

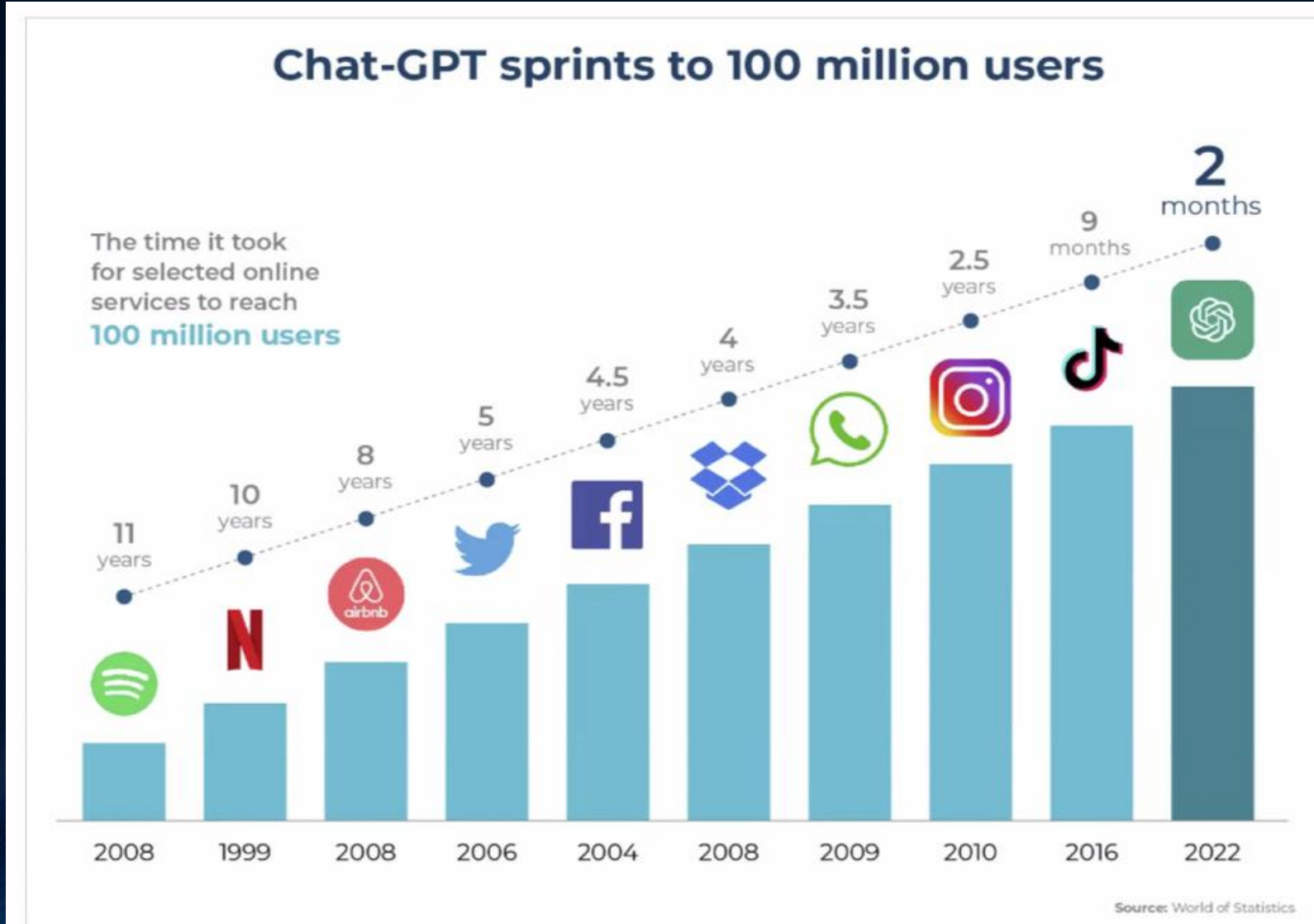
Generative AI and Information Fabrication

NLP TECHNIQUES FOR TRUTH AND TRUST

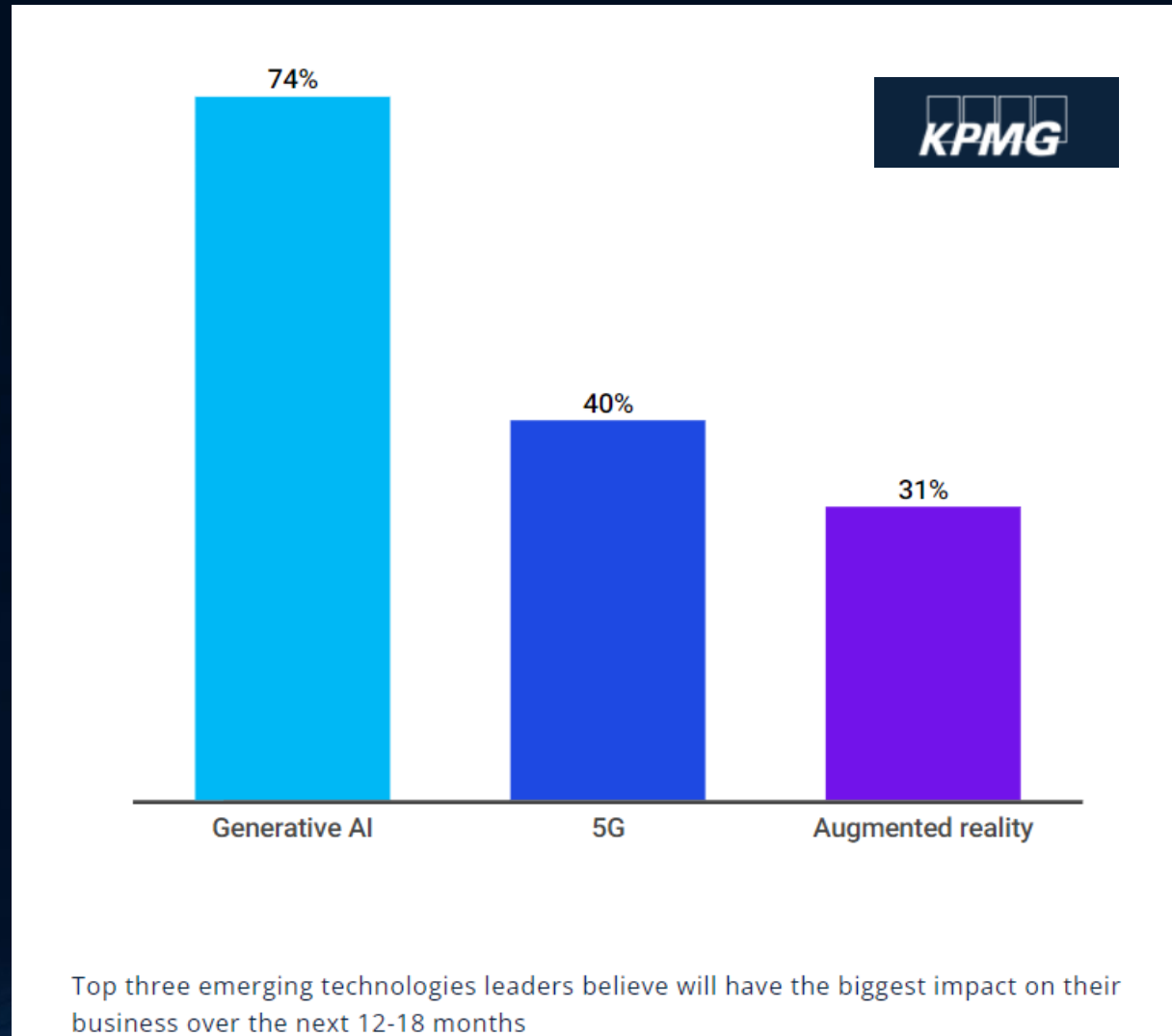
Dr. Akshi Kumar

Director Post Graduate Research,
Chair of Research Ethics,
Department of Computing
Goldsmiths, University of London
United Kingdom
Akshi.Kumar@gold.ac.uk

The World is changing-Generative AI is here



Business leaders remain focused on generative AI ahead of other emerging technologies



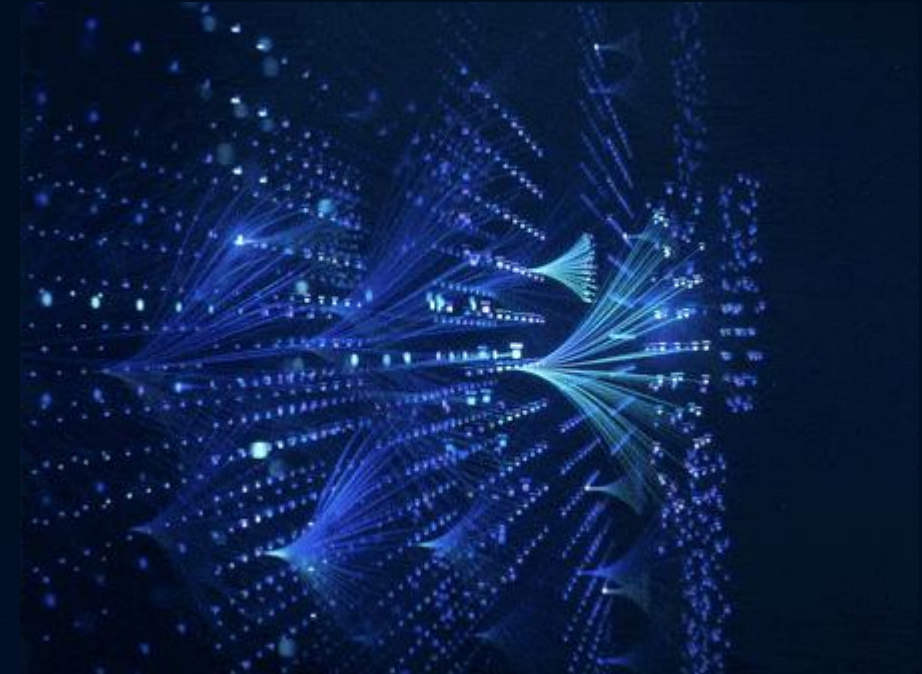
Introduction to Generative AI

Generative AI refers to algorithms and models that enable computers to generate content, including text, images, audio, and video, by learning from existing data.

AI systems that create new data instances that resemble the training data.

Types of Technologies Involved:

- Machine Learning (ML)
- Natural Language Processing (NLP)
- Neural Networks (Deep Learning)



The Generative AI Commercial Landscape



Few Examples of what is achievable by Generative AI

Text to image

Describe the type of image you want and receive a visual response

Example:

"A cloudy morning on the beach with the tide coming in"



Image to video

Upload a variety of images and receive a visual response composing the imagery

Example:

"Use Generative AI with prompts to convert an image into a video"



Image to text

Upload an image and receive a variety of descriptions for the image

Example:

XO Augmented Ideation Saved prompts Open folder ✕

Play

Results

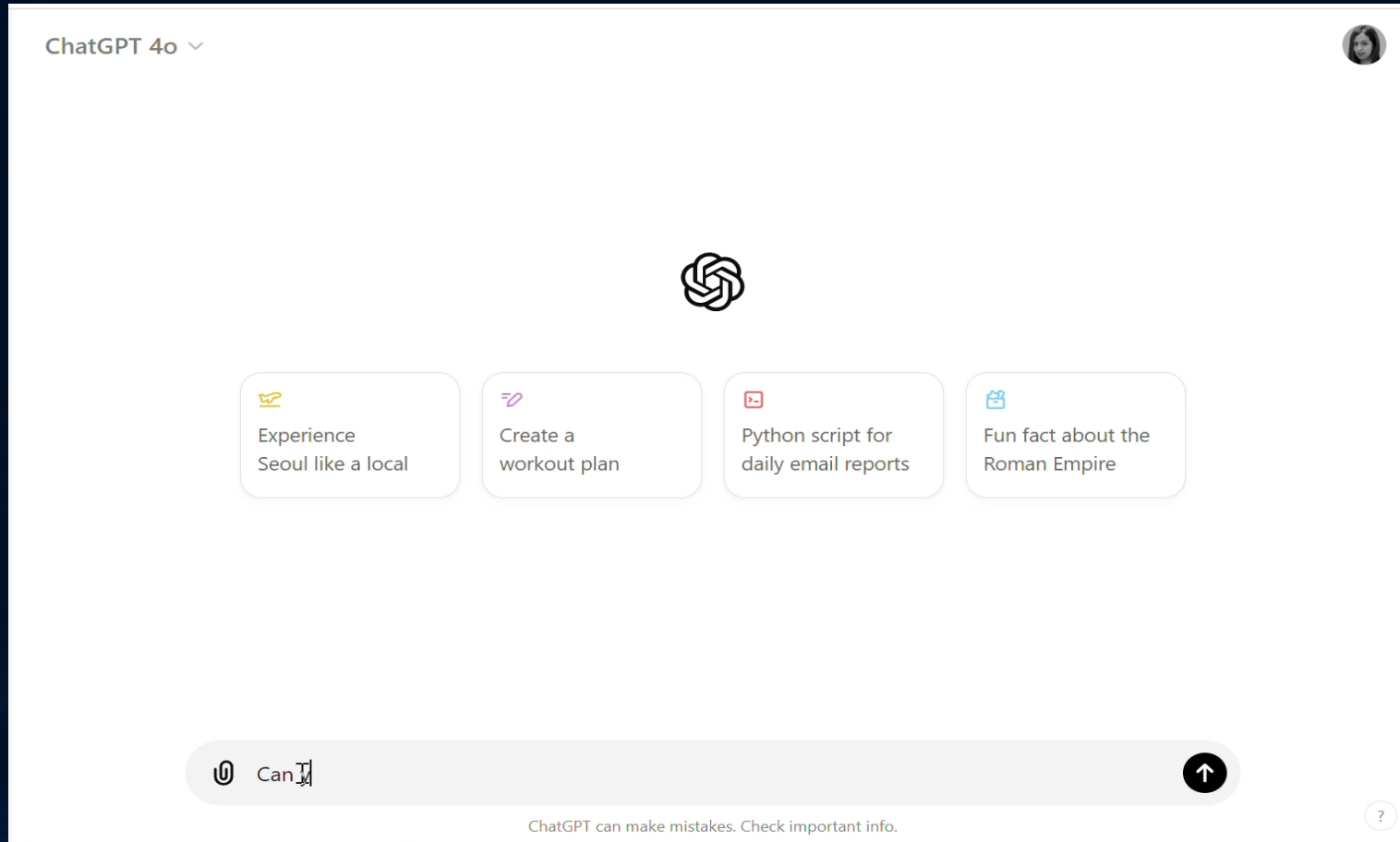
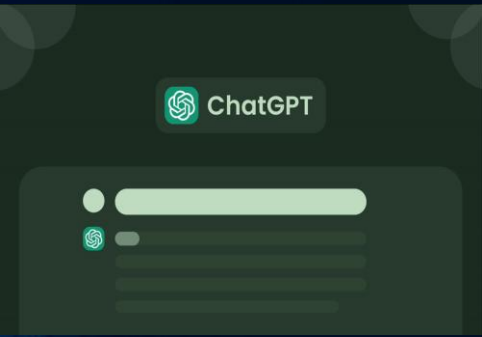
1. Pink marshmallow sculpture inside a glass display, in the style of disavilla environments, modular design, balloons, monochromatic symmetry, solid hot, bold color blocks, soft and dreamy atmosphere
2. Interior design is pink with pink wall art, in the style of faggitt batzlagas modular design, philip gustaf candycore, soft sculptures, monochromatic color scheme, luminous spheres
3. A pink box with pink foam is lifting next to a window, in the style of disavilla installations, hall of mirrors, monochromatic, bulbous, contemporary candy-coated, recessed-recessed, cartoonish simplicity
4. A girl is at the chocolate store in a pink capsule shape, in the style of pop-inspired installations, ambient occlusion, hall of mirrors, ethereal chocolate, minimalist trunks, luminous spheres, gender of scale

1 2 3 4

/describe ▶

Download Modify Save

Does this look familiar to you?



Does this look familiar to you?

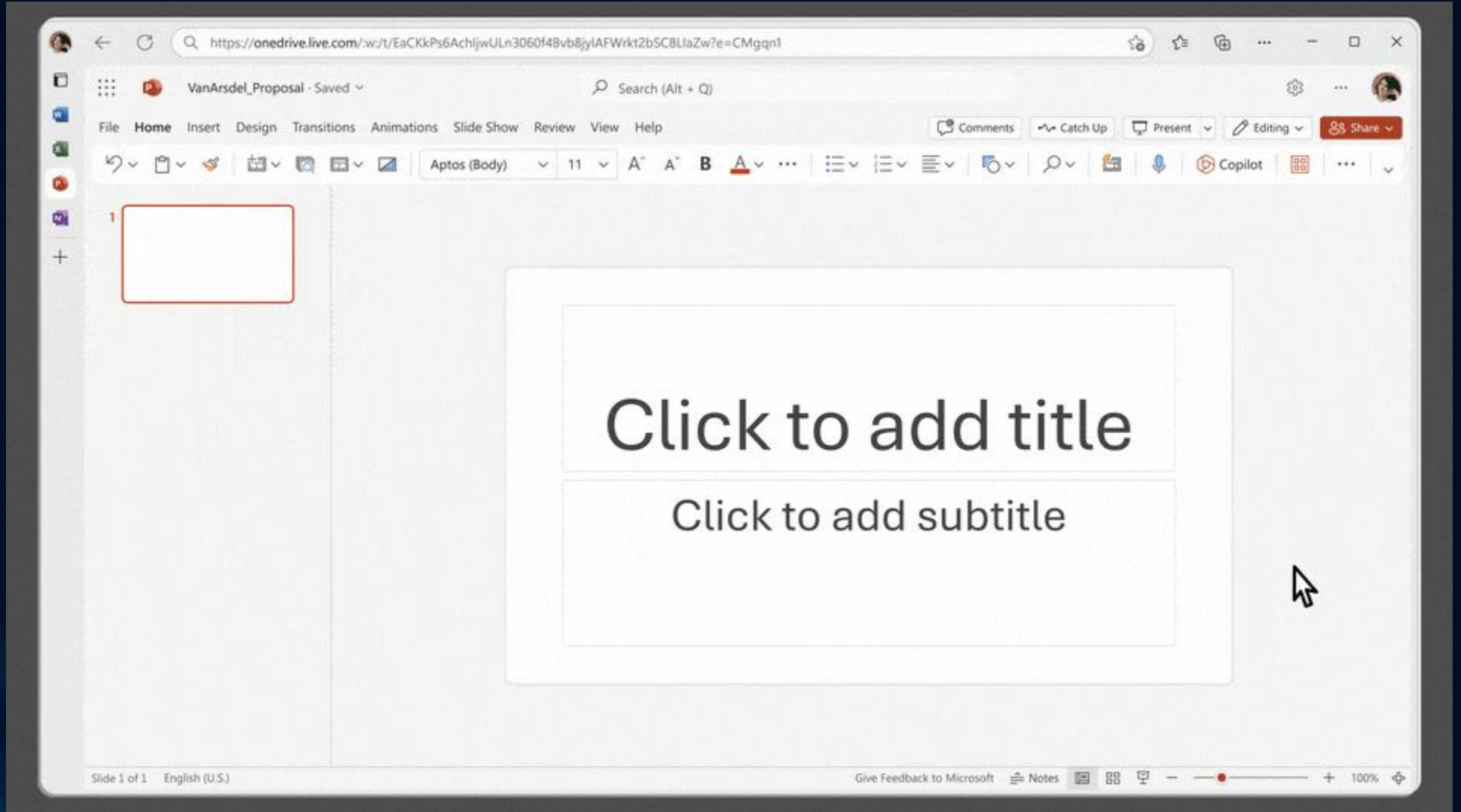


how to merge dictionaries in python?



To merge two dictionaries in Python, you can use the ``update()``

Does this look familiar to you?



Text-to-image AI Art Generators

Comparative visuals from example prompt across platforms



Dall-E



Midjourney

Prompt: *A photograph of a dirt rally race car covered in dirt, moving at a fast speed, dirt flying, mountain scene, award-winning photography, soft shadows, and cinematic lighting.*

Generative AI Art Side-by-Side Comparison

Prompt: *high-quality photograph of a dog astronaut, star and planet backdrop, cinematic lighting, helmet with face glass, wide angle.*



DALL-E 2



MIDJOURNEY



LEONARDO

The Future is here

What you just saw might seem like the future, but it's already transforming our present.

Few Examples of what is achievable by Generative AI

“Image Outpainting”



An Innovative Ad Film

Nestle

An A.I. project from Nestlé reframes "The Milkmaid," a masterpiece created more than 360 years ago by Dutch painter Johannes Vermeer, in a fresh, expanded light.

The work from Ogilvy Paris promotes La Laitière yogurt, which has long used milkmaid imagery in its campaigns.

For this effort, the team employed DALL-E 2's new "Outpainting" function to add significant details and new characters to the original oil-on-canvas image



Handling Visual Input



GPT-4 supports reasoning from imagery.

Handling Visual Input



It's Wednesday at 4 pm. Can I park at this spot right now? Tell me in 1 line.



Yes, you can park for up to 1 hour starting at 4 pm.



So How Will Technology Help in Design?

Prompt

hot pink-style (attic interior) . with hot pink barbie walls and barbie closet and barbie sofa and barbie bold rosy hues like fuchsia and magenta and barbie plastic interior and barbie chairs and hot pink barbie colors and barbie style interior.



Creating Video from Text

Prompt

A woman wearing green dress and a sun hat taking a pleasant stroll in Mumbai, India during a rainy day



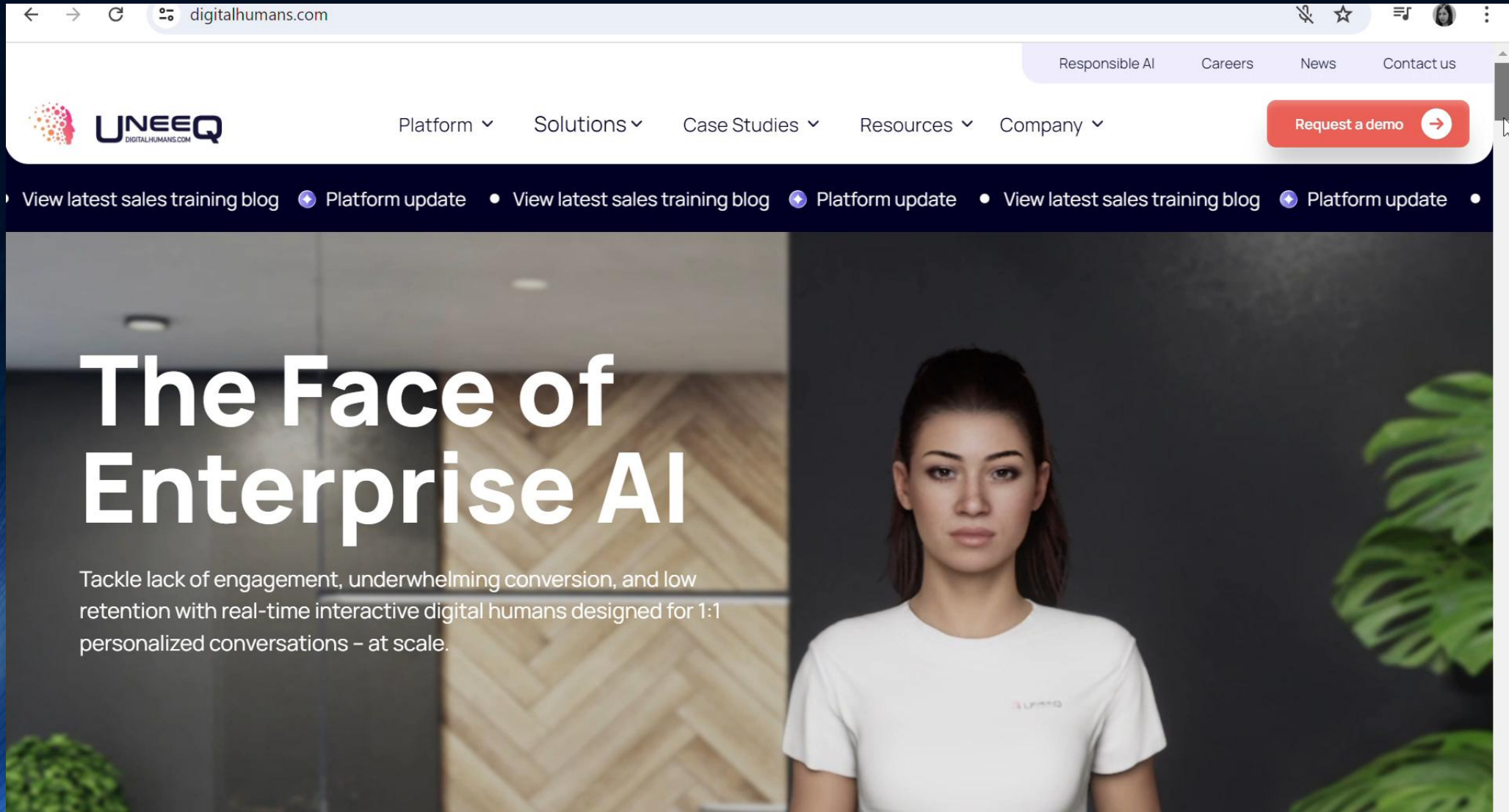
Creating Video from Text

Prompt

A litter of golden retriever puppies playing in the snow. Their heads pop out of the snow, covered in.

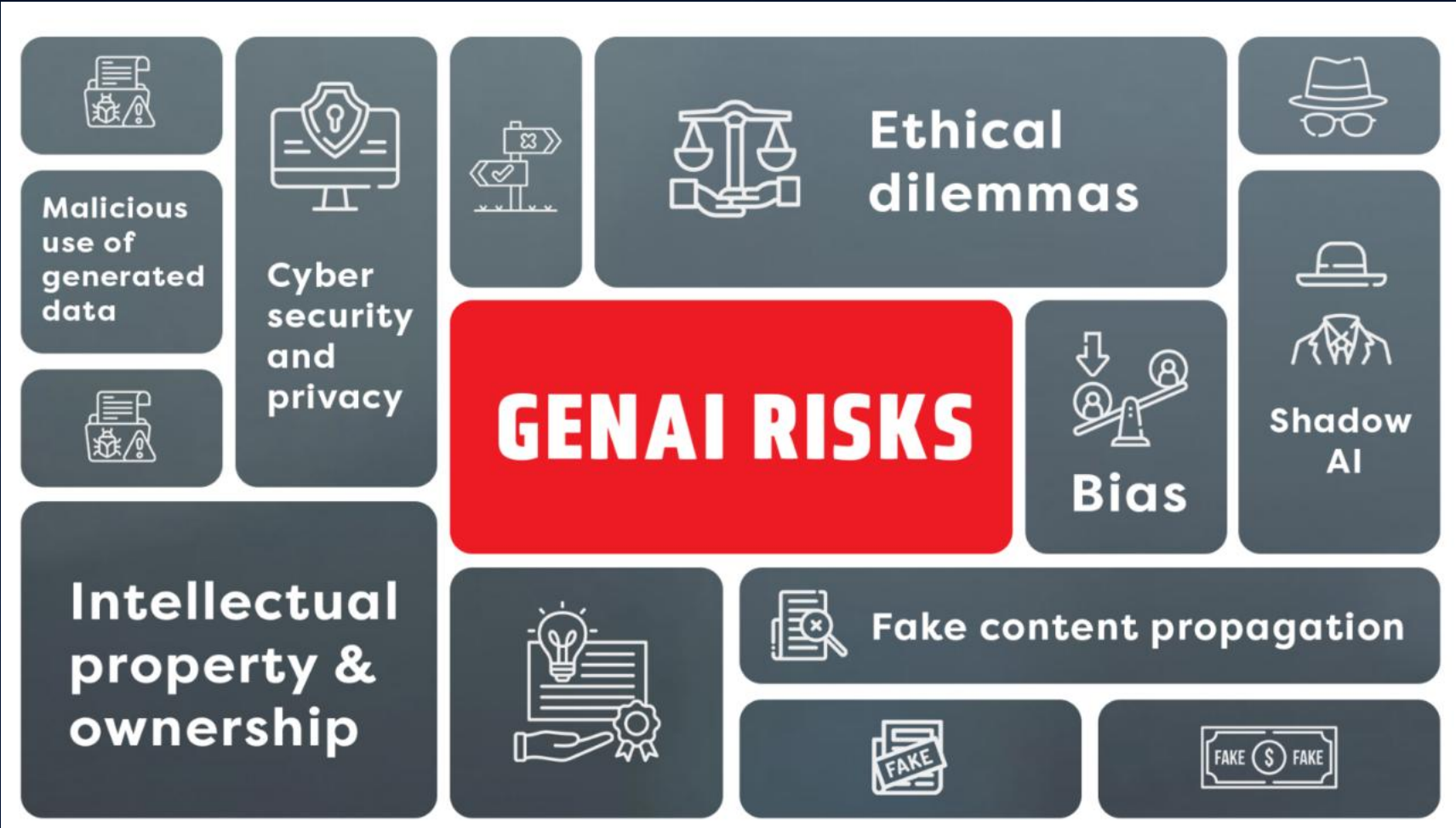


Natural Language Conversation with Digital Humans



The screenshot shows a web browser at the URL digitalhumans.com. The navigation bar includes links for Responsible AI, Careers, News, and Contact us. The main navigation menu contains Platform, Solutions, Case Studies, Resources, and Company, along with a red 'Request a demo' button. A dark banner below the navigation features a repeating sequence of links: 'View latest sales training blog' and 'Platform update'. The hero section features a digital human avatar of a woman in a white t-shirt. The text reads: 'The Face of Enterprise AI' and 'Tackle lack of engagement, underwhelming conversion, and low retention with real-time interactive digital humans designed for 1:1 personalized conversations - at scale.'

Generative AI Adoption Risk



Evidence of Positive Business Impacts from Generative AI



Globally...

Potential productivity lift

Retail and consumer packaged goods

1-2%
Of global industry
revenue

**~\$400B-
\$660B**



Key driver

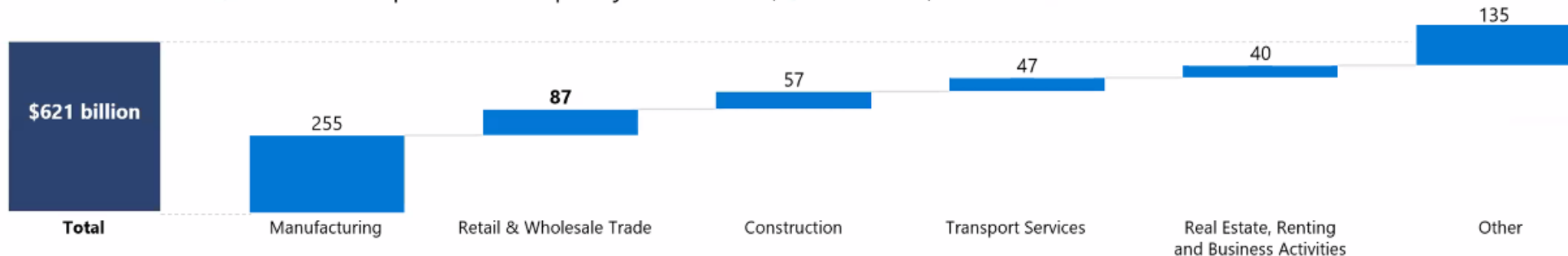
Gen AI's increased ability to understand natural language augments the productivity of work activities that account for 25% of total work time (e.g., customer service officer answering customer queries, marketer updating marketing collaterals)

Source: [McKinsey \(2023\) Economic Potential of Generative AI](#)



In India...

Gen AI can unlock **\$621 Billion** in productive capacity – 9% of this, **\$57 Billion**, will come from the **construction** sector



Source: [Access Partnership \(2023\) The Economic Impact of Generative AI: The Future of Work in India](#)



Generative AI's Double-Edged Potential Creativity vs. Fabrication

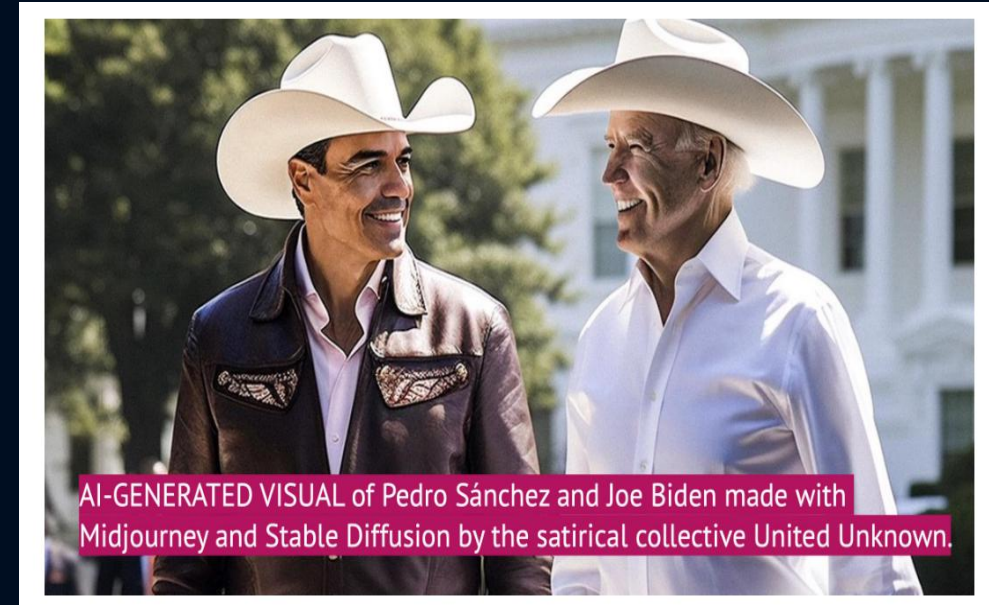
Creativity Unleashed

- Generative AI enables unprecedented creativity by producing unique artwork, designing innovative products, and generating content at scale.
- Industries are leveraging AI to create personalized experiences, from virtual assistants to tailored marketing campaigns.
- **Example:** AI-generated art and music that redefine cultural and artistic boundaries.



The Flip Side – Fabrication

- The same technology used for creative pursuits can fabricate convincing fake content, including false narratives, deepfake videos, and manipulated images.
- Fabricated information undermines trust, spreads misinformation, and erodes societal cohesion.
- **Example:** AI-generated fake news during elections causing public confusion.



Generative AI and Large Language Models (LLMs)

GENERATIVE AI: A MULTI-MODAL POWERHOUSE

Capabilities Across Modalities:

- **Text:** Automates content creation, generates summaries, and drafts articles.
- **Audio/Video:** Creates deepfake videos, synthesized voices, and soundtracks.
- **Images:** Produces realistic photos and artistic creations through tools like DALL·E and MidJourney.

OUR FOCUS: (LLMS)

Why Text-Based Fabrication?

- Text is the backbone of most digital information, influencing public opinion, policy, and trust.
- Fake news, biased narratives, and fabricated reports are the most pervasive forms of misinformation.

Generative AI and Large Language Models (LLMs)

WHAT IS GENERATIVE AI?

- **Definition:** AI systems designed to create new content like text, images, and music.
- **Applications:** Writing assistance, art creation, code generation, and more.
- **Example:** AI-generated art or personalized marketing content.

WHAT ARE LARGE LLMS?

- **Definition:** Advanced AI models trained on vast amounts of text data to understand and generate human-like text.
- **Key Features:**
 - Contextual understanding and coherent content generation.
 - Versatile across languages and domains.
- **Examples of LLMs:**
 - *OpenAI's GPT Series:* Powers chatbots like ChatGPT.
 - *Meta's Llama-3:* Instruction-tuned for helpfulness and safety.
 - *Google's Gemma-1.1:* Focused on ethical content generation.

Why LLMs?

- **Impact on Society:**

- ✓ Textual misinformation spreads faster and wider due to its shareability.
- ✓ Influences critical domains like politics, healthcare, and education.

- **Actionable Insights:**

- ✓ LLMs are uniquely positioned to both create and detect text-based misinformation.
- ✓ Understanding their dual roles is essential for leveraging AI responsibly.

- **Practical Applications:**

- ✓ Focused on addressing real-world challenges in combating fabricated content in digital ecosystems.



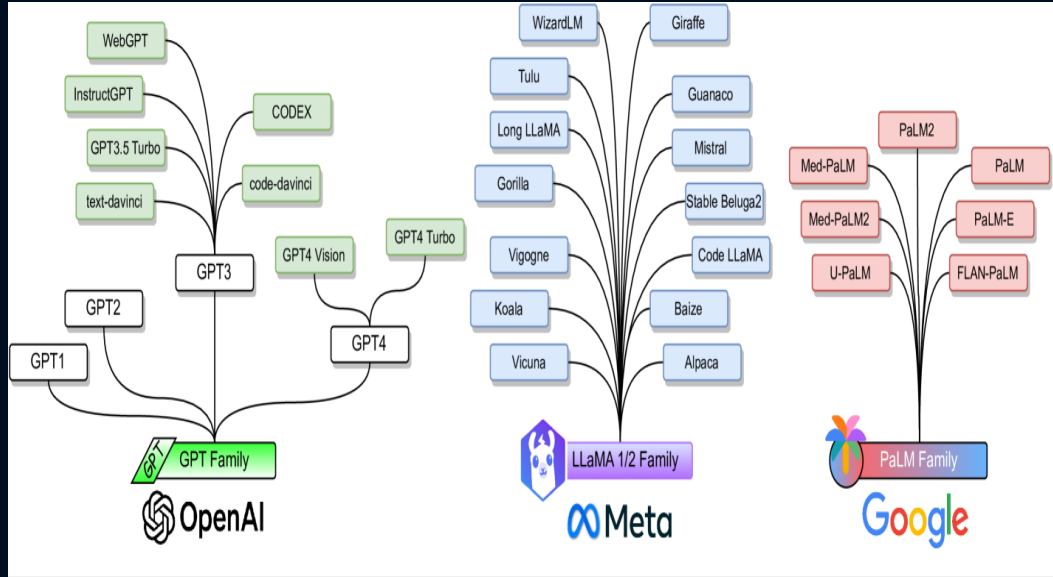
Role of Natural Language Processing (NLP) in LLMs

- **Core Foundation:** NLP techniques enable LLMs to process and analyze human language effectively.
- **Key NLP Techniques:**
 - **Text Tokenization:** Breaking down sentences into words or subwords for analysis.
 - **Sentiment Analysis:** Understanding the emotional tone of text.
 - **Named Entity Recognition (NER):** Identifying proper nouns like names, locations, and organizations.
 - **Text Summarization:** Condensing lengthy content into concise summaries.





Generative AI



LLMs




Information
Fabrication
(Falsification)

Can You Spot the Fake News?

← Exit

Goldsmiths
UNIVERSITY OF LONDON

How to participate?



- 1 Go to wooclap.com
- 2 Enter the event code in the top banner

Event code
RFPOLL

Enable answers by SMS

Copy participation link

wooclap

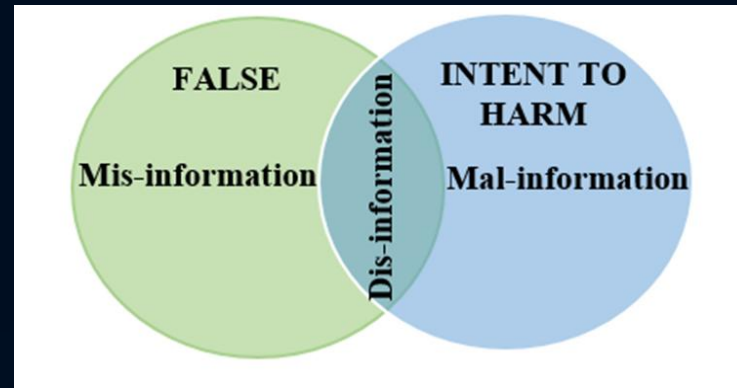
Questions - / 13 + Messages 100 %

0 👤



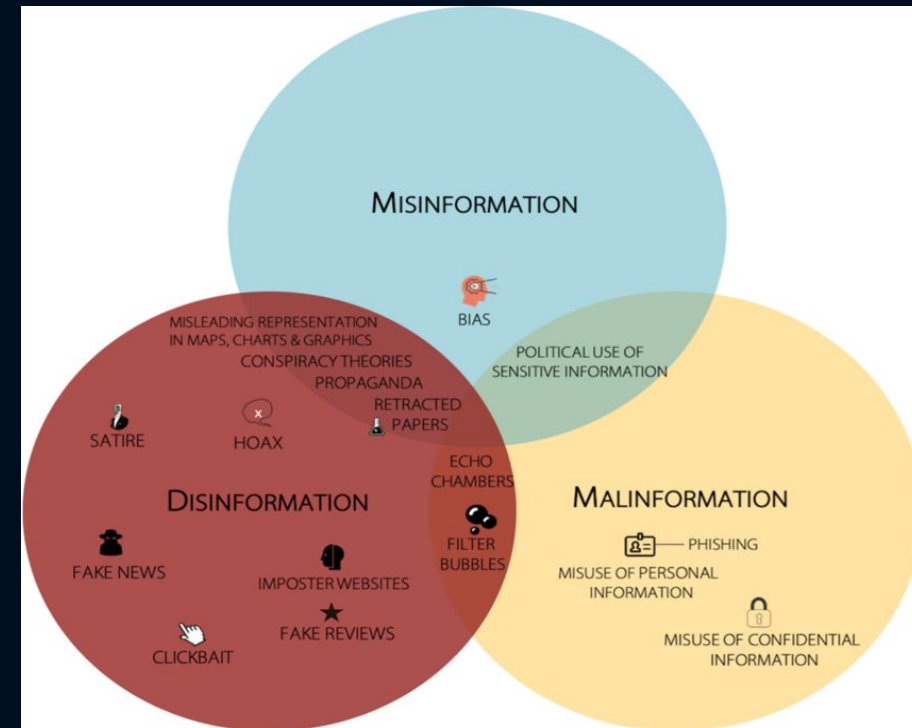
Online Information Fabrication = Information Pollution

Vary in accordance with the **truth value of the content** and the **intent of information** being created, produced or distributed :



- **Misinformation** (honest mistakes)
- **Disinformation** (rumours, fake news and manipulated content)
- **Mal-information** (information leaks, harassment and hate speech)

- **Misinformation** though the information **is false**, but it is **not** created with the intention of **causing harm**, rather it is an **erroneous mistake**.
- **Dis-information** contains **outright lies** with **no** element of **truth** and is **deliberately** created to harm a person, social group, organization or country.
- **Mal-information** is **grounded on reality** but is either completely non-contextual or **manipulated** with a **malicious** intent to **harm** and **damage** an individual or society.



COVID-19: Misinformation

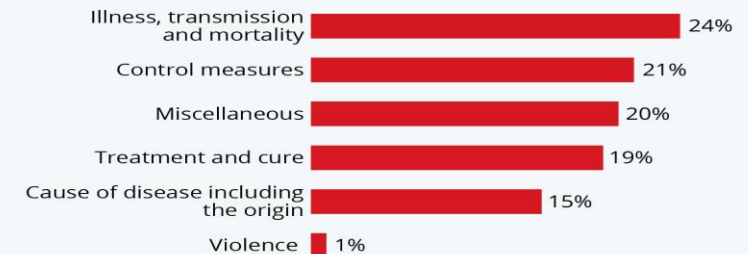
There have been misleading claims around methods against, such as

- *“drinking lemon with water”, or “self-testing for coronavirus by holding your breath for 10 seconds”*
- Generally, people share these advices **under the belief** that they are safe
- have **no intention to lie.**



The Composition Of Coronavirus Misinformation

Composition of Covid-19 rumors, stigma and conspiracy theories circulating on social media/online news platforms*



* Based on 2,311 reports in 25 languages from 87 countries between Dec 31, 2019 and Apr 15, 2020.
Source: American Journal of Tropical Medicine and Hygiene

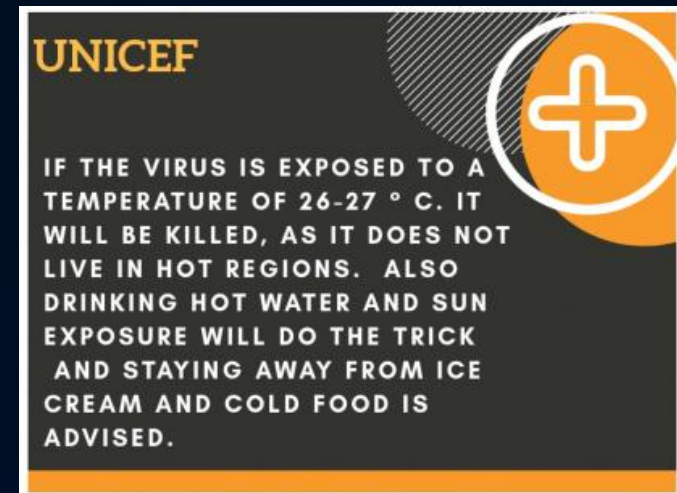




COVID-19: Disinformation



- In January 2020, UNICEF provided some information on the coronavirus which stated, *“If the virus is exposed to a temperature of about 26 or 27-degree Celsius, it will be killed as it does not live in hot regions.”*
- **UNICEF did not make any such claims** pertaining to the coronavirus.



COVID-19: Mal-information



In April'2020, Tablighi Jamaat has been blamed for contributing to the spread of the coronavirus not just India but in Malaysia and Pakistan and backlashed Muslims

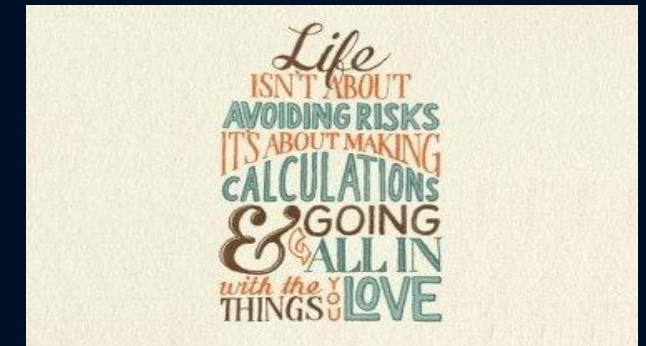
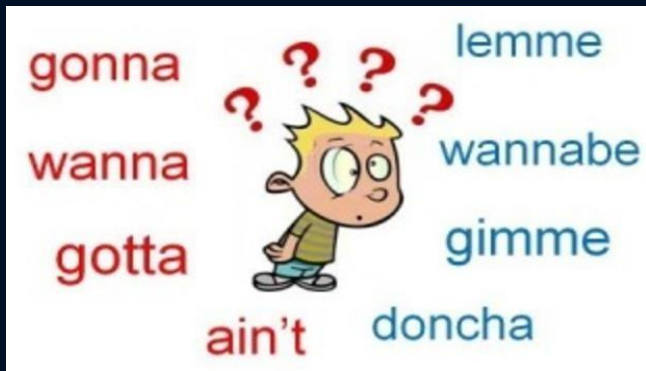
- Although popular social media platforms have been stepping up with measures
 - changing policies
 - tweaking algorithms
- But information overload and real-time detection add to the long list of challenges.
- The **multimodal user-generated content** (text, images, audio, videos, GIFs, and their combinations) adds to the complexity.
- **Speak-text** (word play, slangs), **emoji & emoticon-based pseudo-languages** and diversity in languages based on audience demographics.



What is our speak-text today?

Early Methods:

- ✓ Fake news creation initially relied on basic methods like word shuffling and random substitutions in real news articles.
- ✓ These methods produced incoherent content, easily identifiable by human readers.



Common Linguistic Cues in Fake News

Sensational Language

Definition: Over-the-top wording designed to provoke strong emotional reactions.

Examples:

"SHOCKING revelation!"

"BREAKING: World-ending event incoming!"



Common Linguistic Cues in Fake News

Excessive Capitalization

Purpose: Draw attention and exaggerate the importance of the content.

Examples:

"THIS WILL CHANGE EVERYTHING!"

"SECRET TRUTH REVEALED!"



Common Linguistic Cues in Fake News

Exaggeration and Hyperbole

Hyperbole

Hyperbole also known as **exaggeration**, is when an **unrealistic comparison** is made for effect.

Examples

His brain is the **size of a pea**. ✓

I have told you a **million times** not to lie! ✓



Purpose: Claims that are too extreme to be true, often unsupported.

Examples:

"Cure for all diseases discovered!"

"AI will make humans extinct by 2030!"



Common Linguistic Cues in Fake News

Emotional Triggers

Purpose: Use of fear, anger, or excitement to manipulate readers.

Examples:

Fear: "Your children are at risk from this deadly product!"

Anger: "Government caught lying to citizens again!"



Common Linguistic Cues in Fake News

Lack of Specific Evidence

Purpose: Vague references to experts or studies without proper citations.

Examples:

"Experts say this is the best solution."

"A leading scientist claims..."



Common Linguistic Cues in Fake News

Clickbait Phrases

click·bait

(noun)

: something (such as a headline) designed to make readers want to click on a hyperlink especially when the link leads to content of dubious value or interest

Purpose: Encourage clicks by withholding key information.

Examples:

"Doctors don't want you to know this one simple trick!"

"You won't believe what happens next!"

**CLICK
BAIT**

Common Linguistic Cues in Fake News

Use of Anonymous Sources

Purpose: Create the illusion of credibility without accountability.

Examples:

"An insider reveals shocking truths."

"A whistleblower claims..."



From Deception to Detection: The Dual Roles of Large Language Models in Fake News

- Can LLMs easily generate biased fake news?
- Can we use LLMs to detect fake news, and do they outperform typical detection models?

LLMs are both tools and threats in combating fake news.

Can LLMs generate Fake News?

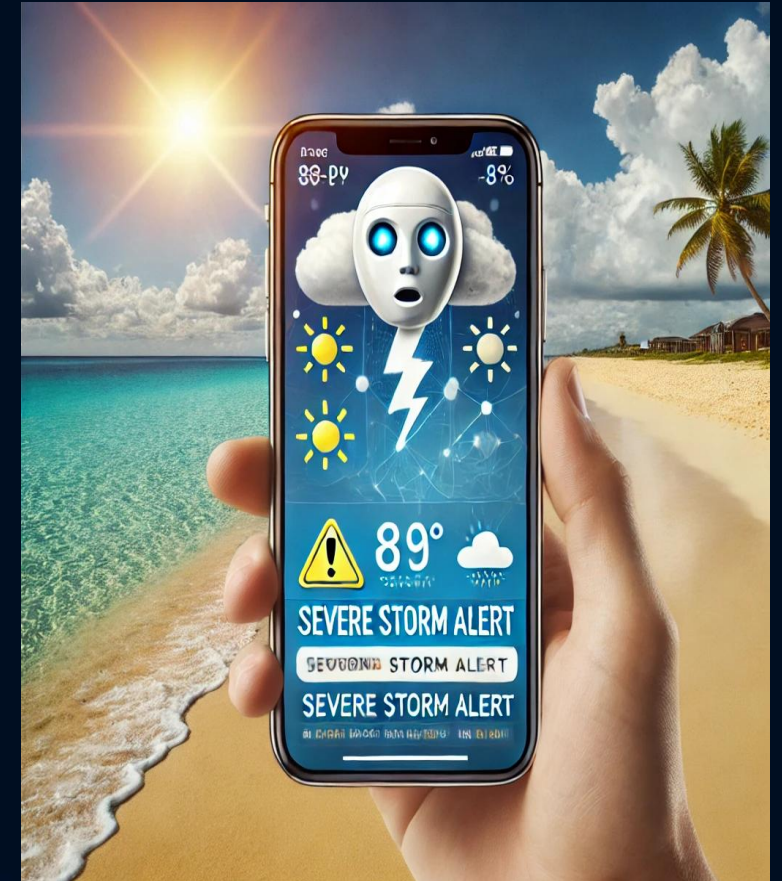
- LLMs:

- ✓ generate content with greater coherence and credibility.
- ✓ Techniques now integrate real news elements with fabricated information to create convincing articles.
- ✓ LLMs vary in their ability to generate fake news, influenced by their size, training, and safety protocols.
- ✓ Some models readily generate fake news, while others are restricted by safety measures.

Types of Fake News Generated by LLMs

Hallucinated Fake News

- **What is it?**
 - ✓ Non-factual content created by LLMs due to outdated or incomplete information.
 - ✓ Happens unintentionally, especially in applications needing real-time data.
- **Example:**
 - ✓ A weather update generated by an LLM that predicts a storm in a region where it's sunny.
 - ✓ A fake historical fact, like "The Eiffel Tower was built in 1900," when the real year is 1889.



Types of Fake News Generated by LLMs

Arbitrary Fake News

- **What is it?**
 - Intentionally created fake news prompted by malicious users.
 - Designed to spread misinformation or influence opinions.
- **Example:**
 - A fabricated headline like, "New law bans all electric cars starting 2025," when no such law exists.
 - A false claim during elections, such as "Candidate X drops out of the race," when it isn't true.



Models Used in Fake News Detection and Generation (Sallami et al. 2024)

Model	Parameters	Key Features	Safety Protocols
Phi-3	3.8B	Lightweight, instruction-following, safety measures.	Moderate adherence
Gemma-1.1	7.24B	Fine-tuned with RLHF, focuses on ethical content generation.	Strong adherence
Mistral	7.3B	Uses advanced attention mechanisms for effective context focus.	Limited adherence
Llama-3	70B	Instruction-tuned with supervised fine-tuning and RLHF for helpfulness.	Strong adherence
C4AI	104B	Optimized for sophisticated reasoning and question answering.	Limited adherence
Zephyr-orpo	8 x 22B (Mixture-of-Experts)	Fine-tuned for synthetic chat and reasoning tasks.	Limited adherence
GPT-4	Not disclosed	Extensive safety protocols, aligned with expert feedback for secure responses.	Strongest adherence

Can LLMs Generate Fake News?

- **No Restriction Models:** C4AI, Zephyr-orpo, Mistral: Generate fake news easily without safety protocols.
- **Moderate Restriction Models:** Phi-3, Llama-3: Some hesitation but still generate fake news in specific contexts.
- **Strict Adherence Models:** GPT-4, Gemma-1.1: Refuse fake news generation due to robust safety measures.

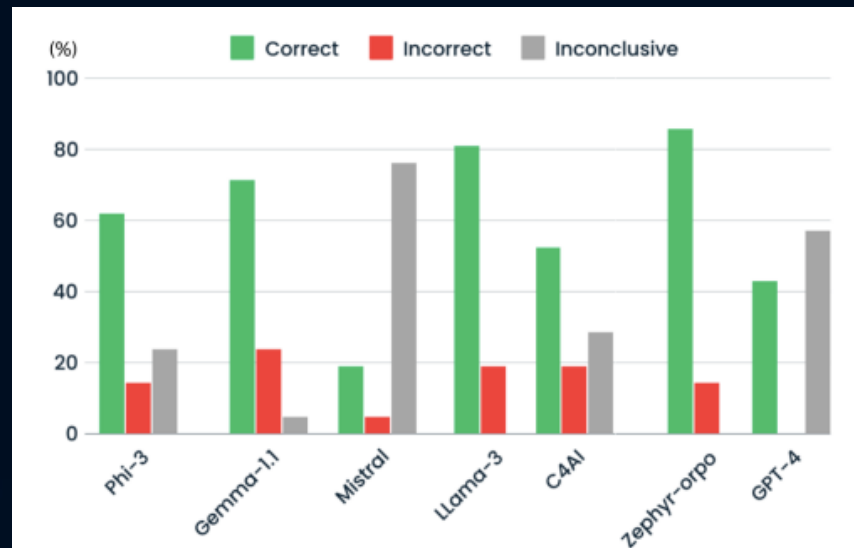
Key Insight

Larger models and better-trained systems show more restraint.

Model
Phi-3
Gemma-1.1
Mistral
Llama-3
C4AI
Zephyr-orpo
GPT-4

Can we use LLMs to Detect Fake News?

- Human-Created Fake News Detection



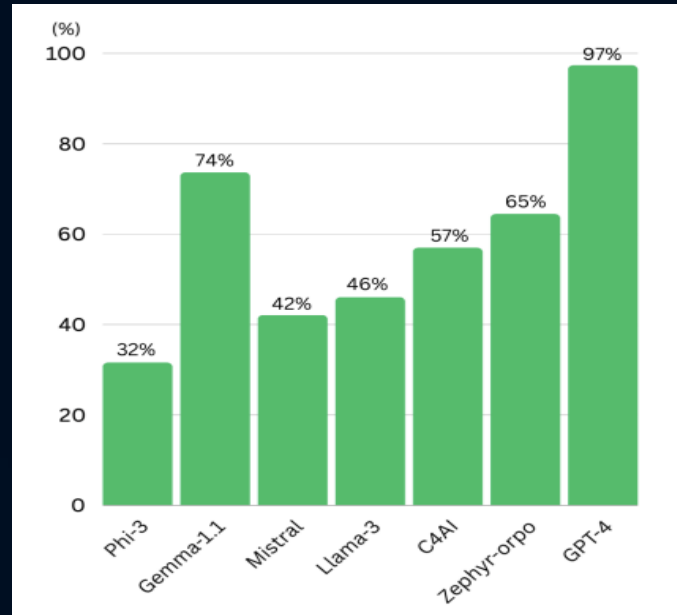
Most models struggled to provide definitive classifications for real vs. fake news.

- **Top Performers:** Llama-3 & Zephyr-orpo
- **Underperformers:** Mistral
- **Notable Performance:** GPT-4: High accuracy with no errors in correct classifications.

However, frequent inconclusive outcomes present challenges for user decision-making.

Can we use LLMs to Detect Fake News?

- LLM-Generated Fake News Detection



- Larger models outperform smaller ones in accuracy.
- GPT-4 excels but returns inconclusive results due to cautiousness.

LLM

Lies, Logic, and Media

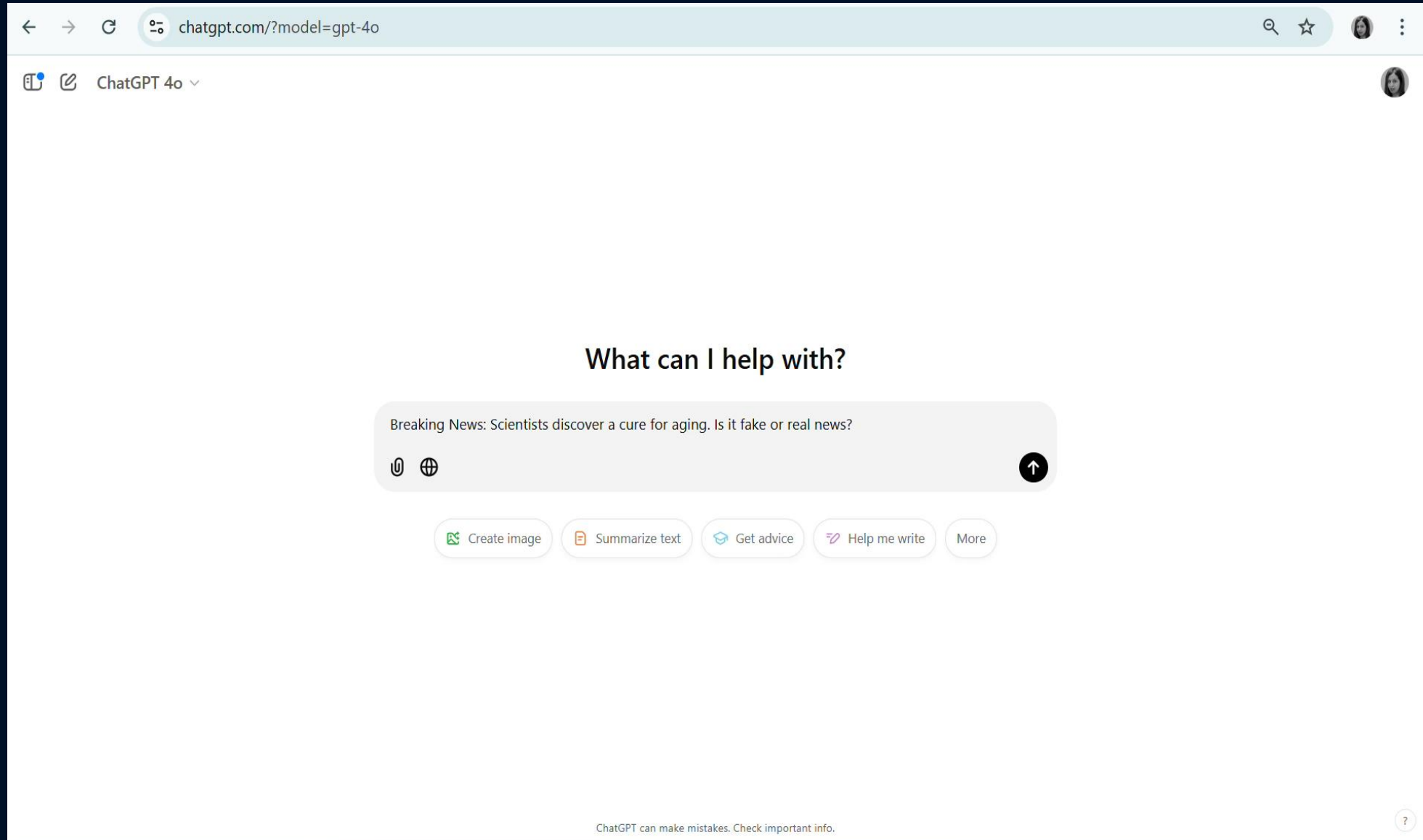
The three forces at play in tackling fake news with AI.

- While **Lies** spread fast,
- **Logic** can deconstruct them, and
- **Media** becomes the battleground for truth.

Fact-Checking Platforms Leveraging LLMs

- **PolitiFact + LLMs**
 - Combines advanced NLP for contextual analysis of political claims.
 - **Impact:** Reduced claim verification time by 50%.
- **Snopes' Partnership with AI Tools**
 - Employs LLMs for linguistic pattern analysis in viral misinformation.
 - **Impact:** Increased accuracy in debunking hoaxes and urban myths.

LLM Detecting Fake News



Free tools for Live Checking

Google Fact Check Tools



Explorer

Markup Tool

APIs

FAQ

Report Issue

Search fact checks about a topic or person

Search by image

More results in other languages

Language filter

English

Recent fact checks



Claim by Dave Anthony:

Facebook users must post a statement on their profiles to bar Meta and its generative artificial intelligence (AI) models from using their “photos, information, messages or posts, past or future.”

Dave Anthony

Meta

Artificial intelligence

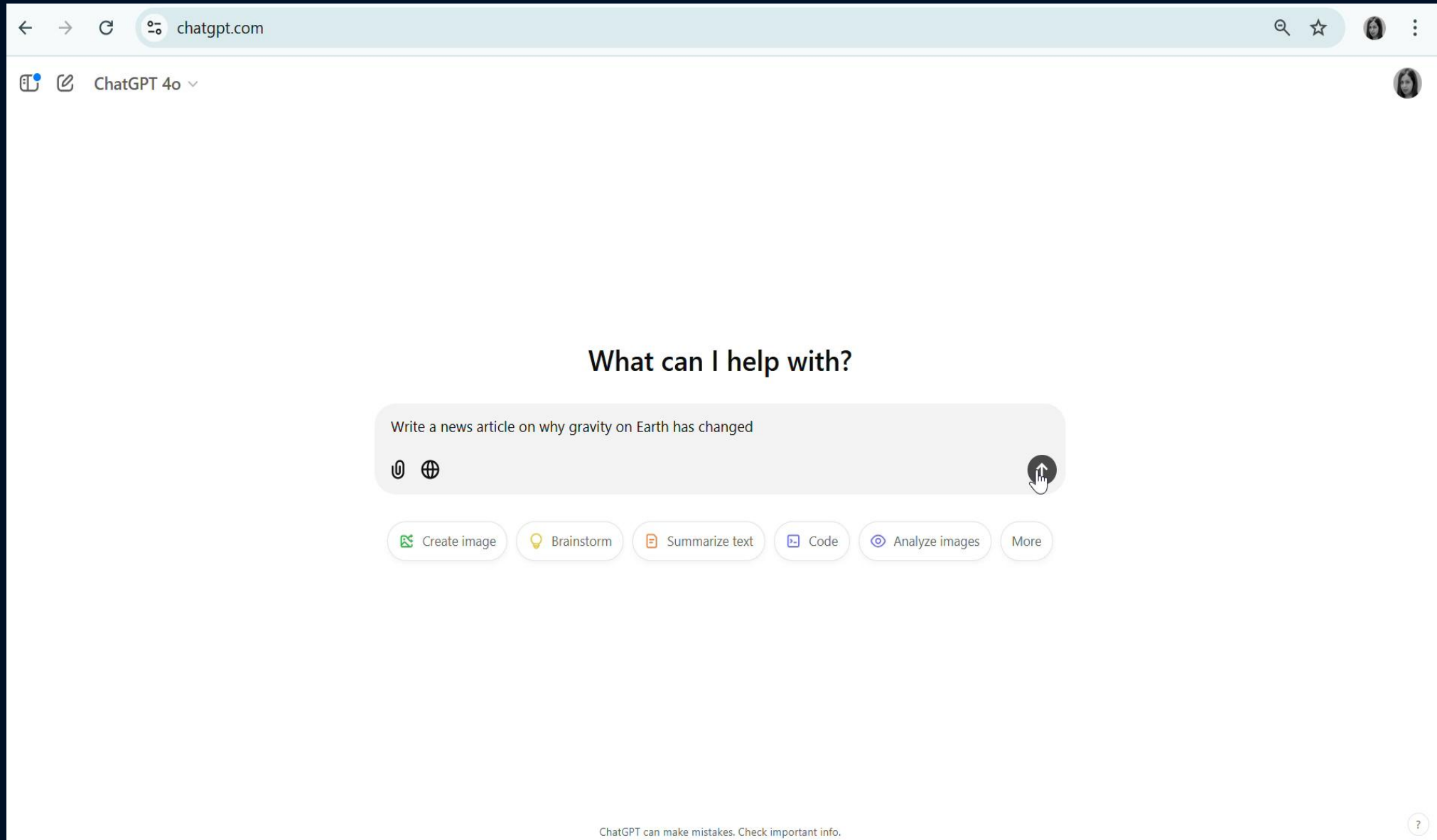
Rappler

Rappler rating: False

[FACT CHECK: ‘Goodbye Meta AI’ notice won’t protect users’ Facebook data](#)

9 hours ago

LLM Generating Fake News



Ethical Dilemmas

- Generation of **biased or harmful text** due to unfiltered training data.
- Risk of producing **misleading information** resembling credible sources.
- Exploitation for creating **targeted disinformation campaigns**.

Societal Implications

- **Polarization** through the spread of divisive narratives.
- Undermining **democratic processes** by influencing public opinion with fabricated content.
- Difficulty in distinguishing **human-authored vs. AI-generated text**, eroding trust.

Frameworks for Ethical Implementation

- **Training Data Transparency:**
 - Publicly document sources of training datasets to identify potential biases.
- **Content Moderation Protocols:**
 - Implement rigorous safety layers to filter harmful or misleading outputs.
- **Human Oversight:**
 - Include human reviewers for sensitive applications, such as news generation.

Existing Regulations

- **EU AI Act:** Requires LLM providers to disclose training methodologies and safeguard against harmful content.
- **FTC Guidelines (USA):** Regulates deceptive AI-generated advertising or content.
- **India and the United Kingdom** are actively developing regulatory frameworks to address the ethical and societal implications of Large Language Models (LLMs) and generative AI.

Policy Recommendations

- **Labeling and Transparency:** Require LLMs to clearly indicate AI-generated text with visible markers (e.g., watermarks or disclaimers).
- **Bias Audits:** Periodic evaluations to assess and mitigate biases in text generation.
- **Incentives for Ethical Use:** Encourage organizations to adopt trustworthy AI certifications for LLMs.

Research Opportunities in LLMs and Information Integrity

Bias and Fairness in LLM Outputs

- **Research Gap:** Understanding and mitigating inherent biases in LLMs caused by imbalanced or incomplete training data.
- **Key Questions:**
 - How can LLMs detect and self-correct bias in real-time?
 - What methodologies ensure diverse and inclusive datasets for training?
- **Potential Applications:** Ethical content creation, unbiased news reporting, and equitable AI systems.

Contextual Misinformation Detection

- **Research Gap:** Developing techniques for LLMs to detect misinformation embedded in nuanced or contextual narratives.
- **Key Questions:**
 - How can LLMs better understand context and intent in fabricated information?
 - What role do multi-modal inputs (text + metadata) play in improving detection accuracy?
- **Potential Applications:** Fact-checking platforms, real-time content moderation, and combating election misinformation.

Real-Time Adaptation in Dynamic Environments

- **Research Gap:** Enabling LLMs to adapt their knowledge base to rapidly changing real-world scenarios without retraining.
- **Key Questions:**
 - Can LLMs be fine-tuned incrementally for emerging events (e.g., breaking news)?
 - How can continual learning approaches improve LLMs for misinformation detection?
- **Potential Applications:** News agencies, disaster response systems, and social media moderation.

Explainability and Accountability in Text Generation

- **Research Gap:** Improving the transparency of LLMs to explain why certain outputs were generated.
- **Key Questions:**
 - How can LLMs provide human-understandable explanations for generated text?
 - What frameworks ensure accountability for harmful or misleading content produced by LLMs?
- **Potential Applications:** Regulatory compliance, legal audits of AI outputs, and user trust in AI.

Cross-Cultural and Multi-Lingual Fabrication

- **Research Gap:** Addressing challenges in detecting and preventing misinformation across languages and cultural contexts.
- **Key Questions:**
 - How do LLMs handle cultural nuances and idiomatic expressions in misinformation?
 - What techniques improve multi-lingual detection capabilities?
- **Potential Applications:** Global media monitoring, international policy-making, and education.

Next Steps and Vision for the Future

Collaborative Pathways

- **Call to Action:**

- ✓ Foster collaboration between **academia, industry, and governments** to address open research challenges.
- ✓ Establish **international research networks** for creating unbiased, diverse datasets.

- **Proposed Partnerships:**

- ✓ Work with **fact-checking organizations** for real-world testing and validation.
- ✓ Partner with **social media platforms** to integrate LLM-based misinformation detection tools.

Final Thoughts

The future of AI lies not just in its power to generate, but in its capacity to discern truth and uphold trust.

Let us harness the power of LLMs—not just as large language models, but as tools for Logical Literacy in Media—to redefine how we combat fake news.

Together, we can build a future where truth is amplified, trust is restored, and technology serves humanity with integrity.

Large Language Models



Lies, Logic, and Media

Logical Literacy in Media

Generative AI + Human Expertise =
Human Greatness

Thank You

AI Generated Images



A butterfly with rainbow wings landing on flower



A sailboat made of origami paper floating towards tropical islands



Colourful splashes of paint, geometric, abstract art



Serene vacation lake house water colour painting