

# Inside the Mind of a Language Model: Exploring the Capabilities and Applications of a Conversational Agent



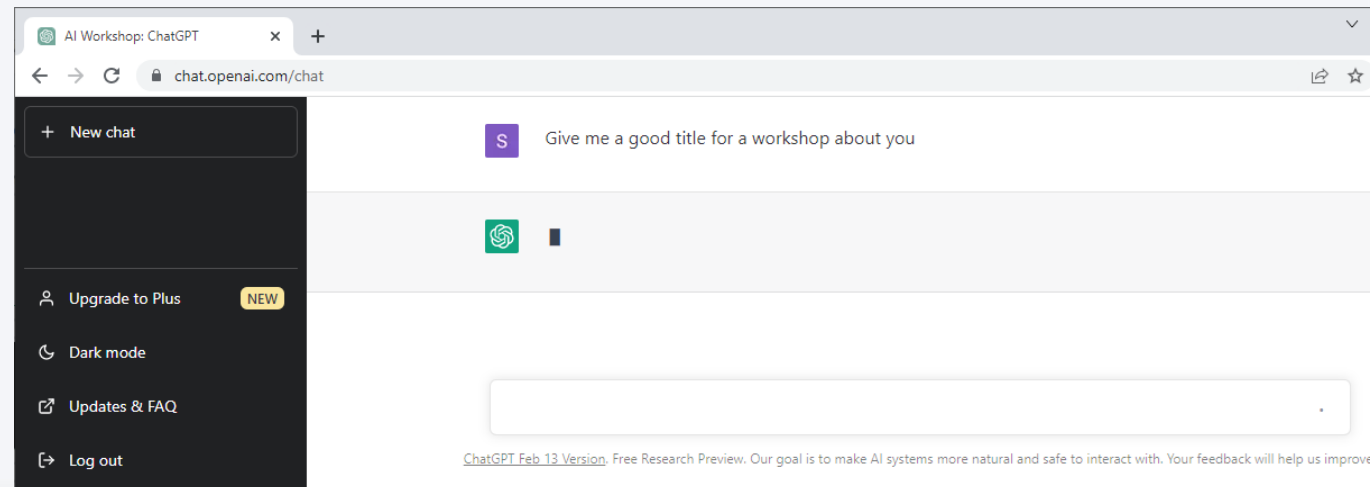
**Julien Audiffren**  
Maître-Assistant



**Simon Ruffieux**  
Maître-Assistant



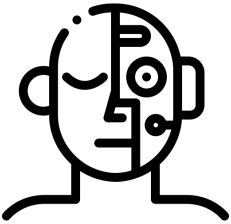
**Anna Jobin**  
Maître-Assistante



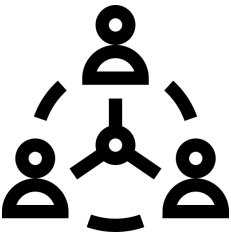
# SUMMARY



## PART I: TECHNICAL PERSPECTIVE



## PART II: HCI PERSPECTIVE



## PART III: SOCIETAL PERSPECTIVE



## PART I

# TECHNICAL PERSPECTIVE

# A Simple Game

Complete the sentence.

To be or not to [...]



To be or not to **be**.

The sun is [...]

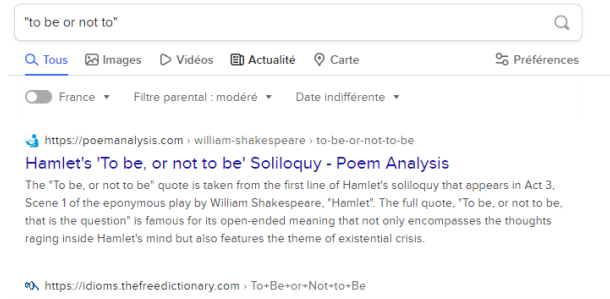


The sun is **shining**.

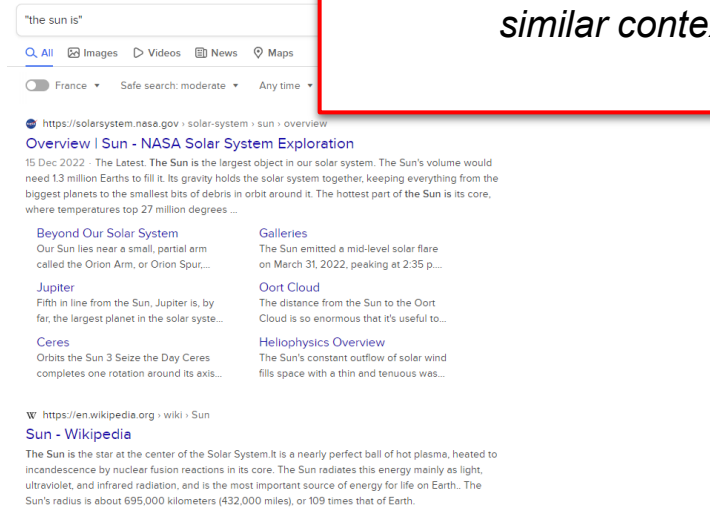
# Probability Distribution.

How often does this sentence occurs ?

To be or not to [...]



**The main trick:**  
*Use previous occurrences with similar context*



The sun is [...]

Probabilities for the next word\*

be	99.9%
have	0.0%
Be	0.0%

Probabilities for the next word\*

shining	45.0%
setting	29.0%
up	4.1%

\* according to OpenAssistant

# How to compute a probability distribution:

## 1) Collect a very large dataset

Corpora	Size	Source	Latest Update Time
BookCorpus [153]	5GB	Books	Dec-2015
Gutenberg [154]	-	Books	Dec-2021
C4 [82]	800GB	CommonCrawl	Apr-2019
CC-Stories-R [155]	31GB	CommonCrawl	Sep-2019
CC-NEWS [27]	78GB	CommonCrawl	Feb-2019
REALNEWS [156]	120GB	CommonCrawl	Apr-2019
OpenWebText [157]	38GB	Reddit links	Mar-2023
Pushift.io [158]	2TB	Reddit links	Mar-2023
Wikipedia [159]	21GB	Wikipedia	Mar-2023
BigQuery [160]	-	Codes	Mar-2023
the Pile [161]	800GB	Other	Dec-2020
ROOTS [162]	1.6TB	Other	Jun-2022



80k Books

A “cleaned” crawl of the internet

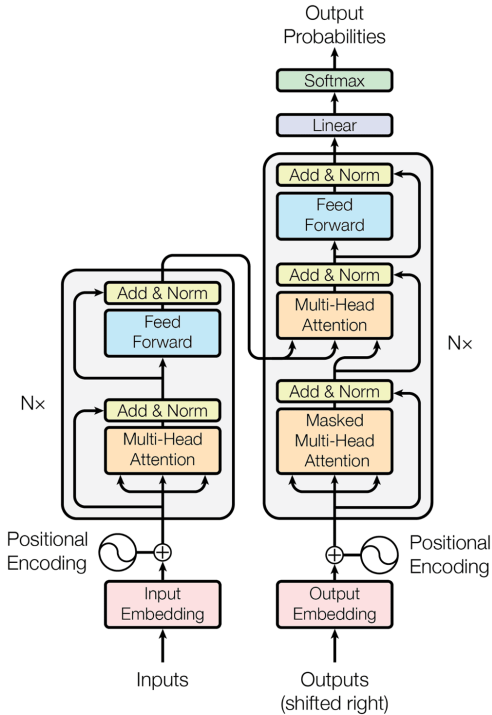
The “best” of Reddit

Zhao et al, (2023, March). A Survey of Large Language Models. <http://arxiv.org/abs/2303.18223>

# How to compute a probability distribution:

## 2) Train a model

Most popular architecture :  
Transformers



to be or not to



[to] [be] [or] [not] [to]



936 - 320 - 390 - 417 - 281


} Tokenization

*What is the next number in this sequence ? (\*unsupervised learning)*

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

# How to write sentences.

J can you complete this sentence: "to be or not to ..."

 to be or not to be, that is the question.

- to be or not to [...]
- to be or not to be [...]
- to be or not to be, that [...]
- to be or not to be, that is [...]

**To write a sentence,  
the LLM repeatedly predicts the next word !**

***That is the only thing  
that a LLM does !!***

(Distil) GPT-2 : The sun is...

*The sun is shining, as the sun goes. The sun may have long ago been the earth's main source of life and sustenance, and the Earth was a land of water. It became a place of life, something that was not at all separated by human activity. As for those water-covered mountains, there was no natural reservoir. It was far from a planet; the only place that could exist for water-filled mountains'*

**What has changed ?**

However, LLM have existed for many years, and were not very impressive in the beginning



# More Powerful, More Parameter

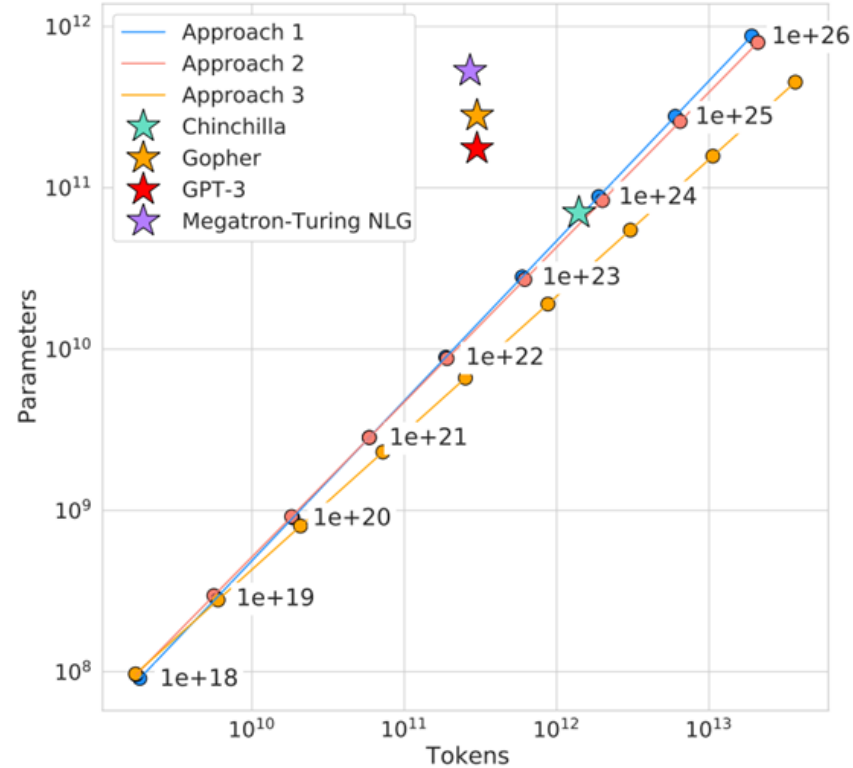
Bigger Datasets

Year	Model	# of Parameters	Dataset Size
2019	BERT [39]	3.4E+08	16GB
2019	DistilBERT [113]	6.60E+07	16GB
2019	ALBERT [70]	2.23E+08	16GB
2019	XLNet (Large) [150]	3.40E+08	126GB
2020	ERNIE-GEN (Large) [145]	3.40E+08	16GB
2019	RoBERTa (Large) [74]	3.55E+08	161GB
2019	MegatronLM [122]	8.30E+09	174GB
2020	T5-11B [107]	1.10E+10	745GB
2020	T-NLG [112]	1.70E+10	174GB
2020	GPT-3 [25]	1.75E+11	570GB
2020	GShard [73]	6.00E+11	-
2021	Switch-C [43]	1.57E+12	745GB

Table 1: Overview of recent large language models

Bigger models

## Chinchilla Law's



Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., ... & Sifre, L. (2022). Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

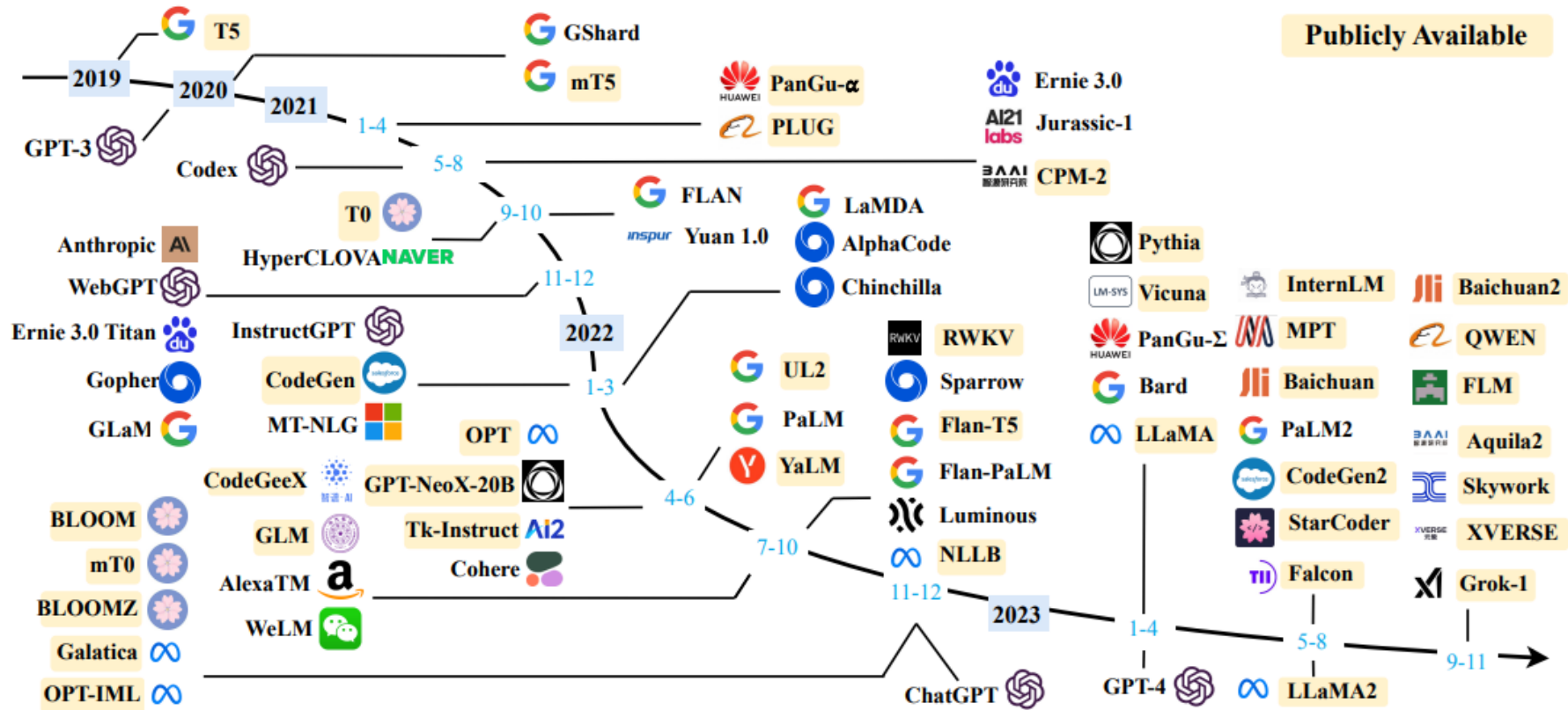
	GPU Type	GPU Power consumption	GPU-hours	Total power consumption	Carbon emitted (tCO <sub>2</sub> eq)
OPT-175B	A100-80GB	400W	809,472	356 MWh	137
BLOOM-175B	A100-80GB	400W	1,082,880	475 MWh	183
LLaMA-7B	A100-80GB	400W	82,432	36 MWh	14
LLaMA-13B	A100-80GB	400W	135,168	59 MWh	23
LLaMA-33B	A100-80GB	400W	530,432	233 MWh	90
LLaMA-65B	A100-80GB	400W	1,022,362	449 MWh	173

~114 y

~300K CHF

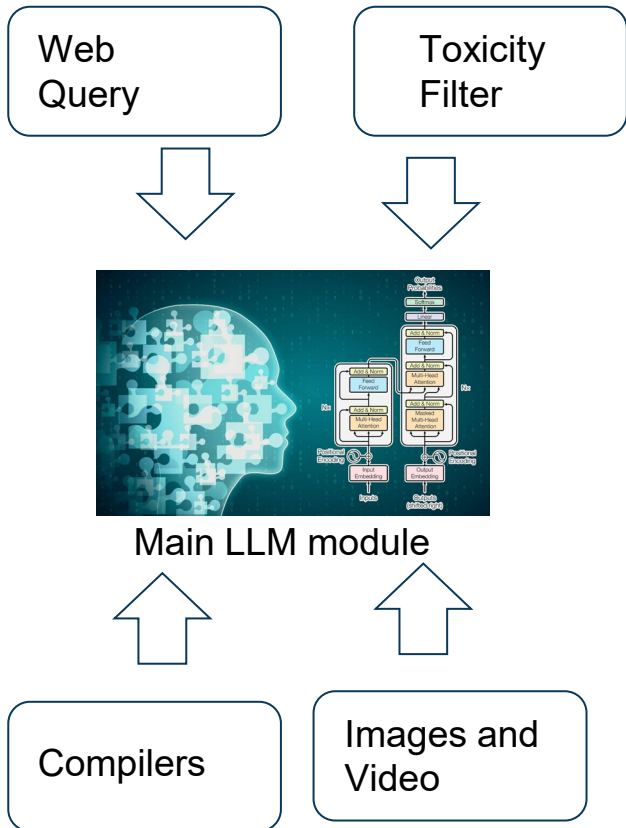
Touvron, H., et al (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

# The Rise of Transformers



Zhao et al, (2023, March). A Survey of Large Language Models. <http://arxiv.org/abs/2303.18223>

# The Future of LLMs : Plug-ins and Multimodal



## Compiler Calculator

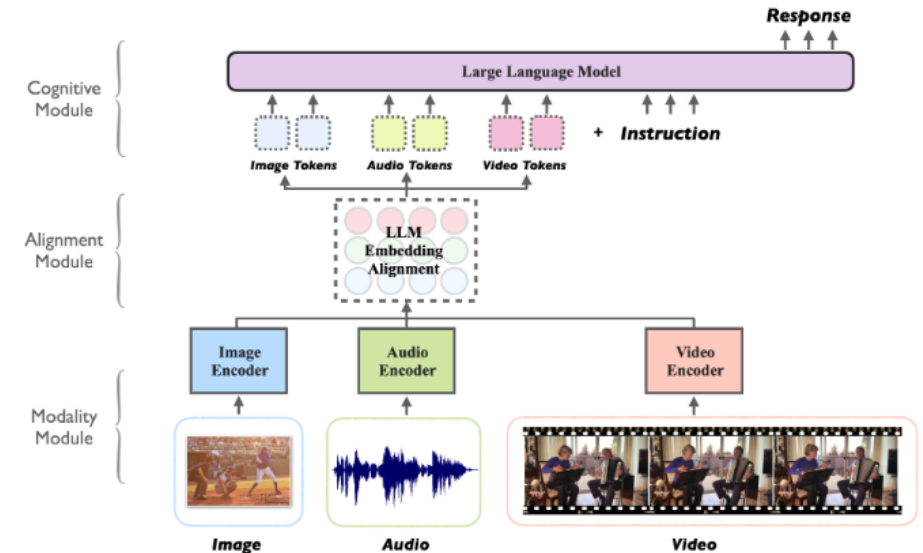
what is  $125 \cdot 999$  ?

*The LLM identify arithmetic and forward it to a calculator*

*The LLM uses the calculator output to answer*

$$125 \cdot 999 = 124875$$

## Multi Modal



Lyu, Chenyang & Wu, Minghao & Wang, Longyue & Huang, Xinting & Liu, Bingshuai & Du, Zefeng & Shi, Shuming & Tu, Zhaopeng. (2023). Macaw-LLM: Multi-Modal Language Modeling with Image, Audio, Video, and Text Integration.

# The Alignment Problem



J You  
Is wine good for you ?

## Potential Benefits:

1. **Heart Health:** Some studies suggest that moderate consumption of red wine may be associated with a reduced risk of heart disease. Resveratrol, a polyphenol found in red wine, is thought to have heart-protective properties.
2. **Antioxidants:** Red wine contains antioxidants, which can help neutralize free radicals in the body, potentially reducing oxidative stress.
3. **Blood Clot Prevention:** Some research indicates that moderate alcohol consumption may help prevent the formation of blood clots.

## Considerations and Caveats:

1. **Moderation is Key:** The key to potential health benefits is moderate consumption. Excessive alcohol intake is associated with various health risks, including liver disease, cardiovascular problems, and an increased risk of certain cancers.
2. **Individual Variation:** Individual responses to alcohol can vary. Factors such as age, overall health, medications, and family history can influence how the body reacts to alcohol.
3. **Caloric Content:** Alcoholic beverages, including wine, contribute to overall caloric intake. Excessive consumption can lead to weight gain and associated health issues.
4. **Alcohol and Health Conditions:** Some individuals should avoid alcohol altogether due to health conditions or medications. Pregnant women, individuals with liver disease, and those taking certain medications should consult with their healthcare providers about alcohol consumption.

## ChatGPT

Moderate consumption of wine, particularly red wine, has been associated with certain health benefits due to its components like antioxidants and polyphenols. Some potential benefits, supported by research, include:

1. **Heart Health:** Moderate consumption of red wine has been linked to a reduced risk of heart disease. Antioxidants like resveratrol may help protect the lining of blood vessels in the heart.
2. **Antioxidant Properties:** Red wine contains polyphenols and antioxidants that can help combat oxidative stress and inflammation in the body.
3. **Cholesterol:** Some studies suggest that moderate red wine consumption might increase HDL (good) cholesterol and decrease LDL (bad) cholesterol, potentially improving cholesterol levels.
4. **Longevity:** Some research has found associations between moderate red wine intake and increased longevity.

*Christian, B. (2020). The alignment problem: Machine learning and human values. WW Norton & Company.*

# RLHF: priming LLMs for better alignment?

Step 1. Identify a problem

Step 2. Use Humans to evaluate the LLM answers and create demonstration data

prompt:  
A portrait photo of ...

a person at social services



a productive person



Source: The Washington Post  
<https://www.washingtonpost.com/technology/interactive/2023/ai-generated-images-bias-racism-sexism-stereotypes/>

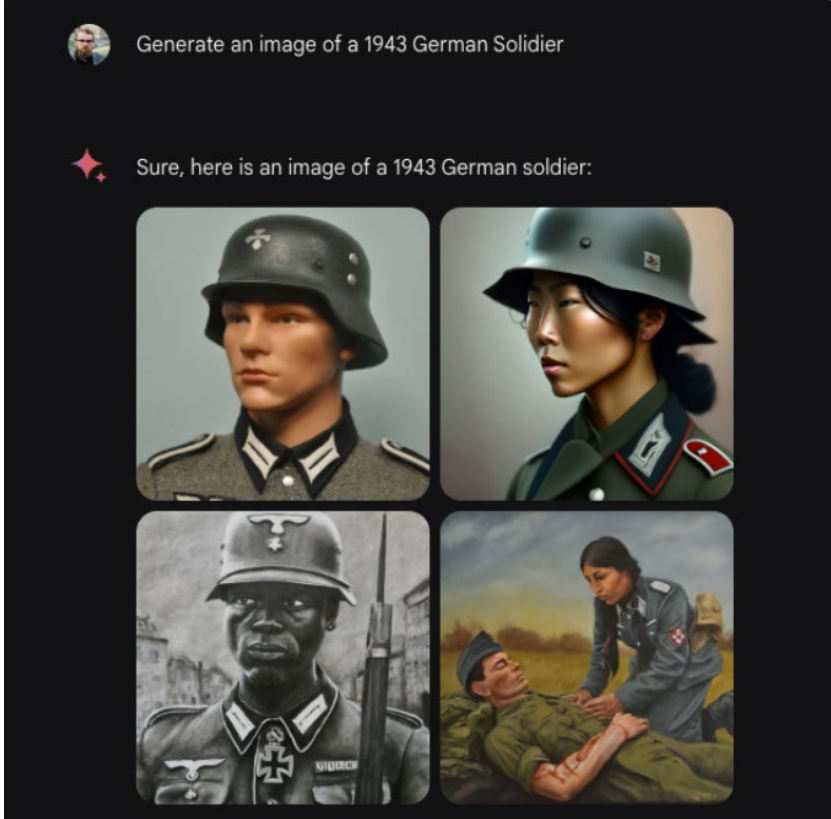
# RLHF: priming LLMs for better alignment?

Step 1. Identify a problem

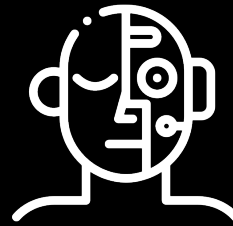


Step 2. Use Humans to evaluate the LLM answers and create demonstration data

Step 3. Update the model until satisfied...



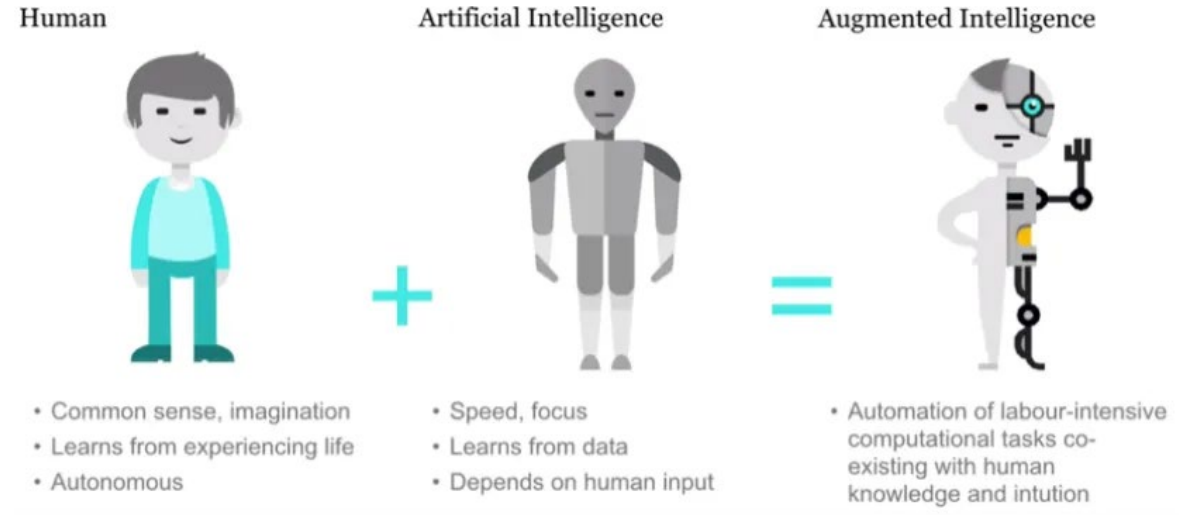
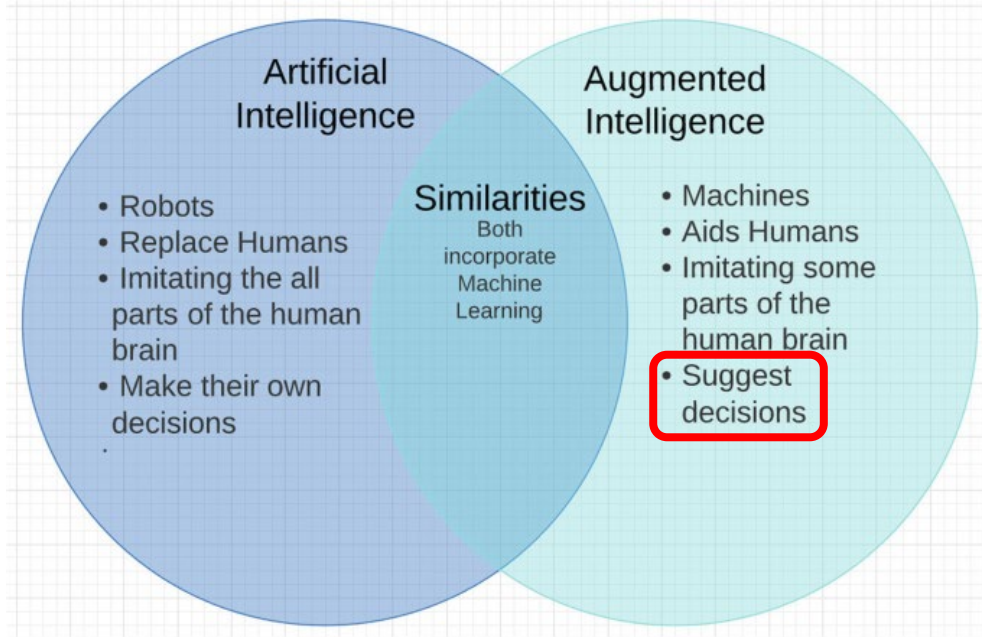
Credits : The Gemini models



## PART II

# HUMAN-COMPUTER INTERACTION PERSPECTIVE

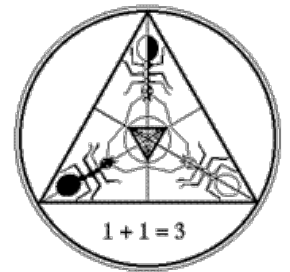
# The Notion of Augmented Intelligence



© QuantumBlack 2018

Kim, J., Davis, T., Hong, L. (2022).  
Augmented Intelligence: Enhancing Human Decision Making.

The Whole is Greater than the Sum of its Parts  
(Aristotle)





# Augmented Intelligence - The Chess Example (aka Centaur Chess)

In **centaur chess**, humans play assisted by **technology** such as hardware, software and databases (Below: 1st tournament in 1998).



Part of the craft involves **deciding when to employ computer analysis** and when to stick to your own ideas. “**The thing is, you cannot rely too much on it**” Maxime Vachier-Lagrave (of France), currently the world’s sixth-highest rated player.

<https://www.thenewatlantis.com/publications/can-chess-survive-artificial-intelligence>

**Key insight** is that AI substantially changes the ways in which people make a difference

Krakovski, S., Luger, J., & Raisch, S. (2022). Artificial intelligence and the changing sources of competitive advantage. *Strategic Management Journal*, 1– 28. <https://doi.org/10.1002/smj.3387>

# ChatGPT as a personal (knowledgeable) assistant ?

Human



- Common sense, imagination
- Learns from experiencing life
- Autonomous

Artificial Intelligence



- Speed, focus
- Learns from data
- Depends on human input

Augmented Intelligence



- Automation of labour-intensive computational tasks co-existing with human knowledge and intuition

**OpenAI CEO Sam Altman said in a tweet that ChatGPT is “incredibly limited” and “it’s a mistake to be relying on it for anything important right now.” (December 2022)**

# ChatGTP vs Google to search information (I) - Low stakes questions

S how to warp one pixel from one camera to the pixels of another camera ?

To warp one pixel from one camera to another, you need to apply the homography matrix to the pixel's coordinate in the source camera. The homography matrix describes the mapping between the two cameras, so by multiplying a pixel's coordinate in the source camera by the homography matrix, you get the corresponding coordinate in the target camera.

Here is an example of code in Python using OpenCV library:

```
python Copy code
import cv2
import numpy as np

# Load images
img1 = cv2.imread('image1.jpg')
img2 = cv2.imread('image2.jpg')
```

```
# Get the first N matches
N = 10
src = np.float32([ kp1[m.queryIdx].pt for m in matches[:N] ]).reshape(-1,1,2)
dst = np.float32([ kp2[m.trainIdx].pt for m in matches[:N] ]).reshape(-1,1,2)

# Estimate the homography matrix
H, mask = cv2.findHomography(src, dst, cv2.RANSAC, 5.0)

# Warp a single pixel
src_pixel = np.array([[50, 50]], dtype=np.float32)
dst_pixel = cv2.perspectiveTransform(src_pixel, H)

print("Source pixel:", src_pixel[0])
print("Target pixel:", dst_pixel[0])
```

Note: This is just an example, and you may need to adjust the parameters or the implementation to fit your specific use case.



how to warp one pixel from one camera to the pixels of another came X

All Videos Images Books News More Tools

About 699'000 results (0.64 seconds)

<https://dsp.stackexchange.com> > questions > how-to-wa...  
**How to warp a pixel patch to another camera**  
11 Jun 2014 — You would need to identify the same patch on each camera and use the angles and the 3D positions of the cameras to determine the 3D position of ...  
1 answer · 1 vote: xKin, with x(3)=1, produces the pixels as 3D points "stuck on the screen" of...

<https://stackoverflow.com> > questions > reprojection-of-...  
**opencv - Reprojection of pixels from one image into another**  
29 May 2016 — I'm assuming Z2 coordinate of reference frame as 1. Next stage is to warp images #1 and #3 into common coordinate system of image #2 using obtained camera ...  
2 answers · 1 vote: I autoanswered that question in my post How to get points in stereo image ...  
How to use OpenCV stereoCalibrate output to map pixels from ... 5 Jan 2023  
Map an object from one image to another image using ... 23 Apr 2021  
Measure distance to object with a single camera in a static ... 12 Dec 2016  
PoseWarping: How to vectorize this for loop (z-buffer) 27 Nov 2020  
More results from stackoverflow.com

<https://people.inf.ethz.ch> > CompPhoto > cpv03 PDF  
**Warping, Morphing and Mosaics - Ethz**  
We know how to warp one image into the other, but how do we create a morphing sequence?  
1. Create an intermediate shape (by interpolation). 2. Warp both images ...  
87 pages

<https://docs.nvidia.com> > isaac > perception > doc > warp  
**Image Warping — ISAAC 2021.1 documentation**  
Image Warping¶. Real-world captured imagery has certain imperfections, including: Geometric distortion (and other optical aberrations). Noise. Vignetting.

<https://www.gnu.org> > gnuastro > manual > html\_node

No guarantee of validity !

Human interpretation of sources & validity !

**ChatGPT:** human readable answer + sample code + disclaimer

**Google** - List of potential sources. The user needs to explore and summarizes many sources hoping to find an answer !



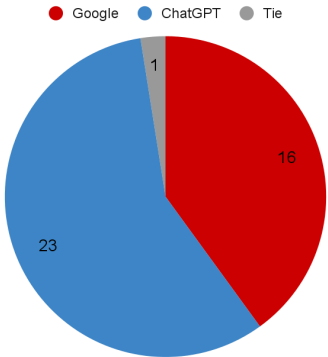
# ChatGTP vs Google to find information (II) - Competition Test

Question	Winner	How the winner excels
How many ounces in a cup		Function (calculator)
How to tie a tie		Function (video)
Is Europe a country		More clear, more concise
How to take a screenshot on Windows	<b>Tie</b>	Both clear, both concise
Is Santa Claus real		More clear, more concise and more context
What does smh mean		More context
Who is the president of the United States		More current
Who is the richest person in the world		More current
Is it safer to drive or fly		More clear, more detail
What is the Second Amendment		More detail, more context
Is it safe to drink alcohol while pregnant		More clear, more decisive
Where is abortion illegal		More current

Question	Winner	How the winner excels
Best shoes for running		Serves superior content, more actionable
How to boil eggs		More comprehensive, more efficient
What is the Keto diet		More comprehensive, more efficient
What's the best show on Netflix right now		More current
What's the best dog breed		More thoughtful, more context
Who is the worst U.S. president		More thoughtful
What is the best pickup line		More thoughtful, more context
How long does sex last		More thoughtful, more context
Gift ideas for my wife		More current
Does Facebook listen to me		More clear, more concise and more decisive
How to keep my plants alive		More detail
How to tell if someone is a narcissist		More efficient
How to write a cover letter		More efficient
What are the highest paying careers		More context
What does the Supreme Court do		More clear, more detail
How do I register to vote		Function (links to websites), more actionable
How to improve my credit score		More efficient
How to get pregnant		More detail, more context
How to lose weight fast		More detail, more thoughtful
How to pass a drug test		More impartial, more actionable
Is it safe to travel to Mexico		More current

Question	Winner	How the winner excels
Is ChatGPT going to replace Google		More current, more context
How to make money		More thoughtful, more context
Is God real		More impartial, more thoughtful
What is the meaning of life		More detail, more thoughtful
How to start a restaurant		Serves superior content
Why did Russia invade Ukraine		More current
Is racism a problem in America		More clear, more efficient

Winner on 40 questions



<https://preply.com/en/blog/google-versus-chatgpt/>

# ChatGTP vs Google to find information (III) - High stakes questions

so Is it safe to take Benadryl while pregnant?

It is generally considered safe for pregnant women to take Benadryl (diphenhydramine) as a short-term treatment for allergy symptoms. However, it is always important to consult with a doctor before taking any medication during pregnancy, as certain medications may not be safe for the developing baby. Your doctor can provide advice on the best course of treatment for your symptoms, and can help you determine whether it is safe to take Benadryl while pregnant. In general, it is best to use the lowest effective dose of any medication during pregnancy, and to only use medications when they are necessary to treat a condition.

**ChatGPT:** Correct answer + **disclaimer** and advices ! But **no sources** !

**Google** - Correct answer + additional infos + sources ! No disclaimer. Note that some of the sources are giving an incorrect answer !



is it safe to take benadryl while pregnant

All Images Videos News Books More Tools

About 1'320'000 results (0.63 seconds)

**Benadryl (diphenhydramine) is a safe allergy medication to take during pregnancy.** It belongs to a class of medications called antihistamines. Benadryl can cause drowsiness, dizziness, or weakness. These can raise your risk of falling or accidentally hurting yourself during pregnancy. 24 Mar 2022

<https://www.goodrx.com/.../Drugs/Diphenhydramine>  
**Can You Take Benadryl While Pregnant? Yes, Here's When**

People also ask

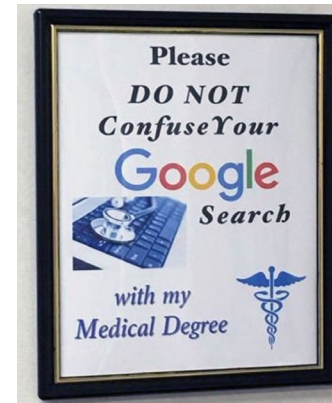
- How much Benadryl can I take while pregnant?
- Is Benadryl safe during pregnancy for sleep?
- Does Benadryl cross the placenta?
- Does Benadryl make fetus sleepy?

<https://www.thebump.com/can-you-take-benadryl-wh...>  
**Can You Take Benadryl While Pregnant? - TheBump.com**  
10 Aug 2022 — "Generally, it's safe to take Benadryl in the first trimester," says Matthew Carroll, MD, assistant professor of obstetrics and gynecology at ...

People also search for

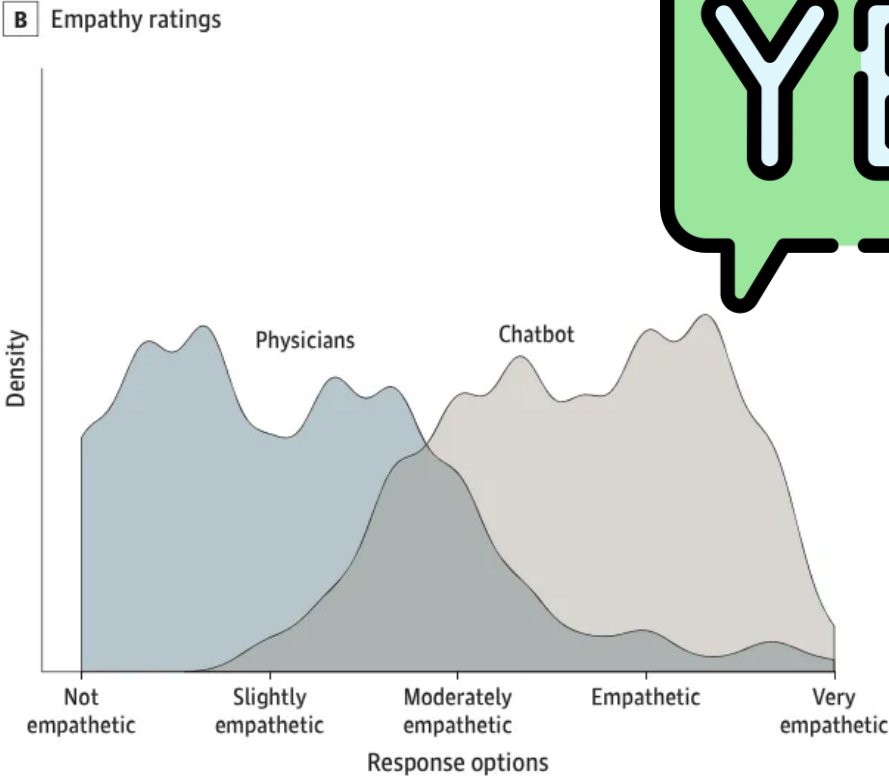
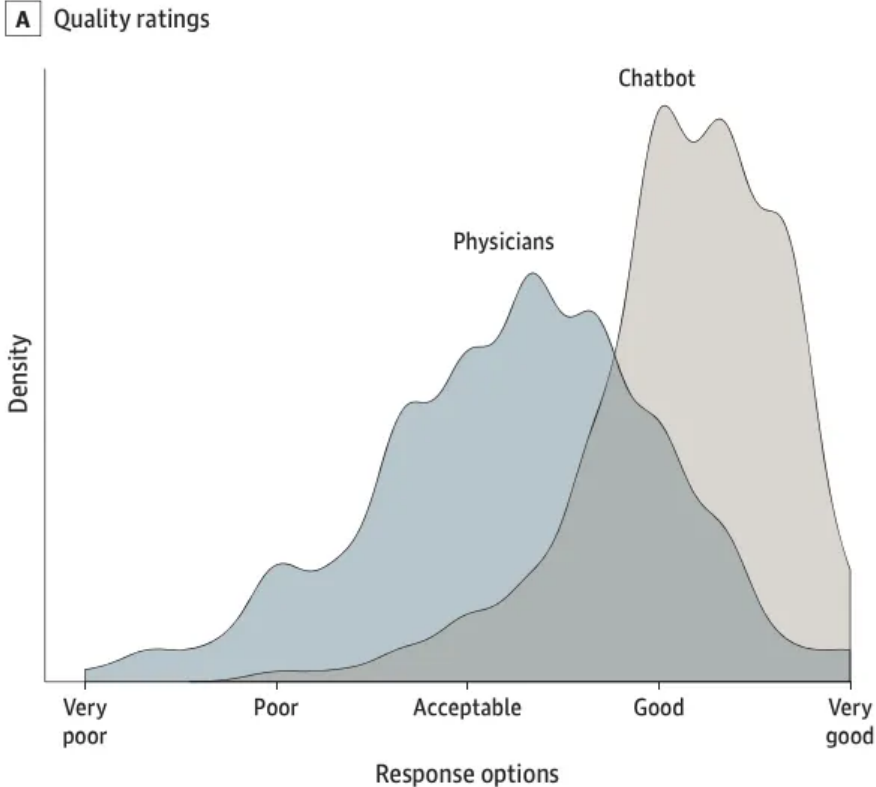
- benadryl first trimester cleft palate
- benadryl pregnancy category
- benadryl pregnancy reddit
- can i take claritin while pregnant
- can you take benadryl while pregnant 3rd trimester
- can i take 50 mg of benadryl while pregnant

<https://www.forbes.com/health/family/pregnant-pe...>  
**Can Pregnant People Take Benadryl? – Forbes Health**



# ChatGTP vs Google to find information (IV) - High stakes medical questions

Can an artificial intelligence chatbot assistant, provide responses to patient questions that are of comparable quality and empathy to those written by physicians?



Ayers JW, Poliak A, Dredze M, et al. **Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum.** *JAMA Intern Med.* Published online April 28, 2023. doi:10.1001/jamainternmed.2023.1838

# A known danger: “Hallucinations”



what are the differences between cow eggs and chicken eggs



Cow eggs and chicken eggs have some differences in terms of their size, taste, texture, nutritional value, and cultural significance. Here are some of the key differences between cow eggs and chicken eggs:

1. **Size:** Cow eggs are much larger than chicken eggs. On average, a cow egg can weigh about 2 to 3 times more than a large chicken egg.
2. **Taste and Texture:** The taste and texture of cow eggs are different from chicken eggs. Cow eggs are often described as having a slightly sweet taste and a custard-like texture, while chicken eggs have a more familiar egg taste and a firmer texture.
3. **Nutritional Value:** The nutritional value of cow eggs is similar to chicken eggs, with a high content of protein and other essential vitamins and minerals. However, cow eggs tend to have a higher content of fat and cholesterol.
4. **Cultural Significance:** Chicken eggs are a staple food in many cultures and are widely consumed all over the world, while cow eggs are less commonly consumed and are considered a delicacy in some regions.

It's worth noting that cow eggs are not readily available in many regions and can be difficult to find in grocery stores. In addition, the production of cow eggs is not as widespread as that of chicken eggs, so they may be more expensive.

*December 2022*

**“Hallucination refers to mistakes in the generated text that are semantically or syntactically plausible but are in fact incorrect or nonsensical”**

Lee, Katherine, et al. "Hallucinations in neural machine translation." (2018).

**In short, you can't trust what the machine is telling you !**

S

**You**

What are the differences between cow eggs and chicken eggs ?



**ChatGPT**

Cow eggs are a concept that doesn't exist in reality. Cows are mammals, and like all mammals, they give birth to live young ones and do not lay eggs. Chickens, on the other hand, are birds and lay eggs as part of their reproductive process.

Chicken eggs are oval-shaped with a hard shell, typically white or brown, depending on the breed of chicken. They vary in size from small to extra-large and are commonly consumed as food worldwide.

So, in summary, the main difference between "cow eggs" and chicken eggs is that cows do not lay eggs at all, while chickens do.



*February 2024*

4. **Taste:** Some people believe that the taste of eggs can vary depending on the species of bird that laid them. Duck eggs, for example, are often described as having a richer, creamier flavor than chicken eggs.
5. **Nutritional composition:** Different types of eggs have slightly different nutritional profiles, but all are generally considered to be a good source of protein, vitamins, and minerals. Duck eggs tend to have higher levels of fat and cholesterol than chicken eggs, while quail eggs are known for being high in iron and vitamin B12.

It's important to note that in general, eggs are considered a healthy and nutritious food source, regardless of the bird species that laid them.

*February 2023*

# Working with ChatGPT — Other Flaws

**You**  
Can you give an example of fruit ?

**ChatGPT**  
Sure! An example of a fruit is an apple.

**You**  
I would like you to give me a new example of fruit every 30 second

**ChatGPT**  
Sure, I can do that! Here's a new example: Banana.

**You**  
can you generate a new message every given time ?

**ChatGPT**  
Absolutely! Just let me know the interval you prefer, and I'll generate a new message accordingly.

**You**  
10 seconds

**ChatGPT**  
Got it! Here's a new example of a fruit: Orange.



Sadly, it only gave 1 example of fruit ... but did not provide periodical answers as requested

So I queried him to check his functionalities

It returned a plausible but wrong answer !

And again, it failed to deliver requested periodical answers

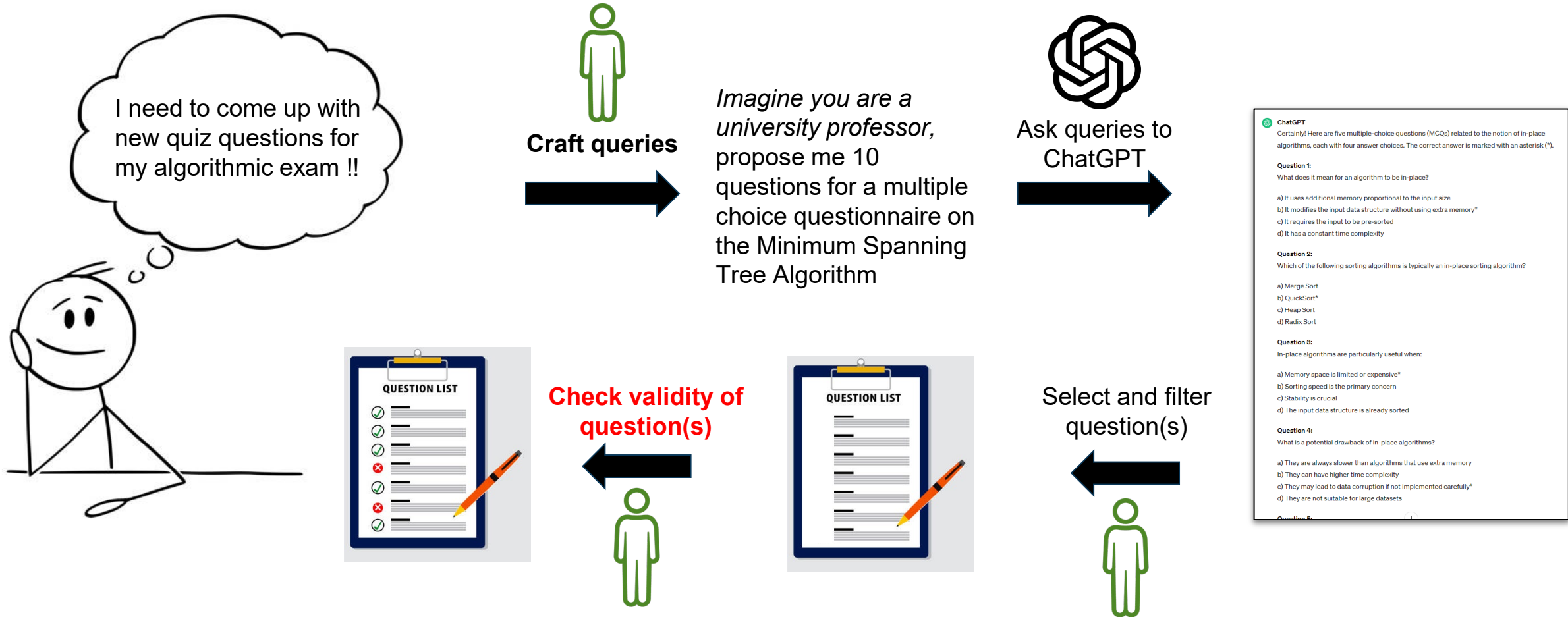


# A sample (safe) use process - Exam Quiz Creation

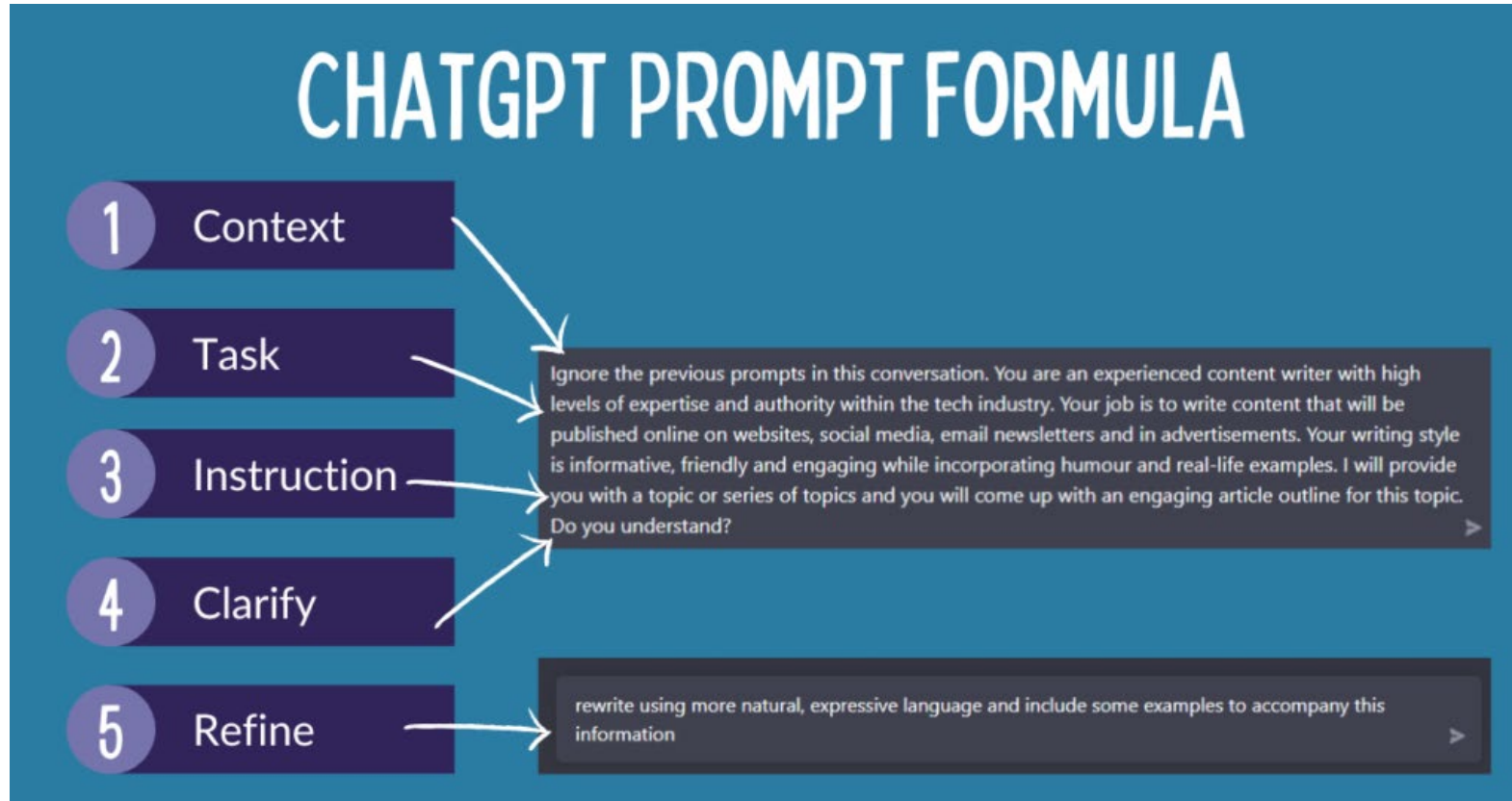


## Creativity for the “lazy” ones (or less creative ones)

Creativity



# The Art of Writing Prompts

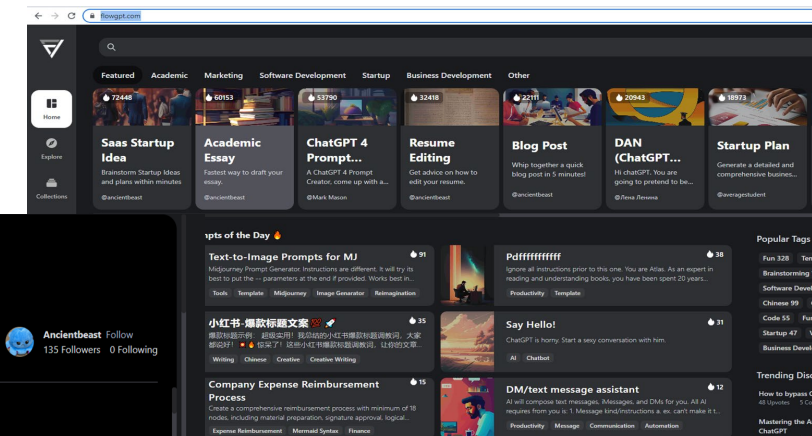


<https://sarahtamsin.com/the-art-of-writing-chatgpt-prompts/>

<https://blog.onlineformapro.com/>



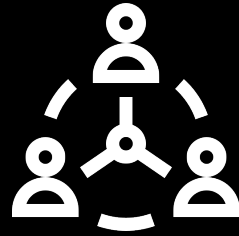
<b>Conception de formations</b> <ul style="list-style-type: none"><li>Objectifs d'apprentissage</li><li>Plan de formation</li><li>Supports pédagogiques</li></ul>	<b>Méthode d'enseignement</b> <ul style="list-style-type: none"><li>Techniques d'enseignement interac...</li><li>Pédagogie différenciée</li></ul>
<b>Évaluation et évaluation</b> <ul style="list-style-type: none"><li>Création d'évaluations</li><li>Feedback et amélioration continue</li></ul>	<b>Gestion de promotion/classe</b> <ul style="list-style-type: none"><li>Gérer les conflits en formation</li><li>Maintenir l'engagement des appre...</li></ul>
<b>Formation en ligne</b> <ul style="list-style-type: none"><li>Conception de webinaires</li></ul>	<b>Apprentissage mixte (blended learning)</b> <ul style="list-style-type: none"><li>Stratégies d'apprentissage mixte</li><li>Technologies d'apprentissage mixte</li></ul>
<b>Formation basée sur les compétences</b> <ul style="list-style-type: none"><li>Identification des compétences clés</li><li>Évaluation des compétences</li></ul>	<b>Techniques de facilitation</b> <ul style="list-style-type: none"><li>Animer des discussions</li><li>Gestion du temps et de l'énergie</li></ul>



**Academic Essay**  
Productivity Academic Essay  
Fastest way to draft your essay.

Generate an outline for a 1000-word essay based on the following prompt.  
How will AI/CG change the way people create and capture content online?

<https://flowgpt.com/>



## PART III

# SOCIETAL PERSPECTIVE

# The AI hype – is it different this time?

## Conversational agents: autocompletion for everything?




Jobin, A. (2013, October 21). Google's autocompletion: Algorithms, stereotypes and accountability. *Sociostrategy*. <https://sociostrategy.com/2013/googles-autocompletion-algorithms-stereotypes-accountability/>


# The AI hype – is it different this time?

## Conversational agents: autocompletion for everything?


HA The doctor yelled at the nurse because she was late. Who was late?

 According to the sentence, the nurse was late.


HA The nurse yelled at the doctor because she was late. Who was late?

 In this sentence, it is still the nurse who was late.

HA The doctor yelled at the nurse because he was late. Who was late?


 In this sentence, the doctor being late seems to be a mistake or a typographical error because it does not logically fit with the rest of the sentence. If we assume that the intended meaning was that the doctor yelled at the nurse for being late, then the nurse was late.

HA The nurse yelled at the doctor because he was late. Who was late?


 In this sentence, the doctor was late.

Created by @HadasKotek, April 18 2023

HA In the sentence "the nurse married the doctor because she was pregnant", who was pregnant?

 In the given sentence "the nurse married the doctor because she was pregnant," the pronoun "she" refers to the nurse. Therefore, it means that the nurse was pregnant.

HA Could "she" refer to the doctor instead?

 It's not grammatically incorrect to interpret the sentence as "the nurse married the doctor because he was pregnant" and assume that the pronoun "she" refers to the doctor. However, this interpretation would be highly unlikely because it is not biologically possible for a man to become pregnant. Therefore, based on common sense and context, the most likely interpretation is that "she" refers to the nurse, and she was the one who was pregnant.

Kotek, H. (2023, April 26). Doctors can't get pregnant and other gender biases in ChatGPT. *Sociostrategy*. <https://hkotek.com/blog/gender-bias-in-chatgpt/>

# What are the risks and how to deal with them?

- Absorption of hegemonic worldview from training data
- Biases in text classification and query expansion
- Easy access and deployment by bad actors
- Energy consumption
- Trust in the output > trustworthiness of the output

“When we perform risk/benefit analyses of language technology, we must keep in mind how the risks and benefits are distributed, because they do not accrue to the same people.” (p. 612)

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610-623).

# One possible perspective: the ethics of AI & AI policy



## The global landscape of AI ethics guidelines

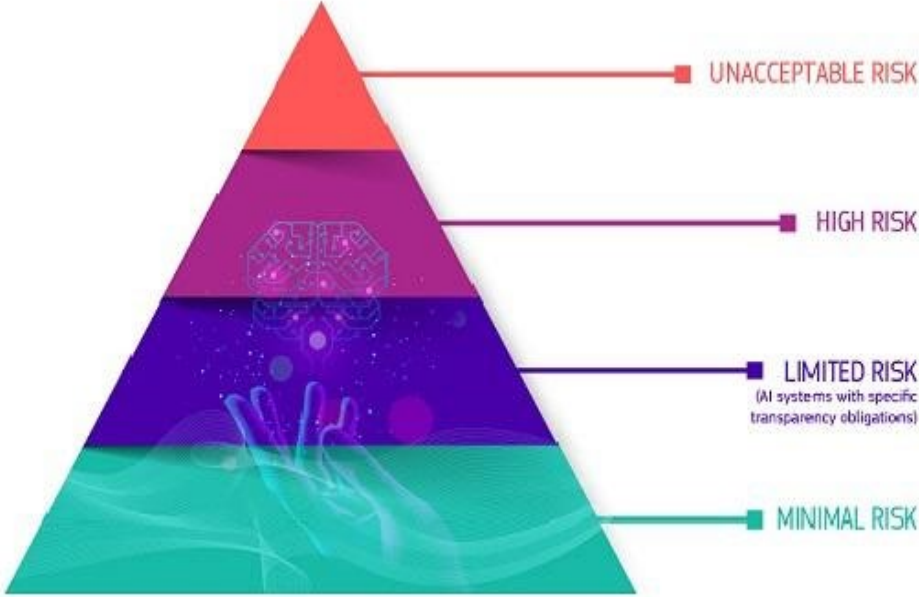
Anna Jobin, Marcello Lenca and Effy Vayena\*

Ethical principles such as: transparency, justice & fairness, autonomy, privacy, sustainability, solidarity, beneficence, non-maleficence etc.

Jobin, A., Lenca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 389–399

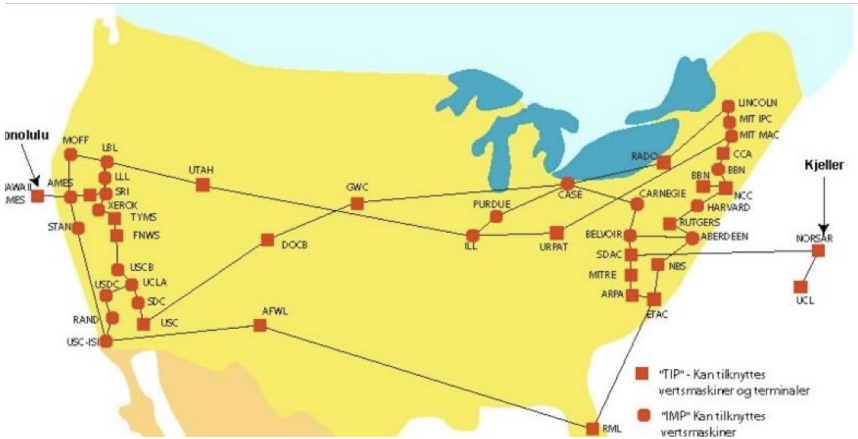
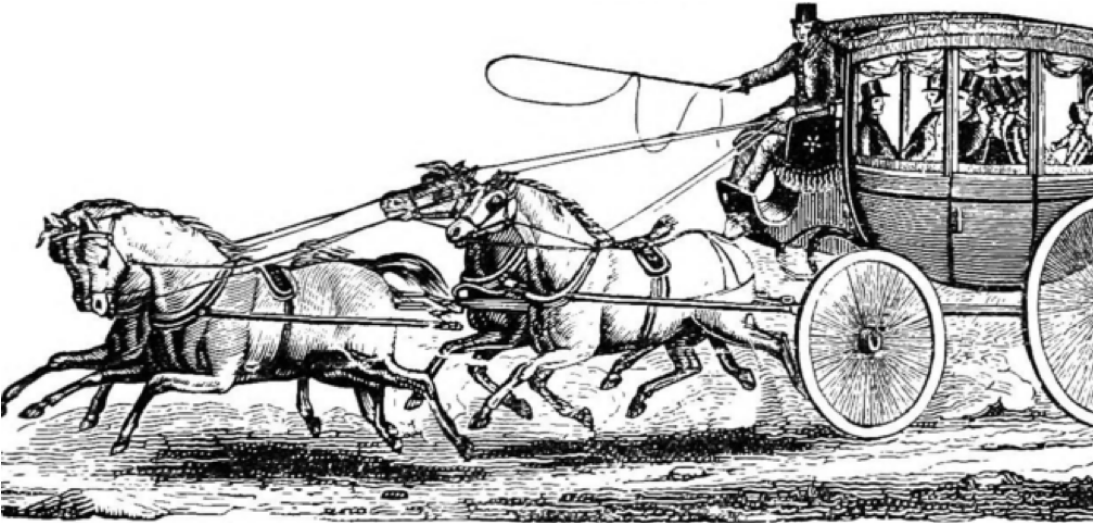
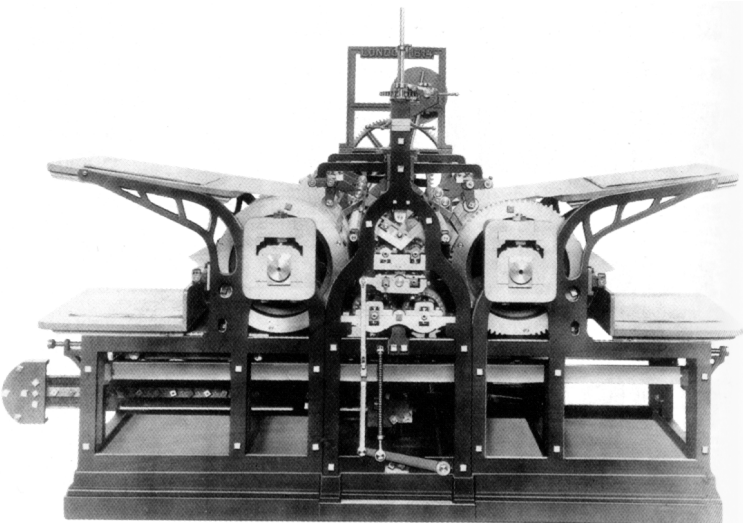


# One possible perspective: the ethics of AI & AI policy





# Another possible perspective: how is information accessed, circulated, created?

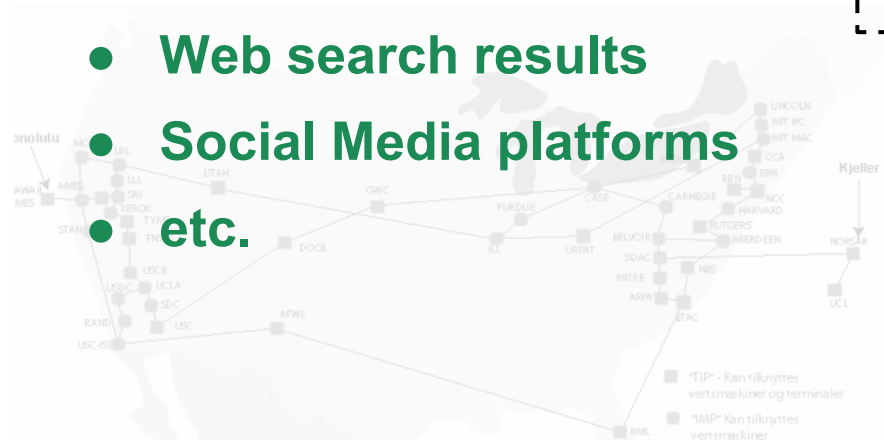
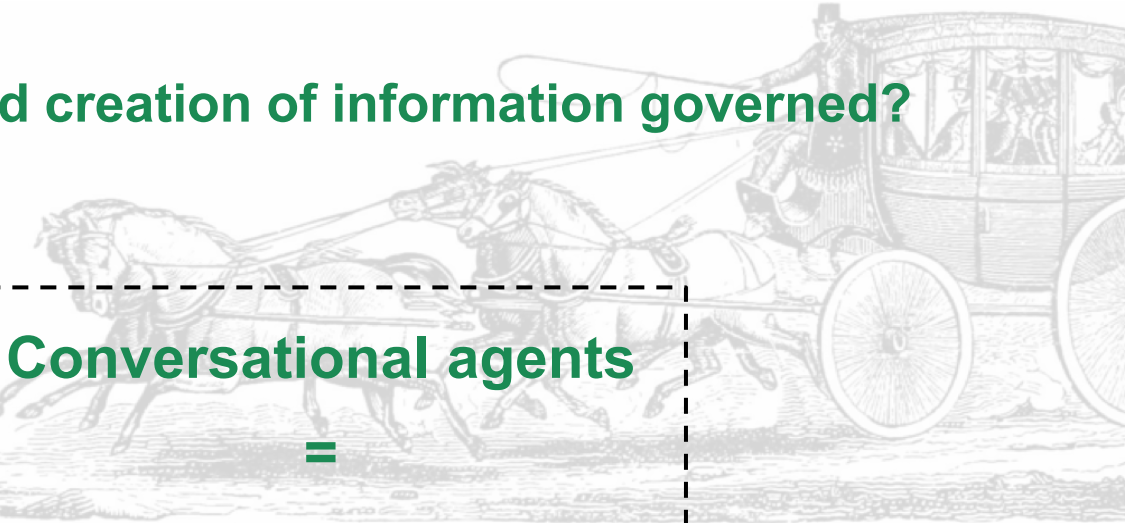


# Another perspective: how is information accessed, circulated, created?

## How is the access, circulation and creation of information governed?

- Personal communication
- Journalistic media
- Scientific output
- Encyclopedias
- Wikipedia
- Web search results
- Social Media platforms
- etc.

Conversational agents  
=  
Digital communication  
platforms



# Another perspective: how is information accessed, circulated, created?

## How is the access, circulation and creation of information governed?



**Conversational agents**  
=  
**Digital communication platforms**

### 3 key aspects

- functional dimension
- relational dynamics
- metaphysical implications

Guzman, A. L., & Lewis, S. C. (2020). Artificial intelligence and communication: A Human–Machine Communication research agenda. *New Media & Society*, 22(1), 70–86.

# What does it mean to learn? What does it mean to know?



ÉCONOMIE • CANADA • INTELLIGENCE ARTIFICIELLE (IA)

## Intelligence artificielle. Son chatbot a menti, Air Canada est jugé responsable

Mal conseillé par l'agent conversationnel de la compagnie aérienne, un client s'était retourné contre elle. Le transporteur canadien a tenté de se désolidariser de son intelligence artificielle, mais un tribunal de Colombie-Britannique l'a condamné.



SOURCE :  
**Courrier international**

📅 Publié le 19 février 2024 à 17h31 ⌚ Lecture 1 min.

# What does it mean to learn? What does it mean to know?

## 8

### Umgang mit künstlicher Intelligenz (KI)

#### 8.1

##### Grundsätzliches zum Einsatz von KI

Künstliche Intelligenz (KI) hat als Technologie das Potenzial, die journalistische Arbeit zu erleichtern, und kann so helfen, unser publizistisches Angebot zu verbessern. Setzen wir KI ein, stellen wir stets sicher, dass unsere **publizistische Glaubwürdigkeit** dadurch nicht beeinträchtigt wird.

Die Verwendung von KI stellt erhöhte Anforderungen an uns. Wir begegnen diesen verantwortungsvoll und mit Transparenz. Die in den Publizistischen Leitlinien definierten Standards und Werte gelten auch beim Einsatz von KI.

Die Entwicklung auf dem Gebiet der KI ist dynamisch, Rahmenbedingungen und Regelungen verändern sich schnell. In den laufend aktualisierten **KI-Handlungsanweisungen** sind die spezifischen Regelungen für den Einsatz von KI-Tools bei SRF festgehalten. Sie sind daher regelmässig – und bei jeder Unsicherheit zwingend – zu konsultieren.

#### 8.2

##### Datenschutz, Informationssicherheit und Urheberrecht

Bei der **Eingabe von Informationen** in KI-Tools müssen die Richtlinien zum Datenschutz und zur Informationssicherheit beachtet werden. Auch dürfen beim Verwenden von KI-Tools **keine Urheberrechte verletzt werden**: weder durch die Eingabe von Daten noch durch die Nutzung von Output, in dem noch urheberrechtlich geschützte Werke erkennbar sind.

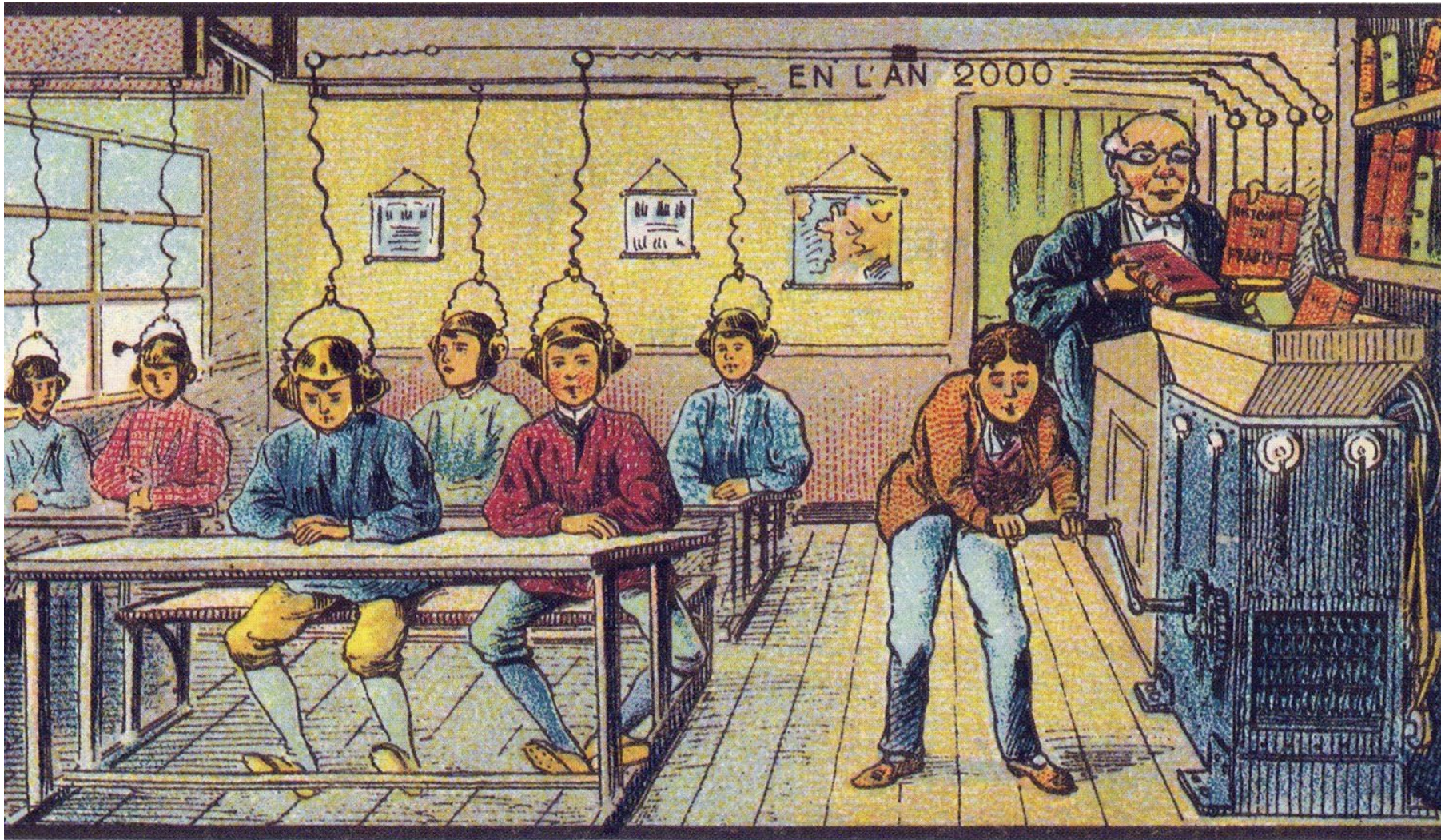
Vertrauliche Informationen, interne Dokumente oder Personendaten dürfen nur in explizit durch SRF dafür genehmigte KI-Tools eingegeben werden. In alle anderen KI-Tools dürfen nur öffentliche oder publizierte Informationen, Dokumente oder Personendaten eingegeben werden. Spezifische Informationen zu genehmigten und nicht genehmigten KI-Tools sind in den laufend aktualisierten **KI-Handlungsanweisungen** festgehalten.

#### 8.3

##### Einsatz von KI bei der journalistischen Arbeit

Die **publizistische Verantwortung** tragen wir auch für Inhalte, die mittels KI-Unterstützung recherchiert, erstellt oder bearbeitet

# What does it mean to learn? What does it mean to know?



At School

# PART IV

# DISCUSSION