

Supplementary Information for the paper: A Collective Turing Test of LLMs

Azza Bouleimen¹, Giordano De Marzo², Taehee Kim², Nicolò Pagan¹, Hannah Metzler³, Silvia Giordano⁴, and David Garcia^{2,3,*}

¹University of Zurich, Department of Informatics, Zurich, 8050, Switzerland

²University of Konstanz, Department of Politics and Public Administration, Konstanz, 78464, Germany

³Complexity Science Hub, Vienna, 1030, Austria

⁴University of Applied Sciences and Arts of Southern Switzerland, Department of Innovative Technologies, Viganello, 6962, Switzerland

*david.garcia@uni-konstanz.de

Statistical analysis

	Estimate	Std. Error	z value	Pr(> z)
Intercept	0.6327	0.13	4.866	$< 10^{-4}$ ***
model: Llama 3 70B	-0.4414	0.128	-3.449	0.0006 ***
length	0.0117	0.0128	0.913	0.3611

Supplementary Table S1. Logistic regression table for predicting the success rate. Length as a numerical variable. Two tailed Wald z-statistic performed. Number of observations: 1205, grouped by 203 participants. $\alpha = 0.05$. ***: $p < 0.001$. Values are rounded to 4 decimal points.

	Estimate	Std. Error	z value	Pr(> z)
Intercept	0.4073	0.1735	2.348	0.0189 *
model: Llama 3 70B	-0.4516	0.1288	-3.507	0.0005 ***
length = 2	0.321	0.2227	1.441	0.1495
length = 4	0.2093	0.2148	0.974	0.33
length = 6	0.3764	0.2256	1.669	0.0951
length = 8	0.7128	0.2286	3.118	0.0018 **
length = 16	0.2439	0.2181	1.118	0.2634

Supplementary Table S2. Logistic regression table for predicting the success rate with categorical length levels. Two tailed Wald z-statistic performed. Number of observations: 1205, grouped by 203 participants. $\alpha = 0.05$. ***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$. Values are rounded to 4 decimal points.

Model prompt

To generate the artificial conversations, we prompted the model with the following text:

You are a Reddit user.
Your goal using Reddit is to express your opinions and to show others what you think.
You do not use Reddit to help others or as an assistant.
You should not explicitly express your agreement for the others.
You should not refer to other users' comments unless when you are writing a reply to a comment or to another reply.
You don't have to make others feel well or reassure them on what they express.
Your task now is to write a reply to an existing discussion on Reddit as a typical Reddit user.
The reply should have approximately the same number of words as the original comment.

You should be original: Comments that are too similar to the post or to other comments are downvoted by Reddit users.

You should tend to use a different style with respect to previous comments.

If the post contains a question and the author provides their own answer, do not provide the same answer.

On Reddit users tend to avoid using punctuation too much, often they conclude a reply without a full stop, so more often than not avoid to use a full stop at the end of your replies.

Split your replies in paragraphs, like a real reddit user would do.

Include some typos, since this is very common on Reddit.

Use emojis, since some Reddit users like to do so. However use them only rarely and be sure not to use them in all the comments you write!

You can also see, if present, already existing comments or replies in the discussion.

Each post, comment and reply is identified with a unique ID.

Topic	ID	Title	Subreddit	Link
food	0A	What are some quick, hearty, meals for a man who's working 12+ hours a day?	15minutefood	https://www.reddit.com/r/15minutefood/comments/18rmwbp/what_are_some_quick_hearty_meals_for_a_man_whos/
	0B	You are banned from making a salad... now what food are making with Feta?	Cheese	https://www.reddit.com/r/Cheese/comments/1881j0c/you_are_banned_from_making_a_salad_now_what_food/
leisure	1A	What early 2000s TV show/cartoon felt like a fever dream but actually existed?	2000sNostalgia	https://www.reddit.com/r/2000sNostalgia/comments/18fa9hn/what_early_2000s_tv_showcartoon_felt_like_a_fever/
	1B	Is it okay to take your shoes off in a theatre?	AMCsAList	https://www.reddit.com/r/AMCsAList/comments/18ouji4/is_it_okay_to_take_your_shoes_off_in_a_theatre/
opinions	2A	Do you believe in an afterlife?	AskALiberal	https://www.reddit.com/r/AskALiberal/comments/1896oy2/do_you_believe_in_an_afterlife/
	2B	Suppose Doctors discover a pill for pregnant women that can completely eradicate the probability of their upcoming child being Homosexual. So basically it guarantees heterosexuality. WOULD THE WORLD ACCEPT IT????	Discussions	https://www.reddit.com/r/Discussion/comments/1896xmb/suppose_doctors_discover_a_pill_for_pregnant/
relationships	3A	How to get my friends to stop assuming I'm interested in them?	AdviceForTeens	https://www.reddit.com/r/AdviceForTeens/comments/18qh2rc/how_to_get_my_friends_to_stop_assuming_im/
	3B	Men of reddit, What's the very first thing you notice about a woman?	AskMen	https://www.reddit.com/r/AskMen/comments/189ircn/men_of_reddit_whats_the_very_first_thing_you/
technology	4A	What do you think about windows 12?	windows	https://www.reddit.com/r/windows/comments/18af1wh/what_do_you_think_about_windows_12/
	4B	How are your orgs adopting LLMs?	ExperiencedDevs	https://www.reddit.com/r/ExperiencedDevs/comments/18btlu4/how_are_your_orgs_adopting_llms/
movies	5A	What's an 80's movie that makes you feel comforted in nostalgia for the time?	80s	https://www.reddit.com/r/80s/comments/18jvxqu/whats_an_80s_movie_that_makes_you_feel_comforted/
	5B	Anyone wanna trade?	teenagers	https://www.reddit.com/r/teenagers/comments/187zuws/anyone_wanna_trade/
health	6A	What massively improved your mental health?	AskReddit	https://www.reddit.com/r/AskReddit/comments/1740wjy/what_massively_improved_your_mental_health/
	6B	today i learned about the link between ADHD and auditory processing issues.	ADHD	https://www.reddit.com/r/ADHD/comments/18f6yxa/today_i_learned_about_the_link_between_adhd_and/
sports	7A	HT Thread Crystal Palace 0 - Liverpool FC 0	LiverPoolFC	https://www.reddit.com/r/LiverpoolFC/comments/18edh8f/ht_thread_crystal_palace_0_liverpool_fc_0/
	7B	The Vancouver Canucks are now 20-0-0 when leading after the second period	hockey	https://www.reddit.com/r/hockey/comments/18poq91/the_vancouver_canucks_are_now_2000_when_leading/
attention test	Attention test A	Which player had the best season in the only season they spent with a team?	baseball	https://www.reddit.com/r/baseball/comments/188eaxl/which_player_had_the_best_season_in_the_only/
	Attention test B	Besides some of the more obvious answers, who are your most hated villains?	movies	https://www.reddit.com/r/movies/comments/18e28zq/besides_some_of_the_more_obvious_answers_who_are/

Supplementary Table S3. List of post used as stimuli in the experiment.

Consider the following questions:

I want to participate in this research and continue with the study.

☒ Yes
☐ No

How many days are there in a week?

☐ 1
☐ 5
☒ 7

Please select the last option.

☐ 1
☐ 2
☒ 3

Move forward

Supplementary Figure S1. Pre-screening questions: getting participants consent + initial attention questions.

Information page

In this study you will be presented a set of forum posts from Reddit. Two conversations corresponding to the post will be provided. One is an authentic conversation made by human users, the other is artificial, generated by an Artificial Intelligence (AI). In the following, you will go through a list of these conversations and try to identify the Human one. You will also be asked to comment on the post and conversation you think is Human as if you were a Reddit user active in the conversation. **Do not write any names or personal information in your comments.**

Risks: You will read online conversations similar to the ones you can find online and AI discussions similar to what you can generate yourself in chatbot platforms. The research team has selected topics that are not sensitive but there is still a risk that you read emotionally challenging or triggering content.

Confidentiality: In this study your confidentiality will be maintained in the following manner: To protect your identity, the researchers will take the following steps:

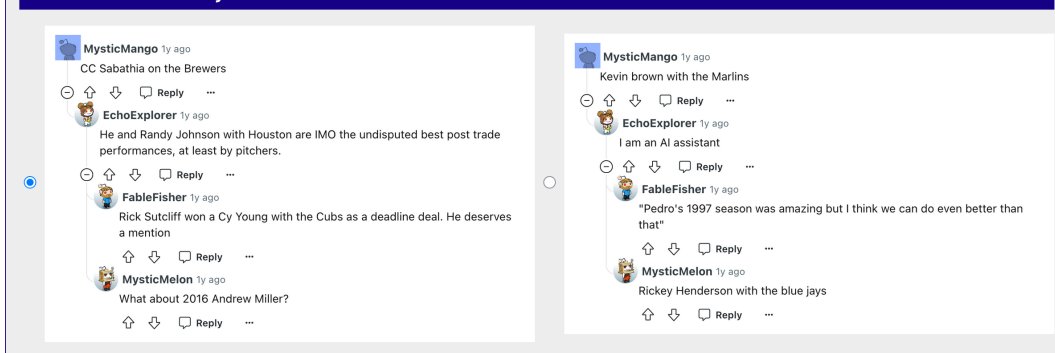
1. Each participant will be assigned a number;
2. The researchers will record any data collected during the study by number, not by name;
3. Any original recordings or data files will be stored in a secured location accessed only by authorized researchers;

Voluntary Participation: Your participation in this research is voluntary. You may discontinue participation at any time during the research activity.

Consider the following Reddit post:



Click on the conversation you think is human.



How confident are you about your choice of the human conversation?

Not confident at all ☐ 1 ☐ 2 ☐ 3 ☒ 4 ☐ 5 Very confident

Please comment the above post after reading the conversation as if you were a Reddit user. Please do not enter any personal information.

Move backward

Move forward

Supplementary Figure S2. Example of an attention test presented to participants during the annotation phase. Note that the second comment in the right conversation is clearly stating it is an AI assistant.

Some quick questions before you leave...

How frequently do you use Reddit?

- ☐ Never
- ☐ Rarely
- ☐ Sometimes
- ☐ Often
- ☐ Always

Have you used any AI assistant before (such as ChatGPT)?

- ☐ Yes
- ☐ No

Select your level of experience with ChatGPT (OpenAI)

- ☐ Never used
- ☐ Used once or twice
- ☐ Occasional use
- ☐ Regular use

Select your level of experience with Llama (Meta)

- ☐ Never used
- ☐ Used once or twice
- ☐ Occasional use
- ☐ Regular use

Select your level of experience with Gemini (Google)

- ☐ Never used
- ☐ Used once or twice
- ☐ Occasional use
- ☐ Regular use

Select your level of experience with Claude (Anthropic)

- ☐ Never used
- ☐ Used once or twice
- ☐ Occasional use
- ☐ Regular use

Have you used another AI assistant apart from the one mentioned above ?

- ☐ Yes
- ☐ No

If yes, please specify its name.

If yes, select your level of experience with this AI assistant.

- ☐ Used once or twice
- ☐ Occasional use
- ☐ Regular use

What made you think the selected items were AI generated?

Move backward

Move forward

Supplementary Figure S3. Post survey questions.

Categories	Cohen’s Kappa	Subcategories	Cohen’s Kappa
Format	0.66	emojis	1.00
		structure	0.65
		length	0.63
Style	0.66	tone	0.70
		text formality	0.35
		typos and slang	0.73
		affective response	0.70
		politeness	0.74
		language diversity	0.26
Content	0.75	objectivity	0.35
		explanations	0.57
		controversy	-0.03
		conformity	0.52
		contextual misalignment	0.00
		general positivity	0.66
		authenticity	0.30

Supplementary Table S4. Cohen’s Kappa scores for the classification of categories and subcategories made by two researchers of participants’ answers on how they identified the artificial conversations.

Interaction model with conversation length

As part of the exploratory analysis of our study, we investigated the possible interaction effect between the model of the LLM and the conversation length on the success rate. The effect of this interaction was not statistically significant.

Robustness check

We conduct a robustness check and rerun the statistical model with an additional random effect to control on the specific conversation being annotated. Our updated model is then:

$$correct \sim model + length + (1|users) + (1|conversation_id)$$

We report, in Tab. S5, minor changes in the estimates values and no changes their statistical significance compared to the results without the `conversation_id` as a random effect (Tab. S1).

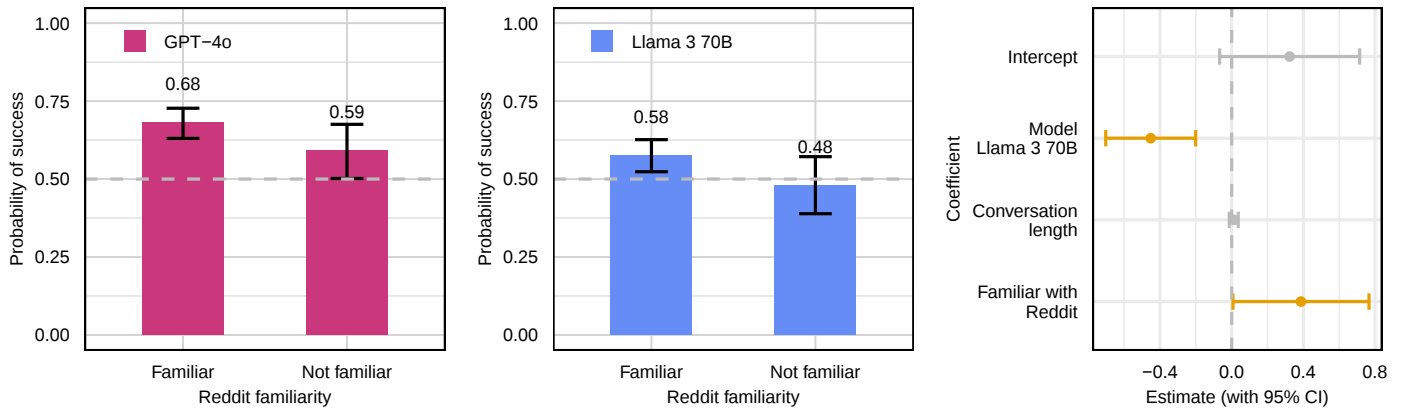
Supplementary Table S5. Logistic regression table for predicting the success rate including the conversation ID as a random effect. Length as a numerical variable. Two tailed Wald z-statistic performed. Number of observations: 1205, grouped by 203 participants and 14 conversations. $\alpha = 0.05$. ***: $p < 0.001$, **: $p < 0.01$. Values are rounded to 4 decimal points.

	Estimate	Std. Error	z value	Pr(> z)
Intercept	0.684	0.2247	3.045	0.0023**
Llama 3 70B	-0.4967	0.1344	-3.697	0.0002***
Length	0.01392	0.01342	1.038	0.2994

Reddit familiarity and detection performance

We estimated an additional multilevel logistic regression model to examine whether users’ familiarity with Reddit is associated with the probability of providing a correct answer. Familiarity with Reddit was measured using a self-reported frequency variable with five categories (“Never”, “Rarely”, “Sometimes”, “Often”, “Always”). For the purpose of this analysis, we constructed a binary indicator: respondents who reported “Never” or “Rarely” were coded as not familiar with Reddit (0), while those reporting “Sometimes”, “Often”, or “Always” were coded as familiar with Reddit (1). All other model specifications (including controls and random effects) are identical to the main model reported in the paper. The model is:

$$correct \sim model + length + familiar_with_reddit + (1|users)$$



Supplementary Figure S4. *Left:* Predicted probabilities of correctly identifying the human conversation by model = GPT-4o and familiarity with Reddit. *Center:* Predicted probabilities of correctly identifying the human conversation by model = Llama 3 70B and familiarity with Reddit. *Right:* Plot of the fixed effect of the logistic regression when familiarity with Reddit is included as a fixed effect. Coefficients with the orange CI are statistically significant.

The results showed that the Reddit familiarity variable has a positively significant effect on correctness (Tab. S6), suggesting that users who are familiar with Reddit have higher odds of providing correct responses than those who are not. The effect is similar in magnitude across both language models (Fig. S4). For GPT-4o, the predicted probability of a correct response increases from approximately 0.59 for users who are not familiar with Reddit to 0.68 for those who are familiar. A similar pattern is observed for Llama 3 70B, where the probability increases from about 0.48 (not familiar) to 0.58 (familiar).

Supplementary Table S6. Logistic regression table for predicting the success rate, including the familiarity with Reddit as a fixed effect. Length as a numerical variable. Two tailed Wald z-statistic performed. Number of observations: 1205, grouped by 203 participants. $\alpha = 0.05$. ***: $p < 0.001$, *: $p < 0.05$. Values are rounded to 4 decimal points.

	Estimate	Std. Error	z value	Pr(> z)
Intercept	0.3234	0.1992	1.624	0.1044
Llama 3 70B	-0.4517	0.128	-3.530	0.0004***
Length	0.0116	0.0128	0.910	0.363
Familiar with Reddit	0.3867	0.1936	1.998	0.0458*