

Consent form

Research on the quality of Natural Language Generation systems

Dear participant,

Tilburg University investigates the quality of texts written by computers. We would like to know how good these produced texts are and what the influence is of different rating scales. Therefore, we ask you to read the generated texts and to judge them by answering a couple of questions. The answers you provide may be used for scientific publication. Naturally, if this is the case, your answers will be presented anonymously and it will be impossible to connect your answers to your identity. For your participation in this research, you will receive £3.00 (approximately 30 minutes of participation time). You will get 50 conversations along with generated responses. We ask you to rate these responses.

Thank you for your time and participation!

Kind regards, Anouck Braggaar, Saar Hommes, Peter Blomsma, Emiel Krahmer, Emiel van Miltenburg, Chris van der Lee, Nadine Braun, Frédéric Tomas, Martijn Goudbeek
Tilburg University

Completing this survey will take approximately 30 minutes and is entirely voluntary. This means that you are allowed to stop participation at any moment without having to provide a reason and without any repercussions. No risks are involved in participating. Collection of all data happens in accordance with the new General Data Protection Regulation (GDPR) rules and the Ethical Committee of TSHD has given permission to conduct this research. Data will be processed and saved anonymously. The anonymous data can, when requested, be shared with other researchers.

If you agree to participate, you confirm that:

- you are over 16 years old;
- you agree that your anonymous data will be saved and stored for 10 years;
- you agree that your anonymous data can be used for future research or scientific publications;
- you agree that your anonymous data can be shared with other researchers;
- you understand that you are allowed to stop participation at any moment. Regardless of reason, without having to provide a reason, and without any negative consequences.

If you would like to contact one of the researchers regarding this study, you can send an e-mail to:

A.R.Y.Braggaar@tilburguniversity.edu For comments on—or complaints about—this research it is also possible to contact the

Tilburg University School of Humanities and Digital Sciences Research Ethics and Data Management Committee via
tshd.redc@tilburguniversity.edu.

Thank you for your time and participation!

Kind regards,

Anouck Braggaar, Saar Hommes, Peter Blomsma, Emiel Krahmer, Emiel van Miltenburg, Chris van der Lee, Nadine Braun, Frédéric
Tomas, Martijn Goudbeek
Tilburg University

Do you agree with participating in this study?

Yes

No

Prolific ID

What is your Prolific ID?

Please note that this response should auto-fill with the correct ID

Introduction

Please read the following instructions:

You will be given a short snippet of conversation between two persons containing two sentences. You will be asked to evaluate a third utterance (**response**) on two metrics (readability and coherence).

The response may come from different automated systems. The overall goal of this experiment is to improve chat bots in their ability to converse with humans.

You will be judging the responses on two metrics:

Metric 1: Readability: Is the response easy to understand, fluent and grammatical and does not have any consecutive repeating words.

Metric 2: Coherence: Is the response relevant to the topic of conversation.

Instructions

You will rate the response utterance on metrics of readability and coherence.

Sample conversation 1.

Person A: Today I learned that 4 US states have completely banned billboards.

Person B: Living in Los Angeles, I would vote for this so hard. I'm so sick of billboards.

Note -> 1:Best; 4:Worst *

If this catches on, it would be safe to bet to say that Ia would be the last place to do this if they ever ever consider it

yup .super excited

i have no no opinion on billboard

Rank
Ordered



Note -> 1:Best; 4:Worst *

1 If this catches on, it would be safe to bet to say that Ia would be the last place to do this if they ever ever consider it

2 yup .super excited

3 i have no no opinion on billboard

[Explanation for Readability:](#) The responses are ranked in this order as “Response 1” seems highly readable and is fluent and does not have any repeating words unlike “response 3”. “Response 2” seem readable but it ends abruptly without being completed.

Note -> 1:Best; 4:Worst *

If this catches on, it would be safe to bet to say that Ia would be the last place to do this if they ever ever consider it

yup .super excited

i have no no opinion on billboard

Rank
Ordered



Note -> 1:Best; 4:Worst *

1 If this catches on, it would be safe to bet to say that Ia would be the last place to do this if they ever ever consider it

2 i have no no opinion on billboard

3 yup .super excited

[Explanation for Coherence:](#) The responses are ranked in this order as “Response 1” seems highly coherent with regards to the context of the conversation as they are talking about billboards. Even though response 3 has repeating words, it still makes a comment about the topic of the conversation which is “billboards” and response 2 has no relevance to the what is going in the conversation.

You will start your task from next page

Convo1

Here is the conversation:

Person A: **And at least one from a housekeeper.**

Person B: **I really want to see this Maury episode.**

Your task is to rank the following responses from best to worst on the metrics of **readability and coherence**

Please rank the four responses below from **best to worst on the metric of readability.**

Note -> 1:Best; 4:Worst *

I would pay money to see someone edit together such a thing.

it is a great show .

I don't get it

I want to see this episode

Please rank the four responses below from **best to worst on the metric of coherence.**

Note -> 1:Best; 4:Worst *

I would pay money to see someone edit together such a thing.

it is a great show .

I don't get it

I want to see this episode

Convo2

Here is the conversation:

Person A: **What is you most successful unethical lifehack?**

Person B: **If you have siblings, blame them for everything**

Your task is to rank the following responses from best to worst on the metrics of **readability and coherence**

Please rank the four responses below from **best to worst on the metric of readability.**

Note -> 1:Best; 4:Worst *

I don't know what you are talking about.

Unless they are more successful than you, then have a good relationship and take the blame for everything.

I don't know if you are being sarcastic or not.

what if you don't have kids?

Please rank the four responses below from **best to worst on the metric of coherence.**

Note -> 1:Best; 4:Worst *

I don't know what you are talking about.

Unless they are more successful than you, then have a good relationship and take the blame for everything.

I don't know if you are being sarcastic or not.

what if you don't have kids?

Convo3

Here is the conversation:

Person A: **I am not convinced he 's coming back**

Person B: **not going to happen. that would be crazy if it did though . what would happen?**

Your task is to rank the following responses from best to worst on the metrics of **readability and coherence**

Please rank the four responses below from **best to worst on the metric of readability.**

Note -> 1:Best; 4:Worst *

I am not sure if that is a thing or not

I don't know, but I think he 's going to be a lot worse.

I'm sure he will be fine.

if the president refused to return? the GOP would definitely impeach him and accept pence.

Please rank the four responses below from **best to worst on the metric of coherence.**

Note -> 1:Best; 4:Worst *

I am not sure if that is a thing or not

I don't know, but I think he 's going to be a lot worse.

I'm sure he will be fine.

if the president refused to return? the GOP would definitely impeach him and accept pence.

Convo4

Here is the conversation:

Person A: **what fact are you tired of explaining to people?**

Person B: **no, I am not cheating on my husband, and no, neither of us pressured the other into this. Healthy and consensual nonmonogamy is a thing.**

Your task is to rank the following responses on the metrics of **readability and coherence**

Please rank the four responses below from **best to worst on the metric of readability**.

Note -> 1:Best; 4:Worst *

sounds more like you are cheating on your husband.

I'm sorry to hear that.

I'm sure he will be fine.

I don't think you understand what you are talking about.

Please rank the four responses below from **best to worst on the metric of coherence**.

Note -> 1:Best; 4:Worst *

sounds more like you are cheating on your husband.

I'm sorry to hear that.

I'm sure he will be fine.

I don't think you understand what you are talking about.

Convo5

Here is the conversation:

Person A: **What supporting character from a tv show would you definitely watch in their own spinoff?**

Person B: **the adventures of creed bratton**

Your task is to rank the following responses on the metrics of **readability and coherence**

Please rank the four responses below from **best to worst on the metric of readability**.

Note -> 1:Best; 4:Worst *

eh . creed worked because they made sure not to overuse him . i think his shock value would get pretty old in a show of nothing but him .

I don't get it.

i would love to see him as a villain .

creed bratton is the best actor of all time .

Please rank the four responses below from **best to worst on the metric of coherence**.

Note -> 1:Best; 4:Worst *

eh . creed worked because they made sure not to overuse him . i think his shock value would get pretty old in a show of nothing but him .

I don't get it.

i would love to see him as a villain .

creed bratton is the best actor of all time .

Convo6

Here is the conversation:

Person A: **His tweets are really entertaining tbh. scary and absurd, but entertaining.**

Person B: **I simply can not look away. it just keeps getting weirder.**

Your task is to rank the following responses on the metrics of **readability and coherence**

Please rank the four responses below from **best to worst on the metric of readability**.

Note -> 1:Best; 4:Worst *

I'm not sure if you 're being sarcastic or not.

bizarro America is a strange show.

I don't get it.

I think it is a bit of a stretch.

Please rank the four responses below from **best to worst on the metric of coherence**.

Note -> 1:Best; 4:Worst *

I'm not sure if you 're being sarcastic or not.

bizarro America is a strange show.

I don't get it.

I think it is a bit of a stretch.

Convo7

Here is the conversation:

Person A: **I am stealing this thank you very much.**

Person B: **I made this.**

Your task is to rank the following responses on the metrics of **readability and coherence**

Please rank the four responses below from **best to worst on the metric of readability**.

Note -> 1:Best; 4:Worst *

well then I'm still stealing it, thank you very much.

I third this

i made this .

you made this ?

Please rank the four responses below from **best to worst on the metric of coherence**.

Note -> 1:Best; 4:Worst *

well then I'm still stealing it, thank you very much.

I third this

i made this .

you made this ?

Convo8

Here is the conversation:

Person A: **Reminds me of how there are more plastic flamingos than real ones**

Person B: **Plastic flamingos in the front yard is a sign that swingers live there. so they are more swingers than real flamingos?**

Your task is to rank the following responses on the metrics of **readability and coherence**

Please rank the four responses below from **best to worst on the metric of readability.**

Note -> 1:Best; 4:Worst *

plastic flamingos?

no, it is just a plastic bag.

I have never heard of that. I have never heard of it.

I guess that explains Florida?

Please rank the four responses below from **best to worst on the metric of coherence.**

Note -> 1:Best; 4:Worst *

plastic flamingos?

no, it is just a plastic bag.

I have never heard of that. I have never heard of it.

I guess that explains Florida?

Convo9

Here is the conversation:

Person A: **keep him as Erick and keep McAvoy as Charles but for the love of god give everyone else a do-over.**

Person B: **would love this, but if I had to pick another to differentiate from the fox , Nikolaj Coster - waldau seems like a great choice**

Your task is to rank the following responses on the metrics of **readability and coherence**

Please rank the four responses below from **best to worst on the metric of readability.**

Note -> 1:Best; 4:Worst *

I would love to see him as a villain.

I would love to see him play the role of doom.

Isn't that Jamie from got? I could see him killing that roll

I am not sure what you are trying to say.

Please rank the four responses below from **best to worst on the metric of coherence.**

Note -> 1:Best; 4:Worst *

I would love to see him as a villain.

I would love to see him play the role of doom.

Isn't that Jamie from got? I could see him killing that roll

I am not sure what you are trying to say.

Convo10

Here is the conversation:

Person A: **Yeah he healed his wound with his spray thing, the same one he used to seal the door.**

Person B: **He would have been if doctor strange did not give Thanos the time stone**

Your task is to rank the following responses on the metrics of **readability and coherence**

Please rank the four responses below from **best to worst on the metric of readability.**

Note -> 1:Best; 4:Worst ★

does this happen on Titan?

that's what I thought, but I don't think that's how it works.

He did not have the power to do that.

I think it is time to rewatch the gauntlet.

Please rank the four responses below from **best to worst on the metric of coherence.**

Note -> 1:Best; 4:Worst ★

does this happen on Titan?

that's what I thought, but I don't think that's how it works.

He did not have the power to do that.

I think it is time to rewatch the gauntlet.

Convo11

Here is the conversation:

Person A: **The inevitable moment**

Person B: **Isn't that the guy that lost the British election against the idiot that initiated the Brexit referendum?**

Your task is to rank the following responses on the metrics of **readability and coherence**

Please rank the four responses below from **best to worst on the metric of readability.**

Note -> 1:Best; 4:Worst *

Ed Miliband

No, it was a joke.

no, that 's the guy who voted for Brexit.

no, that was the referendum

Please rank the four responses below from **best to worst on the metric of coherence.**

Note -> 1:Best; 4:Worst *

Ed Miliband

No, it was a joke.

no, that 's the guy who voted for Brexit.

no, that was the referendum

Convo12

Here is the conversation:

Person A: **what do they taste like?**

Person B: **A camel scotch fillet is marbled like good beef, except triple the size. the flavor is very much like beef - slightly stronger but very delicious.**

Your task is to rank the following responses on the metrics of **readability and coherence**

Please rank the four responses below from **best to worst on the metric of readability.**

Note -> 1:Best; 4:Worst *

Camel is now on my "eat" list. I still have not eaten a venison steak

I've never heard of this before. what's it about?

I've never heard of that. what's it taste like?

What is the difference?

Please rank the four responses below from **best to worst on the metric of coherence.**

Note -> 1:Best; 4:Worst *

Camel is now on my "eat" list. I still have not eaten a venison steak

I've never heard of this before. what's it about?

I've never heard of that. what's it taste like?

What is the difference?

Convo13

Here is the conversation:

Person A: **what will become illegal in the next 50 years?**

Person B: **driving a manual car on a public road.**

Your task is to rank the following responses on the metrics of **readability and coherence**

Please rank the four responses below from **best to worst on the metric of readability.**

Note -> 1:Best; 4:Worst *

I don't think that is illegal.

do you mean manual as in "manual transmission" or manual as in "human operated"?

I don't get it.

what if it is already illegal?

Please rank the four responses below from **best to worst on the metric of coherence**.

Note -> 1:Best; 4:Worst *

I don't think that is illegal.

do you mean manual as in "manual transmission" or manual as in "human operated"?

I don't get it.

what if it is already illegal?

Convo14

Here is the conversation:

Person A: **back when he was the "happy one"**

Person B: **mou has changed so much now. it 's unfortunate that he has.**

Your task is to rank the following responses on the metrics of **readability and coherence**

Please rank the four responses below from **best to worst on the metric of readability**.

Note -> 1:Best; 4:Worst *

He has not changed, he has always had the same mood swings over the years.

I'm not sure what you 're trying to say.

yeah, he has been pretty good in the past few years .

I think he is the best player in the league.

Please rank the four responses below from **best to worst on the metric of coherence**.

Note -> 1:Best; 4:Worst *

He has not changed, he has always had the same mood swings over the years.

I'm not sure what you 're trying to say.

yeah, he has been pretty good in the past few years .

I think he is the best player in the league.

Convo15

Here is the conversation:

Person A: **what two videogames would you like to see in a crossover game?**

Person B: **total war star wars**

Your task is to rank the following responses on the metrics of **readability and coherence**

Please rank the four responses below from **best to worst on the metric of readability.**

Note -> 1:Best; 4:Worst ★

I think star wars and company of heroes or dawn of war would be a better fit. the cover system, unit groups, vehicles, etc . plus the delicious physics.

I'm not sure what you mean by that.

that would be awesome.

star wars?

Please rank the four responses below from **best to worst on the metric of coherence.**

Note -> 1:Best; 4:Worst ★

I think star wars and company of heroes or dawn of war would be a better fit. the cover system, unit groups, vehicles, etc . plus the delicious physics.

I'm not sure what you mean by that.

that would be awesome.

star wars?

Convo16

Here is the conversation:

Person A: **wherever you go, there you are.**

Person B: **what is this a quote from?**

Your task is to rank the following responses on the metrics of **readability and coherence**

Please rank the four responses below from **best to worst on the metric of readability.**

Note -> 1:Best; 4:Worst *

I hear it in aa. I don't know the original origin.

it is a reference to something like this :URL

It is a quote from arrested development.

it is a quote from the office.

Please rank the four responses below from **best to worst on the metric of coherence.**

Note -> 1:Best; 4:Worst *

I hear it in aa. I don't know the original origin.

it is a reference to something like this :URL

It is a quote from arrested development.

it is a quote from the office.

Convo17

Here is the conversation:

Person A: **what are some real facts that sound fake at first?**

Person B: **roughly 70 % of the world's oxygen is produced by marine algae.**

Your task is to rank the following responses on the metrics of **readability and coherence**

Please rank the four responses below from **best to worst on the metric of readability.**

Note -> 1:Best; 4:Worst *

26 % of the worlds hot air is produced by politicians.

that is not true at all.

I'm pretty sure that is a myth.

I don't believe you.

Please rank the four responses below from **best to worst on the metric of coherence.**

Note -> 1:Best; 4:Worst *

26 % of the worlds hot air is produced by politicians.

that is not true at all.

I'm pretty sure that is a myth.

I don't believe you.

Convo18

Here is the conversation:

Person A: **what is the most NSFW thing you have seen at an office gathering or party?**

Person B: **A coworker once stood on a swivel chair**

Your task is to rank the following responses on the metrics of **readability and coherence**

Please rank the four responses below from **best to worst on the metric of readability**.

Note -> 1:Best; 4:Worst *

officer buckle and Gloria would not be happy

that is a good one.

how did you find out?

What was on the chair?

Please rank the four responses below from **best to worst on the metric of coherence**.

Note -> 1:Best; 4:Worst *

officer buckle and Gloria would not be happy

that is a good one.

how did you find out?

What was on the chair?

Convo19

Here is the conversation:

Person A: **homemade my sister's specialty, carrot cake**

Person B: **recipe?**

Your task is to rank the following responses on the metrics of **readability and coherence**

Please rank the four responses below from **best to worst on the metric of readability**.

Note -> 1:Best; 4:Worst *

here is the recipe

here you go

I'll give it a try!

remind me ! 2 days

Please rank the four responses below from **best to worst on the metric of coherence.**

Note -> 1:Best; 4:Worst *

here is the recipe

here you go

I'll give it a try!

remind me ! 2 days

Convo20

Here is the conversation:

Person A: **I ran 3 miles today thinking that was 5 km I was so hyped because I did not walk at all then I look at my Nike run achievement and 5 km equal 3.1 miles**

Person B: **this just gave me a weird thought, is there a metric version of the term mileage ? or is that a universal term ? or am I just an idiot?**

Your task is to rank the following responses on the metrics of **readability and coherence**

Please rank the four responses below from **best to worst on the metric of readability.**

Note -> 1:Best; 4:Worst *

this is a good question

that is what I was going to say.

I think it is called "measurement"

it is a unit of measurement.

Please rank the four responses below from **best to worst on the metric of coherence**.

Note -> 1:Best; 4:Worst *

this is a good question

that is what I was going to say.

I think it is called "measurement"

it is a unit of measurement.

Convo21

Here is the conversation:

Person A: **literally the best team possible. I will never forget how good that show is.**

Person B: **I just watched it for the first time a few weeks ago. I wish I could re-experience it over again.**

Your task is to rank the following responses on the metrics of **readability and coherence**

Please rank the four responses below from **best to worst on the metric of readability**.

Note -> 1:Best; 4:Worst *

my husband and I found it on amazon prime. we binge watched it with the self-control of weight watchers meeting next to a bakery.

I have not watched it yet, but I have never seen it. I'll check it out.

it is a great show, but it is a great show.

I watched the first season. it was amazing.

Please rank the four responses below from **best to worst on the metric of coherence**.

Note -> 1:Best; 4:Worst *

my husband and I found it on amazon prime. we binge watched it with the self-control of weight watchers meeting next to a bakery.

I have not watched it yet, but I have never seen it. I'll check it out.

it is a great show, but it is a great show.

I watched the first season. it was amazing.

Convo22

Here is the conversation:

Person A: **what kind of bear is Paddington?**

Person B: **spectacled bear. he does n't really look like one, but he 's from Peru, and that 's the only species of bear that lives there.**

Your task is to rank the following responses on the metrics of **readability and coherence**

Please rank the four responses below from **best to worst on the metric of readability.**

Note -> 1:Best; 4:Worst *

thank you very much for the response.

I don't know what that is, but I'm not sure what you 're trying to say.

that is not a polar bear.

I thought he was a bear?

Please rank the four responses below from **best to worst on the metric of coherence.**

Note -> 1:Best; 4:Worst *

thank you very much for the response.

I don't know what that is, but I'm not sure what you 're trying to say.

that is not a polar bear.

I thought he was a bear?

Convo23

Here is the conversation:

Person A: **are you living in some other timeline? and if so, can you smuggle me over there?**

Person B: **they are talking about the popular vote. in terms of the popular vote, Hillary had about a million more votes than Donald Trump**

Your task is to rank the following responses on the metrics of **readability and coherence**

Please rank the four responses below from **best to worst on the metric of readability.**

Note -> 1:Best; 4:Worst *

over 2 million more you mean?

that is not what I said.

the popular vote does not matter.

that is not how elections work.

Please rank the four responses below from **best to worst on the metric of coherence.**

Note -> 1:Best; 4:Worst *

over 2 million more you mean?

that is not what I said.

the popular vote does not matter.

that is not how elections work.

Convo24

Here is the conversation:

Person A: **let it cool, it 'll be the only metal colored thing there**

Person B: **actually, it will probably be black and covered with soot. it wo n't really stand out that much among all the ash.**

Your task is to rank the following responses on the metrics of **readability and coherence**

Please rank the four responses below from **best to worst on the metric of readability.**

Note -> 1:Best; 4:Worst *

magnet. already posted above this.

that is what I was thinking, but I'm not sure what you mean by that.

that is what I was thinking, but I'm not sure what you mean by that.

yeah, but it 's not black and white.

Please rank the four responses below from **best to worst on the metric of coherence.**

Note -> 1:Best; 4:Worst *

magnet. already posted above this.

that is what I was thinking, but I'm not sure what you mean by that.

that is what I was thinking, but I'm not sure what you mean by that.

yeah, but it 's not black and white.

Convo25

Here is the conversation:

Person A: **tabs or spaces?**

Person B: **the real question is tab - size.**

Your task is to rank the following responses on the metrics of **readability and coherence**

Please rank the four responses below from **best to worst on the metric of readability.**

Note -> 1:Best; 4:Worst *

I think you meant spaces.

I thought it was the same thing.

what is the difference between the two?

yeah, is it common to have something else than 4 spaces?

Please rank the four responses below from **best to worst on the metric of coherence.**

Note -> 1:Best; 4:Worst *

I think you meant spaces.

I thought it was the same thing.

what is the difference between the two?

yeah, is it common to have something else than 4 spaces?

Convo26

Here is the conversation:

Person A: **but it worked for Turris! Jesus, I miss that guy he was supposed to be my sweater this Christmas.**

Person B: **I miss him too but he was n't signing with the team so we had to trade him. Hoffman is under contract . does not make sense to move him.**

Your task is to rank the following responses on the metrics of **readability and coherence**

Please rank the four responses below from **best to worst on the metric of readability**.

Note -> 1:Best; 4:Worst *

what does not make sense is letting Dorion be in charge of a rebuild?

he was the only one who did not get the joke.

I think it was a joke, but I don't think it's worth it.

yeah, but he is still on the bench.

Please rank the four responses below from **best to worst on the metric of coherence**.

Note -> 1:Best; 4:Worst *

what does not make sense is letting Dorion be in charge of a rebuild?

he was the only one who did not get the joke.

I think it was a joke, but I don't think it's worth it.

yeah, but he is still on the bench.

Convo27

Here is the conversation:

Person A: **who are you and what have you done today?**

Person B: **I'm just a lowly security guard, sitting in a guard shack since 6 am. two more hours. two more hours.**

Your task is to rank the following responses on the metrics of **readability and coherence**

Please rank the four responses below from **best to worst on the metric of readability**.

Note -> 1:Best; 4:Worst *

that is a cool job, man . don't put yourself down. you 're not lowly. that is an awesome job to have.

what kind of job?

good luck!

what is your job?

Please rank the four responses below from **best to worst on the metric of coherence**.

Note -> 1:Best; 4:Worst *

that is a cool job, man . don't put yourself down. you 're not lowly. that is an awesome job to have.

what kind of job?

good luck!

what is your job?

Convo28

Here is the conversation:

Person A: **he would be fine in the right setting, North Korea, Saudi Arabia, Alabama.**

Person B: **can we vote to send him somewhere else?**

Your task is to rank the following responses on the metrics of **readability and coherence**

Please rank the four responses below from **best to worst on the metric of readability**.

Note -> 1:Best; 4:Worst *

like Texas?

no, he will be president.

I don't think so.

he is already in office.

Please rank the four responses below from **best to worst on the metric of coherence**.

Note -> 1:Best; 4:Worst *

like Texas?

no, he will be president.

I don't think so.

he is already in office.

Convo29

Here is the conversation:

Person A: **can't run as fast on a bike.**

Person B: **the bike was two - tired.**

Your task is to rank the following responses on the metrics of **readability and coherence**

Please rank the four responses below from **best to worst on the metric of readability**.

Note -> 1:Best; 4:Worst *

I thought it was funny.

that is not how it works.

that is what I was thinking.

finally, all of those bazooka joe comics, and 1980 's joke books paid off ! this was brilliant btw.

Please rank the four responses below from **best to worst on the metric of coherence**.

Note -> 1:Best; 4:Worst *

I thought it was funny.

that is not how it works.

that is what I was thinking.

finally, all of those bazooka joe comics, and 1980 's joke books paid off ! this was brilliant btw.

Convo30

Here is the conversation:

Person A: **a dominos employee stood outside of recently closed papa john in my neighborhood and started selling some pizzas.**

Person B: **wait wtf , since when did the papa john's down in bay ridge close?**

Your task is to rank the following responses on the metrics of **readability and coherence**

Please rank the four responses below from **best to worst on the metric of readability.**

Note -> 1:Best; 4:Worst *

it's been a while since I've seen it.

I don't know, I've never heard of papa johns.

I'm not sure what you 're trying to say.

it is been like a week

Please rank the four responses below from **best to worst on the metric of coherence.**

Note -> 1:Best; 4:Worst *

it's been a while since I've seen it.

I don't know, I've never heard of papa johns.

I'm not sure what you 're trying to say.

it is been like a week

Convo31

Here is the conversation:

Person A: **do it, bro. get stoned and swim with the sharks.**

Person B: **do you think they 'll know I'm high like cats do?**

Your task is to rank the following responses on the metrics of **readability and coherence**

Please rank the four responses below from **best to worst on the metric of readability.**

Note -> 1:Best; 4:Worst ★

ehhh. They might have their suspicions.

don't worry, you'll be fine.

do you have cats?

if you don't like them, you'll be fine.

Please rank the four responses below from **best to worst on the metric of coherence.**

Note -> 1:Best; 4:Worst ★

ehhh. They might have their suspicions.

don't worry, you'll be fine.

do you have cats?

if you don't like them, you'll be fine.

Convo32

Here is the conversation:

Person A: **then he got super hunky towards the end. basically like trying to believe that Channing Tatum could ever possibly play someone who can't get laid.**

Person B: **you got a picture? I can't imagine hunky Urkel.**

Your task is to rank the following responses on the metrics of **readability and coherence**

Please rank the four responses below from **best to worst on the metric of readability.**

Note -> 1:Best; 4:Worst ★

google?

I don't think I've ever seen him in a movie.

here you go

I don't think I've ever seen a picture of him before.

Please rank the four responses below from **best to worst on the metric of coherence.**

Note -> 1:Best; 4:Worst ★

google?

I don't think I've ever seen him in a movie.

here you go

I don't think I've ever seen a picture of him before.

Convo33

Here is the conversation:

Person A: **still my favourite release this year**

Person B: **overkills > these. awesome silhouette though**

Your task is to rank the following responses on the metrics of **readability and coherence**

Please rank the four responses below from **best to worst on the metric of readability**.

Note -> 1:Best; 4:Worst *

honesty i don't agree, the pk looks like a brillo pad

thanks ! i'll have to check it out!

these are my favorite pair of all time.

I'm not sure if this is a typo or not.

Please rank the four responses below from **best to worst on the metric of coherence**.

Note -> 1:Best; 4:Worst *

honesty i don't agree, the pk looks like a brillo pad

thanks ! i'll have to check it out!

these are my favorite pair of all time.

I'm not sure if this is a typo or not.

Convo34

Here is the conversation:

Person A: **first time watching f1!**

Person B: **enjoy the show! unfortunately, Monaco can be a bit of a parade sometimes, I hope it gets interesting so you have a nice first race!**

Your task is to rank the following responses on the metrics of **readability and coherence**

Please rank the four responses below from **best to worst on the metric of readability**.

Note -> 1:Best; 4:Worst *

chucked \$5 on Ricciardo top 3 finish for more fun haha

I have not seen it yet, but I'll check it out. thanks!

thanks! I'll check it out!

thanks!

Please rank the four responses below from **best to worst on the metric of coherence**.

Note -> 1:Best; 4:Worst *

chucked \$5 on Ricciardo top 3 finish for more fun haha

I have not seen it yet, but I'll check it out. thanks!

thanks! I'll check it out!

thanks!

Convo35

Here is the conversation:

Person A: **we should collect all of the trash in the world, turn it into a ball, and fling it all into space.**

Person B: **what if it comes back in 1000 years?**

Your task is to rank the following responses on the metrics of **readability and coherence**

Please rank the four responses below from **best to worst on the metric of readability**.

Note -> 1:Best; 4:Worst *

that's 10 generations away, who cares!

then we can build a wall.

then you would have to go to the moon.

what if the earth is flat?

Please rank the four responses below from **best to worst on the metric of coherence**.

Note -> 1:Best; 4:Worst *

that's 10 generations away, who cares!

then we can build a wall.

then you would have to go to the moon.

what if the earth is flat?

Convo36

Here is the conversation:

Person A: **I could say the same to you man**

Person B: **you must be fun at parties**

Your task is to rank the following responses on the metrics of **readability and coherence**

Please rank the four responses below from **best to worst on the metric of readability**.

Note -> 1:Best; 4:Worst *

"just leave it" continues to reply

I'm not your buddy, pal.

you must be fun at parties

I don't think you're fun at parties

Please rank the four responses below from **best to worst on the metric of coherence**.

Note -> 1:Best; 4:Worst *

"just leave it" continues to reply

I'm not your buddy, pal.

you must be fun at parties

I don't think you're fun at parties

Convo37

Here is the conversation:

Person A: **what picture did you use to get the chassis itself? did you draw it or is there a load of blank formula 1 car photos I don't know about haha**

Person B: **I used a launch photo of the rs 17, removed the sponsors and color balanced the layers**

Your task is to rank the following responses on the metrics of **readability and coherence**

Please rank the four responses below from **best to worst on the metric of readability.**

Note -> 1:Best; 4:Worst *

thanks for the response haha I'm trying to learn how to make concepts like this so thank you!

I don't think I've ever heard of that. I'll have to check it out.

thanks!

I don't think you're fun at parties

Please rank the four responses below from **best to worst on the metric of coherence.**

Note -> 1:Best; 4:Worst *

thanks for the response haha I'm trying to learn how to make concepts like this so thank you!

I don't think I've ever heard of that. I'll have to check it out.

thanks!

I don't think you're fun at parties

Convo38

Here is the conversation:

Person A: **congratulations op! since you already have gone to community college what field are you in / aiming for?**

Person B: **hi there and thank u. I am in the stem field. studying for a degree in it and comp sci.**

Your task is to rank the following responses on the metrics of **readability and coherence**

Please rank the four responses below from **best to worst on the metric of readability.**

Note -> 1:Best; 4:Worst *

nice! when things become tough to remember URL is always rooting for you!

I'm sorry to hear that. I hope you have a great day!

what are you studying?

what field?

Please rank the four responses below from **best to worst on the metric of coherence.**

Note -> 1:Best; 4:Worst *

nice! when things become tough to remember URL is always rooting for you!

I'm sorry to hear that. I hope you have a great day!

what are you studying?

what field?

Convo39

Here is the conversation:

Person A: **DVD released?**

Person B: **not till Oct 17th digital release is tomorrow**

Your task is to rank the following responses on the metrics of **readability and coherence**

Please rank the four responses below from **best to worst on the metric of readability.**

Note -> 1:Best; 4:Worst *

thanks! I'll check it out. thanks!

what is this 'digital release'?

that is not how it works.

I thought it was released?

Please rank the four responses below from **best to worst on the metric of coherence.**

Note -> 1:Best; 4:Worst *

thanks! I'll check it out. thanks!

what is this 'digital release'?

that is not how it works.

I thought it was released?

Convo40

Here is the conversation:

Person A: **today 25th march is Tolkien reading day**

Person B: **unknownst to me, I finished watching the return of the king today at 1 am.**

Your task is to rank the following responses on the metrics of **readability and coherence**

Please rank the four responses below from **best to worst on the metric of readability**.

Note -> 1:Best; 4:Worst *

URL: wasn't aware you were watching it?

I have not read it yet. I'll check it out. thanks!

same here. I loved it.

I've been meaning to watch it for a while, but I have not seen it yet.

Please rank the four responses below from **best to worst on the metric of coherence**.

Note -> 1:Best; 4:Worst *

URL: wasn't aware you were watching it?

I have not read it yet. I'll check it out. thanks!

same here. I loved it.

I've been meaning to watch it for a while, but I have not seen it yet.

Convo41

Here is the conversation:

Person A: **what do you think of bill nye 's new show on Netflix?**

Person B: **it's terrible. I only watched the first episode though, but it was terrible.**

Your task is to rank the following responses on the metrics of **readability and coherence**

Please rank the four responses below from **best to worst on the metric of readability**.

Note -> 1:Best; 4:Worst *

I tried to like it, but it was so cringeworthy at parts like he was trying too hard to appeal to a young audience.

I'm not sure if you 're being sarcastic or not.

I liked it better than the first one.

I thought it was supposed to be the first season.

Please rank the four responses below from **best to worst on the metric of coherence**.

Note -> 1:Best; 4:Worst *

I tried to like it, but it was so cringeworthy at parts like he was trying too hard to appeal to a young audience.

I'm not sure if you 're being sarcastic or not.

I liked it better than the first one.

I thought it was supposed to be the first season.

Convo42

Here is the conversation:

Person A: **I really like the white shoe with a black sole, I think it 's maybe the Gel Lyte iii "yin yang"**

Person B: **that's a pretty dope colorway. I have a pair of the Gel Lyte iii 's myself, they 're incredibly comfortable and super versatile**

Your task is to rank the following responses on the metrics of **readability and coherence**

Please rank the four responses below from **best to worst on the metric of readability**.

Note -> 1:Best; 4:Worst *

awesome dude, what CW do you have?

I think it's a bit of a stretch, but I don't think I've ever seen a pair of shoes in my life.

thanks! I'll look into it!

do you have a link?

Please rank the four responses below from **best to worst on the metric of coherence.**

Note -> 1:Best; 4:Worst ★

awesome dude, what CW do you have?

I think it's a bit of a stretch, but I don't think I've ever seen a pair of shoes in my life.

thanks! I'll look into it!

do you have a link?

Convo43

Here is the conversation:

Person A: **absolutely. he at first agreed to let Russia question former us ambassadors and others. URL:**

Person B: **he would be impeached soon after he handed over those troops.**

Your task is to rank the following responses on the metrics of **readability and coherence**

Please rank the four responses below from **best to worst on the metric of readability.**

Note -> 1:Best; 4:Worst ★

he would not be impeached.

that is not how impeachment works.

that is not what he said.

your faith in republican congress may be misplaced. you did not see the Nunes story?

Please rank the four responses below from **best to worst on the metric of coherence**.

Note -> 1:Best; 4:Worst *

he would not be impeached.

that is not how impeachment works.

that is not what he said.

your faith in republican congress may be misplaced. you did not see the Nunes story?

Convo44

Here is the conversation:

Person A: **humans built a space rocket with stuff we found in the dirt.**

Person B: **humans built almost everything with things we found in the dirt.**

Your task is to rank the following responses on the metrics of **readability and coherence**

Please rank the four responses below from **best to worst on the metric of readability**.

Note -> 1:Best; 4:Worst *

I thought we were talking about the dirt.

we are all humans on this blessed day.

I'm not sure what you are trying to say here.

I made pasta today with stuff found in the water.

Please rank the four responses below from **best to worst on the metric of coherence**.

Note -> 1:Best; 4:Worst *

I thought we were talking about the dirt.

we are all humans on this blessed day.

I'm not sure what you are trying to say here.

I made pasta today with stuff found in the water.

Convo45

Here is the conversation:

Person A: **I don't remember that scene . was that in part 1/2? I just watched it.**

Person B: **it is the opening sequence of episode 3. long surreal section taking place somewhere between the black lodge and non - existence.**

Your task is to rank the following responses on the metrics of **readability and coherence**

Please rank the four responses below from **best to worst on the metric of readability.**

Note -> 1:Best; 4:Worst ★

episode 3 is already out?

ah, I see. thanks for the heads up!

I thought it was the black mirror episode.

ah, I see. thanks.

Please rank the four responses below from **best to worst on the metric of coherence.**

Note -> 1:Best; 4:Worst ★

episode 3 is already out?

ah, I see. thanks for the heads up!

I thought it was the black mirror episode.

ah, I see. thanks.

Convo46

Here is the conversation:

Person A: **what is your favorite band?**

Person B: **led zeppelin, Greenday, Coldplay, one republic**

Your task is to rank the following responses on the metrics of **readability and coherence**

Please rank the four responses below from **best to worst on the metric of readability.**

Note -> 1:Best; 4:Worst *

hello my new friend.

I have never heard of this one. what is it about?

I have never heard of them. i will check them out.

led zeppelin?

Please rank the four responses below from **best to worst on the metric of coherence.**

Note -> 1:Best; 4:Worst *

hello my new friend.

I have never heard of this one. what is it about?

I have never heard of them. i will check them out.

led zeppelin?

Convo47

Here is the conversation:

Person A: **still, she should speak English because it is an English country.**

Person B: **wales is, by definition, the part of England & Wales that is not English.**

Your task is to rank the following responses on the metrics of **readability and coherence**

Please rank the four responses below from **best to worst on the metric of readability.**

Note -> 1:Best; 4:Worst *

Welsh people are certainly English.

no, it is not. it is a British thing.

Wales isn't a country.

it is not a part of the English language, it is the UK.

Please rank the four responses below from **best to worst on the metric of coherence.**

Note -> 1:Best; 4:Worst *

Welsh people are certainly English.

no, it is not. it is a British thing.

Wales isn't a country.

it is not a part of the English language, it is the UK.

Convo48

Here is the conversation:

Person A: **do these come with the white laces as a spare? I see everyone posting with the white lace swap**

Person B: **nope I bought the laces off of rope lace supply**

Your task is to rank the following responses on the metrics of **readability and coherence**

Please rank the four responses below from **best to worst on the metric of readability**.

Note -> 1:Best; 4:Worst *

what laces are you talking about?

where did you get them?

thanks! I'll check it out! thanks!

ok dope thanks. I think the white boosted stripes would look really clean on this color way too.

Please rank the four responses below from **best to worst on the metric of coherence**.

Note -> 1:Best; 4:Worst *

what laces are you talking about?

where did you get them?

thanks! I'll check it out! thanks!

ok dope thanks. I think the white boosted stripes would look really clean on this color way too.

Convo49

Here is the conversation:

Person A: **tons of stuff including Shawshank. the mist. die hard 2. tales from the crypt: demon knight. he 's a great character actor.**

Person B: **oh yeah, that guy. he looks so different as death.**

Your task is to rank the following responses on the metrics of **readability and coherence**

Please rank the four responses below from **best to worst on the metric of readability**.

Note -> 1:Best; 4:Worst *

he also plays the British father at the end of the bogus journey (says "my word!" as the wyld stallyns play)

I don't think I've ever seen that movie.

I think he was the only one who died.

he looks like he is in the movie.

Please rank the four responses below from **best to worst on the metric of coherence**.

Note -> 1:Best; 4:Worst *

he also plays the British father at the end of the bogus journey (says "my word!" as the wyld stallyns play)

I don't think I've ever seen that movie.

I think he was the only one who died.

he looks like he is in the movie.

Convo50

Here is the conversation:

Person A: **lpt: settle your scores irl and outside. parking lots are a good solution; you 'll get more oxygen and this tangibly improves the taste of victory.**

Person B: **from how far up should I drop my opponent?**

Your task is to rank the following responses on the metrics of **readability and coherence**

Please rank the four responses below from **best to worst on the metric of readability**.

Note -> 1:Best; 4:Worst *

depends on where you live.

depends on what you're doing.

I don't get it.

as high up as you feel comfortable. first time droppers can still get noticeable results from as low as 6 inches

Please rank the four responses below from **best to worst on the metric of coherence**.

Note -> 1:Best; 4:Worst *

depends on where you live.

depends on what you're doing.

I don't get it.

as high up as you feel comfortable. first time droppers can still get noticeable results from as low as 6 inches

Demographics

Have you taken part in any previous studies apart from this on evaluating conversation?

Yes

No

Do you have any previous experience of taking part in studies that involve interacting with a chatbot?

Yes

No

Thank you!

Thank you for taking part in the study.

Debriefing

This is the end of this study. Thanks again for participating! This study looked at the quality of texts that were automatically generated by a neural network. Neural networks need to be "trained" (i.e., learn from examples) in order to be able to produce texts and the training examples that we give them can have a big impact on the texts that these text-generating systems produce. In the

experiment you participated in, you saw text generated by three systems and a reference utterance. You were asked to rate these responses. With this, we wanted to investigate if the manner of rating had any influence on the ratings. For example, we compared rating with a Likert scale to best-worst scaling (ranking results from best to worst).

Do you have any questions or comments? Please contact researcher Anouck Braggaar (a.r.y.braggaar@tilburguniversity.edu) For any comments or complaints about this research, you can also contact the Research Ethics and Data Management Committee of Tilburg School of Humanities and Digital Sciences via tshd.redc@tilburguniversity.edu.

Click on proceed to be redirected to Prolific.

Powered by Qualtrics