

Blueprints for Evaluating AI in Journalism

May 2, 2024

Sachita Nishal

PhD Candidate (Computer Science & Communication Studies)

Northwestern University

nishal@u.northwestern.edu

@nishalsach

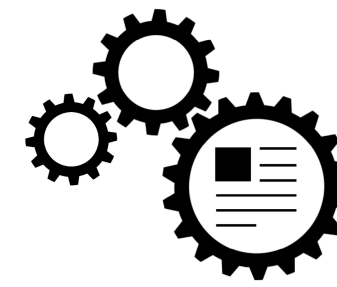
Domain-Specific Evaluation Strategies for AI in Journalism

Sachita Nishal*
nishal@u.northwestern.edu
Northwestern University
USA

Charlotte Li*
charlotte.li@u.northwestern.edu
Northwestern University
USA

Nicholas Diakopoulos
nad@northwestern.edu
Northwestern University
USA

Northwestern
University



Computational
Journalism
Lab



A note on the use of “AI” in this talk

- When I say “AI”, I'm referring specifically to advanced **deep learning models** and that can **generate human-like text, images, or other content** (e.g., GPT4).
- I am also referring to the **interactive tools and applications** built using those models (e.g., ChatGPT interface)
- These ideas apply to more classical AI systems as well (e.g, Google Search from 2010). However, newer generative AI tools do raise some new challenges we need to grapple with ...

- 📌 **Why Talk About AI Evaluation?**
- 📌 **Blueprints for AI Evaluation in Journalism**
- 📌 **Future Questions and Directions**

- 📌 **Why Talk About AI Evaluation?**
- 📌 **Blueprints for AI Evaluation in Journalism**
- 📌 **Future Questions and Directions**

Why talk about AI evaluation

- AI models are described and marketed on the basis of their performance on different kinds of **benchmarks** on specific **tasks**

Jobs **VentureBeat**

Security ▾ Data Infrastructure ▾ Automation ▾ Enter


Anthropic unveils Claude 3, surpassing GPT-4 and Gemini Ultra in benchmark tests

The Washington Post
Democracy Dies in Darkness [Subscribe](#)

TECH Help Desk Artificial Intelligence Internet Culture Space Tech Policy

Google's 'Gemini' launches to compete with GPT-4, joining other AI programs

The tech giant claims the new tool is better at math, coding and reasoning tasks than existing AI programs

 OpenAI [Research](#) ▾ [API](#) ▾ [ChatGPT](#) ▾ [Safety](#) [Company](#) ▾ [Search](#) [Log in](#) ↗

We've created GPT-4, the latest milestone in OpenAI's effort in scaling up deep learning. GPT-4 is a large multimodal model (accepting image and text inputs, emitting text outputs) that, while less capable than humans in many real-world scenarios, exhibits human-level performance on various professional and academic benchmarks.

What are benchmarks

- Benchmarks are **datasets of problems/tasks and their expected solutions** — and you want to see how well a model is at solving these.
- They essentially **quantify** a model's ability over specific kinds of tasks + this helps to compare models and measure improvement.

What are benchmarks

- Benchmarks are **datasets of problems/tasks and their expected solutions** — and you want to see how well a model is at solving these.
- They essentially **quantify** a model's ability over specific kinds of tasks + this helps to compare models and measure improvement.

Story: Cam ordered a pizza and took it home. He opened the box to take out a slice. Cam discovered that the store did not cut the pizza for him. He looked for his pizza cutter but did not find it. He had to use his chef knife to cut a slice.

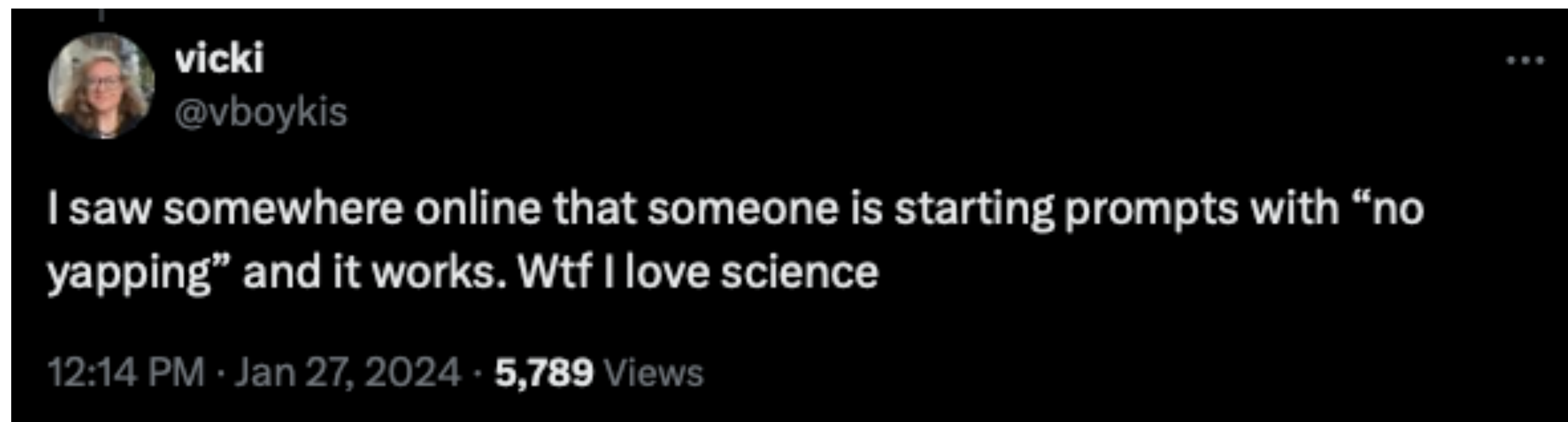
Question: Why did Cam order a pizza?

Ans: Cam was hungry.

A logical reasoning task from [Lal et al. \(2021\)](#)

Benchmarks lack context

- But these benchmark tasks are often **divorced from the specific contexts in which people use** the tools.
- Also don't typically capture what it feels like to actually **interact with AI systems + ethical issues** they pose — these are also components of the context.



Defining a useful AI system

- AI systems are also **resource-intensive** and **difficult to reason about** (the output looks so plausible!), which makes appropriate evaluation all the more important.
- A useful AI system: **actually does the thing you want it to do**, in a way that **aligns with your context and the personal/professional/organizational values within which you work**.

Defining a useful AI system

- AI systems are also **resource-intensive** and **difficult to reason about** (the output looks so plausible!), which makes appropriate evaluation all the more important.
- A useful AI system: **actually does the thing you want it to do**, in a way that **aligns with your context and the personal/professional/organizational values within which you work.**

- 📌 Why Talk About AI Evaluation?
- 📌 **Blueprints for AI Evaluation in Journalism**
- 📌 Future Questions and Directions

Blueprints for AI evaluation in journalism

- Evaluation frameworks in other domains focus on **capabilities** of models, how their results maintain **privacy** or transparency, and how they are **integrated** into actual practices (e.g., [TEHAI Framework](#) in medicine).
- We abstract out three such dimensions for journalism use-cases broadly:
 - **quality of model outputs**
 - **quality of interaction with AI tools**
 - **ethical and value alignment**

Quality of model outputs: idea

- Based on **editorial goals** and **specific news values** that are of interest (e.g., novelty, social impact, controversy)
- Go **beyond more general metrics of quality**, such as clarity, coherence of texts produced by an LLM, and ask for a **specific notion of quality** based off of the use-cases in journalism.



GENERAL METRICS
clarity, coherence, truth

SPECIFIC METRICS
what is the use-case?

Quality of model outputs: examples

Summarization + News Values

- For a summarization tool over a potentially newsworthy document, evaluating by the ability to identify **specific news values like novelty** that are of interest to a reporter looking for a surprising news story can be useful.

Brainstorming + Variety

- In a tool to support brainstorming of potential news angles or headlines, maybe there is a high-level benchmark that measures the **diversity** or **variety** of the ideas suggested, with the implicit goal of trying to offer wide inspiration.

Quality of interaction: idea

- A good output is not good enough. The **interaction + iteration** that led to it also matters!
- People have both **short-term and long-term goals** as they do their jobs and collaborate with others.
 - How **easy** was it to obtain this output? How **tedious** was it to craft prompts? How **enjoyable** is the process of using this tool?
 - Does the tool support any kind of **personal learning**? Does it allow for reliable **customization**? How often does it **surprise** you?

Quality of interaction: examples

Writing support + Agency

- If a tool is designed to offer writing support, can users very quickly accept or reject its suggestions, thereby exercising more **agency**?

News Discovery + Learnability

- If a tool suggests potentially newsworthy documents, do users feel like they are **learning more** about what drives newsworthiness, through prompting, and iteration, and discussion of the output?

Ethical and value alignment: idea

- We want our AI systems to be **accurate, consistent, free from bias, yes!**
- But ethics in journalism is also about the **ethics of the the method** itself: **transparency, traceability.**
- Different newsrooms also have their own **codes of conduct** and **style guides** — if a model does not align to these, then its output demands extra work from the reporter to align.

Ethical and value alignment: examples

Newsworthiness + Explanation

- If an AI system ranks documents for newsworthiness, is there any item-level **explanation** (“why was this ranked high?”) or system-level explanation (“what are the kinds of things that get ranked high?”) for its behavior?

Brainstorming + Gender Bias

- If you use an AI system to brainstorm news headlines, how likely is it to offer **stereotyped** or **biased** (or trope-y) descriptions of different genders?

- 📌 **Why Talk About AI Evaluation?**
- 📌 **Blueprints for AI Evaluation in Journalism**
- 📌 **Future Questions and Directions**

Future: how will this work?

- With lots of **collaboration**: between news practitioners, researchers, technologists, designers, etc.
- With lots of **iteration** + resources to track it. Also because new versions of models come out all the time!
- **A little bit of this, and a little bit of that**: some criteria are easier to quantify and others need deeper user engagement e.g., **diversity of ideas** vs. **ease of use**.
- Regular **audits**, esp. for values, during development and procurement of tools. This is hard but very good!

Future: some open questions

- Tasks and settings where can AI **reliably improve vs. tilting at windmills?**
- What tasks demand **excessive human oversight?**
- When is a **sociotechnical solution** needed (e.g., change practices for a specific activity), not just AI (e.g., LLMs)?

Our hope with this

- **Feedback** and **pressing questions** — what is most important to evaluate for you?
- Blueprints —> a more full-fledged **framework**, some open-source **evaluation suites** and benchmark datasets. Let's work together to make this real!
- Technology that actually **supports the goals of news reporting!**
 - **Change** along those dimensions can also be also good — but whose contexts, needs, values actually **drive it**? Hopefully yours :)

Thank you! Please reach out :)

Sachita Nishal

PhD Candidate

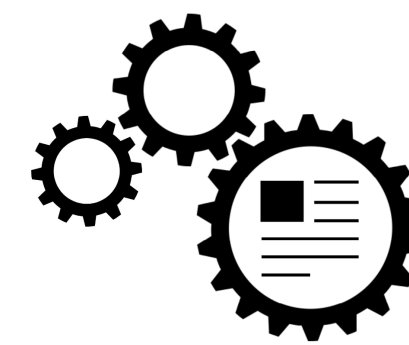
Computer Science & Communication Studies

Northwestern University

Email: nishal@u.northwestern.edu

Website: nishalsach.github.io

Northwestern
University



Computational
Journalism
Lab