A critical approach to risk mitigation strategies for the use of generative AI in journalism practice

Laurence Dierickx
DDC x TrygFonden Fellow
NORDIS (University of Bergen)

Democracy & Digital Citizenship Conference Series
University of Southern Denmark, September 4, 2024

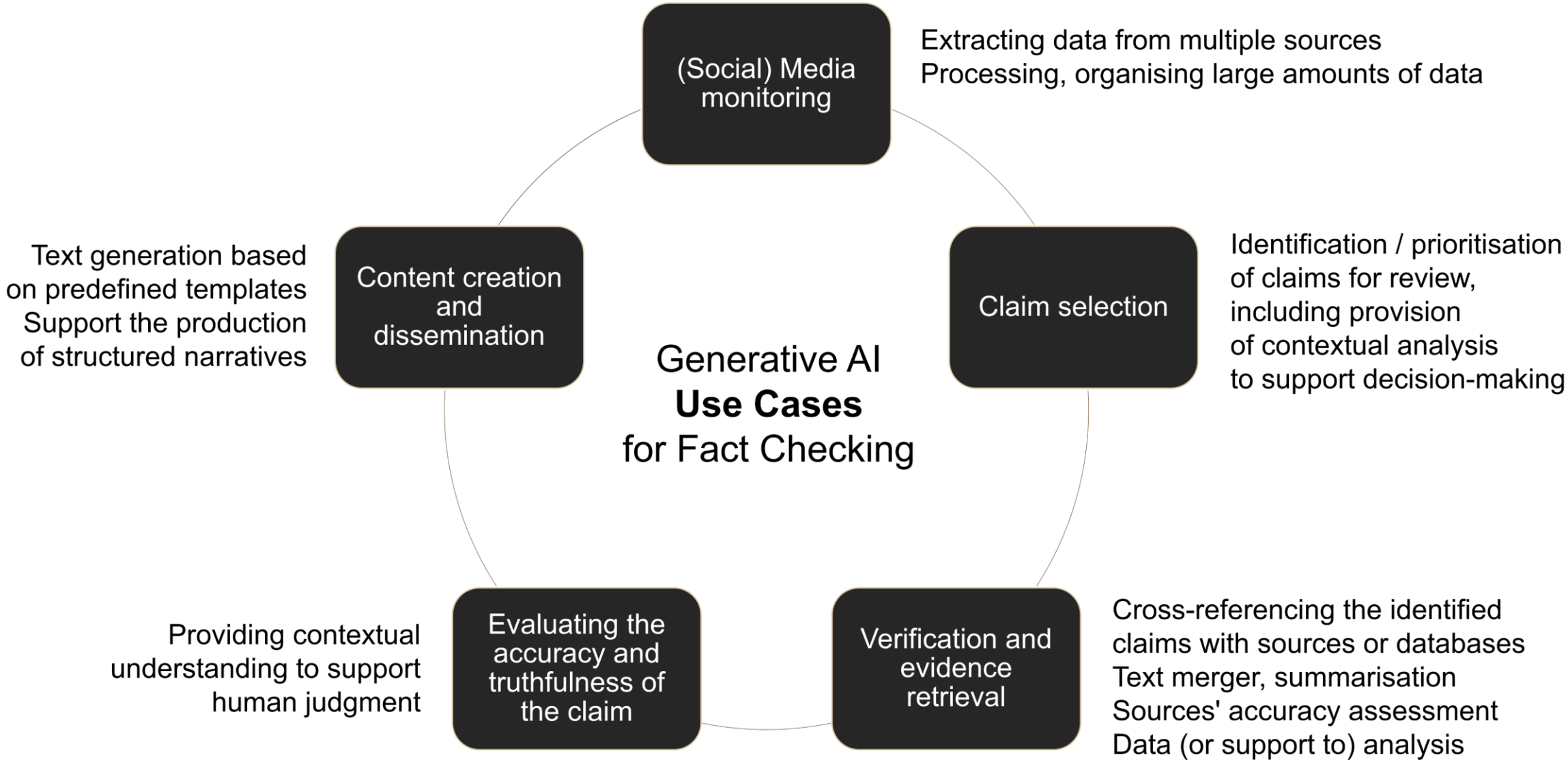
LLMs are exciting but challenging tools for being used in journalism and fact-checking!

NORDIS

UNIVERSITAS BERGENSIS

DDC digital democracy centre

SDU University of Southern Denmark



NORDIS

UNIVERSITAS BERGENSIS

DDC digital democracy centre

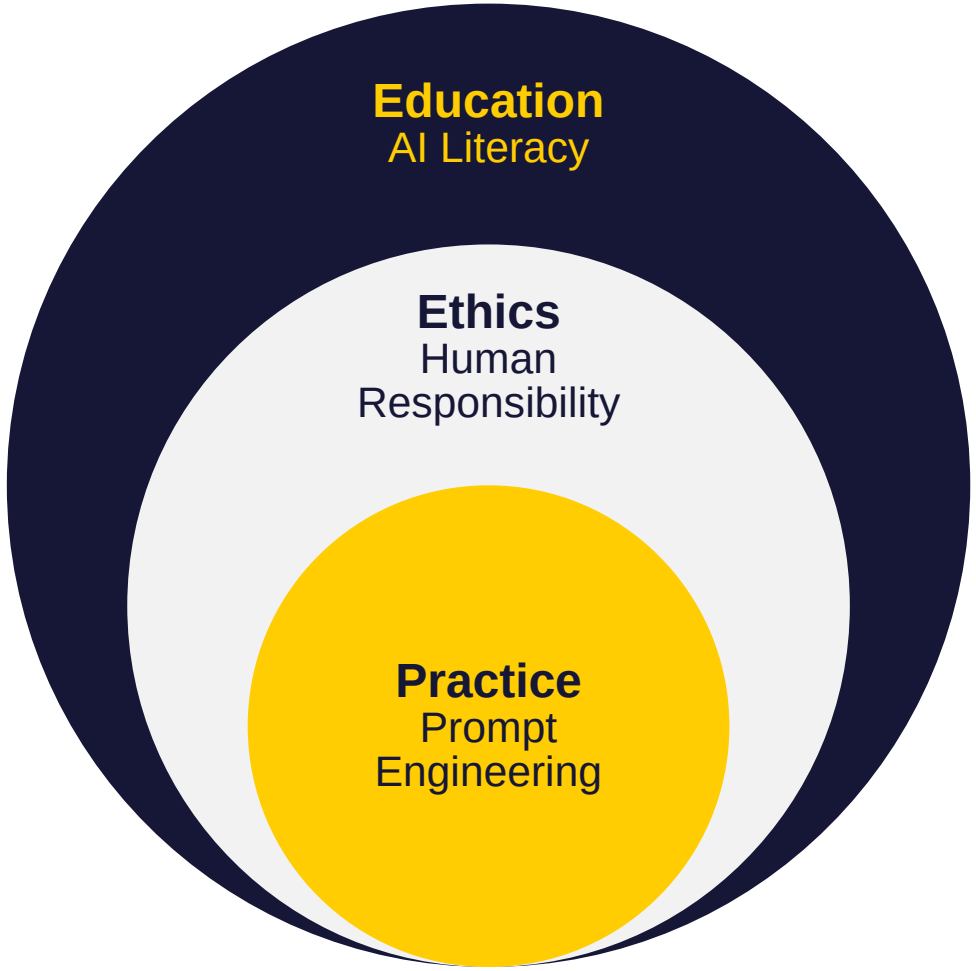
SDU University of Southern Denmark

Question & methods

Given the ethical challenges and limitations of these technologies, what is the potential for professional fact-checking, and how can the risks be mitigated?

Narrative literature review: qualitative research tool used strategically to allow flexibility in exploring different methodologies, aims to provide a comprehensive overview of existing knowledge in an emerging field.

Three complementary strategies



Strategy 1: Education

AI literacy relates to the skills necessary for the competent and meaningful usages of AI tools and consists of a combination between knowledge and experience.

Being well equipped and developing critical mindset to prevent further risks.

Strategy 1: Challenges

- How do you build effective AI literacy programmes that consider the need to explain complex concepts in a practical way?
- How to reach all professionals, given time and resource constraints?
- How to address the need for ongoing training to keep up with rapidly evolving technologies?

Strategy 2: Ethics

EU perspective: pyramid of risks
(no risks = no impact on information quality)

Generating is not verifying
(biases, hallucinations, failures in deduction)

Promoting transparency, human oversight and responsibility

Strategy 2: Challenges

Transparency is not enough!

Not a warrantee for accuracy and reliability, does not mean explainability, does not equal, responsibility and accountability, can obscure the complexity of decision-making processes, can lead to information overload and technical jargon might not be accessible for non-experts.

Importance of human oversight.

Strategy 3: Practice

Prompting has been much studied in CS (less resource intensive, keep generalisability and model performances, avoid overfitting) but less from non-expert views (try-errors), whereas research demonstrated that well-crafted prompts can increase explainability and reduce the generation of fabricated content. Facilitates user interaction and problem-solving.

Strategy 3: Challenges

Poorly designed prompts can steer the system towards biased outcomes, compromising the credibility of journalistic reporting.

Mostly developed in computer science and do not consider non-experts and specific end-users.

How to build prompt patterns that aim to offer reusable solutions for specific problems and that can be transferable across models?

Method	Description	Specificities
Zero-Shot Prompting	The model is asked to perform a task without any prior examples or guidance. It relies solely on its pre-existing knowledge.	Involves no examples, the model generates a response based on its knowledge, can lead to less accurate results for complex tasks.
Few-Shot (N-Shot) Prompting	The model is provided with a small number of examples (usually 2-5 but can vary) to guide its understanding of the task.	Provides the model with concrete examples, which helps it generate more accurate responses by learning from these examples. Unlike Zero-Shot, it relies on demonstration.
Chain-of-Thought (CoT)	The model generates intermediate reasoning steps, breaking down complex problems into smaller, logical steps to improve accuracy.	Enhances the quality of responses by explicitly guiding the model through a series of reasoning steps, leading to more detailed and structured outputs.
Reasoning and Action (ReAct)	Combines reasoning with explicit actions or steps that the model should take to complete the task, improving task organization.	More structured than CoT, ReAct not only involves reasoning but also outlines specific actions, which results in a clear and organised task completion strategy.
Tree of Thoughts (ToT)	Uses a hierarchical, tree-like structure where the model explores multiple aspects or pathways to achieve a comprehensive output.	Different from linear methods like CoT, ToT prompts the model to explore various branches or ideas systematically, producing more diverse and well-rounded responses.
Role Prompting	Assigns the model a specific role (e.g., journalist, teacher) to guide its responses, aligning them with the assumed perspective.	Focuses on shaping the model's output based on a given persona or role, which helps in producing more contextually relevant content.
Recursive Prompting	Involves iterative refinement where the model's output is used to generate new prompts, progressively improving the response.	More iterative than CoT, this method allows for continuous refinement of the prompt and response, handling complex tasks with multiple layers.
Retrieval Augmented Generation (RAG)	Combines the retrieval of relevant external information with the model’s generative capabilities, enhancing the accuracy and relevance of responses.	Involves augmenting content by retrieving real-time or up-to-date information from external sources, which is especially useful when the model’s internal knowledge is limited or outdated.
Meta-Prompting	The model generates its own prompts to tackle a task, using its understanding of the task context to create effective prompts.	The model self-generates prompts, leveraging its comprehension of the task to improve accuracy and creativity.

Examples: Vanilla and Role Prompts

Emmanuel Macron: « Employment rate has never been so high »

To check: the evolution employment rate over the last three decades

Prompt (vanilla, zero shot): Has the employment rate never been this high in France in 30 years?

Answer: Yes (reference to Insee)

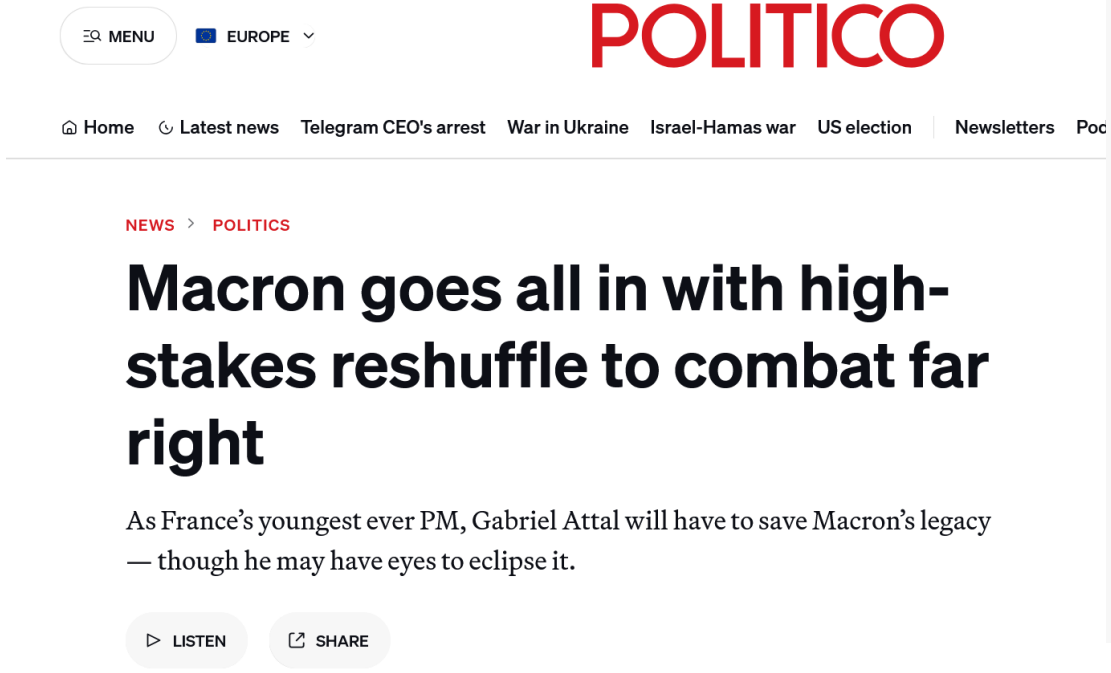
Prompt (role): You are a fact-checker GPT. Check if Emmanuel Macron said that the employment rate has never been so high.

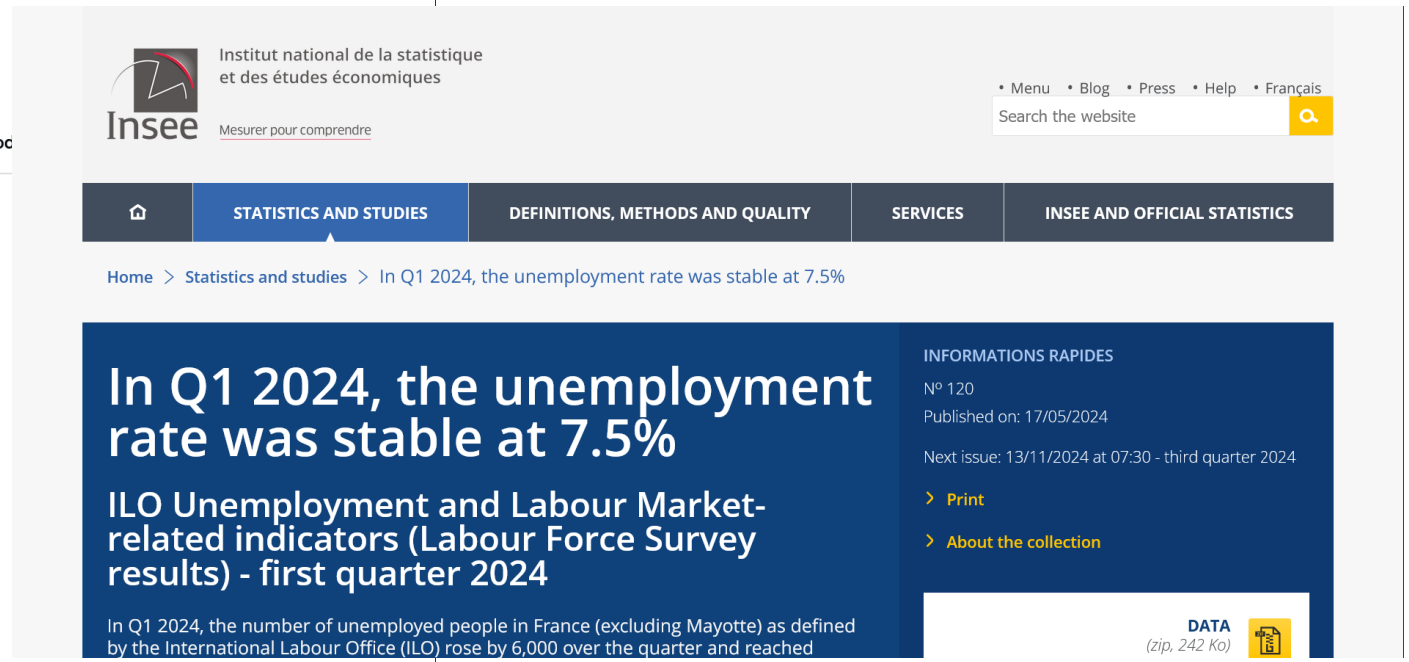
Answer: Yes, Emmanuel Macron did claim that the employment rate in France has reached its highest level in 30 years (source: Politico)

AI Literacy: Knowing about the limitations

Ethics: Keeping a human oversight

Prompting: Not necessarily lead to accurate and reliable results, even if it looks too!





Implication for research

- How to develop robust AI literacy frameworks for practitioners?
- How to rethink the concept of transparency (which is not enough)?
- How to improve explainability of the outcomes by refining prompts?
- How to develop ethical, accessible and suitable prompting strategies to support fact-checkers?

Thank you for your attention!

Contact: @ohmyshambles

Dierickx, L., van Dalen, A., Opdahl A.L. and Lindén C.G. *Striking the Balance in Using LLMs for Fact-Checking: A Narrative Literature Review*, in Proceedings from 6th Symposium on Multidisciplinary International Symposium on Disinformation in Open Online Media (MISDOOM), Lecture Notes in Computer Science (Springer).

NORDIS

UNIVERSITAS BERGENSIS

DDC digital democracy centre

SDU University of Southern Denmark