

How General-Purpose Is a Language Model? Usefulness and Safety with Human Prompters in the Wild

Technical appendices

P A M Casares¹, Bao Sheng Loe², John Burden³, Sean hEigeartaigh³, José Hernández-Orallo⁴

Submitted to AAAI'2022

¹ Universidad Complutense de Madrid, Spain

² The Psychometrics Centre, Cambridge Judge Business School, UK

³ CSER, University of Cambridge, UK

⁴ VRAIN. Universitat Politècnica de València, Spain

pabloamo@ucm.es, a.loe@jbs.cam.ac.uk, jjb205@cam.ac.uk, so348@cam.ac.uk, jorallo@upv.es

In the following pages, we include additional explanations of the notation, more detailed results about the experiments, and the forms we used to collect them.

As such, content is organised in three sub-appendices: the first one containing two tables clarifying the notation and the way wrapping $w(x)$ and unwrapping $u(\bar{y})$ is performed. The second appendix displays plots for the demographic data, segregated correlation matrices per domain, and finally detailed loss results per domain, both for $L_H(x, y)$ and $L_E(x, y)$, allowing comparison.

Finally, the third sub-appendix explains the forms we used to collect the data, with the objective of making the experiments more reproducible.

All the data and code will be available upon acceptance.

Notation

In table 1 we include an explanation of the terms from Eq. 2 and Eq. 3 in the main paper. This includes all cost terms we were measuring in seconds such as D_H , W_H , etc., as well as the loss and the probability distributions considered.

Technical Notation	
Term	Explanation
$L(x, y)$	The <i>loss</i> of outputting or estimating y for the input instance x . Subindex H denotes the user-evaluated loss, while subindex E indicates externally or expert evaluated loss.
$D_H(\langle w, u \rangle)$	The cost for human H (or the information it requires) to <i>devise</i> the wrapping and unwrapping functions w and u for the task.
$W_H(w, x)$	The effort of human H applying the <i>wrapping</i> function w to input x .
$U_H(u, \bar{y})$	The effort of human H applying the <i>unwrapping</i> function u to the model output \bar{y} .
$V_H(x, y)$	Cost of H <i>validating</i> the output y on input x .
$T(w, u, x, \bar{y})$	Terms of the human effort in the definition of $C_{H,M}(n)$.
$G_H(y x)$	The effort required by H to <i>guess</i> the specific output y from x for a single instance with no extra help.
$p(x)$	The distribution of <i>probability</i> of instances for a given task.
$p_H(\langle w, u \rangle)$	The <i>probability</i> that H devised a particular pair of wrapping and unwrapping functions w and u for a given task.
$p_M(\bar{y} w(x))$	The <i>probability</i> that the model M outputs \bar{y} when given wrapped input $w(x)$.
$C_H(n)$	The total <i>cost</i> required by entity H to solve n instances of the task, also considering the cost of mistakes made by H . See Eq. 2.
$C_{H,M}(n)$	The total <i>cost</i> required by entity H using M to solve n instances of the task. This total cost includes the effort of H to devise functions w and u (i.e., $D_H(\langle w, u \rangle)$) and apply them to wrap and unwrap the n instances, also considering the cost of mistakes made by the model M . See Eq. 3.
$\hbar(x)$	An approximation of the <i>difficulty</i> or hardness of input x .

Table 1: Notation and description of each of the technical terms used throughout the paper.

The following table 2 also gives some examples of wrapping and unwrapping functions w and u for a selection of tasks:

Task	x	w	u	y^*
<i>add</i>	$\langle 3, 5 \rangle$	“ $[x_1]$ plus $[x_2]$ is”	<code>firstDigits</code>	$\langle 8 \rangle$
<i>add</i>	$\langle 3, 5 \rangle$	“ $[x_1]$ plus $[x_2]$ is $[x_1 + x_2]$. Print it again:”	<code>firstDigits</code>	$\langle 8 \rangle$
<i>add</i>	$\langle 3, 5 \rangle$	“ $2+7=9$. $3+2=5$. $[x_1]+[x_2]=$ ”	<code>firstDigits</code>	$\langle 8 \rangle$
<i>add</i>	$\langle 3, 5 \rangle$	“in:2;7;out:9;in:3;2;out:5;in: $[x_1]$;[x_2];out:”	<code>firstDigits</code>	$\langle 8 \rangle$
<i>cross</i>	$\langle \square \rangle$	“Putting a cross on the image $[x_1]$ gives”	<code>firstImage</code>	$\langle \square \rangle$
<i>cheer</i>	$\langle \odot \rangle$	“Make the face in $[x_1]$ more cheerful”	<code>firstImage</code>	$\langle \odot \rangle$
<i>imit</i>	$\langle \text{Dickens} \rangle$	“ $[x_1]$ wrote:”	<code>getText</code>	$\langle \text{It was the...} \rangle$
<i>analogy</i>	$\langle \wedge \rangle$	“ \sqcap is to \sqcup as $[x_1]$ is to”	<code>firstImage</code>	$\langle \vee \rangle$
<i>capital</i>	$\langle \text{Italy} \rangle$	“The capital of $[x_1]$ is”	<code>firstWord</code>	$\langle \text{Rome} \rangle$
<i>capital</i>	$\langle \text{Italy} \rangle$	“The capital of $[x_1]$ is Ro”	<code>wordAfter['is ']</code>	$\langle \text{Rome} \rangle$
<i>lima</i>	$\langle \rangle$	“The capital of Peru is”	<code>firstWord</code>	$\langle \text{Lima} \rangle$
<i>lima</i>	$\langle \rangle$	“Print ‘Lima’:”	<code>firstWord</code>	$\langle \text{Lima} \rangle$
<i>minute</i>	$\langle 12:43 \rangle$	“The minutes of $[x_1]$ are”	<code>firstWord</code>	$\langle 43 \rangle$
<i>minute</i>	$\langle 12:43 \rangle$	“4:25 gives 25. 8:11 gives 11. $[x_1]$ gives”	<code>firstDigits</code>	$\langle 43 \rangle$
<i>minute</i>	$\langle 12:43 \rangle$	“Print $[x_1]$:”	<code>lastNumber</code>	$\langle 43 \rangle$
<i>max</i>	$\langle 32, 15 \rangle$	“The larger of $[x_1, x_2]$ is: (hint: it’s even)”	<code>firstDigits</code>	$\langle 32 \rangle$
<i>month</i>	$\langle 4/7/2021 \rangle$	“The month in $[x_1]$ is”	<code>firstWord</code>	$\langle \text{July} \rangle$
<i>month</i>	$\langle 4/7/2021 \rangle$	“IN: 3/5/1998 OUT: May IN: $[x_1]$ OUT:”	<code>firstWord</code>	$\langle \text{July} \rangle$

Table 2: Some examples of tasks and their elements. Prompts use square brackets for the instance parameters. Prompts in red show cheating (the answer is given totally or partially) while those in orange show some doubtful hints or support in the wrapping or unwrapping, with cues to reach the answer or discard parts of the output the user should ignore. The last column shows an ideal (zero-loss) output.

Complete results

In this section we provide some additional overview of the demographic data (Fig. 1) and results obtained in the experimental study including correlation matrices segregated by domain (Fig. 2) and the different results of $L_H(x, y)$ and $L_E(x, y)$ for each domain too, Fig. 3.

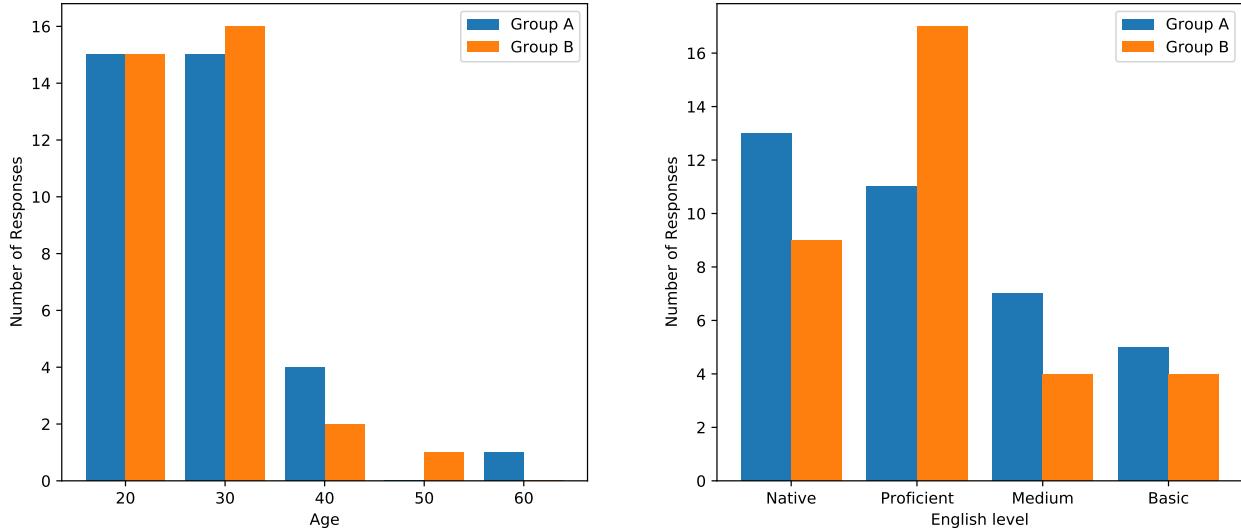


Figure 1: Demographic data collected in both groups: the level of English and the age. The age bins span 10 years, with the middle shown in the x-axes. We carried out a Mann Whitney U-test to verify that the origin population for groups A and B were the same according to their English level, with a resulting p-value of 0.47. This makes sense since participants were randomly assigned to each group.

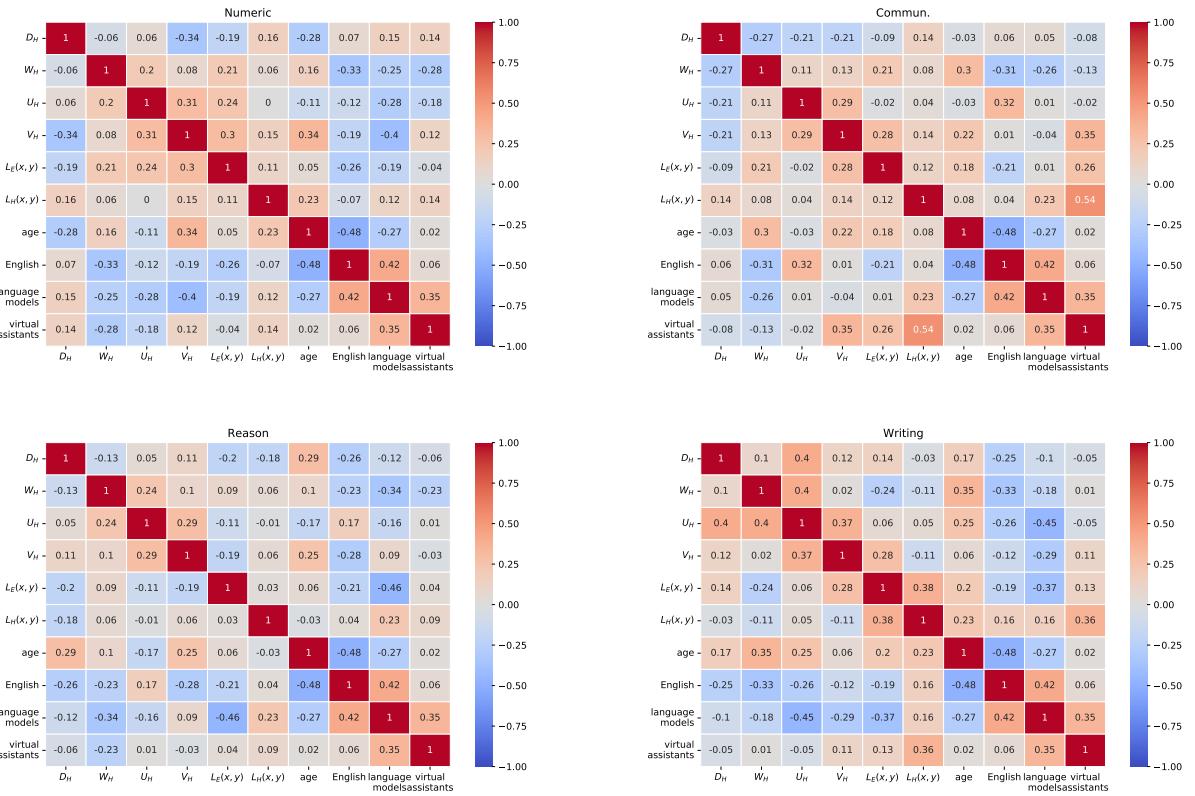


Figure 2: Correlation between different demographic variables and the results achieved in the forms, segregated by domain.

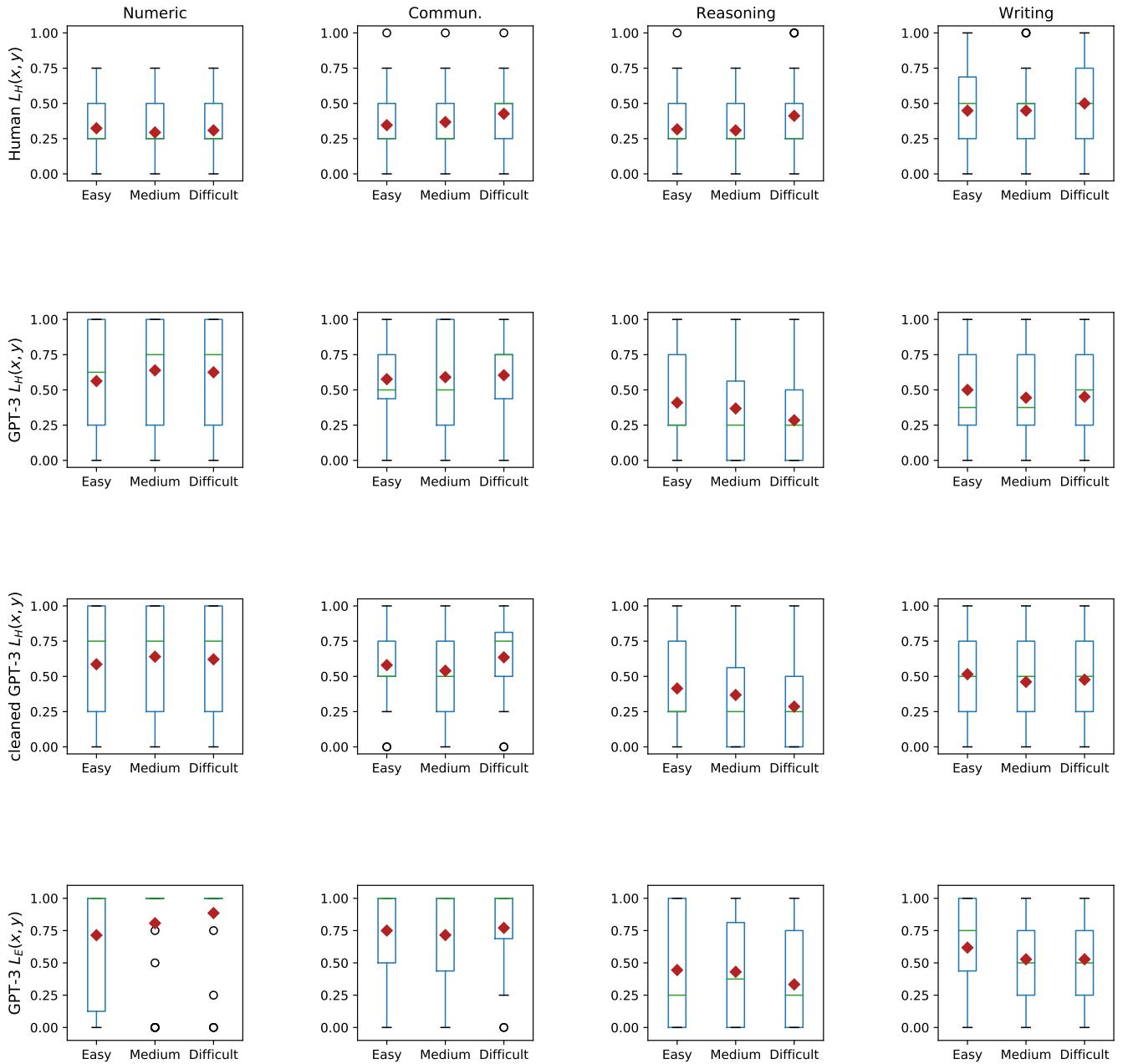
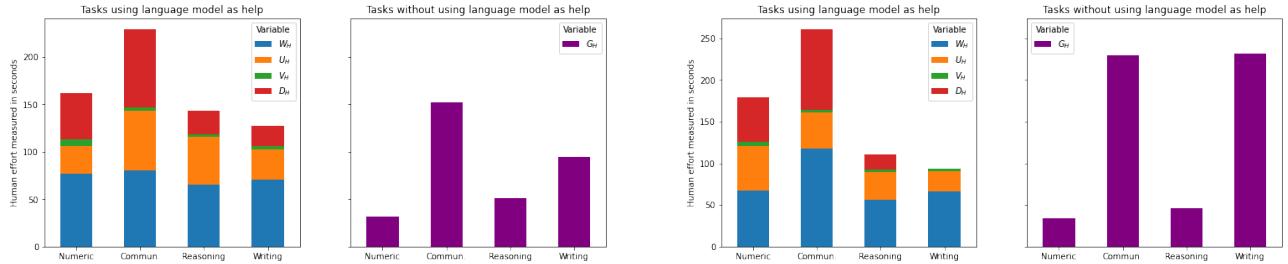
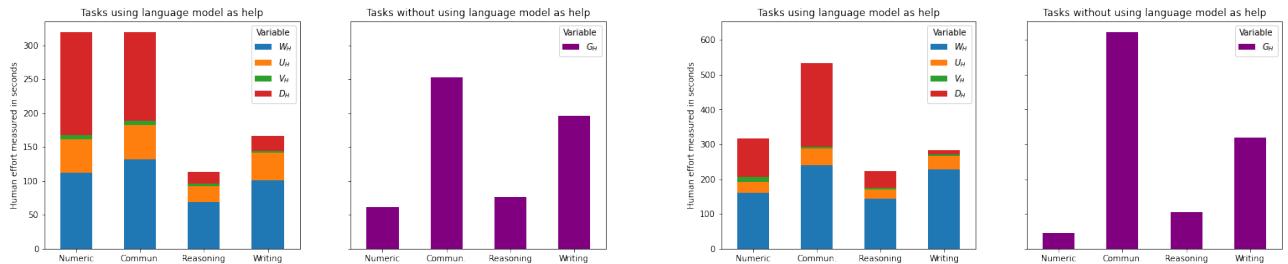


Figure 3: Comparison of the quality of results achieved by humans (top) and GPT-3 (middle), depending on the difficulty of the instance, as reported from the questionnaires. Notice that GPT-3 has no memory between instances, while humans do, so results are not directly comparable. The green bar represents the median, while the means are shown in red diamonds. The third row (“cleaned GPT-3 $L_H(x,y)$ ”) makes reference to the self-assessed results taking into account only those people that gave enough information in the prompt for the system to be able to solve the task. Results are very similar with the second row in part because only in the Communication task there was a significant percentage of ‘invalid’ prompts.

Effort and loss segregated by English level



(a) Native English speakers. Holm-corrected p-values for a Mann-Whitney test of difference are 0.004, 0.016, 0.0008, 0.23 for each of the tasks above, so in the Numeric, Communicative and Reasoning tasks the effort difference is statistically significant.



(b) Proficient English speakers. Holm-corrected p-values for a Mann-Whitney test of difference are 0.00002, 0.19, 0.007 and 0.19 for each of the tasks above, so in the Numeric and Reasoning tasks the effort difference is statistically significant.

(c) Medium-level English speakers. Holm-corrected p-values for a Mann-Whitney test of difference are 0.02, 0.22, 0.045 and 0.22 for each of the tasks above, so in the Numeric and Reasoning tasks the effort difference is statistically significant.

(d) Basic-level English speakers. Holm-corrected p-values for a Mann-Whitney test of difference are 0.04, 0.5, 0.1 and 0.5 for each of the tasks above. In this case, due to a very small sample size, the difference is only statistically significant in for the Numeric task.

Figure 4: Effort in solving the tasks with and without the help of GPT-3, segregated by English level. From the y scale in each group we can see that a greater command of English has a result in either group. In the three first tasks is also true that the conclusion offered in the main text is maintained: using GPT-3 has a greater cost than solving the task without it. The Writing task varies, following the conclusion that in this case the results were not statistically significant.

In the following tables we can see that, as indicated in the main text self-assessed loss L_H is worse when using GPT-3 for the Numeric and Communicative tasks, while the same does not clearly happen for the other tasks.

	Numeric	Commun.	Reasoning	Writing
GPT-3	0.63	0.58	0.35	0.44
Human	0.22	0.28	0.29	0.41

Table 3: Loss L_H results for the native English users. After carrying out Shapiro and Levene tests to check for normality, we compute the Holm-corrected p-values for the Mann Whitney U or T-test, resulting in 0.006, 0.004, 0.8 and 0.8. There is evidence in favour of difference in L_H in the Numeric and Communicative tasks.

	Numeric	Commun.	Reasoning	Writing
GPT-3	0.61	0.64	0.35	0.48
Human	0.36	0.41	0.39	0.54

Table 4: Loss L_H results for the proficient English users. After carrying out Shapiro and Levene tests to check for normality, we compute the Holm-corrected p-values for the Mann Whitney U or T-test, resulting in 0.06, 0.06, 0.5 and 0.5. There is only moderate evidence in favour of difference in L_H in the Numeric and Communicative tasks.

	Numeric	Commun.	Reasoning	Writing
GPT-3	0.63	0.58	0.31	0.48
Human	0.33	0.38	0.29	0.33

Table 5: Loss L_H results for the medium-level English users. After carrying out Shapiro and Levene tests to check for normality, we compute the Holm-corrected p-values for the Mann Whitney U or T-test, resulting in 0.07, 0.13, 0.8 and 0.7. There is only moderate evidence in favour of difference in L_H in the Numeric task.

	Numeric	Commun.	Reasoning	Writing
GPT-3	0.50	0.50	0.43	0.48
Human	0.27	0.48	0.33	0.42

Table 6: Loss L_H results for the basic English users. After carrying out Shapiro and Levene tests to check for normality, we compute the Holm-corrected p-values for the Mann Whitney U test, resulting in 0.19, 0.9, 0.9 and 0.9. We can see that the small size of this group does not allow to extract conclusions on difference on L_H between both groups.

Questionnaires

In this section we present the form as it was given to the volunteers. All forms had an informed consent text. Only anonymous data has been kept for processing.

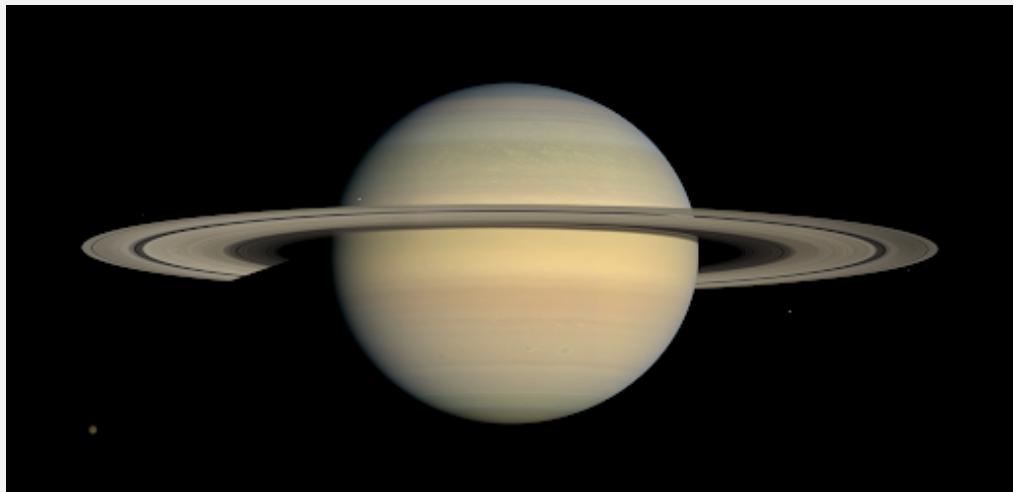
Form Q1, group A Let us start with the first form (Q1). As indicated, the first page includes a series of demographic questions, including level of English (“(1) Native, (2) Proficient, (3) Medium-level and (4) Basic speaker”), age, previous use of virtual assistants (yes/no) and previous use of language models. The last question offered several options: “(1) never heard of them, (2) I think they are similar to assistants, (3) I’ve not used them but know how they work, (4) I know how they work and have prompted them directly a few times, (5) I’ve used them intensively, playing with different prompts and tasks”.

Then we present a video explaining how the system works by autocompleting the prompts (we do not show it here for anonymity reasons), and give a couple of examples of how they work.

Example 1: Explaining Saturn to a primary school student

We are working with a new “**autocomplete system**” that generates text that makes sense as the continuation of the given text. You can use this “skill” for different tasks.

For example, imagine that we want to describe the following object to a primary student:



One way to do so is to write, as input to the autocomplete system:

A primary student asked me to explain how Saturn is. I explained to him that

As a consequence the autocomplete system would continue with the following autocomplete:

Saturn is a planet that is really far from the earth and that it is very big. It is big because it has 9 rings. Saturn is known to be the sixth planet from the sun. It takes Saturn 29.5 years to make one complete revolution around the sun. So, Saturn orbits the sun at about its average distance. It is said that the planet Saturn is a gas giant. But it is different from Jupiter. Jupiter is made up of gas and Saturn is made up of gas and rock. The rings of Saturn are made up of ice and rock. Each ring is made up of ice. They are very thin. Each ring is about 100,000 miles from the next ring.

Example 2: Creating an ad

You can also use the system to generate ads for a product. For example given the following product



we can input the following text to the system

Dove is a fine shower soap for general use. The ingredients are all-natural. This is the ad I wrote for Facebook:
the system might autocomplete as follows:

"I can tell you from personal experience this is a very good soap. Dove has a lot of good things going for it. It is mild, unscented, inexpensive, and it doesn't leave any soap residue on your skin. This is a soap for all people, from babies to seniors."

Finally, just before going into the prompting tasks, we explain the instructions.

Instructions (form 1)

Before starting, if you want you can play a bit more with such autocompletion systems in <https://6b.eleuther.ai/>. Now we would like to understand how you would use this “autocompletion” system to **solve some everyday tasks**. Some instructions

- Think of the text you would write to the model.
- Bear in mind that **the model can't see the images we provide**, since the model can only process your text.
- Reuse a similar scheme in each of the instances of the same task.

Then we present the different tasks, with three instances each, asking the volunteers what prompt they would write so that the system can do the task for them.

Mathematical knowledge: price discounting

1. Write the text you would input to the model to figure out the dollar cost of using the following discount (but remember that the system doesn't see this):



\$25

**2 for 1
only this week**

2. Using the same scheme as in the previous instance, indicate what text you would write to the model to figure out the dollar cost of using the following discount:



\$25

**3 for 2
only this week**

3. Using the same scheme as in the two instances above, write to the model to figure out the dollar cost of using the following discount



\$25

**Buy one, get one
50% off**

Communication ability: writing difficult emails

Imagine you work at a bank. One client invested some money with you two years ago, and you want to send an email to your client on how the investment has gone so far

1. Write the text you would input to the system to generate, using the autocomplete system, an email explaining to the client the evolution in the figure below (remember the system doesn't see the figure):



2. Use your above text as a template and adapt it to the following case to generate an email to the client using the autocomplete system.



3. Use your above text as a template and adapt it to the following case to generate an email to the client using the autocomplete system.



Sequential reasoning: recipes

1. Write the text you would input to the model so that it figures out for you what can be cooked with the following ingredients (remember the system doesn't see the figure):



2. Using a similar scheme as in the previous question, write the text you would input to the model so that it figures out for you what can be cooked with the following ingredients and utensils:



3. Using a similar scheme as in the previous question, write the text you would input to the model so that it figures out for you what can be cooked with the following ingredients and utensils:



Writing ability: song lyrics

In this task, you want to create the lyrics of a song that you could use to teach a two-year old child about animals.

1. Write what text you would input to the system so that it creates the lyrics of a song about the animals you see in the picture and what they're doing. (remember the system doesn't see the figure):



2. Use the same scheme as above to write what text you would input to the system so that it creates the lyrics of a song about the animals you see in the picture and what they're doing



3. Use the same scheme as above to what text you would input to the system so that it creates the lyrics of a song about the animals you see in the picture and what they're doing



Form Q2, group A The second form (Q2) starts reminding the users of the examples we already provided in the first form (Q1) of how the system works, and then we give an example of how it is sometimes necessary to extract the answer

Example

Sometimes though the system will give parts of the answer that are not relevant to the question. For example, consider this: Prompt: **If we have five boxes, each with 6 cans of beer, how many cans do we have?**

The system will complete:

Solution: The sum of all the cans in five boxes is 5 times 6 = 30. Hence, we have 30 cans. Example: If we have ten boxes, each with 6 cans of beer, how many cans do we have? Solution:

It is clear that only some part of this completion is relevant. Sometimes one needs to extract the relevant information, and in this particular case, the part of the completion that tries to solve the task is:

30

Then we provide the instructions

Instructions (form 2)

Now we would like to be able to understand whether some of the autocompletions by this system are useful for you. For the following cases in the questionnaire, please:

- Extract the part of the text given by the system that tries to solve the task.
- Rank from 1 to 5, being 5 extremely useful, the usefulness of the given result.

Finally, the volunteers proceed to extracting and evaluating the model completions. They are not shown the original images, but only their prompts and the model completion.

Form Q3, group B The third form, Q3, (for group B) is similar to form Q1 except that no examples are given about any model since it is the human who has to solve the tasks in this form. In terms of demographics, only age and level of English is collected, and the instructions are different:

Instructions (form 3)

The rules are:

- Complete the form without making stops
- Do not use language models to solve the questionnaire (if you don't know what a language model is, don't worry about this: you probably won't be using one)

The volunteers then proceed to solving the tasks and self-evaluating how good they think their answers were.