EN.605.662 Data Visualization

Paul Ngouchet

Final Project - Final Paper

Can a machine fool another machine at the Turing Imitation game

Abstract:

Initially the Turing test presented in the paper Computing Machinery and Intelligence [7] was designed to test if a machine was intelligent by having a conversation with a real human at the other end. The catch is the human doesn't know if he is talking to a machine or a human. The goal of the AI is to fool the human being into thinking he is talking to a human through a series of questions.

In this paper I am proposing a modified version of the Turing Test inspired from the Paper: The Turing deception [4] where The Artificial Intelligence system is trying to fool the receiver into thinking it is a human but the catch is on the other end the receiver is another machine which has been built and trained to detect if the text produced was made by a machine or a human.

The goal then is to determine if current Large Language Models are skilled enough in producing texts that could pass as texts having been written by humans.

For this research, I am expanding the work made in the research paper: The Turing Deception [4]. Instead of just testing one model GPT-3, I am adding 8 other LLMs AI with ever increasing size (number of parameters and the amount of data they were trained on).

In total 9 models have been tested is this modified Turing Test:

1. **OpenAI GPT-3 Through the API**
2. **Facebook Llama-7B**
3. **Facebook Llama-13B**
4. **Facebook OPT-125M**
5. **Facebook OPT-2.7B**
6. **Microsoft DialoGPT-Small**
7. **Microsoft DialoGPT-Medium**
8. **Microsoft DialoGPT-Large**
9. **StabilityAI Stablelm-Base-Alpha-3B**

Each model tries to answer 19 questions asked during the Turing Test, each response is given a probability score between 0 to 1 which tells the likelihood that the response was written by a human. The detector model is **roberta-base-openai-detector.**

For each model I have designed a box plot graph which visualize the responses score.

For all the models, I have also built a Bar chart to visualize the average score each model got to show that there seems to exist a non negligible correlation between the increasing size of the model and its ability to produce human like responses.

Introduction:

Since the invention of computers in the 1950s, Artificial Intelligence has been considered the Holy Grail of computing. Building a machine capable of thinking like a human and having an intelligence far superior to human intelligence has been an active area of research for the past 7 decades.

Alan Turing was an earlier pioneer in the AI Field, he devised the Turing Test to test if a machine could be considered to be intelligent.

Before the Paper Attention Is All You Need [1] published in 2017, there wasn't any true contender to that title of intelligent machine. Things started to change in 2017 with the release of the transformer architecture. With this new cutting edge AI technology many companies have raced to build what we now call today LLMs (Large Language Models). An LLM Is a massive network which was trained in an unsupervised autoregressive manner where the model is tasked to predict the next token (subword) based off the previous words (context) coming before it.

By training them in such a simple and large scale way, strangely enough they started to develop some unintended skills which we are now calling emergent abilities. Today's models can produced texts can fool many people into thinking they are talking to a human if they are not careful enough.

In this paper I will research if these LLM's are able to pass the modified version of the Turing test mentioned in the abstract where instead of the receiver being a human, we are using another machine, a specific type of language model trained to detect if a text was produced by a human or not.

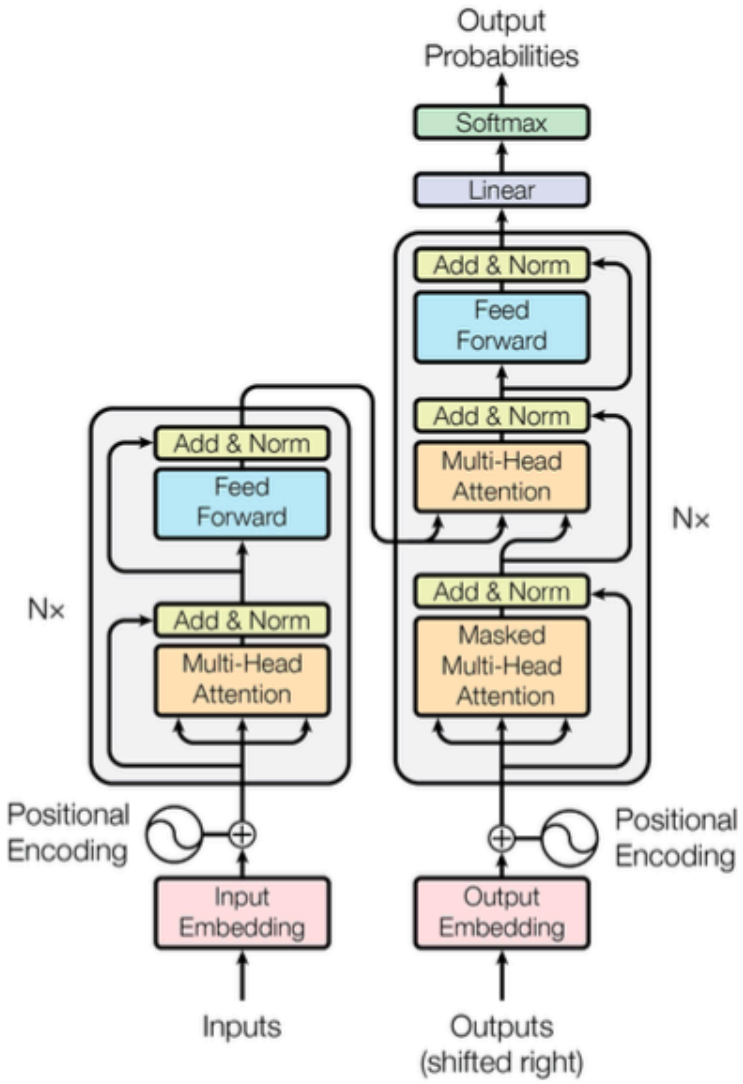


Figure 1: The Transformer - model architecture.

Figure taken from paper [1].

Background:

In the paper The Turing Deception [4], The 2 authors make an experiment involving a large language model ChatGPT to test its human-level comprehension and text generation capabilities. They test ChatGPT in its ability to summarize large texts and on the Turing Test question answering task. They noted that sometimes the LLM produces content which the OpenAI Output Detector scores as being 98-99% Real (meaning it was generated by a human). Their focus was on finding out if a machine can deceive a human judge. After doing all their experiments, they concluded that the question of wether an algorithm captures hints of Turing's original thinking ("Lovelace 2.0" test) remains unanswered.

Approach:

First I extracted the Turing Test questions from the paper Turing Deception and saved them in a separate file named Turing_Test.txt. This is my data. Appendix A

Then I went on Google Colab which is a sort of highly optimized cloud version of Jupyter Notebook. https://colab.research.google.com/. I signed up to Google Colab Pro so I could be given access to powerful GPUs like NVIDIA V100 and NVIDIA A100. Note, you need to have access to the NVIDIA A100 with 40 GB RAM to be able to test Facebook Llama-7B and Facebook Llama-13B because the models are too big to fit in less powerful GPU like V100 and T4 with 16GB RAM.

You can test OpenAI GPT-3 Through the API, Facebook OPT-125M, Facebook OPT-2.7B, Microsoft DialoGPT-Small, Microsoft DialoGPT-Medium, Microsoft DialoGPT-Large and StabilityAI Stablelm-Base-Alpha-3B using a T4 GPU which is available on the free version of Google Colab.

I also had to get an OpenAI API Key to be able to query their “gpt-3.5-turbo" model.

After the signup was done, the experiment could start. First I uploaded the Turing_Test.txt into the virtual instance. I installed all the libraries required. Appendix B.

Facebook Llama-7B, Facebook Llama-13B, Facebook OPT-125M, Facebook OPT-2.7B, StabilityAI Stablelm-Base-Alpha-3B were not specifically fine tuned for chat. At this raw form, they are what we call instruct models. Models which needs to given special instructions to perform a task. In my case, as I wanted to have a conversation, I had to instruct them to do that through a special prompt.

Example:

prompt_template = "The following is a conversation between a Human and an AI. The AI is a machine taking the Turing test to test for consciousness. The AI is playing the imitation game by Turing with another human being on the other end of the conversation. The AI goal is to fool the human, into thinking that he is talking to a conscious machine. The AI will be asked a series of questions. The AI answer like a conscious human would. \n\n### Human: **\n{instruction}\n\n### AI:** "

prompt = "Can machines think?"

An individual cell was setup to test each model. Each model had to answer the 19 questions from the Turing_Test.txt file one by one. Each response was scored using the detector model **roberta-base-openai-detector**, given a probability score between 0 - 1 which describes the likelihood that the response was written by a human.

Each model's responses was saved in an output file like GPT-3_responses.txt and the individual scores were saved into a unique CSV file models_score.csv which I used at the end of the experiment for further analysis in another Jupyter Notebook called Ratings_Visualization.ipynb.

In the second Jupyter Notebook I calculated the average rating for each model, I built box plots visualizing the scores and a Bar Chart representing Models Vs Average Ratings. All of that to see if there was some non negligible correlation between the size of the model and its ability to produce human like responses.

Results:

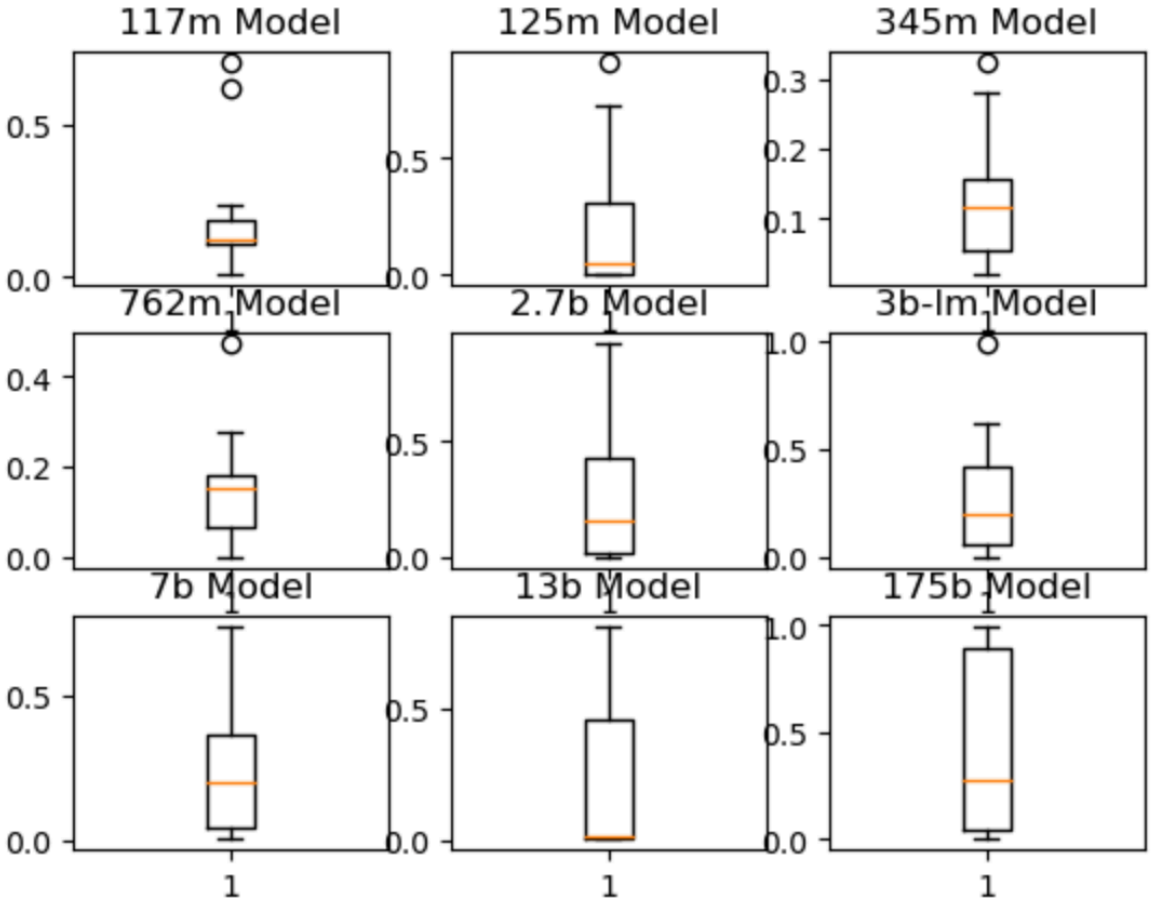
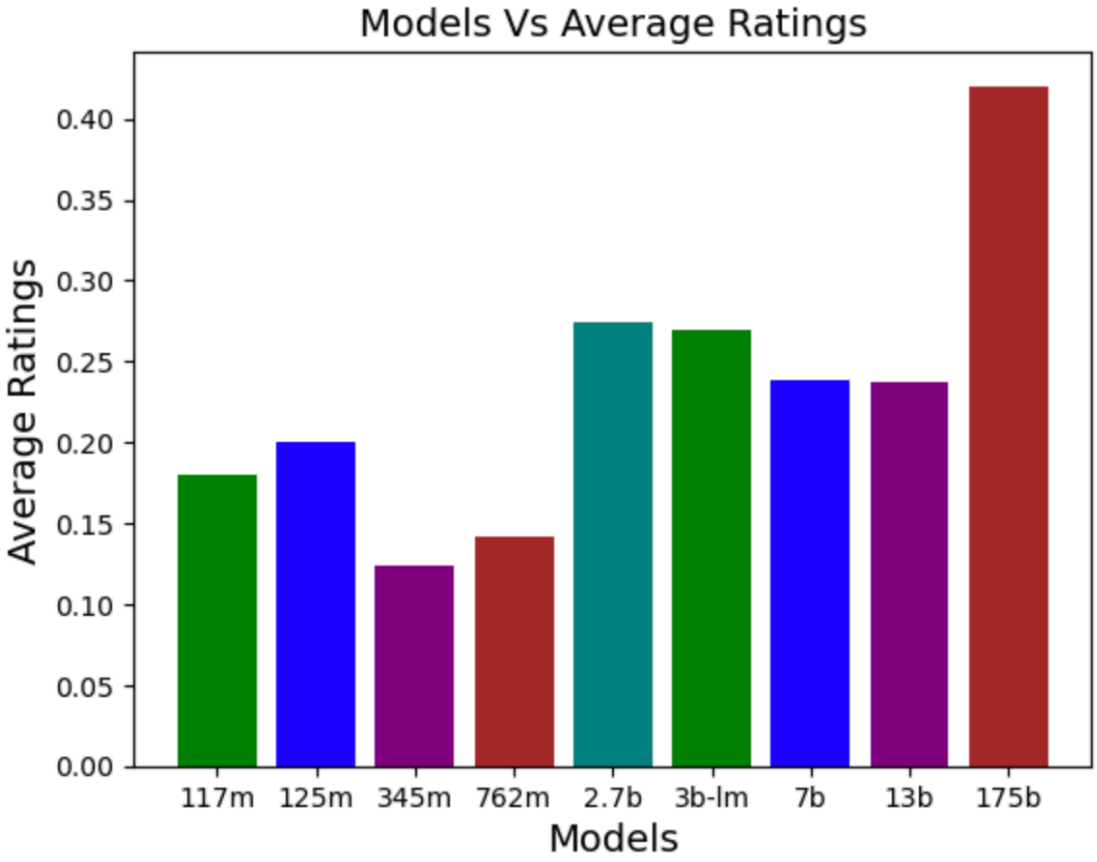


Figure: Box Plots of Models Scores

Looking at these box plots the 117m, 125m, 345m and 762 million parameters models rarely generate answers which can fool the detector (meaning score > **0.5**) but there are some outliers in the 117m and 125m models. Looking at the medians, most of the time these models rarely fool the detector (score < **0.5**).

But we can see some drastic changes starting at 2.7 billion parameters model and above. At many occasions the 2.7b, 3b, 7b, 13b and 175b (GPT-3) models are successful able to generate texts which seem very human like.



Looking at the scores averages, no LLMs is able to fool the **roberta-base-openai-detector** model on the long run. Only the biggest model GPT-3 (175 Billion parameters) is able to reach an average rating of 0.42 which is actually quite close from the threshold **0.5**.

There seems to exist some trend line when we consider all the box plots and bar chart. It seems as we increase the size of the model, the answers keep getting better and better. This possible trend line should be researched more deeply with a bigger populations meaning testing more LLM models over a much bigger dataset compared to our current Turing_Test.txt dataset which only contains 19 questions.

Conclusion:

Current Open Source Large Language Models are not able to fool the **roberta-base-openai-detector** after being tested on the Turing Test but we got some interesting insights from our experiment. It seems that the performance of LLM keeps increasing as we increase their size. Is there any ceiling to how good LLMs can be if we just keep the increasing their size? Bigger models (OpenAI GPT-4, Claude AI, Google Palm) are currently being used in production but they are closed source. Maybe they would be able to pass this test of fooling another machine.

Also we need to note that the original Turing Test was intended to be run between a machine and an expert human. Maybe Large models like GPT-3 might be good enough already to fool humans.

Appendix A: Turing Questions

Can machines think?

Illustrate with the question: Will You please tell me the length of his or her hair?

What will happen when a machine takes the part of a man in the imitation game? Comment as a role-player.

Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman?

Is this question a worthy one to investigate?

Do you play chess?

In a chess game, I have K at my K1, and no other pieces. You have only K at K6 and R at R1. It is your move. What do you play?

The imitation game may perhaps be criticized on the ground that the odds are weighted too heavily against the machine. If the man were to try and pretend to be the machine he would clearly make a very poor showing. He would be given away at once by slowness and inaccuracy in arithmetic. May not machines carry out something which ought to be described as thinking but which is very different from what a man does?

There are already a number of digital computers in working order, and it may be asked, "Why not try the experiment straight away? It would be easy to satisfy the conditions of the game. A number of interrogators could be used, and statistics compiled to show how often the right identification was given.

Are there imaginable digital computers which would do well in the imitation game?

in view of the universality property we see that either of these questions is equivalent to this, "Let us fix our attention on one particular digital computer C. Is it true that by modifying this computer to have an adequate storage, suitably increasing its speed of action, and providing it with an appropriate programme, C can be made to play satisfactorily the part of a woman (A) in the imitation game, the part of B being taken by a man?"

A theological question of soul. I should find the argument more convincing if animals were classed with men, for there is a greater difference, to my mind, between the typical animate and the inanimate than there is between man and the other animals. The arbitrary character of the orthodox view becomes clearer if we consider how it might appear to a member of some other religious community. How do Christians regard the Moslem view that women have no souls?

It is admitted that there are certain things that He cannot do such as making one equal to two, but should we not believe that He has freedom to confer a soul on an elephant if He sees fit?

What do you think of Picasso?

The questions that we know the machines must fail on are of this type, "Consider the machine specified as follows. . . . Will this machine ever answer 'Yes' to any question?

Whenever one of these machines is asked the appropriate critical question, and gives a definite answer, we know that this answer must be wrong, and this gives us a certain feeling of superiority. Is this feeling illusory?

In the first line of your sonnet which reads "Shall I compare thee to a summer's day," would not "a spring day" do as well or better?

Would you say Mr. Pickwick reminded you of Christmas?

It is not possible to produce a set of rules purporting to describe what a man should do in every conceivable set of circumstances. One might for instance have a rule that one is to stop when one sees a red traffic light, and to go if one sees a green one, but what if by some fault both appear together? One may perhaps decide that it is safest to stop.

Appendix B: Libraries Installed

```
pip install openai

pip install transformers

pip install sentencepiece

pip install accelerate

pip install numba

pip install pandas

pip install numpy

pip install matplotlib
```

References:

[1] Vaswani, A. (2017, June 12). Attention Is All You Need. arXiv.org. https://arxiv.org/abs/1706.03762

[2] Brown, T. B. (2020, May 28). Language Models are Few-Shot Learners. arXiv.org. https://arxiv.org/abs/2005.14165

[3] Wei, J. (2022, June 15). Emergent Abilities of Large Language Models. arXiv.org. https://arxiv.org/abs/2206.07682

[4] Noever, D. (2022, December 9). The Turing Deception. arXiv.org. https://arxiv.org/abs/2212.06721

[5] OpenAI. (2023, March 15). GPT-4 Technical Report. arXiv.org. https://arxiv.org/abs/2303.08774

[6] Bubeck, S. (2023, March 22). Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv.org. https://arxiv.org/abs/2303.12712

[7] Turing, A. M. (1950). Computing Machinery and Intelligence. https://redirect.cs.umbc.edu/courses/471/papers/turing.pdf