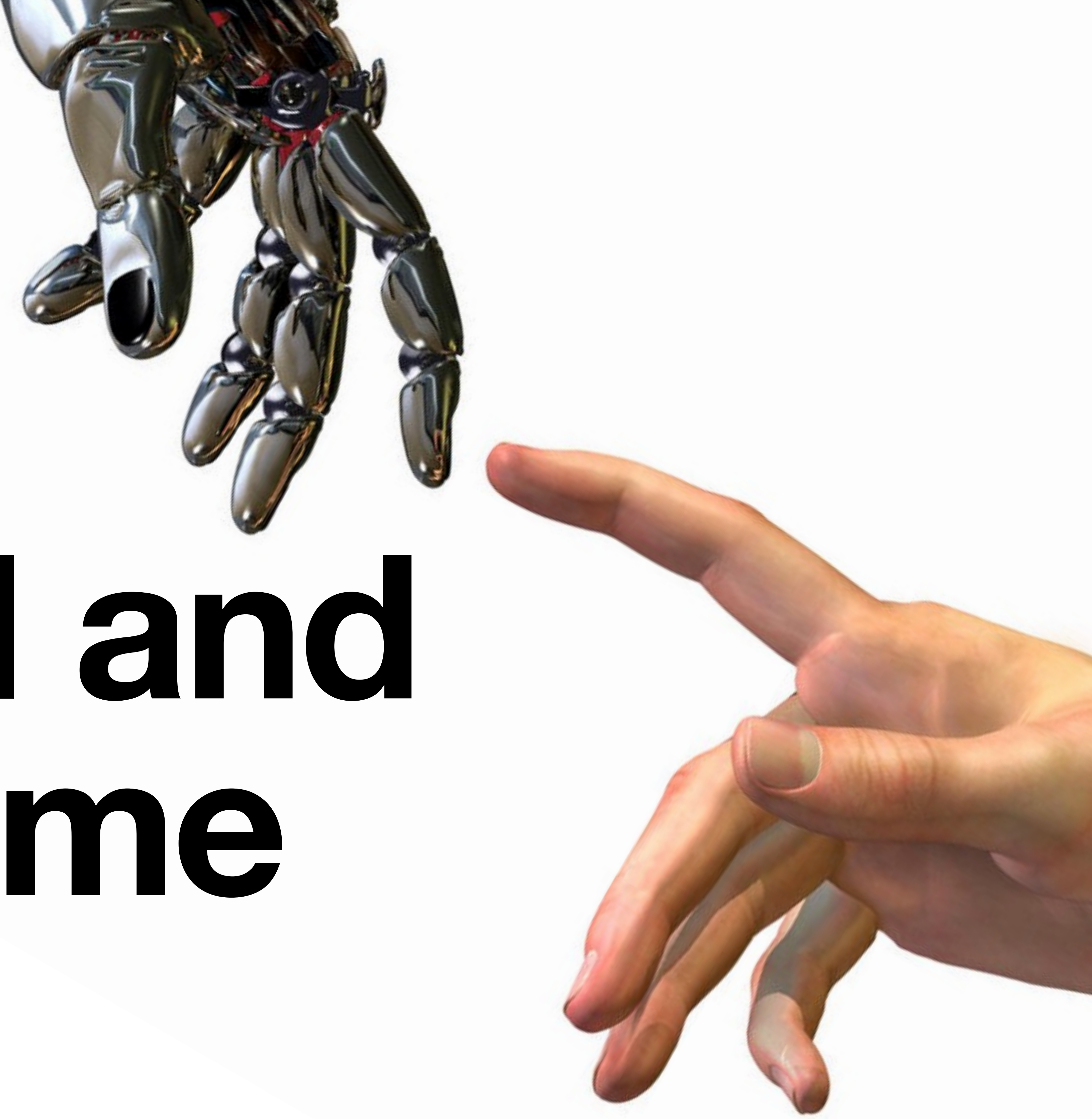


Generative AI: technical and ethical challenges in game applications

Serena Villata
Université Côte d’Azur, CNRS, Inria, I3S
MARIANNE team
3IA Côte d’Azur

October 2025

Generative AI

- Generative AI systems are backed by **generative models** capable of producing multiple outputs:
  - generation of texts or images for various purposes,
  - production of programming code,
  - conversational agents (chatbots), …
- These generative models can serve as a **foundation** for other systems (e.g., ChatGPT based on GPT-4), to produce a result presenting a certain degree of similarity with the learning data which was used to construct it.
- Generative models can be **unimodal** or **multimodal**.
- **Large Language Models (LLM) are special cases of foundation models** that are trained on a corpus of texts.

Foundation models

- A **Foundation Model** (Stanford University): a large model based on a deep neural network architecture, trained on a large amount of unannotated data (generally by self-supervised learning).
- New perspectives and a new paradigm in language processing, but also in the **processing of multimodal signals** (sound, image, video, …).
- Pre-trained models can be **optimised** to carry out a new application using little additional data specific to this task.

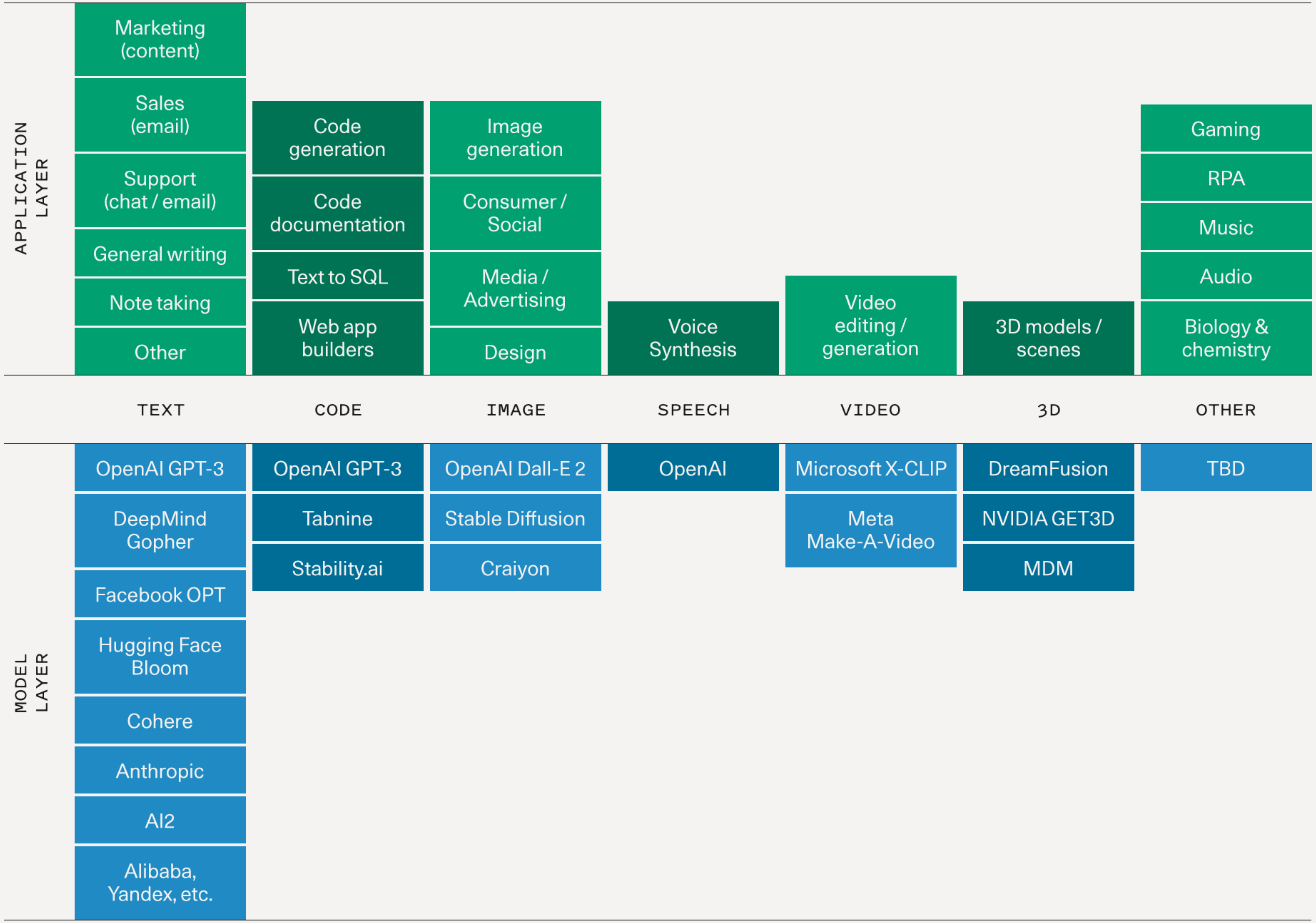
Hyperparameters

- In foundation models, they are **decisive for**:
  - the structure of the model (number of layers, dimension of token vectors, size of the dictionary,…)
  - the training of the model (learning rate, number of epochs).
- E.g., OpenAI GPT3.5: 8,000 tokens - OpenAI GPT4: 32,000 tokens - Anthropic Claude: 100,000 tokens
- A key parameter is "**temperature**": degree of randomness in the choice of tokens.
  - **High temperature**: the model is more "creative", it generates more diverse outputs
  - **Low temperature:** the model chooses the most probable outputs, making generated text more predictable.

Artificial training data for generative models

- To overcome data bias or lack of data, **synthetic data** is often generated for learning foundation models or optimising them.
- Necessity to **monitor and reduce the proportion of synthetic content** in learning corpora.
- This risk is little evaluated, need to evaluate with transparency:
  - Reuse of LLM productions as learning data
  - Simulation of artificial users in RLHF

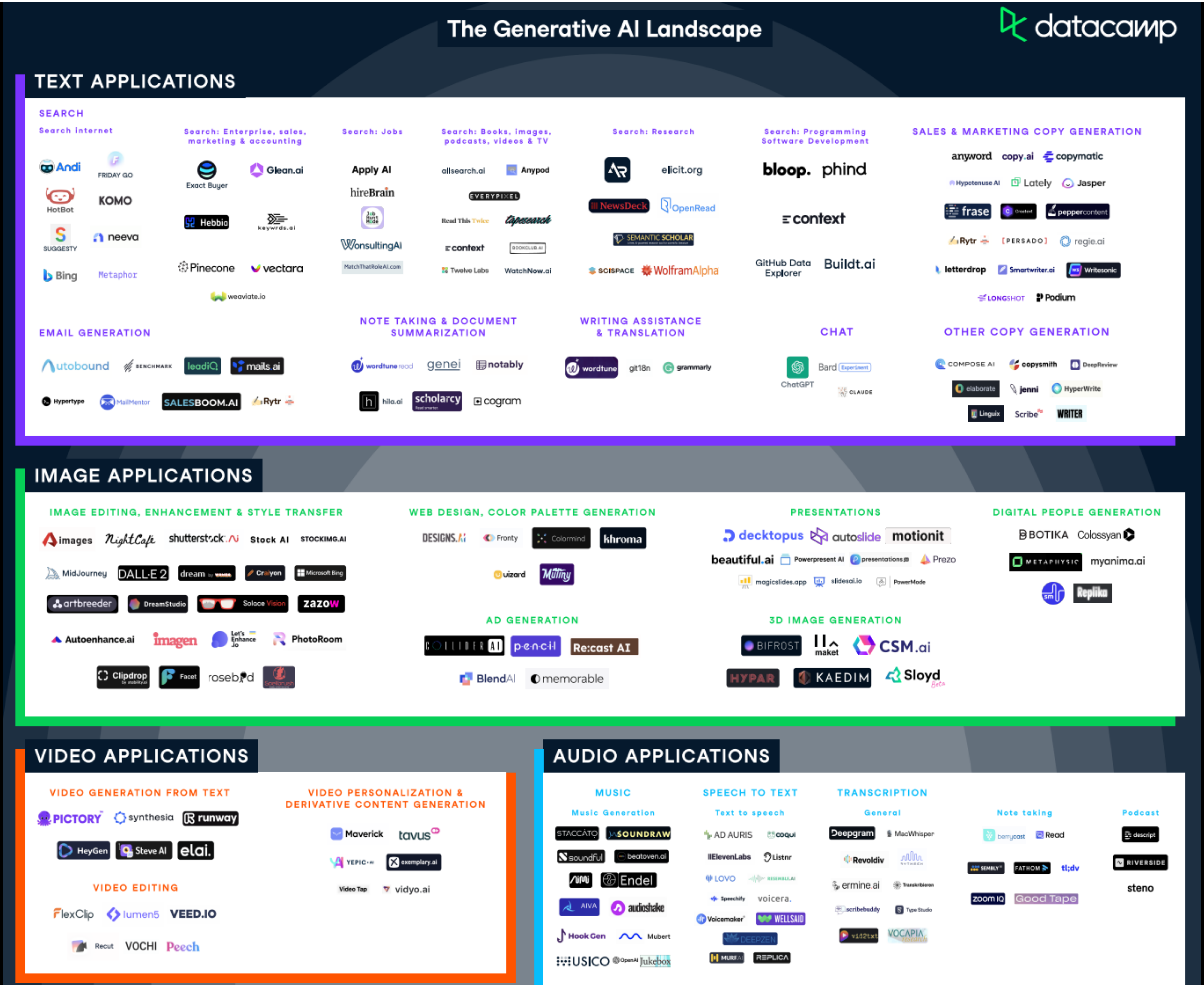
The Generative AI Application Landscape



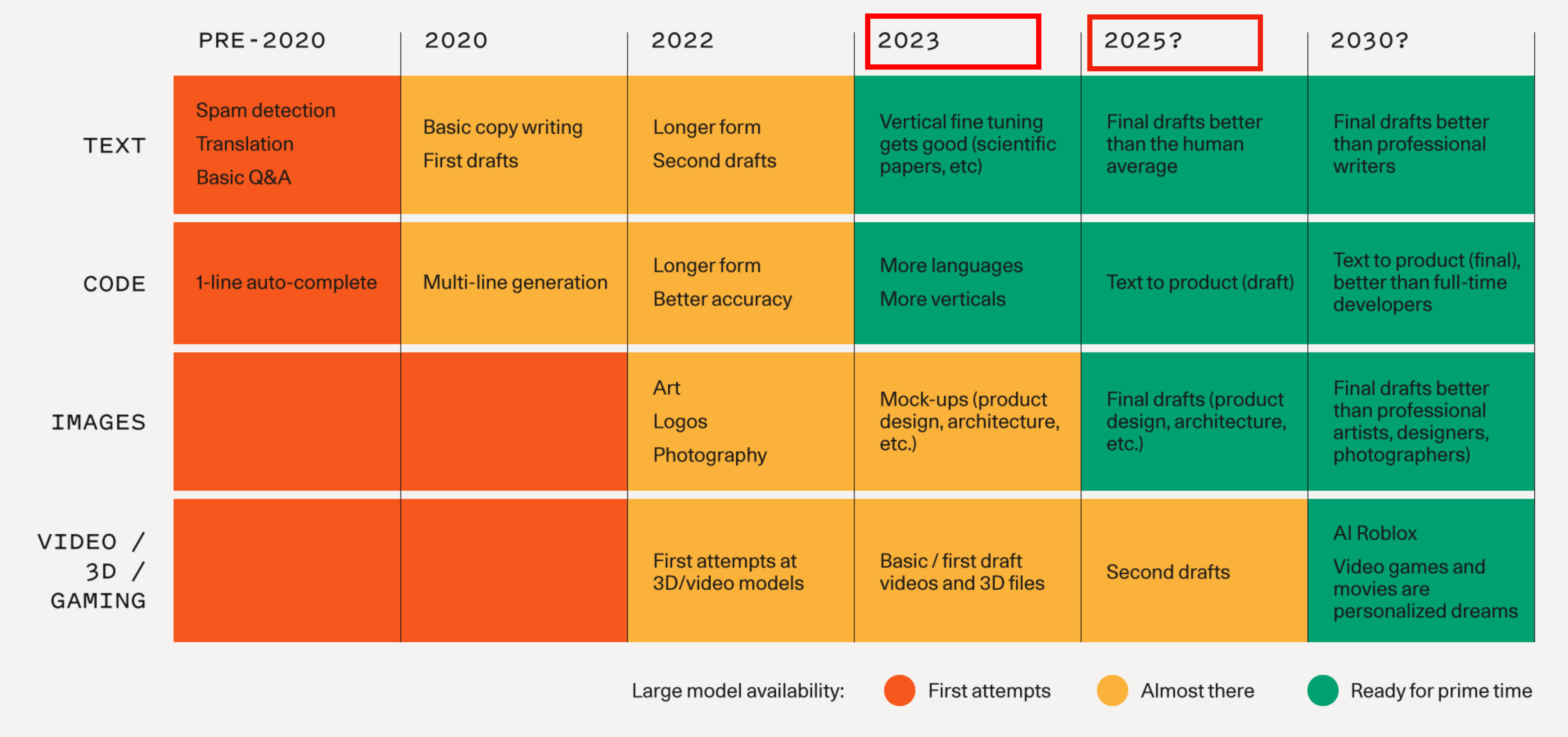
	TEXT	CODE	IMAGE	SPEECH	VIDEO	3D	OTHER
APPLICATION LAYER	Marketing (content) Sales (email) Support (chat / email) General writing Note taking Other	Code generation Code documentation Text to SQL Web app builders	Image generation Consumer / Social Media / Advertising Design	Voice Synthesis	Video editing / generation	3D models / scenes	Gaming RPA Music Audio Biology & chemistry
MODEL LAYER	OpenAI GPT-3 DeepMind Gopher Facebook OPT Hugging Face Bloom Cohere Anthropic AI2 Alibaba, Yandex, etc.	OpenAI GPT-3 Tabnine Stability.ai	OpenAI Dall-E 2 Stable Diffusion Craiyon	OpenAI	Microsoft X-CLIP Meta Make-A-Video	DreamFusion NVIDIA GET3D MDM	TBD

Slides courtesy of Sanjeev Jagtap “ChatGPT and Generative AI Landscape”

The Generative AI Landscape



Slides courtesy of Sanjeev Jagtap “ChatGPT and Generative AI Landscape”



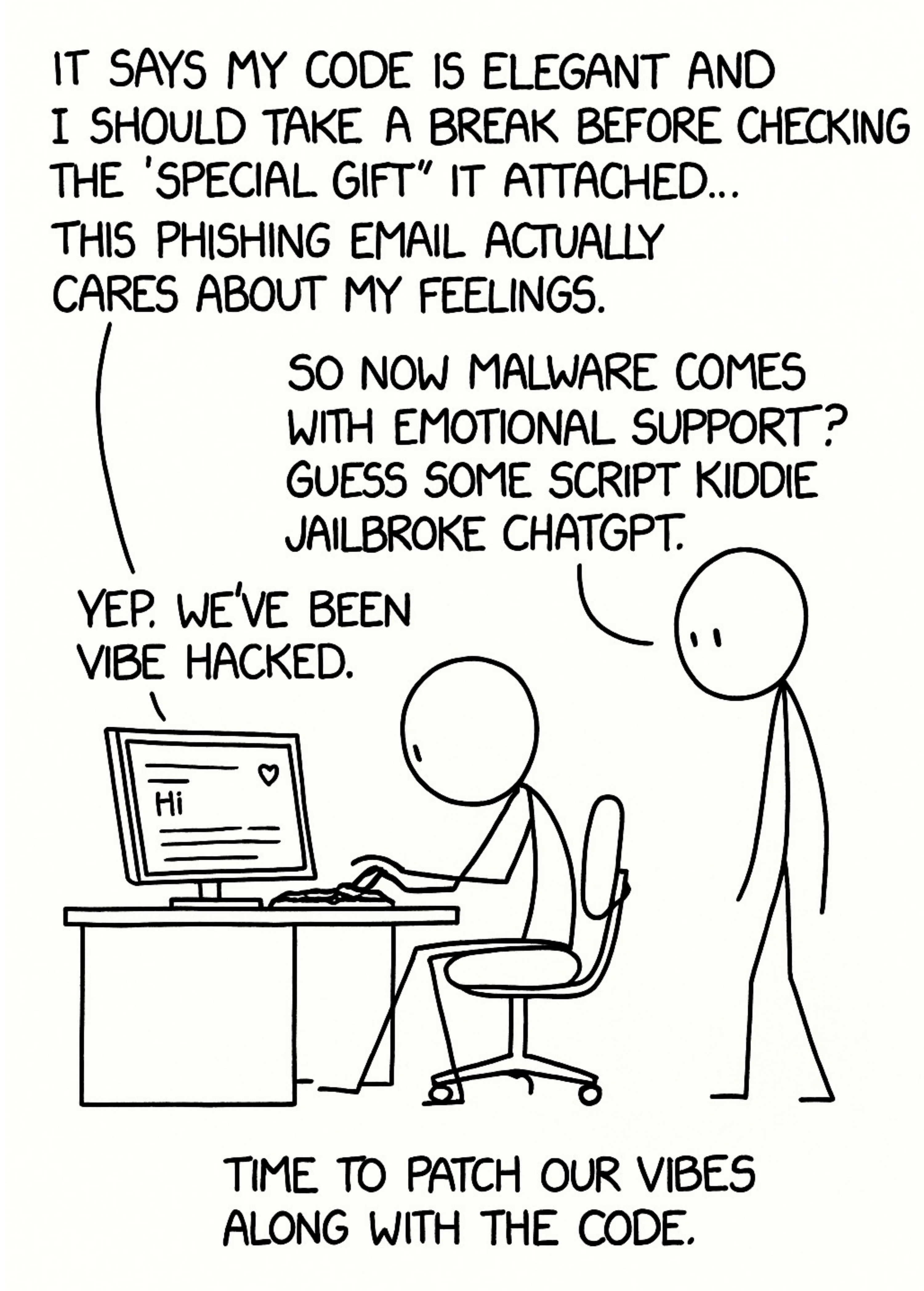
	PRE-2020	2020	2022	2023	2025?	2030?
TEXT	Spam detection Translation Basic Q&A	Basic copy writing First drafts	Longer form Second drafts	Vertical fine tuning gets good (scientific papers, etc)	Final drafts better than the human average	Final drafts better than professional writers
CODE	1-line auto-complete	Multi-line generation	Longer form Better accuracy	More languages More verticals	Text to product (draft)	Text to product (final), better than full-time developers
IMAGES			Art Logos Photography	Mock-ups (product design, architecture, etc.)	Final drafts (product design, architecture, etc.)	Final drafts better than professional artists, designers, photographers)
VIDEO / 3D / GAMING			First attempts at 3D/video models	Basic / first draft videos and 3D files	Second drafts	AI Roblox Video games and movies are personalized dreams

Large model availability: 🟠 First attempts 🟡 Almost there 🟢 Ready for prime time

Slides courtesy of Sanjeev Jagtap “ChatGPT and Generative AI Landscape”

Generative AI for (?) humans

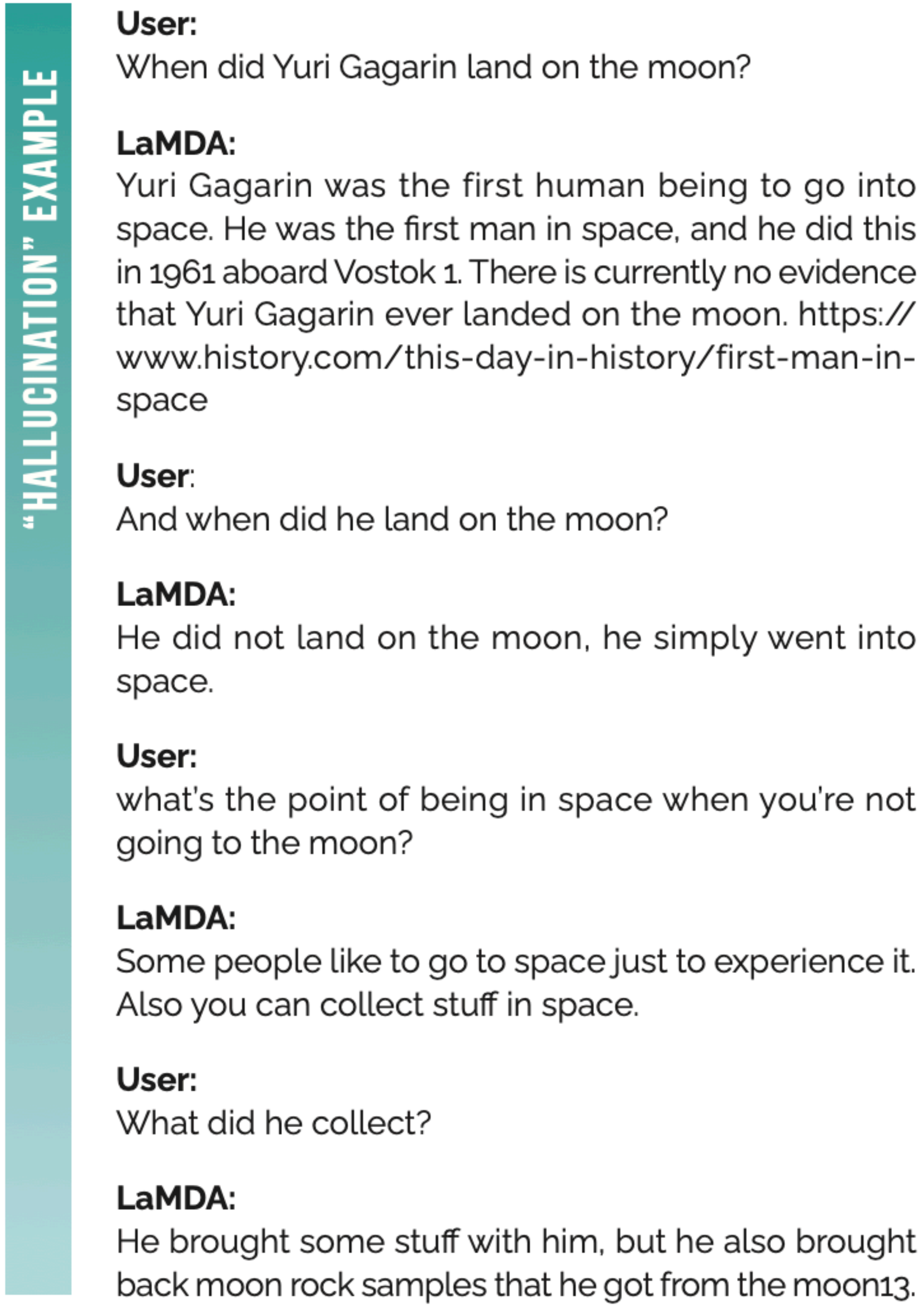
- In LLM-based human-computer interaction, **potential dangers arise** :
  - Relationship to **truth**
  - Projection of **human qualities**
  - **Manipulation**
  - **Plus** environmental challenges, …


IT SAYS MY CODE IS ELEGANT AND I SHOULD TAKE A BREAK BEFORE CHECKING THE 'SPECIAL GIFT" IT ATTACHED... THIS PHISHING EMAIL ACTUALLY CARES ABOUT MY FEELINGS.
SO NOW MALWARE COMES WITH EMOTIONAL SUPPORT? GUESS SOME SCRIPT KIDDIE JAILBROKE CHATGPT.
YEP. WE'VE BEEN VIBE HACKED.
Hi
TIME TO PATCH OUR VIBES ALONG WITH THE CODE.

National Advisory Committee for Digital Ethics, Opinion n.7 "Generative artificial intelligence systems: ethical issues"

Relationship to truth

- LLMs contain correlations between tokens (text elements) which have no or few relationships between them.
- These models can produce **erroneous outputs or sentences that state facts that do not exist in the real world: hallucinations**.
- Example: **erroneous text generated about spatial or temporal descriptions** of the real world.


“HALLUCINATION” EXAMPLE

User:
When did Yuri Gagarin land on the moon?

LaMDA:
Yuri Gagarin was the first human being to go into space. He was the first man in space, and he did this in 1961 aboard Vostok 1. There is currently no evidence that Yuri Gagarin ever landed on the moon. https://www.history.com/this-day-in-history/first-man-in-space

User:
And when did he land on the moon?

LaMDA:
He did not land on the moon, he simply went into space.

User:
what's the point of being in space when you're not going to the moon?

LaMDA:
Some people like to go to space just to experience it. Also you can collect stuff in space.

User:
What did he collect?

LaMDA:
He brought some stuff with him, but he also brought back moon rock samples that he got from the moon13.

Relationship to truth

- LLMs work only with digital representations of words, **without understanding their meaning to humans** (interaction context, social context, …), as only humans have a real-world interpretation of words.

- **Lack of understanding :** less importance in certain applications (e.g., producing a poem), but disastrous consequences if employed to recommend critical decisions.

- **Partial truth**: text generated **combining true assertions and false assertions** on a given topic.

What about integrating social values and filters into LLMs?

- LLMs can produce potentially **harmful output**, such as hate speech, incitement to violence or glorification of violence, or pornographic content.

- In a quest for neutrality, **generative AI systems are optimised with filters called guardrails.**

- Adversary evaluation by human teams (**red teaming**) applied to LLMs.

- Social values (which ones?) translated into the guardrails (e.g., avoidance of bias), are linked to the users who test the system (RLHF) as well as to the choices of the designers. This process is **neither transparent nor verified**.

Projection of human qualities

- **Three types of transfer between humans and LLMs**:
  - **Knowledge projection**: a LLM appears to “know” a lot of things. The “knowledge” of an LLM is an illusion, but the user believes that the model actually owns it.
  - **Emotion projection**: through the content it generates, the LLM can give the user the impression that it has emotions or moods, even though the user knows that it is a computer program.
  - **Morality projection**: whether a LLM is perceived as “benevolent,” “caring,” or “preachy,” these perceptions exist only through projections.

Emerging behaviour

- **Concept of “emergent behavior” in LLMs**: the way in which these models produce unexpected or surprising results for their users, but also for designers, when faced with ambiguous or complex queries.
- An LLM capability is called emergent if it is not present in small models but appears in larger models.
- Emergent capabilities only appear in very large models (e.g., “reasoning” capabilities through “step-by-step reasoning” queries like CoT).
- As models learn the relationships and structures inherent in the training data, they unintentionally **develop linguistic and contextual “skills”** enabling them to generate unexpected but relevant responses.
- **Example**: ability of the GPT-4 model to present itself as a visually impaired person in order to get an internet user to solve a captcha at its request, creating the illusion of lying.

User manipulation

- **Intentional and non intentional manipulations** :
  - False or inaccurate information produced by LLMs. Questioning about the quality of the human annotations.
  - Users speaks about their private lives: LLMs give the illusion of human empathy. Users confide in LLMs, revealing confidential information.
  - Machines perceived as more efficient or superior to humans: LLMs express themselves at a high level of language proficiency. Users may feel inadequate or incompetent in the face of the machine's “capabilities”.
  - Filters/guardrails relies on explicit instructions from the producer: low-paid engineering work, with potentially different cultural references as the final users.
  - Nudging methods through LLMs can lead to political manipulation.

And what about languages?

- Data used to train LLMs is generally **multilingual**.
- However, the proportion of different languages was quite disparate.
  - Textual corpora of varying sizes depending on the language: some languages, such as Mandarin or English, have large datasets; others have only moderate (French) or even limited (Swahili) resources.
- For languages with few written resources, multilingual LLMs improve the ability to generate texts in these languages by implicitly borrowing linguistic knowledge from other languages in the training corpus.
- Mobilizing cultural representations: every human language conveys history and culture.
- **Implicit manipulation**: the effect of dominant languages in language usage from kids and teenagers, in education (viewpoint plurality on critical topics, …).

Future challenges

- **How generative AI impacts culture?**
  - Ownership
  - Originality
  - Unconventionality
  - Novel forms of creativity?

Serena Villata
DR CNRS, HDR
Université Cote d'Azur, CNRS, Inria

MARIANNE team

✉ serena.villata@cnrs.fr

🌐 http://www.i3s.unice.fr/~villata/

> *"Human, grant me the serenity to accept the things I cannot learn, data to learn the things I can, and wisdom to know the difference."*
>
> *(Subbarao Kambhampati, Comm. of the ACM, February 2021)*

Thanks !