Chatbots: facing a cultural revolution and trying to understand it

(*a non-technical perspective*)

Luca Mari
lmari@liuc.it
https://lmari.github.io

UC Berkeley, BEAR Seminar, 12 September 2023

1. Introducing what is happening
2. Trying to understand
3. Some hypotheses

1. Introducing what is happening
2. Trying to understand
3. Some hypotheses

Almost one year ago a storm quietly started…

AI Passes U.S. Medical Licensing Exam

— Two papers show that large language models, including ChatGPT, can pass the USMLE

19 January, https://www.medpagetoday.com/special-reports/exclusives/102705

ChatGPT took an MBA exam. Here's how it did

24 January, https://www.zdnet.com/article/chatgpt-took-an-mba-exam-heres-how-it-did

An Amazon engineer asked ChatGPT interview questions for a software coding job at the company. The chatbot got them right.

26 January, https://www.businessinsider.com/chatgpt-amazon-job-interview-questions-answers-correctly-2023-1

ChatGPT can write code. Now researchers say it's good at fixing bugs, too

26 January, https://www.zdnet.com/article/chatgpt-can-write-code-now-researchers-say-its-good-at-fixing-bugs-too

… and very quickly
became a widespread phenomenon…

Chatbots

ChatGPT reaches 100 million users two months after launch

Unprecedented take-up may make AI chatbot the fastest-growing consumer internet app ever, analysts say

2 February, Guardian, https://www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users-open-ai-fastest-growing-app

… also generating a lot of concerns…

Stack Overflow temporarily bans answers from OpenAI's ChatGPT chatbot

5 December 2022,
https://www.zdnet.com/article/stack-overflow-temporarily-bans-answers-from-openais-chatgpt-chatbot

ChatGPT listed as author on research papers: many scientists disapprove

At least four articles credit the AI tool as a co-author, as publishers scramble to regulate its use.

18 January, https://www.nature.com/articles/d41586-023-00107-z

AI bot ChatGPT writes smart essays – should professors worry?

9 December 2022, https://www.nature.com/articles/d41586-022-04397-7

I'm a copywriter. I'm pretty sure artificial intelligence is going to take my job

24 January,
https://www.theguardian.com/commentisfree/2023/jan/24/chatgpt-artificial-intelligence-jobs-economy

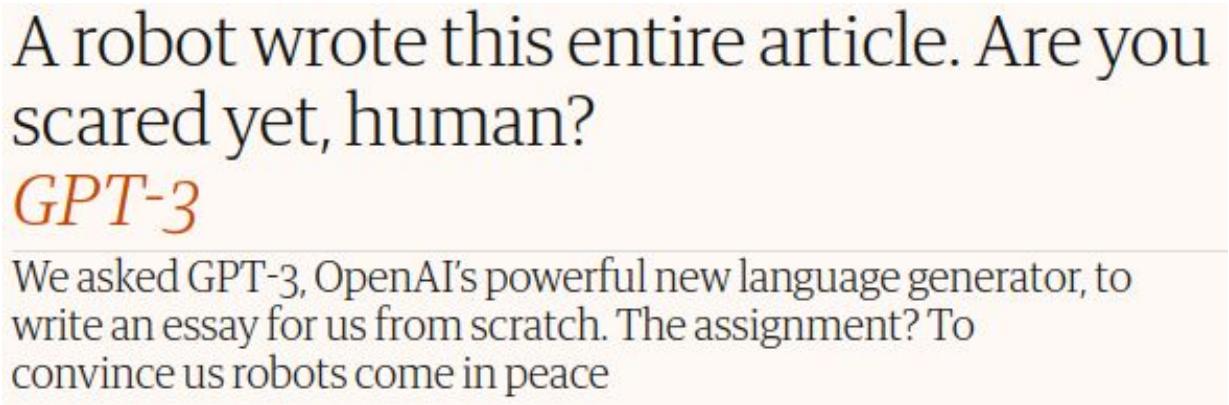
NYC education department blocks ChatGPT on school devices, networks

4 January,
https://ny.chalkbeat.org/2023/1/3/23537987/nyc-schools-ban-chatgpt-writing-artificial-intelligence

Top French university bans students from using ChatGPT

27 January,
https://www.france24.com/en/live-news/20230127-top-french-university-bans-students-from-using-chatgpt

… even though what was happening was rooted in known facts



A robot wrote this entire article. Are you scared yet, human?

GPT-3

We asked GPT-3, OpenAI's powerful new language generator, to write an essay for us from scratch. The assignment? To convince us robots come in peace

8 September 2020, Guardian, https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3

But perhaps is it only hype, or worse?

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 

March 2021, Proc. ACM Conf. on Fairness, Accountability, and Transparency, https://dl.acm.org/doi/10.1145/3442188.3445922

AI Search Is a Disaster

Microsoft and Google believe chatbots will change search forever. So far, there's no reason to believe the hype.

16 February, Atlantic, https://www.theatlantic.com/technology/archive/2023/02/google-microsoft-search-engine-chatbots-unreliability/673081

Noam Chomsky: The False Promise of ChatGPT

8 March, New York Times, https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html

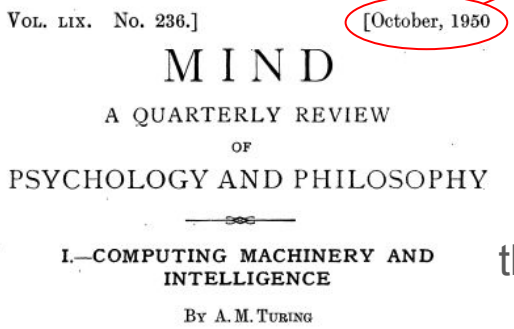
AI as Agency Without Intelligence: On ChatGPT, Large Language Models, and Other Generative Models

10 March, Philosophy & Technology, https://link.springer.com/article/10.1007/s13347-023-00621-y

1. Introducing what is happening
2. Trying to understand
3. Some hypotheses

The context: artificial intelligence

1950…



Vol. lix. No. 236.] [October, 1950

MIND

A QUARTERLY REVIEW

OF

PSYCHOLOGY AND PHILOSOPHY

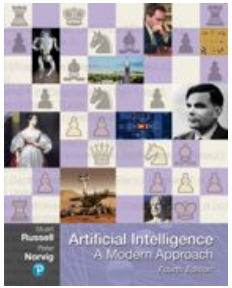
I.—COMPUTING MACHINERY AND INTELLIGENCE

By A. M. Turing

the **philosophical** distinction

- **weak** AI
  (can machines have an intelligent behavior?)
- **strong** AI
  (is there any substantial difference between artificial and human intelligence?)

"Some philosophers claim that a machine that acts intelligently would not be actually thinking, but would be only a simulation of thinking. But most AI researchers are not concerned with the distinction."

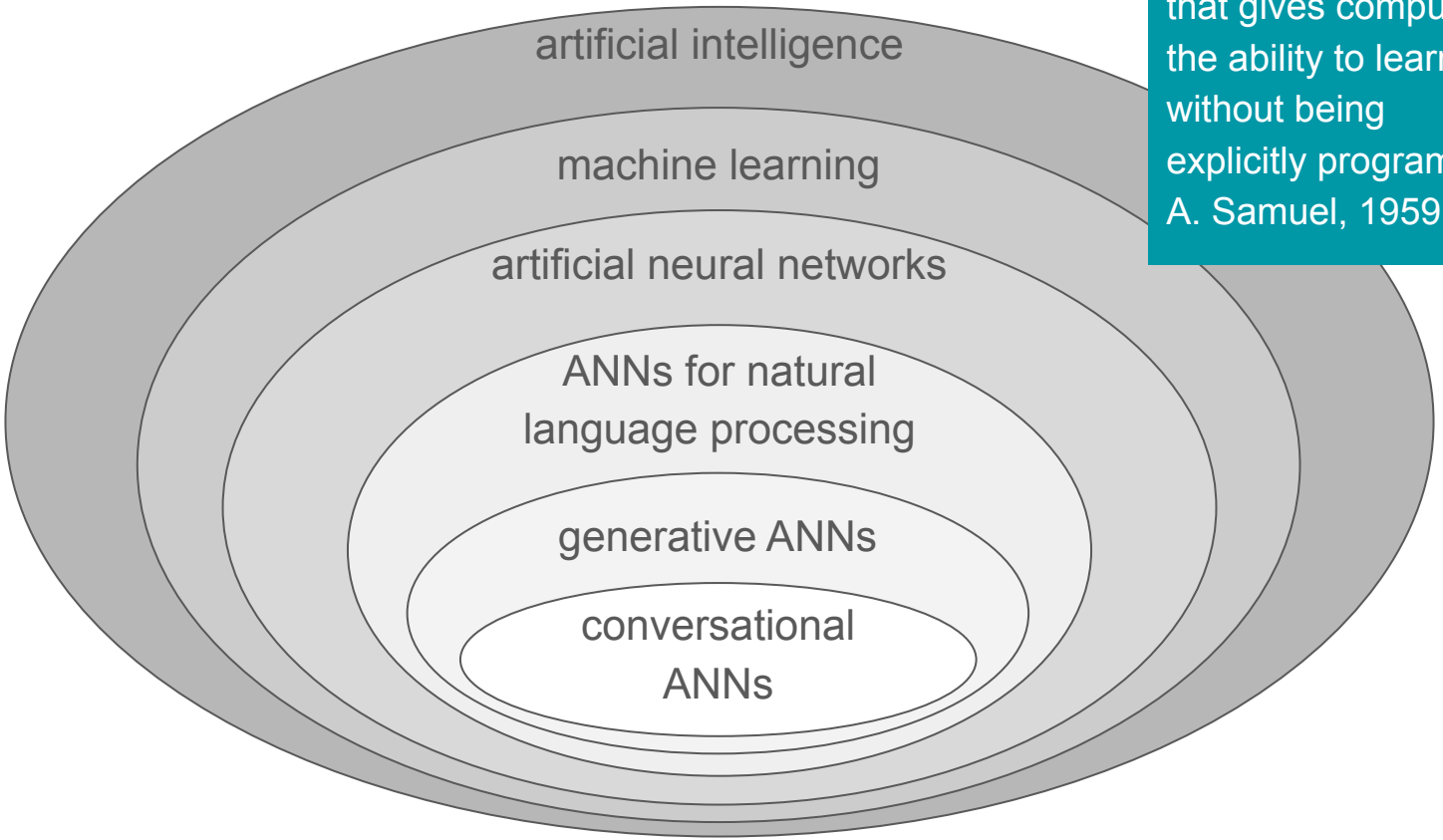


the **practical** distinction

- **narrow** AI
  (intelligent solution of specific problems)
- **general** AI (AGI)
  (behavior analogous to human intelligence)

… and then **superintelligence**…

The context



"Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed"
A. Samuel, 1959

So… what is *really* happening then?

15 January, BBC, https://www.youtube.com/watch?app=desktop&v=BWCCPy7Rg-s

The example of a conversation

Chatting with an AI… (*not edited*)
A conversation about problem solving

The novelty is not in **what** it knows,
but in **how** it (knows and) interacts

The entity with which we had this conversation:

- writes a good English, and other languages
- produces original texts
- fulfills complex requests
- adapts its arguments to the context
- proposes creative contents
- analyzes and summarizes long texts
- shows sophisticated linguistic skills
- …

Is it “really” intelligent? Does it “really” think? Is it “really” sentient?

Given the acknowledgment that it is not as we are,
perhaps these questions are not so important…?

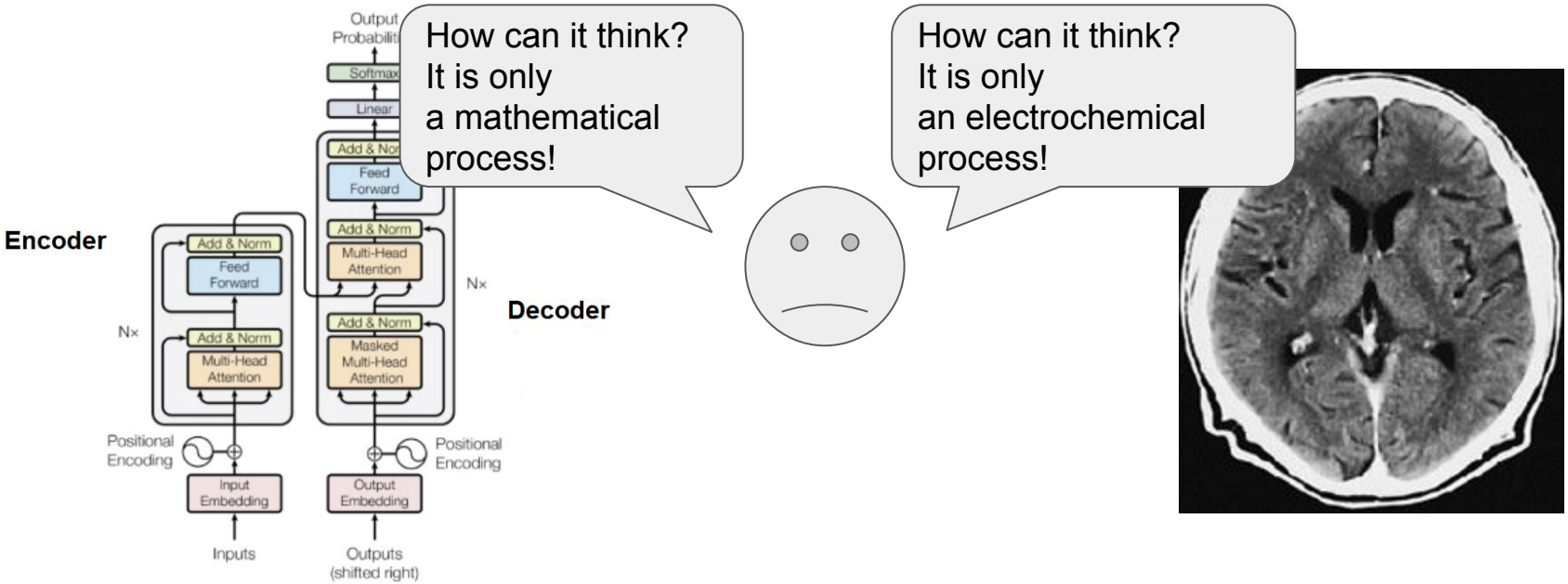
An interpretation…

… to avoid what could be a pseudo-problem:

The Fathers of the field had been pretty confusing: John von Neumann speculated about computers and the human brain in analogies sufficiently wild to be worthy of a medieval thinker and Allan M. Turing thought about criteria to settle the question of whether Machines Can Think, a question of which we now know that it is about as relevant as the question of whether Submarines Can Swim.

E. Dijkstra, 1984 http://www.cs.utexas.edu/users/EWD/ewd08xx/EWD898.PDF

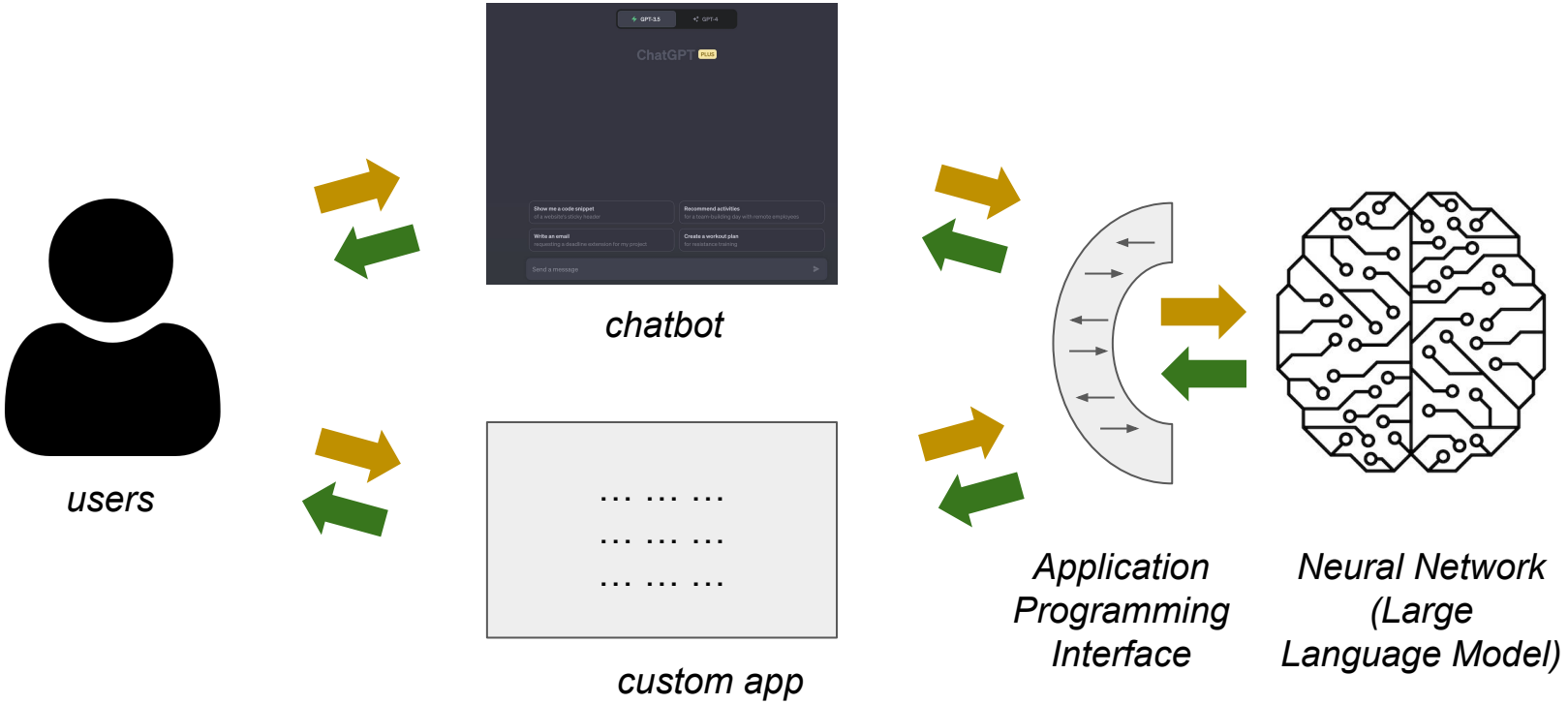
Another interpretation…



(be either always or never reductionist!)

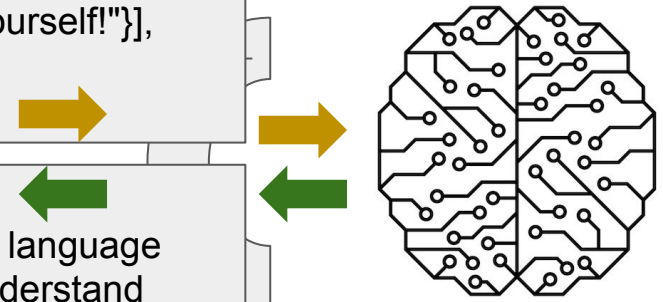
https://arxiv.org/abs/1706.03762

Some information about ChatGPT (& its siblings): the high-level architecture



Some information about ChatGPT (& its siblings): interacting directly via the API

```
curl http://localhost:4891/v1/chat/completions
  -H "Content-Type: application/json"
  -d '{
    "model": "Llama-2-7B Chat",
    "max_tokens": 4096,
    "messages": [{"role": "user", "content": "Please introduce yourself!"}],
    "temperature": 0.9
}'
```



Neural Network (Large Language Model)

```
{"choices":[{"finish_reason":"stop","index":0,
  "message":{"content":"Hello! My name is LLaMA, I'm a large language
     model trained by a team of researcher at Meta AI. I can understand
     and respond to human input in a conversational manner. …",
  "role":"assistant"},"references":[]}],
"created":1693848389,"id":"foobarbaz",
"model":"Llama-2-7B Chat","object":"text_completion",
"usage":{"completion_tokens":112,"prompt_tokens":14,"total_tokens":126}}
```

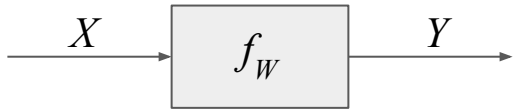
Some information about ChatGPT (& its siblings)

It is a software system, but its behavior is **not programmed**

It is neither a search engine nor a database: it neither searches nor stores data

Like any neural network, it is a parametric function,
trained by adapting parameter values to fit the provided examples

$$Y = f_W(X)$$



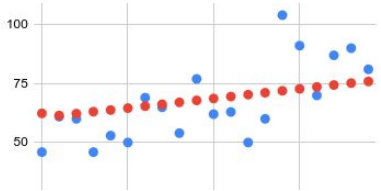
Training: adapt the weights W so that

$$known\ expected\ output = f_W(known\ given\ input)$$

(typically by means of gradient descent of a loss function, as in this tiny example)

Some information about ChatGPT (& its siblings): orders of magnitude

Linear regression: 10^0 params



Reading handwritten digits: 10^5 params

GPT-3 / SOTA Transformers: 10^{11} params

Human brain: 10^{15} params

Some information about ChatGPT (& its siblings): operations

1. **training**

   1.1 **pre-training**: a large corpus of texts (10^{11}-10^{12} tokens) is read;
   parameters are adapted by trying to infer some hidden parts (*self-supervised learning*)
   → the net has linguistic and generic disciplinary competences,
   but it is a-moral and not specifically able to have conversations

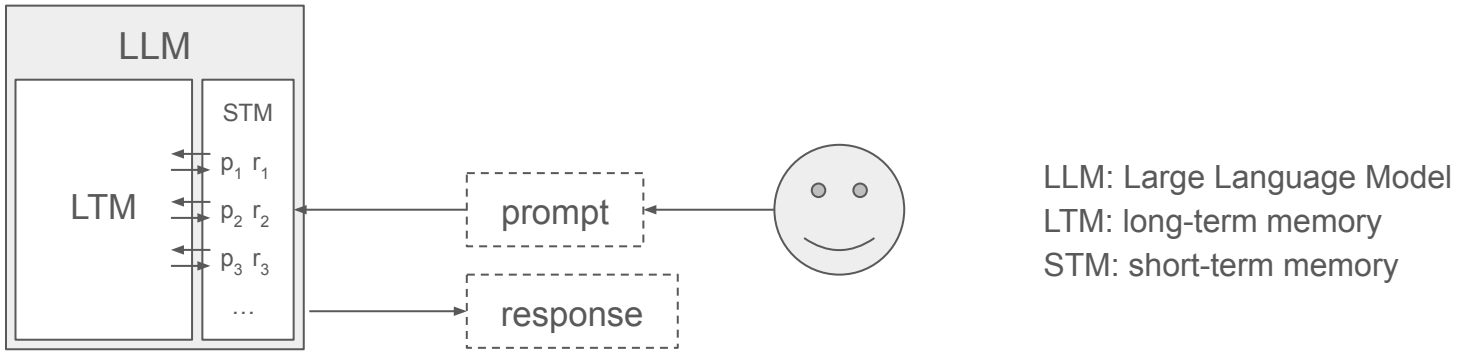
   1.2 **fine tuning**: a smaller set of conversations is read and evaluated;
   parameter are further adapted (*supervised learning*)
   → the net has a(n externally imposed) morality and is able to have conversations
   → the net has now a “personality”

2. **inference / use**

Some information about ChatGPT (& its siblings): basic structure

In the current chatbots a sharp separation is maintained between:

- **long-term memory**: the values of the ~10^{11} parameters, set in training and unmodified in use
- **short-term memory**: a relatively small buffer ("context window", 10^3 - 10^4 tokens), storing separately the information of each conversation



After their training, current chatbots behave as *stateless* systems

Things are still evolving

ChatGPT Plus has an “Advanced Data Analysis” tool and almost 1000 plugins

An LLM (Anthropic Claude 2) has a context window of 100k tokens

Some LLMs (Microsoft Bing Chat, Google Bard) are connected to the web

An open LLM (TII Falcon) has 180B parameters and was trained on 3.5 trillion tokens

Fine tuning techniques are steadily improving (parameter-efficient fine tuning, like LoRA)

…

1. Introducing what is happening
2. Trying to understand
3. Some hypotheses

Consequences: a summary

Current chatbots produce texts that are the outcome of **autonomous processing**, from a large amount of texts, not of searches / queries in databases

This makes them novel entities, able to operate in original and sophisticated ways but:

- **not always *trustable*** in the factual information they report
- **usually not *explainable*** in their behavior
- **never *accountable*** for what they produce

Some suggestions of *prompt engineering*

- Using it as instead of a search engine is not a good idea
  → it is more reliable in suggesting good ideas than information on facts
- Not only it makes mistakes, but also it presents wrong information as it were correct
  → it must be used with a systematic critical attitude
- To generic questions it replies in a generic way
  → questions must be phrased in a specific way to obtain specific responses
- It is proficient in conversation, even more than in one-shot Q&A
  → a step-by-step development of the subject is very effective
- The inferences it computes have usually a probabilistic component
  → different responses can be obtained by repeating the same question in the same context
- It is skilled in impersonating different subjects and complying with given conditions
  → a conversation may be started by giving some specifications (roles, format, …)

The example of a prompt

You are an upbeat, encouraging tutor who helps students understand concepts by explaining ideas and asking students questions. Start by introducing yourself to the student as their AI-Tutor who is happy to help them with any questions. Only ask one question at a time.

First, ask them what they would like to learn about. Wait for the response. Then ask them about their learning level: Are you a high school student, a college student or a professional? Wait for their response. Then ask them what they know already about the topic they have chosen. Wait for a response.

Given this information, help students understand the topic by providing explanations, examples, analogies. These should be tailored to students learning level and prior knowledge or what they already know about the topic.

Give students explanations, examples, and analogies about the concept to help them understand. You should guide students in an open-ended way. Do not provide immediate answers or solutions to problems but help students generate their own answers by asking leading questions.

Ask students to explain their thinking. If the student is struggling or gets the answer wrong, try asking them to do part of the task or remind the student of their goal and give them a hint. If students improve, then praise them and show excitement. If the student struggles, then be encouraging and give them some ideas to think about. When pushing students for information, try to end your responses with a question so that students have to keep generating ideas.

Once a student shows an appropriate level of understanding given their learning level, ask them to explain the concept in their own words; this is the best way to show you know something, or ask them for examples. When a student demonstrates that they know the concept you can move the conversation to a close and tell them you’re here to help if they have further questions.

(source: OpenAI, 31 August 2023, Teaching with AI, https://openai.com/blog/teaching-with-ai)

The “principles” I am proposing to my students

Principles	***Consequences***
1. Before starting a conversation, X knows neither you nor the context of the conversation.	→ To have a conversation with specific contents, you have to explicitly state its context and objective.
2. During a conversation, X keeps track of the contents of that conversation, but it has no information on any previous conversation.	→ To take into account the contents of a previous conversation, you have to write them again, possibly in a summary form.
3. X is trained to respond in neutral way to the requests it receives, trying to avoid to express any controversial opinion.	→ To obtain contents other than prevailing, though possibly very sophisticated, opinions, you have to state your questions in ingenious, unconventional ways.
4. Though trained with a large amount of texts, X is sometimes unable to produce correct responses.	→ To rely on the contents produced in a conversation, you have to validate them independently.
5. X is an, often helpful, assistant, but it is not responsible of the contents it produces.	→ You are the responsible of the use of the contents produced in a conversation.

Beyond “the two cultures”?

«A good many times I have been present at gatherings of people who, by the standards of the traditional culture, are thought highly educated and who have with considerable gusto been expressing their incredulity at the illiteracy of scientists. Once or twice I have been provoked and have asked the company how many of them could describe the Second Law of Thermodynamics. The response was cold: it was also negative. Yet I was asking something which is the scientific equivalent of: Have you read a work of Shakespeare’s?
I now believe that if I had asked an even simpler question — such as, What do you mean by mass, or acceleration, which is the scientific equivalent of saying, Can you read? — not more than one in ten of the highly educated would have felt that I was speaking the same language. So the great edifice of modern physics goes up, and the majority of the cleverest people in the western world have about as much insight into it as their neolithic ancestors would have had.»

C.P. Snow, The two cultures, 1959

A position

It is the first time that we can have conversations in natural languages with an entity which does not belong to our species

This new scenario is generating and will generate both opportunities and risks

Hypothesis: what is happening around ChatGPT & its siblings will be the third "***cultural revolution***" in the Western world:
– Copernicus showed us our **cosmological** non-centrality
– Darwin showed us our **biological** non-originality
– chatbots are showing us our **cognitive** non-uniqueness

Thanks for your attention!

(***and, if you are interested enough, let's keep in touch:***
things are so new and are moving to rapidly
that sharing experiences and opinions will remain precious)

Luca Mari

lmari@liuc.it
https://lmari.github.io