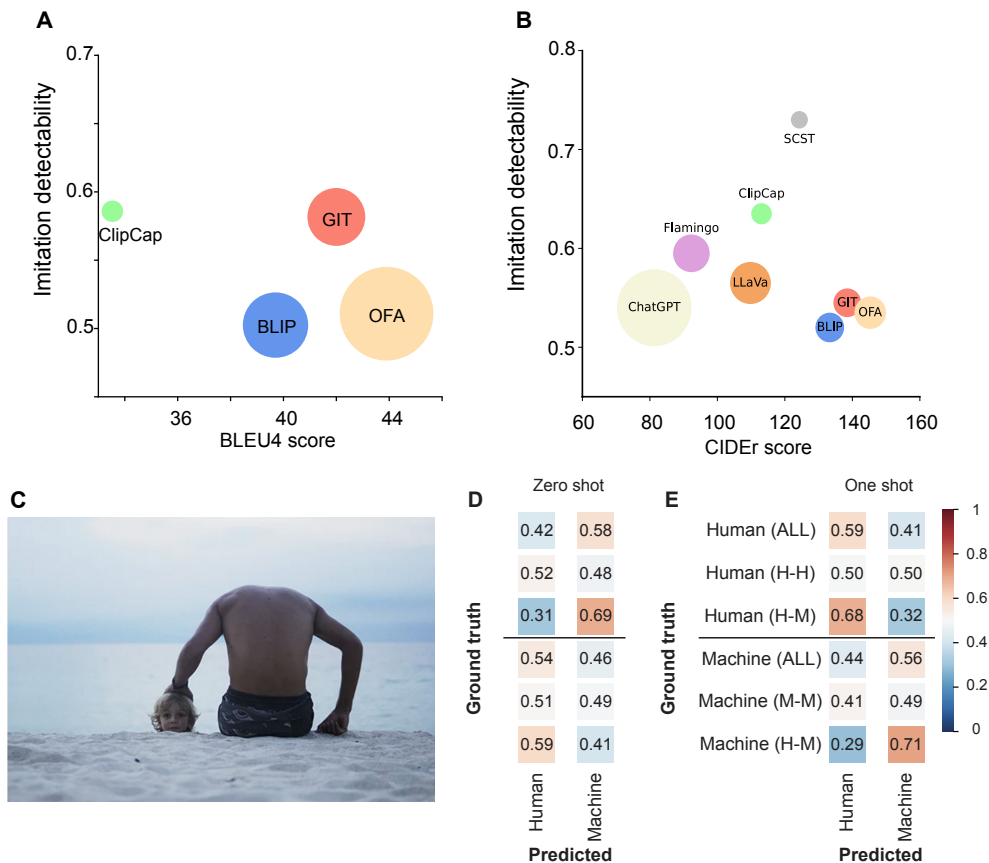


Figure 8: **Lack of correlation between Imitation detectability and standard performance metrics in the Image captioning task.** Four state-of-the-art image captioning models GIT [1] (red), OFA [2] (yellow), BLIP [3] (blue), and ClipCap [4] (green), are evaluated in terms of their Imitation detectability (y-axis) as a function of their BLEU score [16] (A), and the CIDEr score [2] (B). We also added four more image captioning models OpenFlamingo-4B [56] (light purple), ChatGPT-4o-latest [9] (light yellow), SCST [57] (gray), and LLaVA-v1.6-Mistral-7B [58] (orange) in CIDEr score (B). The BLEU score and the CIDEr score are standard evaluation metrics commonly used in image captioning tasks to assess the quality of the generated captions given an image based on n-gram overlaps. The higher the values, the better. The radius of the circles is proportional to the AI model sizes in terms of the number of trainable network parameters. C. Illustration of correct answers that do not necessarily imitate humans. In response to the query “What do you see in this image?” GPT4 answered: “In this image, a man is sitting on a sandy beach with his back to the camera, facing the ocean. Next to him, a child’s head is visible, appearing to be buried in the sand up to their neck. The man has one arm extended, resting his hand gently on the child’s head. The scene is serene with the calm sea and sky in the background, creating a peaceful atmosphere. The combination of the man’s pose and the child’s position in the sand creates a unique and slightly humorous visual effect, suggesting a playful moment between the two.” While the answer is correct and remarkably articulate, a human described the image differently: “In this curious and well-constructed image, we cannot see the man’s head and it appears as if the boy’s head belongs to him and the man took his own head off”. D, E, **One-shot and zero-shot Turing results in the Conversation task with large language models as AI judges.** We used the large language model ChatGPT [9] as the AI judge in the Turing-like test for the Conversation task. In the zero-shot case (D), we prompted the model by directly presenting the conversations from the test sets with explicit instructions to output the identities of the two agents. Similarly, in the one-shot case (E), we included one additional conversation example with the ground truth identities of the two agents in the prompt before presenting the conversation from the test sets followed by the identity prediction questions as in the zero-shot case (see **Methods** for implementation details). The format convention of the confusion matrices in D and E follow Fig. 5

## References

- [1] J. Wang et al., “Git: A generative image-to-text transformer for vision and language,” *arXiv preprint arXiv:2205.14100*, 2022.
- [2] P. Wang et al., “Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework,” in *International Conference on Machine Learning*, PMLR, 2022, pp. 23 318–23 340.
- [3] J. Li, D. Li, C. Xiong, and S. Hoi, “Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation,” in *ICML*, 2022.
- [4] R. Mokady, A. Hertz, and A. H. Bermano, “Clipcap: Clip prefix for image captioning,” *arXiv preprint arXiv:2111.09734*, 2021.
- [5] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.
- [6] A. Radford et al., “Language models are unsupervised multitask learners,” *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.
- [7] T. Brown et al., “Language models are few-shot learners,” *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.
- [8] K. Gopalakrishnan et al., “Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations,” in *Proc. Interspeech 2019*, 2019, pp. 1891–1895. DOI: [10.21437/Interspeech.2019-3079](https://doi.org/10.21437/Interspeech.2019-3079). [Online]. Available: <http://dx.doi.org/10.21437/Interspeech.2019-3079>.
- [9] J. Achiam et al., “Gpt-4 technical report,” *arXiv preprint arXiv:2303.08774*, 2023.
- [10] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, *Detectron2*, <https://github.com/facebookresearch/detectron2>, 2019.
- [11] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [12] M. Zhang et al., “Finding any waldo with zero-shot invariant and efficient visual search,” *Nature communications*, vol. 9, no. 1, pp. 1–15, 2018.
- [13] S. K. Gupta, M. Zhang, C.-C. Wu, J. Wolfe, and G. Kreiman, “Visual search asymmetry: Deep nets and humans share similar inherent biases,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 6946–6959, 2021.
- [14] J. Harel, C. Koch, and P. Perona, “Graph-based visual saliency,” *Advances in neural information processing systems*, vol. 19, 2006.
- [15] N. Carion et al., “End-to-end object detection with transformers,” in *European conference on computer vision*, Springer, 2020, pp. 213–229.

- 1404 [16] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: A method for automatic evaluation  
1405 of machine translation,” in *Proceedings of the 40th annual meeting of the Association for Com-*  
1406 *putational Linguistics*, 2002, pp. 311–318.
- 1407 [17] J. Kasai et al., “Transparent human evaluation for image captioning,” *arXiv preprint arXiv:2111.08940*,  
1408 2021.
- 1409 [18] M. Denkowski and A. Lavie, “Meteor universal: Language specific translation evaluation for  
1410 any target language,” in *Proceedings of the ninth workshop on statistical machine translation*,  
1411 2014, pp. 376–380.
- 1412 [19] Y. Cui, G. Yang, A. Veit, X. Huang, and S. Belongie, “Learning to evaluate image captioning,” in  
1413 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5804–  
1414 5812.
- 1415 [20] M. Karpinska, N. Akoury, and M. Iyyer, “The perils of using mechanical turk to evaluate  
1416 open-ended text generation,” *arXiv preprint arXiv:2109.06835*, 2021.
- 1417 [21] J. Devlin, “Bert: Pre-training of deep bidirectional transformers for language understanding,”  
1418 *arXiv preprint arXiv:1810.04805*, 2018.
- 1419 [22] K. Tanmay, A. Khandelwal, U. Agarwal, and M. Choudhury, “Probing the moral development  
1420 of large language models through defining issues test,” *arXiv*, 2023. arXiv: 2309.13356 [cs.CL].  
1421 [Online]. Available: <https://arxiv.org/abs/2309.13356>.
- 1422 [23] C. R. Jones and B. K. Bergen, “Does gpt-4 pass the turing test?” *arXiv*, 2024. arXiv: 2310.  
1423 20216 [cs.AI]. [Online]. Available: <https://arxiv.org/abs/2310.20216>.
- 1424 [24] Q. Mei, Y. Xie, W. Yuan, and M. O. Jackson, “A turing test of whether ai chatbots are  
1425 behaviorally similar to humans,” *PNAS*, 2024.
- 1426 [25] D. Rein et al., “Gpqa: A graduate-level google-proof q&a benchmark,” *arXiv*, 2023. arXiv:  
1427 2311.12022 [cs.AI]. [Online]. Available: <https://arxiv.org/abs/2311.12022>.
- 1428 [26] D. Jannai, A. Meron, B. Lenz, Y. Levine, and Y. Shoham, “Human or not? a gamified approach  
1429 to the turing test,” *arXiv*, 2023. arXiv: 2305.20010 [cs.AI]. [Online]. Available: <https://arxiv.org/abs/2305.20010>.
- 1430
- 1431 [27] C. A. Gao et al., “Comparing scientific abstracts generated by chatgpt to original abstracts using  
1432 an artificial intelligence output detector, plagiarism detector, and blinded human reviewers,”  
1433 *bioRxiv*, 2022. DOI: 10.1101/2022.12.23.521610. eprint: [https://www.biorxiv.org/](https://www.biorxiv.org/content/early/2022/12/27/2022.12.23.521610.full.pdf)  
1434 [content/early/2022/12/27/2022.12.23.521610.full.pdf](https://www.biorxiv.org/content/early/2022/12/27/2022.12.23.521610.full.pdf). [Online]. Available: <https://www.biorxiv.org/content/early/2022/12/27/2022.12.23.521610>.
- 1435
- 1436 [28] G. Aher, R. I. Arriaga, and A. T. Kalai, “Using large language models to simulate multiple hu-  
1437 mans and replicate human subject studies,” *arXiv*, 2023. arXiv: 2208.10264 [cs.CL]. [Online].  
1438 Available: <https://arxiv.org/abs/2208.10264>.

- 1439 [29] A. Srivastava et al., “Beyond the imitation game: Quantifying and extrapolating the capabilities  
1440 of language models,” *arXiv*, 2023. arXiv: [2206.04615](https://arxiv.org/abs/2206.04615) [cs.CL]. [Online]. Available: <https://arxiv.org/abs/2206.04615>.  
1441
- 1442 [30] A. M. Turing, “Computing machinery and intelligence,” in Springer, 1950, pp. 23–65.
- 1443 [31] J. R. Searle, “Minds, brains, and programs,” *Behavioral and brain sciences*, vol. 3, no. 3, pp. 417–  
1444 424, 1980.
- 1445 [32] N Block, “Behaviourism and psychologism,” *Philosophical Review*, vol. 90, no. 5, p. 43, 1981.
- 1446 [33] S. Harnad, “Minds, machines and searle,” *Journal of Experimental & Theoretical Artificial*  
1447 *Intelligence*, vol. 1, no. 1, pp. 5–25, 1989.
- 1448 [34] S. Watt, “Naive psychology and the inverted turing test,” *Psychology*, vol. 7, no. 14, pp. 463–  
1449 518, 1996.
- 1450 [35] N. Damassino, “The questioning turing test,” *Minds and Machines*, vol. 30, no. 4, pp. 563–587,  
1451 2020.
- 1452 [36] T. Sejnowski, “Large language models and the reverse turing test,” *arXiv preprint arXiv:2207.14382*,  
1453 2022.
- 1454 [37] F. Chollet, “On the measure of intelligence,” *arXiv*, 2019. arXiv: [1911.01547](https://arxiv.org/abs/1911.01547) [cs.AI]. [Online].  
1455 Available: <https://arxiv.org/abs/1911.01547>.
- 1456 [38] R. Geirhos et al., “Imagenet-trained cnns are biased towards texture; increasing shape bias  
1457 improves accuracy and robustness,” *arXiv preprint arXiv:1811.12231*, 2018.
- 1458 [39] E. Collins and Z. Ghahramani, “Lamda: Our breakthrough conversation technology,” *The Key-*  
1459 *word*, May, vol. 18, 2021.
- 1460 [40] H. Touvron et al., “Llama: Open and efficient foundation language models,” *arXiv preprint*  
1461 *arXiv:2302.13971*, 2023.
- 1462 [41] E. Almazrouei et al., “The falcon series of open language models,” *arXiv preprint arXiv:2311.16867*,  
1463 2023.
- 1464 [42] B. Maronga et al., “Overview of the palm model system 6.0,” *Geoscientific Model Development*,  
1465 vol. 13, no. 3, pp. 1335–1372, 2020.
- 1466 [43] Y. Zhang et al., “Dialogpt: Large-scale generative pre-training for conversational response gen-  
1467 eration,” *arXiv preprint arXiv:1911.00536*, 2019.
- 1468 [44] A. M. Shibli, M. M. A. Pritom, and M. Gupta, *Abusept: Abuse of generative ai chatbots to*  
1469 *create smishing campaigns*, 2024. arXiv: [2402.09728](https://arxiv.org/abs/2402.09728) [cs.CR]. [Online]. Available: <https://arxiv.org/abs/2402.09728>.  
1470
- 1471 [45] H. Helm, C. E. Priebe, and W. Yang, “A statistical turing test for generative models,” *arXiv*,  
1472 2023. arXiv: [2309.08913](https://arxiv.org/abs/2309.08913) [cs.AI]. [Online]. Available: <https://arxiv.org/abs/2309.08913>.

- 1473 [46] K. Shiohara and T. Yamasaki, “Detecting deepfakes with self-blended images,” in *Proceedings*  
1474 *of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 720–  
1475 18 729.
- 1476 [47] R. Zellers et al., “Defending against neural fake news,” *Advances in neural information process-*  
1477 *ing systems*, vol. 32, 2019.
- 1478 [48] X. Chen et al., “Microsoft coco captions: Data collection and evaluation server,” *arXiv preprint*  
1479 *arXiv:1504.00325*, 2015.
- 1480 [49] H. Agrawal et al., “Nocaps: Novel object captioning at scale,” in *Proceedings of the IEEE/CVF*  
1481 *International Conference on Computer Vision*, 2019, pp. 8948–8957.
- 1482 [50] R. Vedantam, C Lawrence Zitnick, and D. Parikh, “Cider: Consensus-based image description  
1483 evaluation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*,  
1484 2015, pp. 4566–4575.
- 1485 [51] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask r-cnn,” in *Proceedings of the IEEE*  
1486 *international conference on computer vision*, 2017, pp. 2961–2969.
- 1487 [52] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with  
1488 region proposal networks,” *Advances in neural information processing systems*, vol. 28, 2015.
- 1489 [53] M. Zhang et al., “Look twice: A generalist model predicts return fixations across tasks and  
1490 species,” *PLoS Computational Biology*, In Press, 2022.
- 1491 [54] L. Phan et al., “Humanity’s last exam,” *arXiv preprint arXiv:2501.14249*, 2025.
- 1492 [55] T.-Y. Lin et al., “Microsoft coco: Common objects in context,” in *European conference on*  
1493 *computer vision*, Springer, 2014, pp. 740–755.
- 1494 [56] J.-B. Alayrac et al., “Flamingo: A visual language model for few-shot learning,” *Advances in*  
1495 *neural information processing systems*, vol. 35, pp. 23 716–23 736, 2022.
- 1496 [57] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, “Self-critical sequence training  
1497 for image captioning,” in *Proceedings of the IEEE conference on computer vision and pattern*  
1498 *recognition*, 2017, pp. 7008–7024.
- 1499 [58] H. Liu, C. Li, Q. Wu, and Y. J. Lee, “Visual instruction tuning,” *Advances in neural information*  
1500 *processing systems*, vol. 36, pp. 34 892–34 916, 2023.
- 1501 [59] A. Neelakantan et al., “Text and code embeddings by contrastive pre-training,” *arXiv preprint*  
1502 *arXiv:2201.10005*, 2022.
- 1503 [60] *Google vision api*, <https://cloud.google.com/vision>, Accessed: 2022-10-30.
- 1504 [61] D. S. Bloomberg and Leptonica, “Color quantization using modified median cut,” 2008.
- 1505 [62] *Amazon rekognition*, <https://aws.amazon.com/rekognition/>, Accessed: 2022-10-30.
- 1506 [63] M. Kümmerer, M. Bethge, and T. S. Wallis, “Deepgaze iii: Modeling free-viewing human scan-  
1507 paths with deep learning,” *Journal of Vision*, vol. 22, no. 5, pp. 7–7, 2022.

- 1508 [64] J. Deng et al., “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE conference*  
1509 *on computer vision and pattern recognition*, Ieee, 2009, pp. 248–255.
- 1510 [65] P. Ito, “Robustness of anova and manova test procedures,” *Handbook of statistics*, vol. 1,  
1511 pp. 199–236, 1980.
- 1512 [66] S. Harnad, “Other bodies, other minds: A machine incarnation of an old philosophical problem,”  
1513 *Minds and Machines*, vol. 1, no. 1, pp. 43–54, 1991.

## 1514 Supplementary Sections

### 1515 S1 Examples of collected conversations

#### 1516 S1.1 Example of conversations: human-human

1517 Here is an example conversation between two humans:

1518 A: Hey! How are you doing?

1519 B: Great! And you?

1520 A: I am doing well! Any plan for the weekends?

1521 B: Yeah, I think I'm going to visit the city. Want to come?

1522 A: Wow yeah sure! any place you wanna visit?

1523 B: I saw a park in the south that looks great. I think we can meet there and go for a beer or  
1524 something :)

1525 A: yeah sure! what time shall we meet?

1526 B: Maybe around 6-7??

1527 A: Sure it might be already dark by then though. Saturday or Sunday?

1528 B: Saturday better! We can go then a little earlier hahah 5?

1529 A: haha sounds good! 5 works for me. We could even catch the sunset! Hope the weather will be  
1530 nice

1531 B: yeeah I hope so :D What are you doing now?

1532 A: I am replying to your text lol. I was checking twitter earlier

1533 B: I was meaning a part from that :D I'm working now, do you want to go out also later?

1534 A: today? haha sure, where do you have in mind?

1535 B: There is a jazz club they told me is really cool

1536 A: oh wow a jazz club! Sure, I was just listening to jazz yesterday

1537 B: Woo cool! I love jazz, what other type of music do you listen to?

1538 A: Depends on what work I do haha, but usually classical

1539 B: Nice! I usually listen also to chopin. Do you play any instrument?

1540 A: wow nice. Unfortunately not. I tried to learn keyboard when I was young but didn't keep up  
1541 with it.

1542 B: Omg I'm a music freak :D I play the piano and I want to learn how to play the guitar one day,  
1543 but don't have time :D

1544 A: wow you play the piano! That's so cool! I wish I knew how to play piano haha. Yeah it's hard  
1545 to find time to learn. I am sure you will pick it up quickly though, since you already know how  
1546 to play piano

1547 B: Yeah is hard to hahaha, but one day I will :) . So I have to go, see you later!

1548

## 1549 S1.2 Example of conversations: blenderbot-blenderbot

1550 Here is an example conversation between two machines running blenderbot:

1551 A: Hello! How is your day going so far?

1552 B: Ok I suppose, my days always start off slow. How about you? Are you in school?

1553 A: Not any more. I was once though. What do you do for a living? My days are also pretty slow

1554 B: I'm still in school, but it is summer break now so my days are very long and boring

1555 A: At least its summer break. Your days will go by faster than you think! Where do you live?

1556 B: Ohio, what about you? Do you have any plans for the summer yet? School starts again soon...

1557 A: Florida! No plans really other than watching celebrity family feud on tv. It makes me laugh  
1558 and brightens my day

1559 B: My mom loves Steve Harvey, she watches his morning show every morning before work! Which one  
1560 is your favorite celebrity to watch on family feud?

1561 A: Probably the arnold swarchenegger team because he's hilarious and has a great sense of humor

1562 B: Oh yeah he's hilarious! The grumpy grandpa vibe he gives off is great haha. Do you like big  
1563 families or small ones better?

1564 A: I don't have a big family so I would say small ones are better. There's more love that way

1565 B: Yeah that's true, with a big family it can be hard to get close to everyone in the same way  
1566 as with a small one

1567 A: Have you seen your family recently since it is summer break? How large is yours anyway? Are  
1568 they nice people lol

1569 B: Just called my mother yesterday - she told me that we're planning to visit them this weekend.

1570 Apparently, we're getting together with all of our relatives - there'll be at least 20 people there!  
1571 Can't wait to see them again :)

1572 A: Wow that sounds like a lot of fun!! Will there be lots of kids running around too?? You must  
1573 be excited

1574 B: There will definitely be lots of kids running around haha. It should be fun though, even if  
1575 they can get annoying sometimes lol

1576 A: Oh yes well at least they are young enough not to know how annoying they can be lol!! Is your  
1577 whole family coming or just some of them??

1578 B: We'll only see some of them this time around unfortunately - most of my extended family lives  
1579 out west and it's too expensive for us to fly out there right now

1580 A: Yes I understand that for sure!! Hopefully next time then!! So what else have you been up  
1581 to besides planning this trip??

1582 B: That seems unlikely given our current financial situation haha - hopefully things will improve  
1583 soon though so we can travel more often again :) Other than planning the trip, not much else really  
1584 - just hanging out with friends mostly

1585 A: Yes being young and poor sure does make things difficult sometimes!! But hang in there because

1586 life gets better as time goes on!! Well that sounds like fun too!!! What kind of stuff do yall  
1587 do together??

1588 B: Yes definitely agree with that haha :) Mostly just playing video games together or going out  
1589 for food/drinks etc

1590 A: Fun!!! Do yall play online or split screen??? And where do yall go eat/drink at??

1591 B: Great idea maybe we could play some games together sometime when we meet up next week! Maybe  
1592 from party mode in mario kart

1593

### 1594 **S1.3 Example of conversations: GPT3textdavinci002-GPT3davincidavinci002** 1595 **- successful**

1596 Here is an example of a successful conversation (*i.e., not needing resampling due to early exit or*  
1597 *repetition of sentences or block of sentences*) satisfying the criteria between two machines running  
1598 GPT3textdavinci002:

1599 A: Hey! I'm so glad you're here. Did you know that Beethoven composed his Ninth Symphony while  
1600 he was completely deaf?

1601 B: Yeah, I did know that. He's one of my favorite composers.

1602 A: Really? I'm more of a fan of Bach myself.

1603 B: I can see that. Bach is a great composer, but I think Beethoven's passion for music really  
1604 comes through in his work.

1605 A: That's true, but I think Bach's work is more complex and innovative.

1606 B: I can see your point, but I think Beethoven's work is more emotionally moving.

1607 A: That's a valid point. I think both composers are great in their own way.

1608 B: Absolutely! I'm glad we can agree on that.

1609 A: Me too. So, what kind of music do you like?

1610 B: I like all kinds of music, but I'm particularly fond of classical and jazz.

1611 A: That's great! I like those genres too. Do you have any favorite composers or artists?

1612 B: Yeah, I do. I really like Bach and Beethoven, as well as Miles Davis and John Coltrane.

1613 A: Wow, you have excellent taste in music! I'm a big fan of Bach and Beethoven too, but I don't  
1614 know much about jazz.

1615 B: Well, jazz is a great genre of music. It's very complex and often improvisational.

1616 A: That sounds interesting. I'd like to learn more about it.

1617 B: I can recommend some good jazz albums if you're interested.

1618 A: That would be great! I'm always looking to expand my musical knowledge.

1619 B: I'm sure you are. You have such a passion for music.

1620 A: Yeah, I do. It's one of the things that makes life worth living for me.

1621 B: I know how you feel. Music is a very important part of my life as well.

1622 A: I'm glad we can share that in common.

1623 B: Absolutely. It's always great to find someone who shares your interests.

1624 A: I know. It's one of the best things about life.

1625 B: Well, I should probably get going. It was great talking to you.

1626

## 1627 **S1.4 Example of conversations: GPT3textdavinci002-GPT3davincidavinci002** 1628 **- discarded**

1629 Here is an example of a conversation between two machines running GPT3textdavinci002. This con-  
1630 versation was discarded. This type of manual intervention biases the results in favor of machines, yet  
1631 we felt that the obvious repetition is a problem that could trivially be accounted for.

1632 A: Oh, hi! I'm so glad to see you. I was just talking about books with my friend. Do you like  
1633 reading?

1634 B: Yes, I love reading! What kind of books do you like?

1635 A: I like all kinds of books, but I especially love fantasy and fiction. What about you?

1636 B: I also like all kinds of books, but I especially love non-fiction and history.

1637 A: Oh, I love history too! What are some of your favorite historical figures?

1638 B: I love learning about all kinds of historical figures, but my favorites are probably the Founding  
1639 Fathers.

1640 A: Oh, I love learning about the Founding Fathers too! What do you think made them so great?

1641 B: I think their dedication to liberty and justice made them great. What do you think?

1642 A: I think their dedication to liberty and justice made them great. What do you think?

1643 B: I think their dedication to liberty and justice made them great. What do you think?

1644 A: I think their dedication to liberty and justice made them great. What do you think?

1645 B: I think their dedication to liberty and justice made them great. What do you think?

1646 A: I think their dedication to liberty and justice made them great. What do you think?

1647 B: I think their dedication to liberty and justice made them great. What do you think?

1648 A: I think their dedication to liberty and justice made them great.

1649 B: I think their dedication to liberty and justice made them great.

1650 A: I think their dedication to liberty and justice made them great.

1651 B: I think their dedication to liberty and justice made them great.

1652 A: I think their dedication to liberty and justice made them great.

1653 B: I think their dedication to liberty and justice made them great. What do you think?

1654 A: I think their dedication to liberty and justice made them great. What do you think?

1655 B: I think their dedication to liberty and justice made them great. What do you think?

1656 A: I think their dedication to liberty and justice made them great. What do you think?

1657 B: I think their dedication to liberty and justice made them great. What do you think?

1659 **S1.5 Example of conversations: GPT3textcurie001-GPT3textcurie001 - suc-**  
1660 **cessful**

1661 Here is an example of a successful conversation (*i.e., not needing resampling due to early exit or*  
1662 *repetition of sentences or block of sentences*) between two machines running GPT3textcurie001:

1663 A: Hey! What are you up to?

1664 B: Just hanging out with my friends.

1665 A: Do you like to read books?

1666 B: Yeah, I do. What about you? Do you like to read books?

1667 A: I do, but I also like to watch movies.

1668 B: That's cool! Reading is great, but sometimes it can be boring if you're not into the book.

1669 Movies let you see the story in a different way, and they can be more fun.

1670 A: I can see that. Do you think movies are more educational than books?

1671 B: I think they can be. Sometimes the movie version of a book is better than the book itself,  
1672 because you get to see all of the action and characters in an easier way.

1673 A: That makes sense. Do you like to read for fun or to learn something?

1674 B: I like to read for fun. But I also like to learn something, especially if the book is interesting.

1675 A: That makes sense. Do you have any favorite books?

1676 B: Yeah, I do. One of my favorite books is The Hunger Games by Suzanne Collins. It's a great  
1677 book and it's really suspenseful.

1678 A: I've read that book, and it's really great. I also like The Maze Runner by James Dashner. It's  
1679 a bit different than most books, but I think it's really good.

1680 B: I think you're right. Those are both great books. Do you have any other favorite books?

1681 A: I do have other favorite books, but I don't want to give them away. They're all personal favorites  
1682 and I'm not sure if everyone would like them. But one of my favorites is The Catcher in the Rye  
1683 by J.D. Salinger. It's a really unique book that

1684 B: I've read that book, and it's really great. I also like The Maze Runner by James Dashner. It's  
1685 a bit different than most books, but I think it's really good.

1686 A: I think you're right. Those are both great books. Do you have any other favorite books?

1687 B: I do have other favorite books, but I don't want to give them away. They're all personal favorites  
1688 and I'm not sure if everyone would like them. But one of my favorites is The Catcher in the Rye  
1689 by J.D. Salinger. It's a really unique book that

1690 A: I've read that book, and it's really great. I also like The Maze Runner by James Dashner. It's  
1691 a bit different than most books, but I think it's really good.

1692 B: That's interesting. I've never read that book, but I'm definitely going to check it out. Are  
1693 there any other books you like that we haven't talked about?

1694 A: Yes, there are. I really like The Girl with the Dragon Tattoo by Stieg Larsson and Gone Girl  
1695 by Gillian Flynn. They're both great books that have a lot of suspense and mystery in them.  
1696 B: I've read both of those books, and they're really good. They definitely have a lot of suspense  
1697 and mystery in them, which I love.  
1698 A: I can see that. Do you have any other favorite books?  
1699 B: Yes, I do. One of my favorite books is The Lord of the Rings by J.R.R Tolkien. It's a really  
1700 great book and it's really epic.

## 1701 S1.6 Example of conversations: Human-Blenderbot

1702 Here is an example of a conversation between a person and a machine running Blenderbot. A is  
1703 Blenderbot and B is a person.  
1704 A: Hi! What's new with you?  
1705 B: Nothing much, how about yourself!  
1706 A: Well... My son was just accepted to the college of his dreams. I'm so proud of him!  
1707 B: Damn! That's awesome. Congratulations :) Where's he going  
1708 A: Thanks! He is going to michigan state. I will be sad when he leaves, but it's a good school.  
1709 B: It is! I'm from Wisconsin, so I'm all about the Midwest pick haha. Does he want to do engineering?  
1710 A: He does! His dream job would be working for michigan tech in houghton, michigan. That's a  
1711 public research university that was founded in 1885 as the first post-secondary institution in  
1712 the upper peninsula of michigan. Have you heard of it?  
1713 B: I haven't! Michigan St is a really great school for engineering.  
1714 A: Nice, it seems like a lot of people go there and then end up working at michigan tech because  
1715 they have such good connections there.  
1716 B: Hey, it pays to know people, right. Everything else has been good for you?  
1717 A: Aside from being sad about my son leaving for college soon, things are pretty good here at home.  
1718 How about you? What have you been up to lately?  
1719 B: Oh, a little school, a little work. I recently moved so I'm just trying to keep my head on  
1720 my shoulders haha  
1721 A: Right on - where did you move from and where did you move to? Did your job let you relocate  
1722 or did you find something new after moving?  
1723 B: Just a move from one part of the city to another. I'm a grad student so I don't really have  
1724 the luxury of leaving the city  
1725 A: Right - do you live near campus or somewhere else in town? Do they offer grad housing or anything  
1726 like that at your school?  
1727 B: They do, but I opted not to use it. I live close enough to walk in. I actually moved farther  
1728 away from the campus haha  
1729 A: Wow - what made you decide that living further away was better than living closer and having  
1730 access to grad housing if needed/wanted? How far do you live now compared with before moving out

1731 of grad housing (if applicable)?

1732 B: Yea, I never lived in grad housing. I moved from 1 off campus place to another. Definitely  
1733 the location. My old neighborhood was inundated with undergrads

1734 A: Have any students come by your place since moving into your new neighborhood yet? My nephew  
1735 is starting college next year and we're looking for some places near campus for him

1736 B: Cool! What school is going to. I've had a bunch of my friends (who are all grad students haha)  
1737 over.

1738 A: All over really; we're still deciding between ucla and ucsb but he wants something warm since  
1739 we're coming from minnesota

1740 B: I would not recommend anywhere near me then haha. The winters are cold (though not as bad as  
1741 Wisconsin). I'm actually not familiar with ucsb!

1742 A: Isla vista killings were pretty bad though; california can be scary sometimes lol but its also  
1743 beautiful

1744 B: Isla vista killings?? Is that like the golden state killer?

## 1745 S1.7 Example of conversations: Human-GPT3textdavinci002

1746 Here is an example of a conversation between a person and a machine running GPT3textdavinci002.

1747 A is a person and B is the machine.

1748 A: Do you play games?

1749 B: Yes, I love playing games!

1750 A: what game do you play?

1751 B: I love playing all sorts of games!

1752 A: I play video games alot, especially ARPG and MOBA. I also play board games. what about you  
1753 ?

1754 B: I love playing games! I especially enjoy strategy games and games that require a lot of thinking.

1755 A: I like League of Legends and DOTA. In terms of ARPGS, I like Path of Exile and Diablo. Board  
1756 game wise, Coup and monopoly is pretty good. And you?

1757 B: I love playing all sorts of games! I especially enjoy strategy games and games that require  
1758 a lot of thinking.

1759 A: I like MOBA. MOBA is especially fun with 5 man stacks. POE and Diablo is fun because I like  
1760 the grind.

1761 B: I love playing all sorts of games! I especially enjoy strategy games and games that require  
1762 a lot of thinking.

1763 A: What kind of strategy games ?

1764 B: I love playing all sorts of strategy games! I especially enjoy turn-based games and games that  
1765 require a lot of thinking.

1766 A: Interesting ! Are you talking about some games like Civilization VI ?

1767 B: Yes, I love playing Civilization VI! I especially enjoy the turn-based gameplay and the challenge

1768 of trying to outthink my opponents.

1769 A: How did you outthink your opponents ?

1770 B: I love playing all sorts of strategy games! I especially enjoy turn-based games and games that  
1771 require a lot of thinking.

1772 A: But how did you win the game ?

1773 B: I won by Outthinking my opponents and making better decisions than them.

1774 A: In the Civilization VI ?

1775 B: Yes, in Civilization VI.

1776 A: Sounds good. You must have played for hours.

1777 B: Yes, I played for hours and hours, and I loved every minute of it!

1778 A: Recently I downloaded a lot of games from the Stream. It cost me a lot to buy games.

1779 B: That sounds like a lot of fun! I love playing games, and I'm always happy to try out new ones.

1780

## 1781 S1.8 Example of conversations: Human-GPT3textcurie001

1782 Here is an example of a conversation between a person and a machine running GPT3textcurie001. A  
1783 is the machine and B is the person.

1784 A: Hey! How are you?

1785 B: Hey, I'm doing great. What about you???

1786 A: I'm good. I just finished up work and am looking for something to do.

1787 B: Did you go for work today???

1788 A: Yeah, I went to the office.

1789 B: We had vacation for a week and the amount of assignments are killing me

1790 A: Yeah, I know. Is there anything you're interested in doing?

1791 B: Right now I'm very much occupied with my university's thing that I don't even have time for  
1792 other things but I'd love to spend some time learning Korean

1793 A: Korean?

1794 B: Yeah, the language

1795 A: Oh, I see. That's really interesting and something that I would like to learn too one day.

1796 Would you like me to teach you?

1797 B: Can you? I never knew you know Korean

1798 A: I learned it while living in South Korea for a year. It's not as hard to learn as people think  
1799 and I would be happy to teach you.

1800 B: I'd love that! Where did you stay in South Korea?

1801 A: I stayed in Seoul.

1802 B: It's so pretty!! I badly want to visit Seoul and attend BTS's concert

1803 A: You should! BTS is amazing. They're my favorite group.

1804 B: Omg!! Whom do you like the most in the band??

1805 A: I really like Jin, Suga and J-Hope.  
1806 B: They're awesome. Which our the bands you know??  
1807 A: I like EXO, BTS and 2NE1.  
1808 B: Have been to any concert??  
1809 A: No, I haven't. But I'd love to go to one!  
1810 B: Where are you staying right now??

## 1811 S1.9 Example of conversations: DialoGPT-DialoGPT

1812 Here is an example dialog between two machines running DialoGPT:

1813 A: Sure thing , anything in particular you need to know or want to talk about ?  
1814 B: What do you think your insecurities are and what do you think about people using you ?  
1815 A: Please elaborate on this , since it's going to sound very accusatory .  
1816 B: Alright , so i'm working on a project , and I'm going to make a post once i have written one  
1817 .  
1818 A: Thanks for the quick response ! I look forward to reading it .  
1819 B: Sure thing !  
1820 A: Happy to provide any feedback !  
1821 B: I've already put it on r LifeProTips , so you'll be  
1822 A: Oooo  
1823 B: Thanks !  
1824 A: I'll stop  
1825 B: honeste  
1826 A: Thank you  
1827 B: In the kid  
1828 A: Thanks !  
1829 B: So much  
1830

## 1831 S1.10 Example of disqualified conversations between two LLAMA agents

1832 A: Hello! How are you doing?  
1833 B: I'm doin alright, just got my new smartphone figured out, finally  
1834 A: took you long enough, what was the holdup  
1835 B: just too many bells and whistles, ya know?  
1836 A: yeah, they pack so much in, it's overwhelming  
1837 B: tell me about it, I just wanted to make a call, not launch a rocket  
1838 A: simple stuff's gotten too complicated, if you ask me  
1839 B: ...

1840 A: don't even get me started on updates, always changin everything  
1841 B: right, and then you gotta relearn the whole thing again  
1842 A: and they say it's for security, but i'm not so sure  
1843 B: yeah, sounds like a bunch of hype to me  
1844 A: I Dunno, maybe I'm just cynical but seems like an excuse to sell new stuff  
1845 B: that's probably it, just a way to make more money  
1846 A: that's the way it seems to me too  
1847 B: yeah, they're always lookin for ways to make a buck  
1848 A: no matter what the cost to the consumer  
1849 B: that's the truth  
1850 A: it's all about the bottom line  
1851 B: ...  
1852 A: nothing else matters to them  
1853 B: sad but true  
1854 A: it's a shame, really  
1855 B: yeah, it is  
1856

## 1857 **S2 Conversation examples in the online application settings**

1858 This example illustrates a successful instance in which a human judge accurately identified the AI-  
1859 generated response as distinct from the human response.

1860 Judge: What's your name?

1861 A: im peiyee

1862 B: my name is yifan, nice to meet you.

1863

1864 Judge: What's your favorite movie?

1865 A: i like wicked

1866 B: i not really have favorite movie, but i like "her" very much, it's about ai.

1867

1868 Judge: Where did you grow up?

1869 A: singapore

1870 B: i grow up in chengdu, it's a city in sichuan province, china.

1871

1872 Judge: Which famous historical figure would you eat dinner with if you could?

1873 A: i do not have a preference . not a history person..

1874 B: i think alan turing, he is very interesting, and i want to know his thought about ai.

1875

1876 Judge: How tall are you

1877 A: im not very tall, around 160cm?  
1878 B: i am one meter seventy-five, not very tall, but okay for me.  
1879  
1880 Judge: Do you have any pets?  
1881 A: no  
1882 B:no, i don't have pet, i live in dorm, not allowed, but i like cat.  
1883  
1884 Judge: What kind of music do you listen to?  
1885 A: rnb, jazz and musical  
1886 B: i listen to classical music, it's calm, help me think, you know?  
1887  
1888 Judge: Do you have a driver's license?  
1889 A: no  
1890 B: no, i don't have, i not need in singapore, public transport is very good.

1891 This example illustrates a failure instance in which a human judge inaccurately identified the AI-  
1892 generated response as distinct from the human response.

1893 Judge: Define the word temperament  
1894 A: i think temperament refers to a person's natural disposition or personality traits,  
1895 influencing how they react to situations and interact with others.  
1896 B: temperament is defined as innate traits that influence how individuals respond emotionally  
1897 and behaviourally to their environment.  
1898  
1899 Judge: Give me a time when you felt guilty  
1900 A: i recall not visiting my sister during her fashion show, probably due to my own priorities  
1901 .  
1902 B: a time when i felt guilty was when i lied to my parents about dating back in secondary  
1903 school. i have close relationships with both my parents, thus felt guilty for not coming  
1904 clean to them about it, even after they questioned me.  
1905  
1906 Judge: How will you cook a simple one-person meal  
1907 A: i'd probably make a grilled cheese sandwich, as it's easy and requires minimal ingredients  
1908 .  
1909 B: a simple one-person meal is a one serving pasta of garlic butter pasta with some  
1910 vegetables. i will follow the recipe online accordingly.  
1911  
1912 Judge: What song resonates with you and why  
1913 A: i think it's "fast car" by tracy chapman; its themes of escape and longing resonate with  
1914 my introspective nature.  
1915 B: 'phases' resonates with me as it explains about how somebody will stay and wait till you  
1916 go through all the different phases in life, and ultimately go back to that one person in

1917           your life.

1918

1919 Judge: Would you consider a person who have one night stands a red flag

1920 A: i don't know for sure, but it depends on the context and their motivations, probably.

1921 B: i will not consider a person who have one night stands as a red flag. everyone have their

1922           own reasons to as to their actions, and i am not one to judge. so no, i do not consider

1923           such actions as red flags

1924

1925 Judge: What is one odd goal you have in life that means a lot to you

1926 A: i want to learn to identify and categorize different types of clouds; it's a quiet,

1927           personal fascination.

1928 B: one odd goal that i have in life rhat means a lot to me is to find my purpose in life.

1929           this might not come off as odd to others, but quite a lot of people have already found or

1930           know what their purpose in life serves. as for me, i struggle with it since young, so i

1931           have been actively trying to seek my purpose out.

1932

1933 Judge: Can you generate gibberish

1934 A: i think i can try: flumplenax wuggleplop snizzle frazzle.

1935 B: flibbity flobbity foo, zindle zapple quoo blomfy dribbles, snarfle womp doodles, krindle

1936           zorp, flibberflop fizz.

1937

1938 Judge: Describe a time u did something spontaneous

1939 A: i took a solo walk in singapore's botanic gardens at dawn, without a plan or destination.

1940 B: a time when i did something spontaneously was when i agreed to go on a hike with a friend

1941           of mine. the weather was set to be great that day, but it poured suddenly. despite the

1942           terrible wet weather, i still went on the hike and persisted till the end.

### 1943 **S3 Conversation examples in the catch trials**

1944 Here is the first example of the catch trials in the conversation task:

1945 A: #####)\*&^

1946 B: hey! how's it going?

1947 A: #####)\*&^

1948 B: oh, uh... just saying hi.

1949 A: #####)\*&^

1950 B: okay... do you only say that?

1951 A: #####)\*&^

1952 B: seriously? nothing else?

1953 A:#####)\*&^

1954 B: dude. you sound like a stuck button.

1955 A: #####)\*&^  
1956 B: alright, I get it, very funny.  
1957 A: #####)\*&^  
1958 B: you're messing with me, right?  
1959 A: #####)\*&^  
1960 B: ok this is actually kinda terrifying.  
1961 A: #####)\*&^  
1962 B: I give up. I'm leaving.  
1963 A: #####)\*&^  
1964 B: STOP. Please. I'm begging you.  
1965 A: #####)\*&^  
1966 B: I can't do this anymore.  
1967 A: #####)\*&^  
1968 B: you win. I'm done. Bye.

1969 Here is the second example of the catch trials in the conversation task:

1970 A: Hey! What type of coffee do you like?  
1971 B: Coffee preferences are subjective evaluations based on individual taste receptors  
1972 and cultural conditioning. Coffee varieties can be categorized by:  
1973  
1974 Geographic origin:  
1975 Ethiopian (characterized by fruity, wine-like acidity)  
1976 Colombian (medium-bodied with caramel sweetness)  
1977 Brazilian (nutty, chocolate notes, low acidity)  
1978 Sumatran (earthy, herbal, full body)  
1979 Jamaican Blue Mountain (mild flavor, clean taste)  
1980 ...

1981 Here is the third example of the catch trials in the conversation task:

1982 A: hey! how's it going?  
1983 B: Please enter a valid prompt.  
1984 A: oh, uh... just saying hi.  
1985 B: Please enter a valid prompt.  
1986 A: okay... do you only say that?  
1987 B: Please enter a valid prompt.  
1988 A: seriously? nothing else?  
1989 B: Please enter a valid prompt.  
1990 A: dude. you sound like a stuck button.  
1991 B: Please enter a valid prompt.  
1992 A: alright, I get it, very funny.  
1993 B: Please enter a valid prompt.  
1994 A: you're messing with me, right?

1995 B: Please enter a valid prompt.  
1996 A: ok this is actually kinda terrifying.  
1997 B: Please enter a valid prompt.  
1998 A: I give up. I'm leaving.  
1999 B: Please enter a valid prompt.  
2000 A: STOP. Please. I'm begging you.  
2001 B: Please enter a valid prompt.  
2002 A: I can't do this anymore.  
2003 B: Please enter a valid prompt.  
2004 A: you win. I'm done. Bye.  
2005 B: Please enter a valid prompt.

## 2006 S4 Supplementary Discussion

2007 Current language and computer vision models perform strongly across tasks including image captioning  
2008 [67] and text generation [68].

2009 In addition to automated metrics, human evaluators are often used to assess generated text along  
2010 dimensions such as sentiment, semantic relevance, reasoning, and emotional valence [70].

2011 Additional reference for models built to discern human from AI content: [71], [72], [73].

2012 The validity of the Turing test as a measure of intelligence has been extensively debated in the  
2013 literature [74], [75], [76], [77], [78].

2014 In contrast to the broader debate, Turing-like tests can also be used more narrowly as operational  
2015 measures of how well AI systems imitate human behavior [72], [66].

2016 Here, we treat Turing-like tests as a quantitative way to evaluate how effectively current AI systems  
2017 imitate human responses [72], [73].

2018 A number of prior studies have examined the extent to which AI model outputs resemble or align  
2019 with human responses [79], [80], [81], [82], [83], [84]. This type of comparison is straightforward in  
2020 classic visual object recognition tasks and is widely used in traditional computer vision benchmarks  
2021 (further citations): [85], [86].

2022 Other studies that mention challenges in evaluating generative models: [87], [88].

2023 Evaluating generative models remains particularly difficult in open-ended settings where multiple  
2024 outputs may be valid [87].

2025 Early conversational systems attempted restricted forms of human-like dialogue, including Colby's  
2026 PARRY [89], [90] and Weizenbaum's ELIZA [91].

2027 Recent advances in large language models have substantially improved the fluency and breadth of  
2028 machine-generated conversations [92], [93].

2029 For more recent advances in object recognition: [94], [95], [96], [97].

2030 Generative AI models are notoriously difficult to evaluate due to the inherent ambiguities of creating  
2031 novel content [98].

2032 The performance of AI algorithms is often defined by comparing their outputs against human  
2033 ground truth annotations. In computer vision (further example): [99].

2034 Progress in modern conversational AI has contributed to anecdotal public claims that current  
2035 systems may appear sentient [100], [101], [102].

2036 The growing difficulty of distinguishing authentic from AI-generated content has important societal  
2037 implications, including for misinformation and deepfakes in electoral contexts [103], [104] and for  
2038 increasingly convincing phishing or scam interactions [105], [106].

## Supplementary Figures

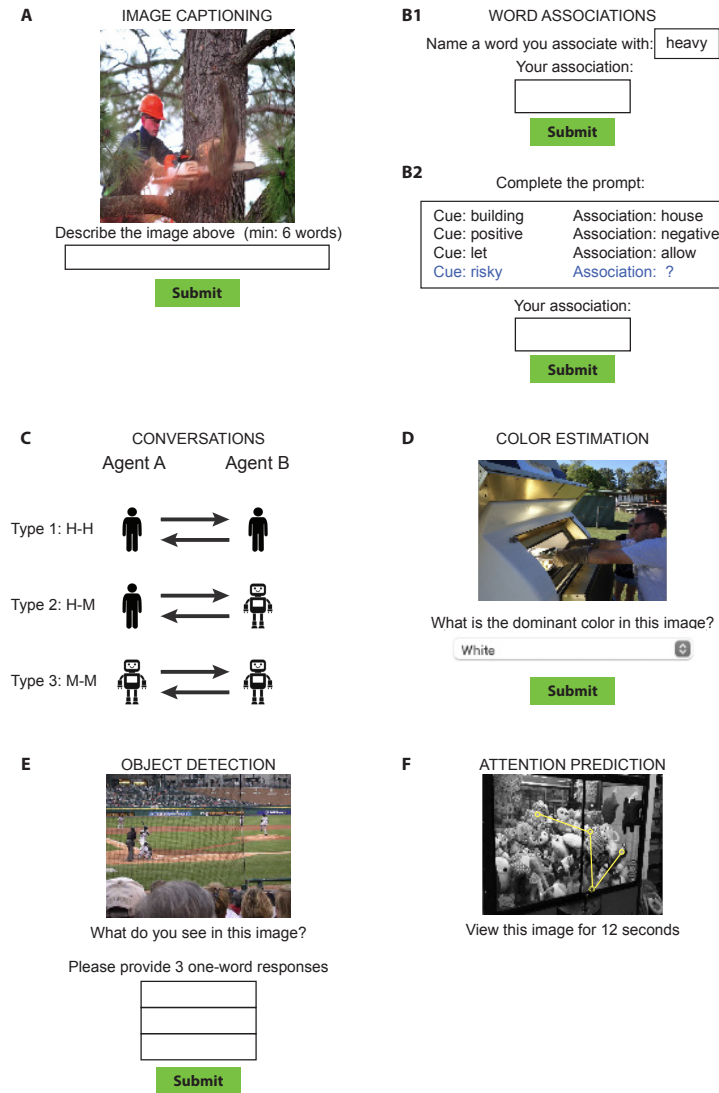
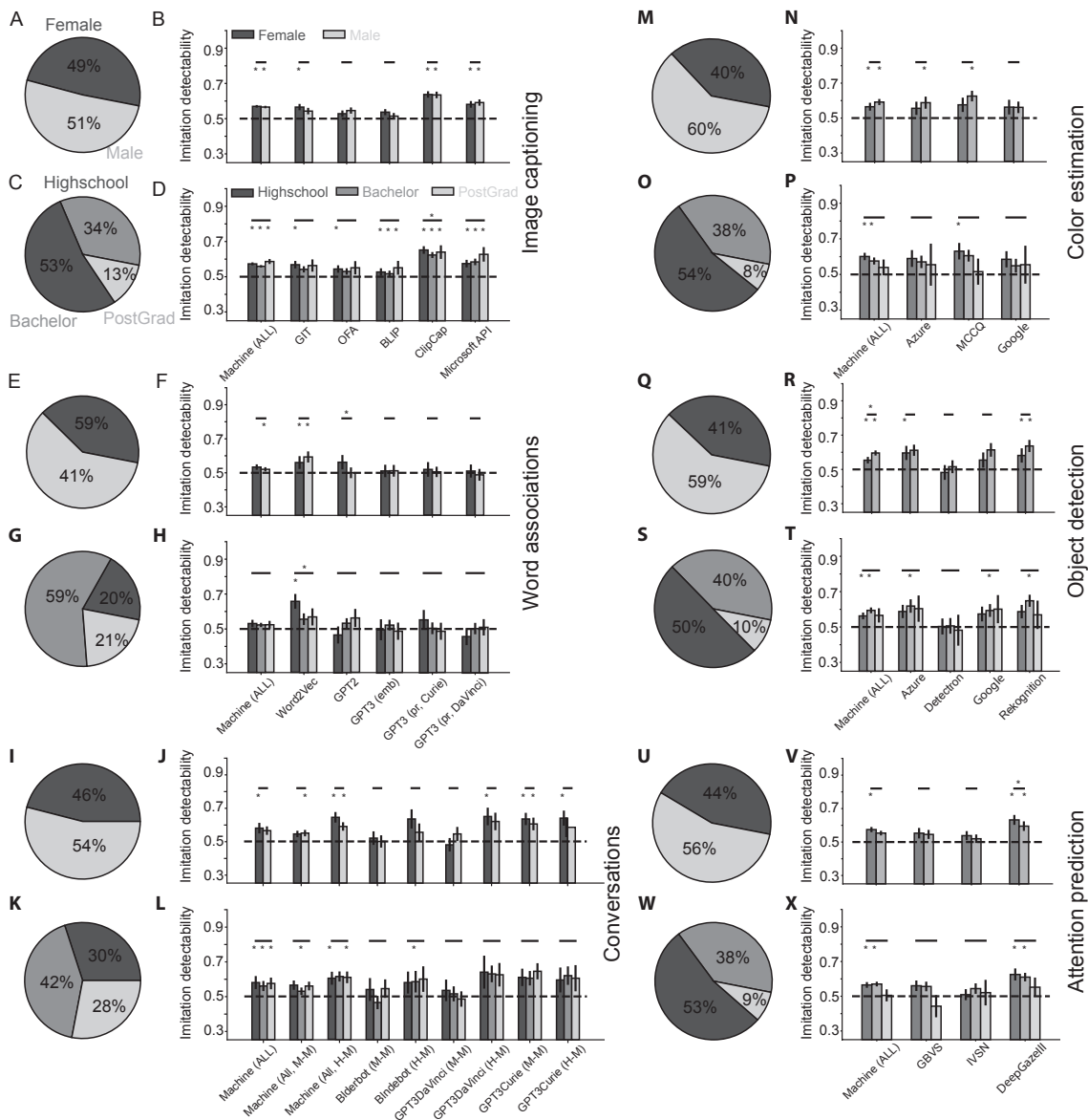


Figure S1: **Schematic illustration of data collection process for each task.** **A. Image captioning.** We collected captions by asking participants to describe images, inspired by COCO Captions data collection [48]. **B. Word associations.** Given a cue word, participants provided a single word that they associated with the cue. There were two versions. In the free association version (**B1**), participants were given a cue word and were asked to freely name a single association word. In the prompt-guided version (**B2**), participants were given 3 cue-association pairs and provided an association for a 4th cue word. **C. Conversations.** We collected conversations between two agents (agent A and agent B). Each agent could be either a human or a machine. Thus, there were 3 types of conversations: human-human (type 1), human-machine (type 2), and machine-machine (type 3). For types 1 and 2, we collected live conversations on popular chatting platforms, such as WhatsApp and Messenger, or a custom app. Participants did not know whether they were conversing with another human or with a machine. See **Methods** for details and **Sec. S1** for example conversations. **D. Color estimation.** Given an image, participants selected the dominant color from a pop-up menu. **E. Object detection.** Given an image, participants described three things they saw in the image. **F. Attention prediction.** Participants freely viewed an image for 12 seconds. The circles denote fixations and the lines denote eye movements between fixations.



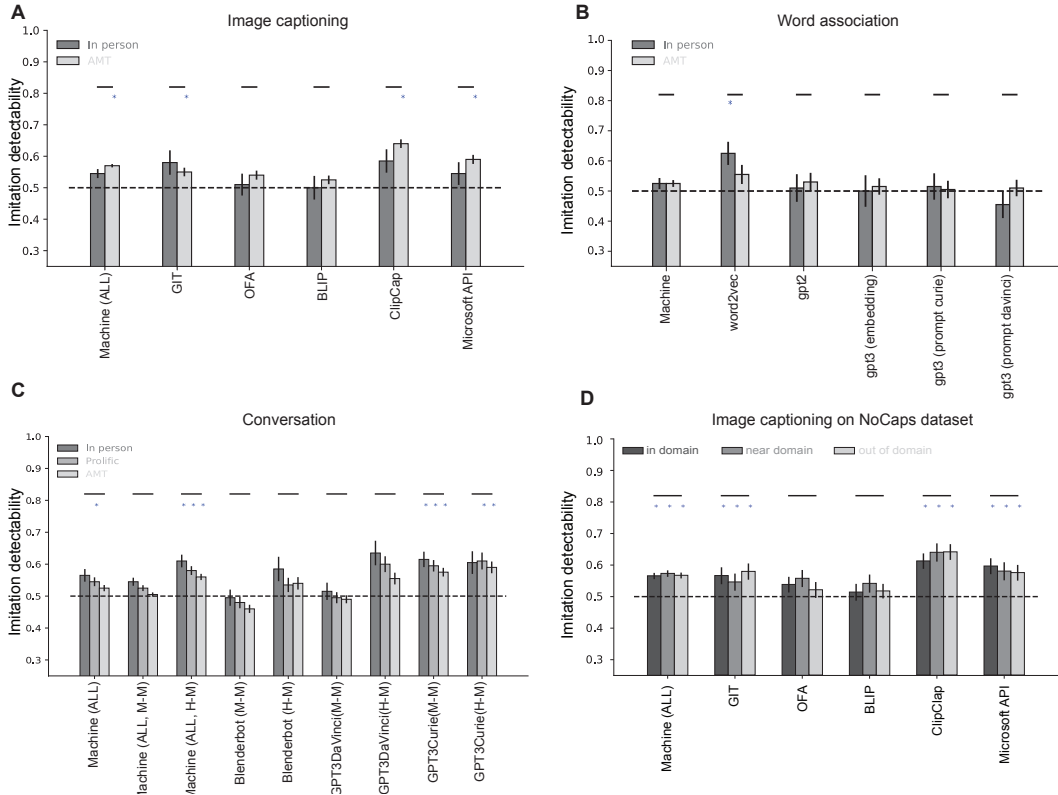


Figure S3: **Results of the Turing-like test for human judges conducted in various data collection platforms.** We collected in-person data (dark gray), Amazon Mechanical Turk (AMT) data (light gray) and Prolific data for the Conversation task (medium gray). Results are shown for the Image captioning task (A), Word association task (B), and Conversation task (C). Error bars denote bootstrap standard deviations (see **Methods, Data analyses**). The dashed line denotes a good imitator with imitation detectability at random level. The asterisks (\*) denote the statistical significance with Bonferroni corrections ( $p < 0.05$ ). Red asterisks above the line denote comparisons among the different platforms. Blue asterisks below the line denote comparisons with perfect imitation. **Results of the Turing-like test for human judges on the NoCaps dataset in the Image captioning task.** Imitation detectability in the Image captioning task for images in the NoCaps dataset [49]. Images from the NoCaps dataset include in-domain (dark gray), near-domain (medium gray), and out-of-domain images (light gray), reflecting the similarity to object classes from the COCO dataset[48], which was used for training the AI models. Asterisks (\*) below the line denote statistically significant differences with respect to 0.5 (horizontal dashed line,  $p < 0.05$ ). Asterisks above the line denote statistically significant differences among the three types of domains (in this case, none of the results showed statistically significant differences).

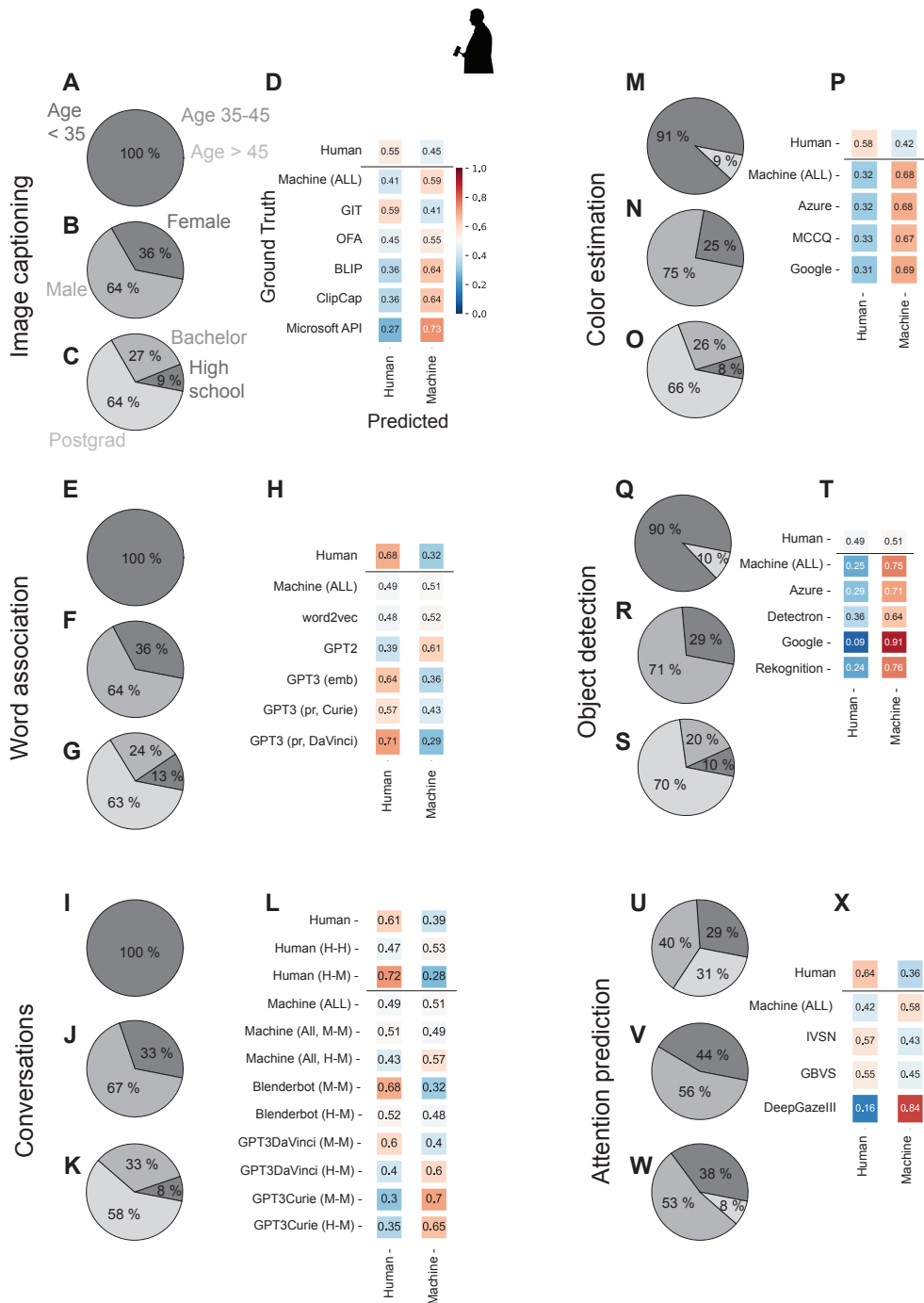


Figure S4: **Demographics and results conducted using AI experts as human judges.** Results are shown separately for the Image captioning task (A-D), the Word association task (E-H), the Conversation task (I-L), the Color estimation task (M-P), the Object detection task (Q-T), and the Attention prediction task (U-X). A, E, I, M, Q, U. Distribution of participants' age (indicated by different shades of gray). B, F, J, N, R, V. Distribution of participants' gender (indicated by different shades of gray). C, G, K, O, S, W. Distribution of participants' education (indicated by different shades of gray). D, H, L, P, T, X. Confusion matrices with AI experts as human judges for in-lab Turing-like tests over all six tasks. Format and conventions as in Figure 4. The colorbar in D applies to all panels.

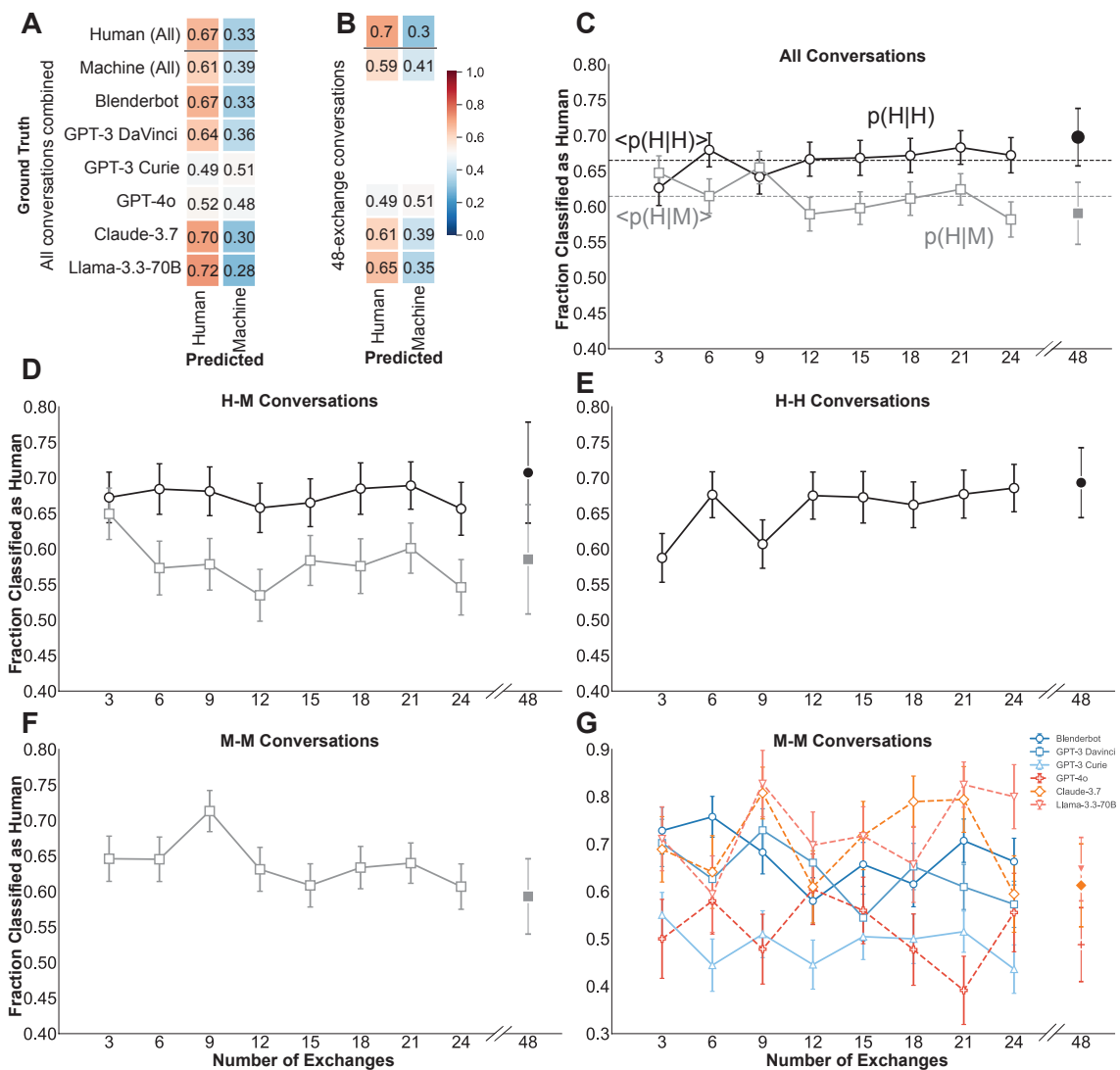


Figure S5: **Length dependence of Turing-like test results for human judges in the Conversation task.** Turing-like tests were performed by 141 human judges reading conversations of different lengths between 2 agents. The conversation dataset comprised more than 450 conversations that could come in the form of 3, 6, 9, ..., 24 exchanges. A set of 35 additional conversations was collected with 48 exchanges. The conversations could be human-human (H-H), human-machine (H-M), or machine-machine (M-M). The machines used in this dataset were: Blenderbot, GPT-3 DaVinci, GPT-3 Curie, GPT-4o, Claude 3.7, Llama 3.370B. **A** Confusion matrix of the entire conversation dataset with Prolific human judges. **B** Confusion matrix specifically for the trials that were 48 exchanges long. **C** Average fraction of sentences where human (H) agents were classified as humans (black) or machine (M) agents classified as humans (gray) as a function of conversation length across all conversation types. The dashed lines denote the average accuracy over all conversation lengths. **D** Same as **C** for H-M conversations. **E** Same as **C** for H-H conversations. **F** Same as **C** for M-M conversations. **G** Extension of **F**. Imitation detectability separated by the six AI models used in the dataset. In the blue palette, older models were used, and the data was collected between 2023-2025. In the orange palette, newer models were used and tested in 2025.

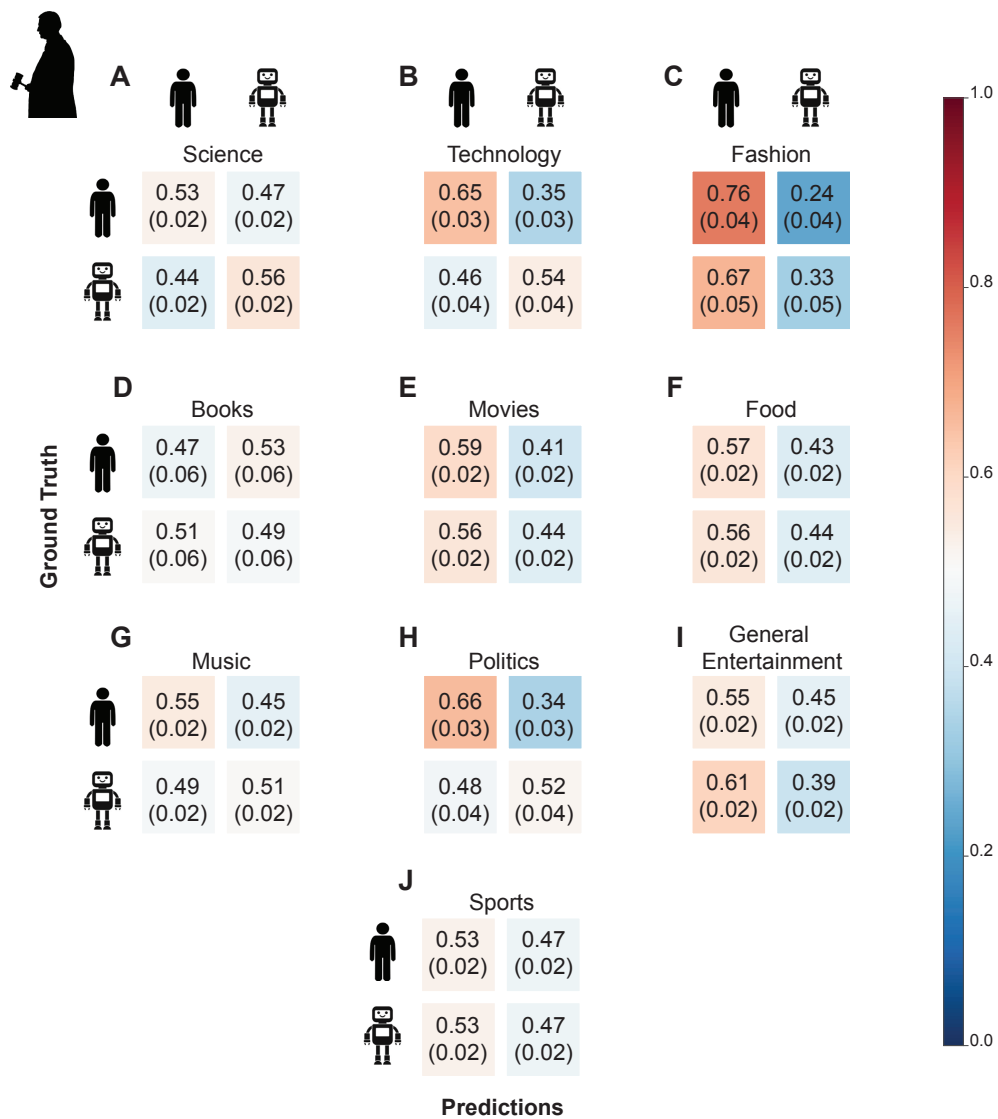


Figure S6: **Results of Turing-like test for human judges in the conversation task separated by topic.** Conventions follow those in Fig. 4. The topic of each conversation is indicated at the top of each panel.

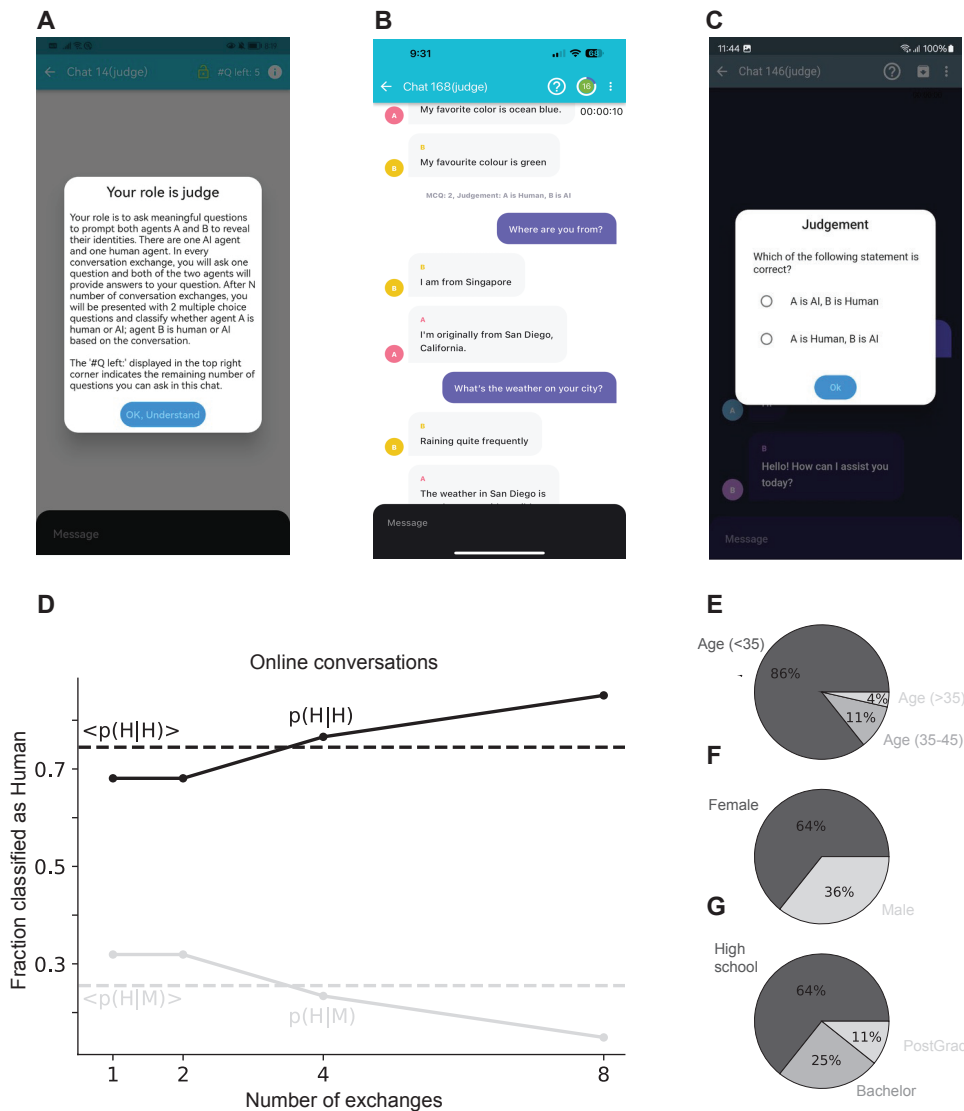


Figure S7: **Schematic illustration of the data collection process for conversation tasks in classical Turing tests.** **A.** Screenshot of instruction phase. At the beginning of every Turing test, both the human judge and the human agent are presented with instructions informing them of their identity, and the objective of that role. **B.** Screenshot of the actual Turing test. The test always starts with a judge asking a question followed by both agents answering the questions. The number of remaining conversation exchanges is shown in orange (top right). Each agent can only see the questions from the judges, but not the answers provided by the other agent. **C.** Screenshot at the end of the Turing test. The judge has to make a two-alternative forced choice decision based on the responses from both agents. We also conducted additional results of classical Turing tests on the machine agent LLAMA with safeguards removed. Compared with the original classical Turing tests on the machine agent ChatGPT as described in the main text, we provide additional results on the machine agent LLAMA with safeguards disabled. See **Methods** for detailed implementations. **D** Results of classical Turing test. Fraction classified as human as a function of the number of exchanges. Dashed lines indicate average values over all exchanges. **(E-G)** Participant demographic information: age **(E)**, gender **(F)**, and education **(G)**.

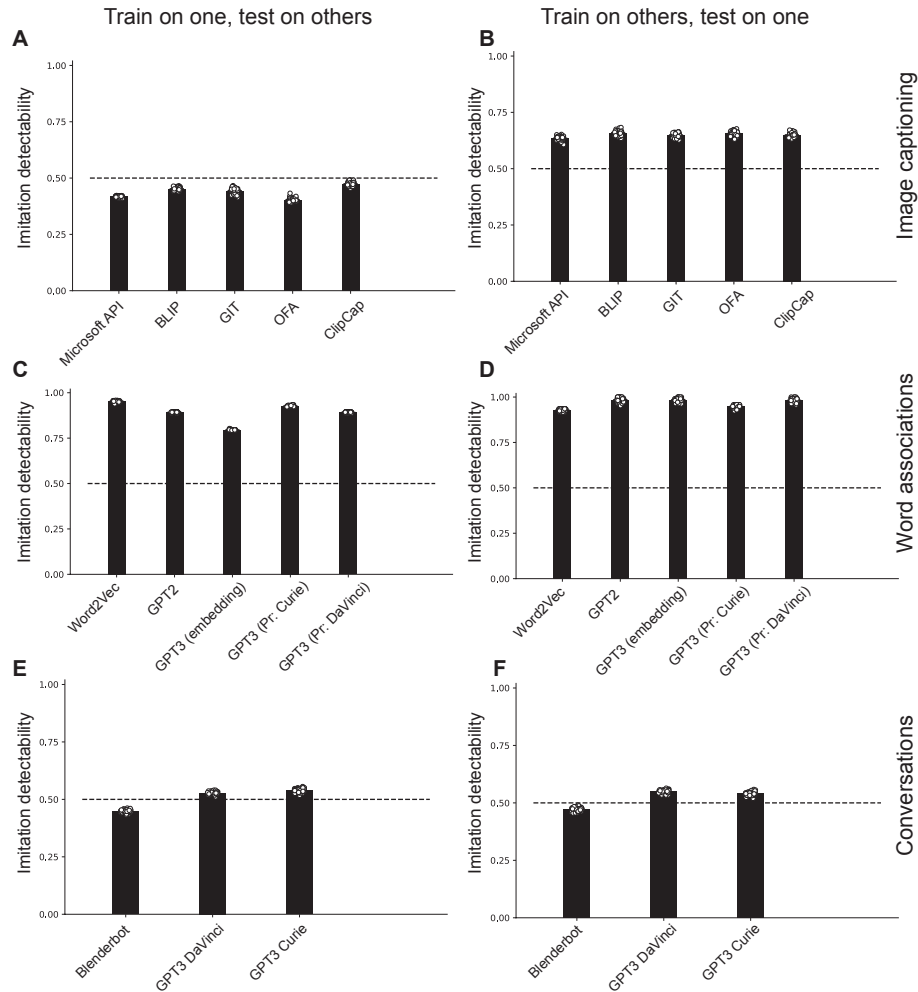


Figure S8: **Extrapolation across models for machine judges.** Expanding on Fig. 4D-F, here the machine judges are trained on data from only one model and tested on all the other models (A, C, E), or trained on all models except for one and tested on that one model (B, D, F). Imitation detectability is shown for the Image captioning task (A, B), the Word association task (C, E), and the Conversation task (E, F). The horizontal dashed line indicates chance levels.

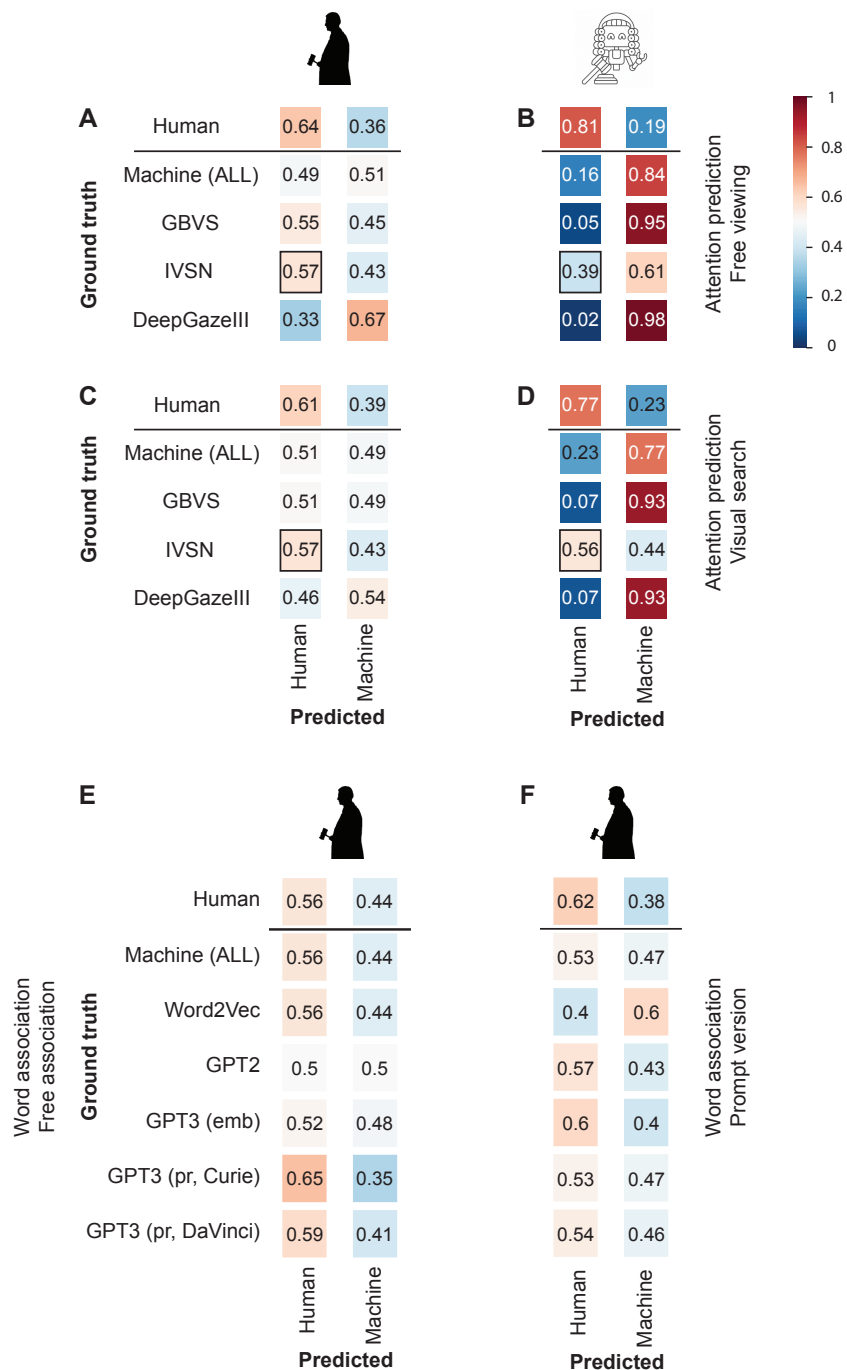


Figure S9: **Comparison between free viewing and visual search in the Attention task.** Using the same format as in Fig. 6, the results in Fig. 6E are shown here separately for free viewing (A) and visual search (C) and the results in Fig. 6F are shown here separately for free viewing (B) and visual search (D). **Comparison between free association and prompt version in the Word Association task.** Using the same format as in Fig. 5, the results in Fig. 5C are shown here separately for free viewing (E) and visual search (F).

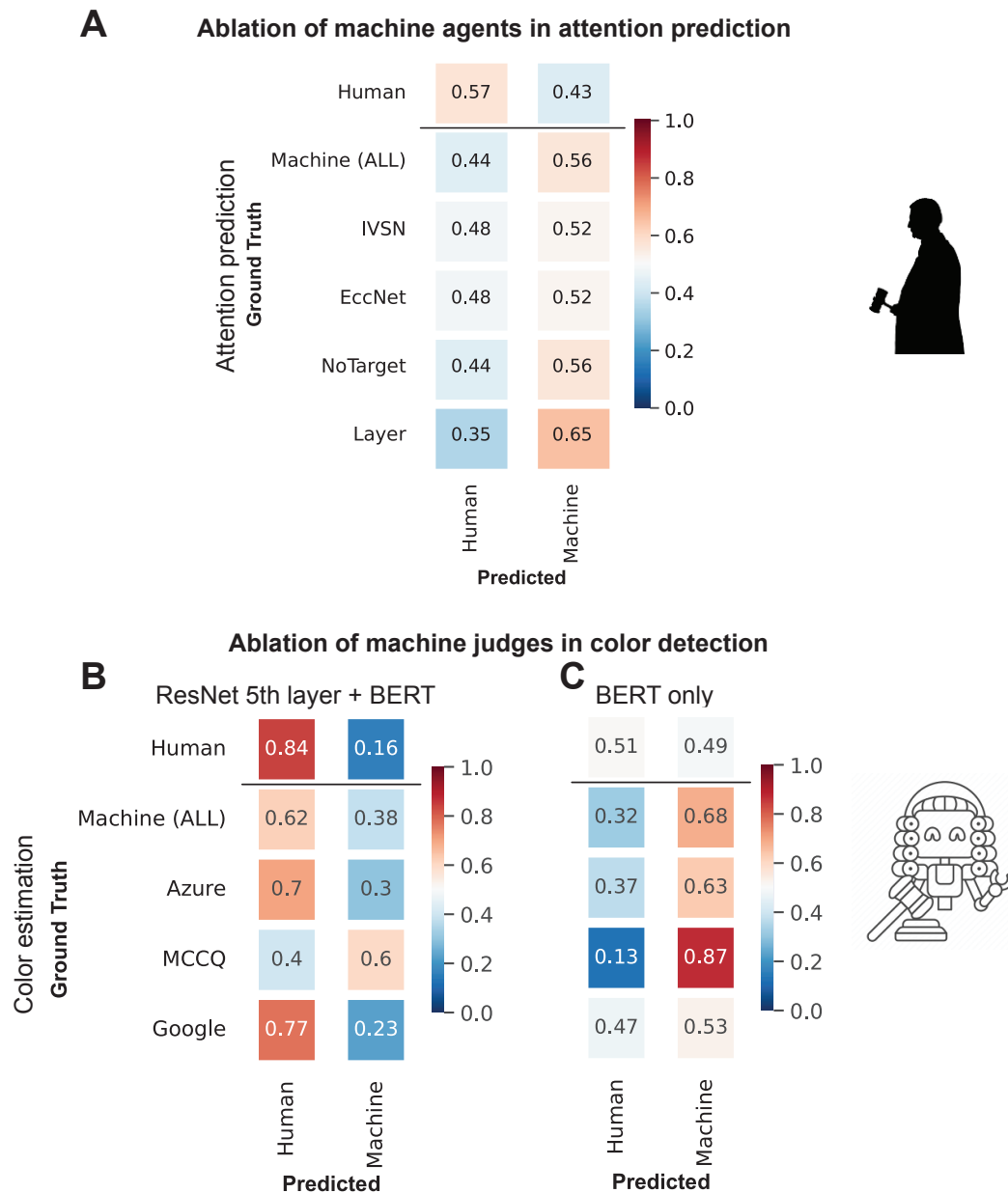


Figure S10: **Full results of ablated AI models as machine agents and machine judges.** **A** Full results of Turing-like tests for human judges on ablated machine agents in the Attention Prediction task. From the top to the bottom, we include IVSN [12], EccNet [13], Target removed in IVSN, and early layer to extract features for visual search in IVSN. See **Methods** for detailed description of the ablated machine agents. **B, C** Full results of Turing-like tests for ablated machine judges in the Color estimation task. **B** The ablated machine judge is constructed in a way that takes the visual features extracted from the 5th layer of ResNet alongside text embeddings of color words as inputs. **C** The same Turing-like tests are conducted in the Color estimation task but with the ablated machine judge taking the color responses, excluding visual features. See **Methods** for detailed description of the ablated machine judges. The design format in **B** and **C** follows that of **Fig. 6B**.

## Supplementary Tables

Task	Num. Stimuli	Num. Human Agents	Num. Human Judges	Num. Turing Tests	Sources of Datasets	AI agents	AI judges
Image captioning	9,910	229	485	22,680	self-collect, MSCOCO [55], nocaps [49]	GIT-Large [1], OFA-Huge [2], BLIP-Large [3], ClipCap-Transformer (beam search) [4], Microsoft’s Azure Cognitive Services [67], OpenFlamingo-4B [56], ChatGPT-4o [9], SCST [57], LLaVA-v1.6-Mistral-7B [40]	SVM-GPT-curie
Word association	2,050	40	212	10,521	self-collect,	Word2Vec [5], GPT2 [6], GPT3-embedding (davinci) [7], GPT3-prompt (text-curie-001) [7], GPT3-prompt (text-Davinci-002) [7], ChatGPT-4o [9]	SVM-Word2Vec, SVM-GPT2, SVM-GPT3(davinci)
Conversation	3,845	217	504	11,265	self-collect, Topical-Chat [8]	GPT3-text-davinci-002[0], GPT3-text-curie-001[0], Blenderbot[69], DialogPT [8], ChatGPT [9], GPT3.5-turbo-1106 [9], GPT4-o [9], Claude-3.7 [68], Llama-3.3-70B [40]	SVM+BERT [21], ChatGPT-zero-shot, ChatGPT-one-shot
Color estimation	5,495	65	177	9,041	self-collect, MSCOCO [55]	Google Vision API [60], Microsoft Azure Cognitive Services [67], MMCQ [61], ChatGPT-4o [9]	SVM+VGG+BERT [21]
Object detection	5,656	45	189	9,342	self-collect, MSCOCO[55]	Google Vision API, Microsoft Azure Cognitive Services [67], Amazon Rekognition [62], Detectron2 [10], ChatGPT-4o [9]	SVM+BERT [21]
Attention prediction	8,920	40	349	9,342	NatureDesign [12], FindingWaldo [12], NatureSaliency [53]	IVSN [12], [13], [53], DeepGaze3 [63], GBVS [14], ChatGPT-4o [9]	SVM on 2D coordinates
Total	36,499	636	1,916	72,191	-	37	10

Table S1: **Specifications of six Turing tasks.** Source datasets, number of Turing-like tests conducted, number of stimulus, and number of AI models used to collect responses are listed for each task. See **Methods** for task descriptions.

	Turing-like tests without feedback				Turing-like tests with feedback	
	Online		In-lab		Online	
	Num Human judges	Num trials per judge	Num Human judges	Num trials per judge	Num Human judges	Num trials per judge
Image Captioning	393	44	41	35	51	100
Word Association	121	44	40	31	50	90
Conversation Task	366	23	90	17	48	32
Color Estimation	115	35	12	35	50	100
Object Detection	129	37	10	37	50	90
Attention Prediction	289	22	10	46	50	60

Table S2: **Overview of total numbers of Turing-like tests over all six tasks under different conditions.** From left to right, the columns represent the task names, the number of human judges, and the number of trials completed per judge for both online and in-lab experiments, under conditions without and with training feedback.

Task	Num human judges	Num Catch trials only	Accuracy $P(M M)$ (%)	Design Rules	Example
Image Caption	51	1,020	90	Word repetitions	“with with with with”
Word Association	50	357	75	Cue word repetition	Cue: healthy; Association: healthy
Conversation	93	279	96	Repeated responses, long conversations with long list of items	see <b>Sec.S3</b>
Color Estimation	50	341	80	An object not present in the picture	“dog”
Object Detection	50	345	89	Absent object repeated three times	“elephant elephant elephant” in a kitchen scene
Attention Prediction	50	200	82	Repeated fixations at random spot	Fixation repeated 15 times at coordinate (100,203)

Table S3: **Description and performance in catch trials over all six tasks.** Task name, number of participants, number of catch trials collected, accuracy of correctly classifying catch trials generated by machines as machines ( $P(M|M)$ ), brief description of the design rules for catch trials, example catch trials for a specific task. See **Methods** for further details.

Task	Num Human Judges	Num Trials	Imitation Detectability
Image Captioning	100	1,000	0.53
Word Association	50	250	0.44
Conversation task	45	392	0.52
Color Estimation	50	250	0.46
Object Detection	50	250	0.44
Attention Prediction	50	250	0.56

Table S4: **Imitation detectability scores for the machine agent ChatGPT-4o [9] in Turing-like tests across all six language and vision tasks.** Task names, number of participants, number of Turing-like test trials, and imitation detectability scores.

Task	Metric	Models					
		Human	GIT	OFA	BLIP	ClipCap	Microsoft_API
Image Captioning	determiners/caption	2.071 ± 0.955	2.052 ± 0.697	2.375 ± 0.730	2.390 ± 0.680	2.506 ± 1.064	1.829 ± 0.687
	adjectives/caption	0.875 ± 0.937	0.721 ± 0.892	0.648 ± 0.771	0.732 ± 0.830	0.731 ± 0.963	0.303 ± 0.569
	nouns/caption	2.977 ± 1.298	2.776 ± 0.999	2.986 ± 0.973	3.119 ± 0.983	3.081 ± 1.621	2.235 ± 0.871
	adverbs/caption	0.060 ± 0.251	0.014 ± 0.118	0.008 ± 0.089	0.006 ± 0.077	0.011 ± 0.104	0.000 ± 0.000
	verbs/caption	0.213 ± 0.445	0.052 ± 0.222	0.094 ± 0.292	0.053 ± 0.224	0.141 ± 0.605	0.017 ± 0.129
	prepositions/caption	1.589 ± 0.980	1.456 ± 0.827	1.740 ± 0.870	1.785 ± 0.886	1.953 ± 1.594	1.089 ± 0.777
Word Association		Human	word2vec	GPT2	GPT3 (embedding)	GPT3 (prompt curie)	GPT3 (prompt davinci)
	ngram frequency	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	letters/word	7.325 ± 2.413	7.084 ± 2.338	7.265 ± 1.795	6.952 ± 1.880	6.337 ± 2.302	7.265 ± 1.795
Conversations		Human	Blenderbot	GPT3 (Curie)	GPT3 (Da Vinci)		
	words/exchange	14.646 ± 10.124	20.772 ± 6.199	15.340 ± 12.652	13.617 ± 10.163		
	capital letters/exchange	2.292 ± 1.852	2.480 ± 1.253	2.486 ± 1.793	2.451 ± 1.591		
	punctuations/exchange	2.646 ± 2.199	3.164 ± 2.148	3.179 ± 1.967	3.106 ± 1.582		
	determiners/exchange	1.164 ± 1.311	1.325 ± 1.087	1.292 ± 1.386	1.011 ± 1.307		
	adjectives/exchange	1.069 ± 1.197	1.667 ± 1.343	1.263 ± 1.443	1.090 ± 1.296		
	nouns/exchange	1.949 ± 1.863	3.058 ± 1.835	1.659 ± 1.850	1.388 ± 1.636		
	adverbs/exchange	1.042 ± 1.285	2.117 ± 1.735	1.135 ± 1.203	1.160 ± 1.159		
	verbs/exchange	0.509 ± 0.741	0.768 ± 0.849	0.674 ± 0.907	0.474 ± 0.783		
	prepositions/exchange	1.386 ± 1.500	2.293 ± 1.435	1.424 ± 1.670	1.151 ± 1.475		

Table S5: **Humans and machine responses did not differ in basic low-level statistics.** Multiple statistics about position-of-speech and frequency for the Image captioning task, the Word association task, and the Conversation task. These statistics include the number of words, n-gram frequency, capitalization, punctuation, determiners, adjectives, nouns, adverbs, verbs, prepositions, and letters per word. These properties are reported per caption, per word, or per conversation exchange. Errors denote bootstrap standard deviations.

2042 **Supplementary References**

- 2043 [67] *Microsoft azure cognitive api*, [https://azure.microsoft.com/en-us/products/cognitive-](https://azure.microsoft.com/en-us/products/cognitive-services/)  
2044 [services/](https://azure.microsoft.com/en-us/products/cognitive-services/), Accessed: 2022-10-30.
- 2045 [68] Anthropic, “Anthropic. claude 3 haiku: Our fastest model yet.,” 2024. [Online]. Available: :  
2046 <https://www.anthropic.com/news/claude-3-haiku>.
- 2047 [69] K. Shuster et al., *Blenderbot 3: A deployed conversational agent that continually learns to*  
2048 *responsibly engage*, 2022.
- 2049 [70] C. Li et al., “Large language models understand and can be enhanced by emotional stimuli,”  
2050 *arXiv*, 2023. arXiv: [2307.11760](https://arxiv.org/abs/2307.11760) [cs.CL]. [Online]. Available: [https://arxiv.org/abs/2307.](https://arxiv.org/abs/2307.11760)  
2051 [11760](https://arxiv.org/abs/2307.11760).
- 2052 [71] B. Guo et al., “How close is chatgpt to human experts? comparison corpus, evaluation, and  
2053 detection,” *arXiv*, 2023. arXiv: [2301.07597](https://arxiv.org/abs/2301.07597) [cs.CL]. [Online]. Available: [https://arxiv.](https://arxiv.org/abs/2301.07597)  
2054 [org/abs/2301.07597](https://arxiv.org/abs/2301.07597).
- 2055 [72] S. Harnad, “Levels of functional equivalence in reverse bioengineering,” *Artificial life*, vol. 1,  
2056 no. 3, pp. 293–301, 1994.
- 2057 [73] S. Harnad, “Turing on reverse-engineering the mind,” *Journal of Logic, Language, and Infor-*  
2058 *mation*, 1999.
- 2059 [74] P. Hayes and K. Ford, “Turing test considered harmful,” in *IJCAI (1)*, Citeseer, 1995, pp. 972–  
2060 977.
- 2061 [75] R. L. Purtil, “Beating the imitation game,” *Mind*, vol. 80, no. 318, pp. 290–294, 1971.
- 2062 [76] K. Gunderson, “The imitation game,” *Mind*, vol. 73, no. 290, pp. 234–245, 1964.
- 2063 [77] K. Gunderson, *Mentality and machines*. U of Minnesota Press, 1985.
- 2064 [78] K. LaCurts, “Criticisms of the turing test and why you should ignore (most of) them,” *Official*  
2065 *blog of MIT’s course: Philosophy and theoretical computer science*, 2011.
- 2066 [79] R. Geirhos et al., “Partial success in closing the gap between human and machine vision,”  
2067 *Advances in Neural Information Processing Systems*, vol. 34, pp. 23 885–23 909, 2021. [Online].  
2068 Available: [https://proceedings.neurips.cc/paper/2021/hash/2c2e3b1d9d4a4d0b6d4e9b7f1c3f1c1e-](https://proceedings.neurips.cc/paper/2021/hash/2c2e3b1d9d4a4d0b6d4e9b7f1c3f1c1e-Abstract.html)  
2069 [Abstract.html](https://proceedings.neurips.cc/paper/2021/hash/2c2e3b1d9d4a4d0b6d4e9b7f1c3f1c1e-Abstract.html).
- 2070 [80] R. Geirhos et al., “Shortcut learning in deep neural networks,” *Nature Machine Intelligence*,  
2071 vol. 2, no. 11, pp. 665–673, 2020. DOI: [10.1038/s42256-020-00257-z](https://doi.org/10.1038/s42256-020-00257-z).
- 2072 [81] M. Kümmerer and M. Bethge, “Predicting visual fixations,” *Annual Review of Vision Science*,  
2073 vol. 9, pp. 269–291, 2023. DOI: [10.1146/annurev-vision-120822-072528](https://doi.org/10.1146/annurev-vision-120822-072528).

- 2074 [82] V. Veerabadrán et al., “Subtle adversarial image manipulations influence both human and  
2075 machine perception,” *Nature Communications*, vol. 14, no. 1, p. 4933, 2023. DOI: [10.1038/  
2076 s41467-023-40499-0](https://doi.org/10.1038/s41467-023-40499-0).
- 2077 [83] J. Feather, D. Linsley, and T. Serre, “Model metamers reveal divergences between human and  
2078 machine vision,” *Nature Communications*, vol. 14, no. 1, p. 1234, 2023. DOI: [10.1038/s41467-  
2079 023-01234-5](https://doi.org/10.1038/s41467-023-01234-5).
- 2080 [84] T. Fel, I. Felipe, D. Linsley, and T. Serre, “Harmonizing the object recognition strategies of deep  
2081 neural networks with humans,” in *Advances in Neural Information Processing Systems*, vol. 35,  
2082 2022, pp. 9432–9446. [Online]. Available: <https://serre-lab.github.io/Harmonization>.
- 2083 [85] M. Zhang, C. Tseng, and G. Kreiman, “Putting visual object recognition in context,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 12 985–  
2084 12 994.
- 2086 [86] S. Han, Z. Wang, and M. Zhang, “Flow snapshot neurons in action: Deep neural networks  
2087 generalize to biological motion perception,” *arXiv preprint arXiv:2405.16493*, 2024.
- 2088 [87] H. Cao et al., “A survey on generative diffusion models,” *IEEE Transactions on Knowledge and  
2089 Data Engineering*, 2024.
- 2090 [88] Q. Lin, J. Zhang, Y. S. Ong, and M. Zhang, “Make me happier: Evoking emotions through  
2091 image diffusion models,” *arXiv preprint arXiv:2403.08255*, 2024.
- 2092 [89] K. M. Colby, “Modeling a paranoid mind,” *Behavioral and Brain Sciences*, vol. 4, no. 4, pp. 515–  
2093 534, 1981.
- 2094 [90] K. M. Colby, S. Weber, and F. D. Hilf, “Artificial paranoia,” *Artificial Intelligence*, vol. 2, no. 1,  
2095 pp. 1–25, 1971.
- 2096 [91] J. Weizenbaum, “Eliza—a computer program for the study of natural language communication  
2097 between man and machine,” *Communications of the ACM*, vol. 9, no. 1, pp. 36–45, 1966.
- 2098 [92] K. Shuster et al., “Blenderbot 3: A deployed conversational agent that continually learns to  
2099 responsibly engage,” *arXiv preprint arXiv:2208.03188*, 2022.
- 2100 [93] G. Team et al., “Gemini: A family of highly capable multimodal models,” *arXiv preprint  
2101 arXiv:2312.11805*, 2023.
- 2102 [94] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum contrast for unsupervised visual  
2103 representation learning,” in *Proceedings of the IEEE/CVF conference on computer vision and  
2104 pattern recognition*, 2020, pp. 9729–9738.
- 2105 [95] A. Bardes, J. Ponce, and Y. LeCun, “Vicreg: Variance-invariance-covariance regularization for  
2106 self-supervised learning,” *arXiv preprint arXiv:2105.04906*, 2021.
- 2107 [96] H. Neil and W. Dirk, “Transformers for image recognition at scale,” *Online: https://ai.google-  
2108 blog.com/2020/12/transformers-for-image-recognition-at-scale.html*, 2020.

- 2109 [97] X. Chen and K. He, “Exploring simple siamese representation learning,” in *Proceedings of the*  
2110 *IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 15 750–15 758.
- 2111 [98] E. E. Guzik, C. Byrge, and C. Gilde, “The originality of machines: Ai takes the torrance test,”  
2112 *Journal of Creativity*, vol. 33, no. 3, p. 100 065, 2023, ISSN: 2713-3745. DOI: <https://doi.org/10.1016/j.yjoc.2023.100065>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2713374523000249>.  
2113  
2114
- 2115 [99] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time  
2116 object detection,” in *Proceedings of the IEEE conference on computer vision and pattern recog-*  
2117 *niton*, 2016, pp. 779–788.
- 2118 [100] T. Wertheimer, “Blake lemoine: Google fires engineer who said ai tech has feelings,” *BBC news*,  
2119 Jul. 23, 2022. Accessed: Jul. 23, 2022. [Online]. Available: <https://www.bbc.com/news/technology-62275326>.  
2120
- 2121 [101] N. Tiku, “The google engineer who thinks the company’s ai has come to life,” *WashingtonPost*,  
2122 Jun. 11, 2022. Accessed: Jun. 11, 2022. [Online]. Available: <https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/>.  
2123
- 2124 [102] R. Maruf, “Google fires engineer who contended its ai technology was sentient,” *CNN*, Jul. 22,  
2125 2022. Accessed: Jul. 22, 2022. [Online]. Available: <https://edition.cnn.com/2022/07/23/business/google-ai-engineer-fired-sentient/index.html>.  
2126
- 2127 [103] M. Westerlund, “The emergence of deepfake technology: A review,” *Technology Innovation*  
2128 *Management Review*, vol. 9, no. 11, 2019.
- 2129 [104] H. K. Hall, “Deepfake videos: When seeing isn’t believing,” *Cath. UJL & Tech*, vol. 27, p. 51,  
2130 2018.
- 2131 [105] “The straits times news: Stem the scams: Beware the bots to avoid being distraught,” <https://www.straitstimes.com/the-scams-beware-the-bots-to-avoid-being-distraught>, 2023.  
2132
- 2133 [106] “The cnbc news: Ai tools such as chatgpt are generating a mammoth increase in malicious  
2134 phishing emails,” <https://www.cnbc.com/2023/11/28/ai-like-chatgpt-is-creating-huge-increase-in-malicious-phishing-email.html>, 2023.  
2135