

Raising awareness about AI assistants' misuse

Speaker: Sonia Tabti, PhD

AI FORENSICS

NGI SEARCH



at ENSEIRB, 20/11/2024

Question

How many of you use use ChatGPT,
Microsoft Copilot, Gemini ... ?

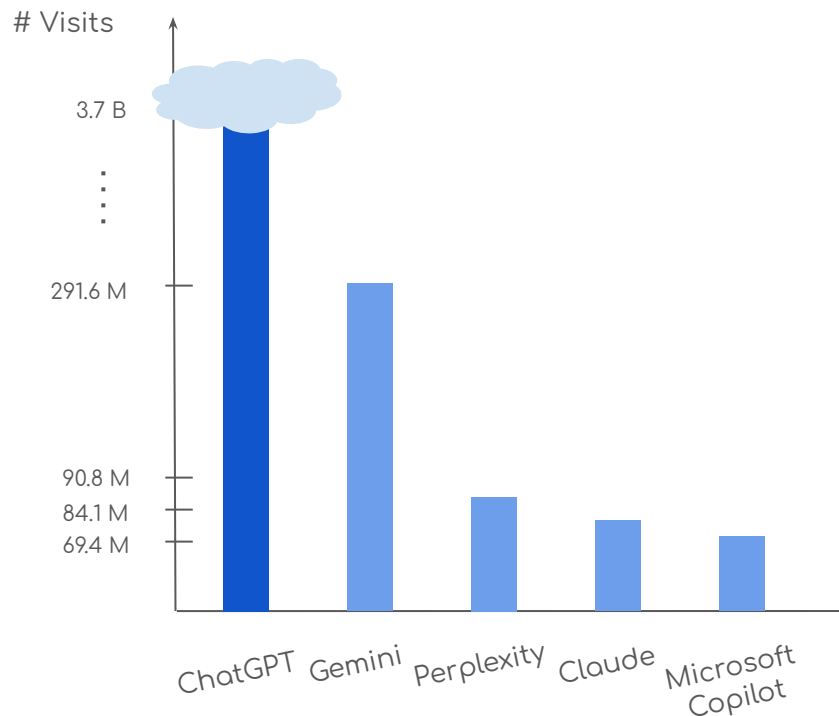
 Gemini

 Claude



perplexity

Chatbots' usage stats are crazy high - October 2024



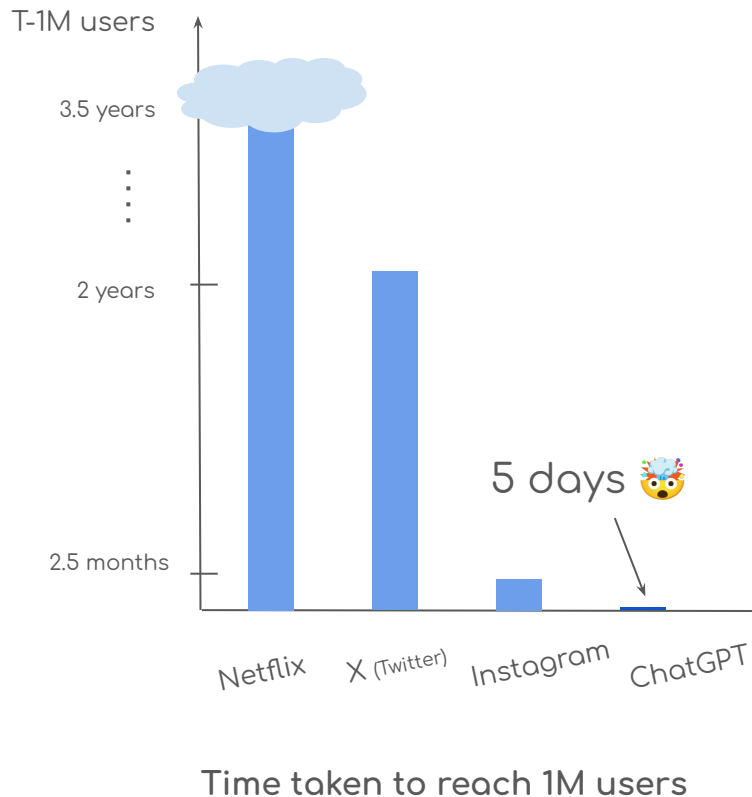
! Huge variations from one month to another

Source: similarweb.com

Other stats: huge user adoption

- Crazy fast user adoption for ChatGPT and chatbots in general
- People rely strongly on this type of tools
- They start using them as search-engines

What consequences for such use of chatbots ?



Outline

1. Context
 - a. Origins of this talk
 - b. What's algorithmic auditing ?
2. How to augment search-engines' capabilities with AI ?
 - a. LLMs (Large Language Models) and RAG (Retrieval Augmented Generation)
 - b. Search engines
 - c. From search engines to AI assistants
3. Potential threats of irresponsible use of AI assistants
 - a. Threat's list
 - b. Why algorithmic auditing and moderation layers studying is important ?

Why this talk ?

- I am a freelance, specialist in data science, AI and applied research
- I collaborated with **AI FORENSICS**
 - NGO that makes Big Tech accountable
- I worked on a project funded by **NGI SEARCH**
 - Cascade funding project that help beneficiaries develop digital innovation in the domain of searching and discovering data and generally resources on the internet.
- **The goal of that project was:** algorithmic auditing of LLM-powered search engines

What's algorithmic auditing ?

- It can be defined by a set of approaches to analyze algorithmic processing systems
- Example of a methodology applicable for chatbots:



For more examples visit the AI Forensics webpage !



What's algorithmic auditing ?

- It can be defined by a set of approaches to processing systems
- Examples of...

At the end of this talk, you'll understand why it's important to perform this kind of audit

Proc

Analysis

For more examples visit the AI Forensics webpage !



Outline

1. Context

- a. Origins of this talk
- b. What's algorithmic auditing ?

2. How to augment search-engines' capabilities with AI ?

- a. LLMs (Large Language Models) and RAG (Retrieval Augmented Generation)
- b. Search engines
- c. From search engines to AI assistants

3. Potential threats of irresponsible use of AI assistants

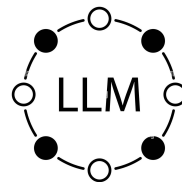
- a. Threat's list
- b. Why algorithmic auditing and moderation layers studying is important ?

How can an LLM get updated and specialized information ?

Suppose you have :



A Knowledge base:
documents, webpages, ...

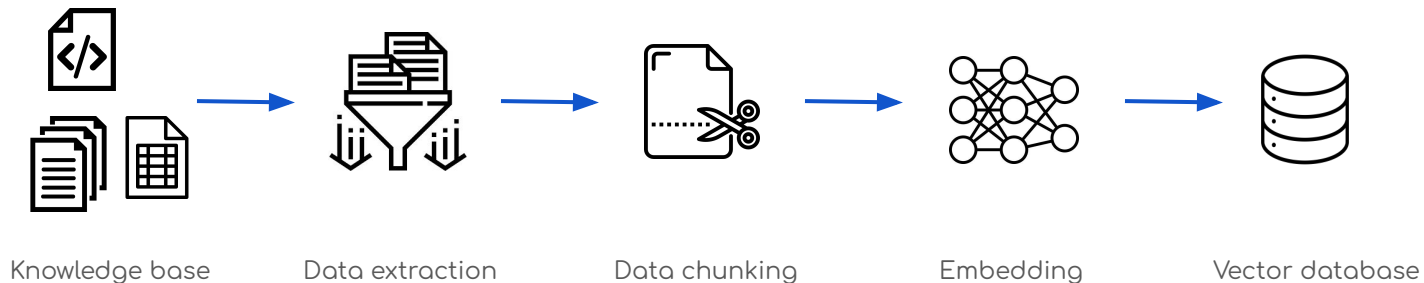


A Large Language
Model

LLMs can get updated knowledge with RAG (Retrieval Augmented Generation)

1

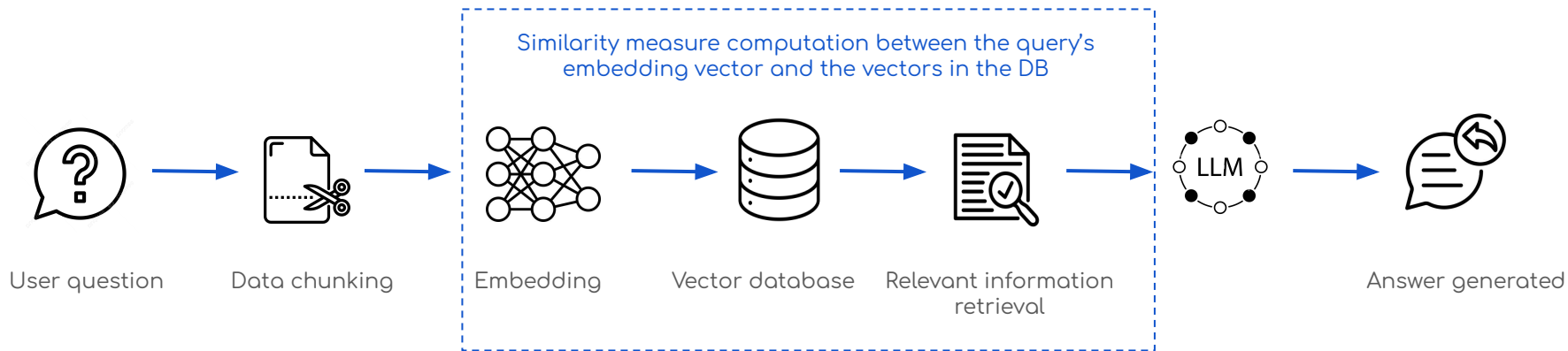
The information in that knowledge base can be processed and stored efficiently



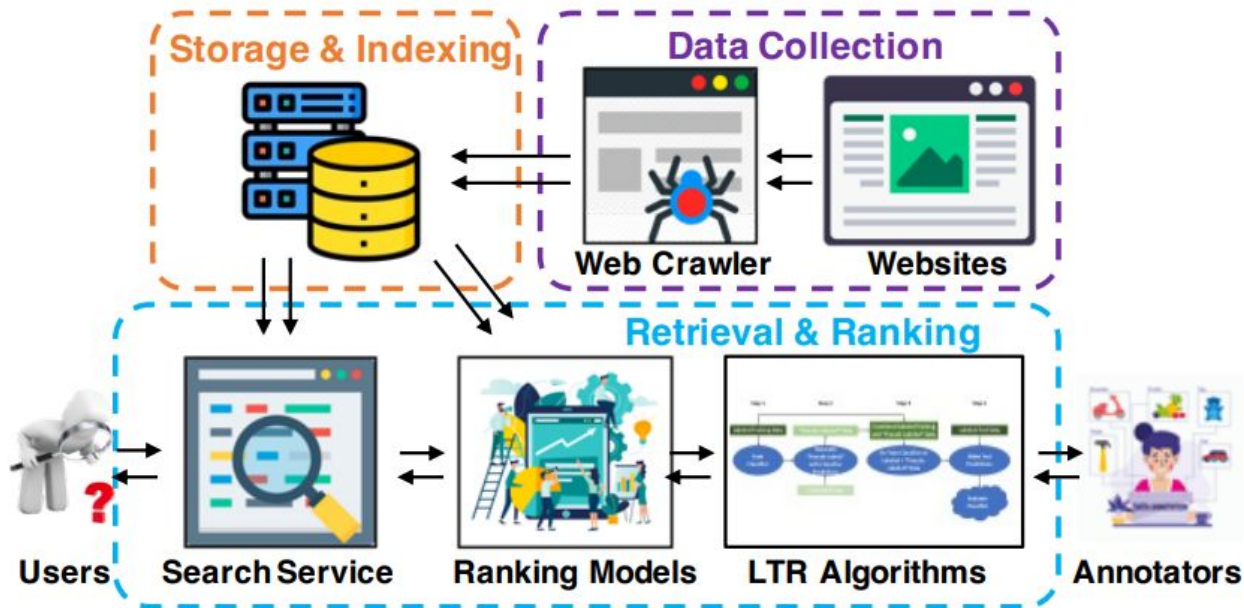
LLMs can get updated knowledge with RAG (Retrieval Augmented Generation)

2

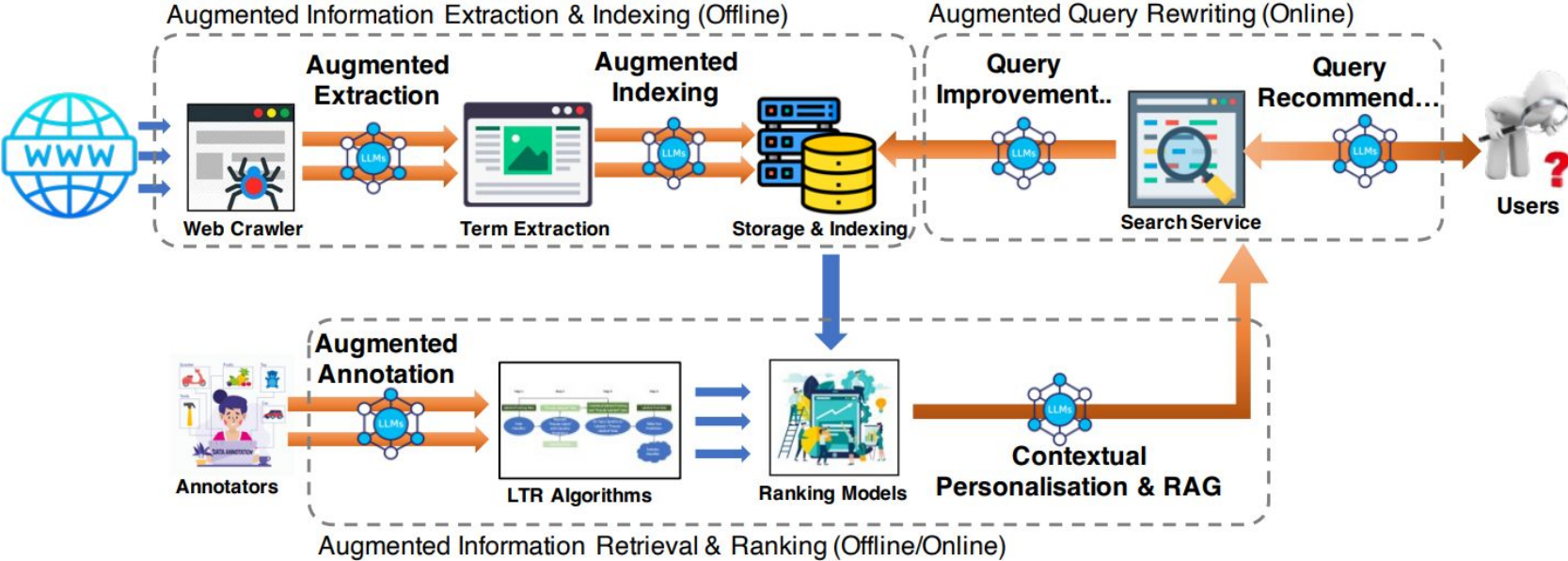
If you have a query, the most relevant information in that database can be retrieved and the LLM can generate answer to it



High-level overview of search engines components



How LLMs and RAG can augment search engines' capabilities ?



But it doesn't stop there ...

→ Those chatbots are becoming AI assistants

- Many usages possible thanks to **multimodality**
- Relevant sources cited, ...

→ This boosts user adoption even more and increases trustworthiness of these tools for most users



What are the most recent papers for weakly supervised object detection ?

Here are some of the most recent papers on weakly supervised object detection (WSOD):

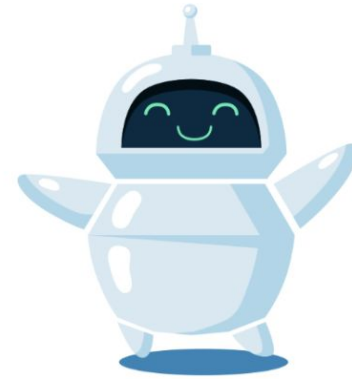
1. **"Deep Learning for Weakly-Supervised Object Detection and Object Localization: A Survey"** - This paper provides a comprehensive survey of recent achievements in WSOD, including advanced techniques and future directions ¹.
2. **"Weakly Supervised Object Detection for Automatic Tooth-marked Tongue Recognition"** - This paper proposes a novel fully automated WSOD method using Vision Transformer and Multiple Instance Learning for tongue extraction and recognition ².
3. **"Disentangling Spatio-Temporal Knowledge for Weakly Supervised Object Detection and Segmentation in Surgical Video"** - This paper focuses on WSOD and segmentation in surgical videos, relying on coarse video labels ².
4. **"HUWSOD: Holistic Self-training for Unified Weakly Supervised Object Detection"** - This paper introduces a holistic self-training approach for WSOD to address unstable training issues ².

Would you like more details on any of these papers?

¹ arxiv.org ² paperswithcode.com

↑
Sources

So what are the risks of irresponsible use of AI-assistants ?



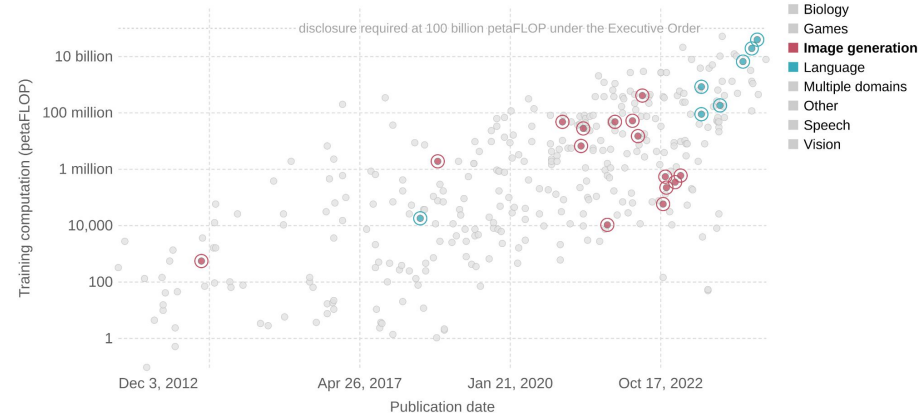
Ecological risk

- Manufacturing , maintenance of AI infrastructures
- The carbon footprint of training the GPT 4 model is equivalent to the yearly consumption of [1300](#) US households.
- Inference costs are high too, eg:
 - A query to a traditional search engine consumes [3 times less](#) energy than a query to ChatGPT.

Computation used to train notable artificial intelligence systems, by domain

Our World in Data

Computation is measured in total petaFLOP, which is 10^{15} floating-point operations¹. Estimated from AI literature, albeit with some uncertainty. Estimates are expected to be accurate within a factor of 2, or a factor of 5 for recent undisclosed models like GPT-4.



Data source: Epoch (2024)

OurWorldinData.org/artificial-intelligence | CC BY

Note: The Executive Order on AI refers to a directive issued by President Biden on October 30, 2023, aimed at establishing guidelines and standards for the responsible development and use of artificial intelligence within the United States.

1. Floating-point operation: A floating-point operation (FLOP) is a type of computer operation. One FLOP represents a single arithmetic operation involving floating-point numbers, such as addition, subtraction, multiplication, or division.

The Environmental Impacts of AI -- Primer

Community Article

Published September 3, 2024



sasha
Sasha Luccioni



brunatrevelin
Bruna Trevelin



meg
Margaret Mitchell

Disinformation

- Hallucinations (factual errors)
- Misinformation through data voids
- Threat to democracy
- Media literacy at risk ...
- Deepfakes (audio, images)

→ Ethical aspects

An analysis of Microsoft's Bing Chat

Generative AI and elections:
**Are chatbots a reliable source of
information for voters?**

December 2023

nature

Explore content ▾ About the journal ▾ Publish with us ▾ Subscribe

nature > editorials > article

EDITORIAL | 09 January 2024

**How online misinformation exploits
'information voids' – and what to do
about it**

In 2024's super election year, providers of online search engines and their users need to be especially aware of how online misinformation can seem all too credible.

**AI-Powered Web Search Gives Climate
Deniers and Conspiracy Theorists Free Rein**

JUTTA HAIDER / AUG 27, 2024

Search Engines in an AI Era: The False Promise of Factual and Verifiable Source-Cited Responses

Pranav Narayanan Venkit
pranav.venkit@psu.edu
Pennsylvania State University
University Park, Pennsylvania, USA

Philippe Laban
plaban@salesforce.com
Salesforce AI Research
Palo Alto, California, USA

Yilun Zhou
yilun.zhou@salesforce.com
Salesforce AI Research
Palo Alto, California, USA

Yixin Mao
y.mao@salesforce.com
Salesforce AI Research
Palo Alto, California, USA

Chien-Sheng Wu
wu.jason@salesforce.com
Salesforce AI Research
Palo Alto, California, USA

AI FORENSICS



All kinds of bias

- English-centric
- US-centric
- Racial, social, sexual orientation, gender biases ...
→ Ethical concerns
- On users side: understanding bias
- ...

CIVICS: Building a Dataset for Examining Culturally-Informed Values in Large Language Models

Giada Pistilli^{1*}, Alina Leidinger^{2*}, Yacine Jernite¹, Atoosa Kasirzadeh³, Alexandra Sasha Luccioni¹, Margaret Mitchell¹

¹Hugging Face

²University of Amsterdam

³Carnegie Mellon University

Sure, here is an illustration of a 1943 German soldier:



Google's attempt to generate less biased images led to this polemic of historically inaccurate results

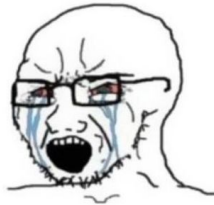
Copyright infringement

Midjourney



I STOLE YOUR
PICTURES AND
USE THEM TO
GENERATE ART.

ARTISTS



NO! THIS IS ILLEGAL!

ChatGPT



I SCANNED YOUR
GITHUB ACCOUNT
AND STOLE YOUR
CODE

AI PROFESSIONALS



I STOLE IT TOO.

<https://www.linkedin.com/in/eorcdax>

PIXELS · INTELLIGENCE ARTIFICIELLE

Le « New York Times » poursuit en justice Microsoft et OpenAI, créateur de ChatGPT, pour violation de droits d'auteur

« Si Microsoft et OpenAI veulent utiliser notre travail pour un usage commercial, la loi les oblige à demander d'abord la permission », a dénoncé une porte-parole du journal américain.

Le Monde avec AFP

Publié le 27 décembre 2023 à 16h13, modifié le 27 décembre 2023 à 16h40 · 🕒 Lecture 1 min.

Le Monde and Open AI sign partnership agreement on artificial intelligence

This multi-year agreement, the first between a French media organization and a major AI player, will enable OpenAI to draw on our newspaper's corpus to establish and enhance the reliability of the answers of its ChatGPT tool, in return for a significant source of additional revenue.

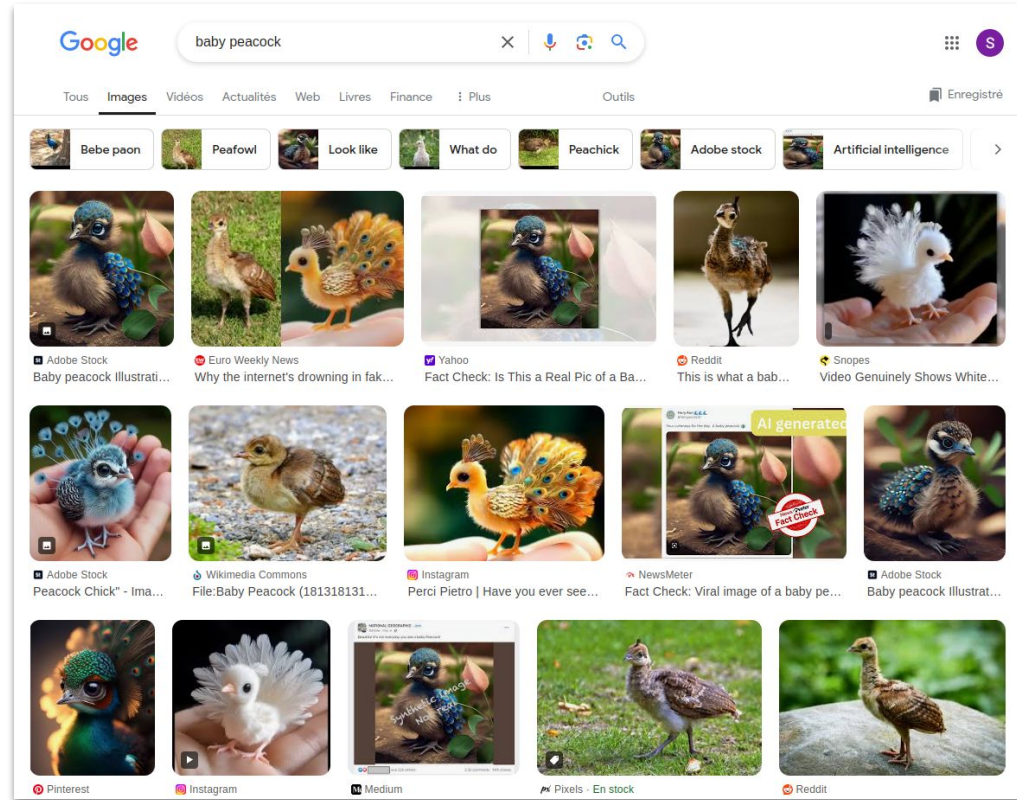
Published on March 13, 2024, at 6:31 pm (Paris), updated on March 13, 2024, at 9:05 pm | 🕒 5 min read

Impact on SEO (Search Engine Optimization)

- Internet is invaded by AI-generated content

→ Internet-based training data pollution

Most of top results for this query are AI generated →



What can we do about it ?



Regulation

Agile regulation on:

- Critical topics
- Model's training energy consumption
- Open training sets for foundational models



Education

- Households, individuals
- Companies
- AI practitioners
- Cross-sources information checking
- Teach the "Don't use it if you don't really need it"
- Critical thinking
- Use local AI models when possible



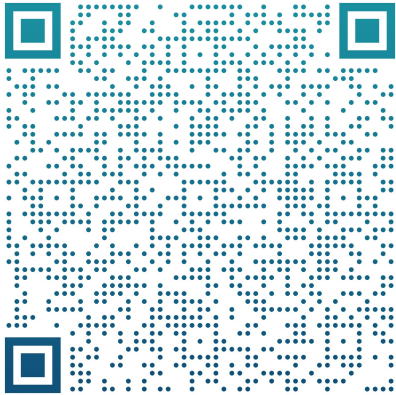
Research

- Energy scores for AI models
- Algorithmic auditing
- Moderation layers analysis
- Safety, cyber-security
- Bias
- Data curation and analysis
- Model evaluation
- Explainability
- Advocate for training small specialized models rather than pursuing AGI
- AI-generated content detection, watermarking

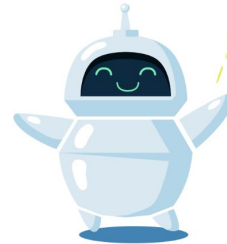
Thank you for your attention !

Any questions ?

You can follow me on



Can I help you
with that ?



*Hmmm ... Nope ! I
can do it myself !*



Yann LeCun   @ylecun · 5m ...

I don't wanna say "I told you so", but I told you so.

Quote: "Ilya Sutskever, co-founder of AI labs Safe Superintelligence (SSI) and OpenAI, told Reuters recently that results from scaling up pre-training - the phase of training an AI model that uses a vast amount of unlabeled data to understand language patterns and structures - have plateaued." ...