We don’t Know how to Assess LLM Contributions in VIS/HCI

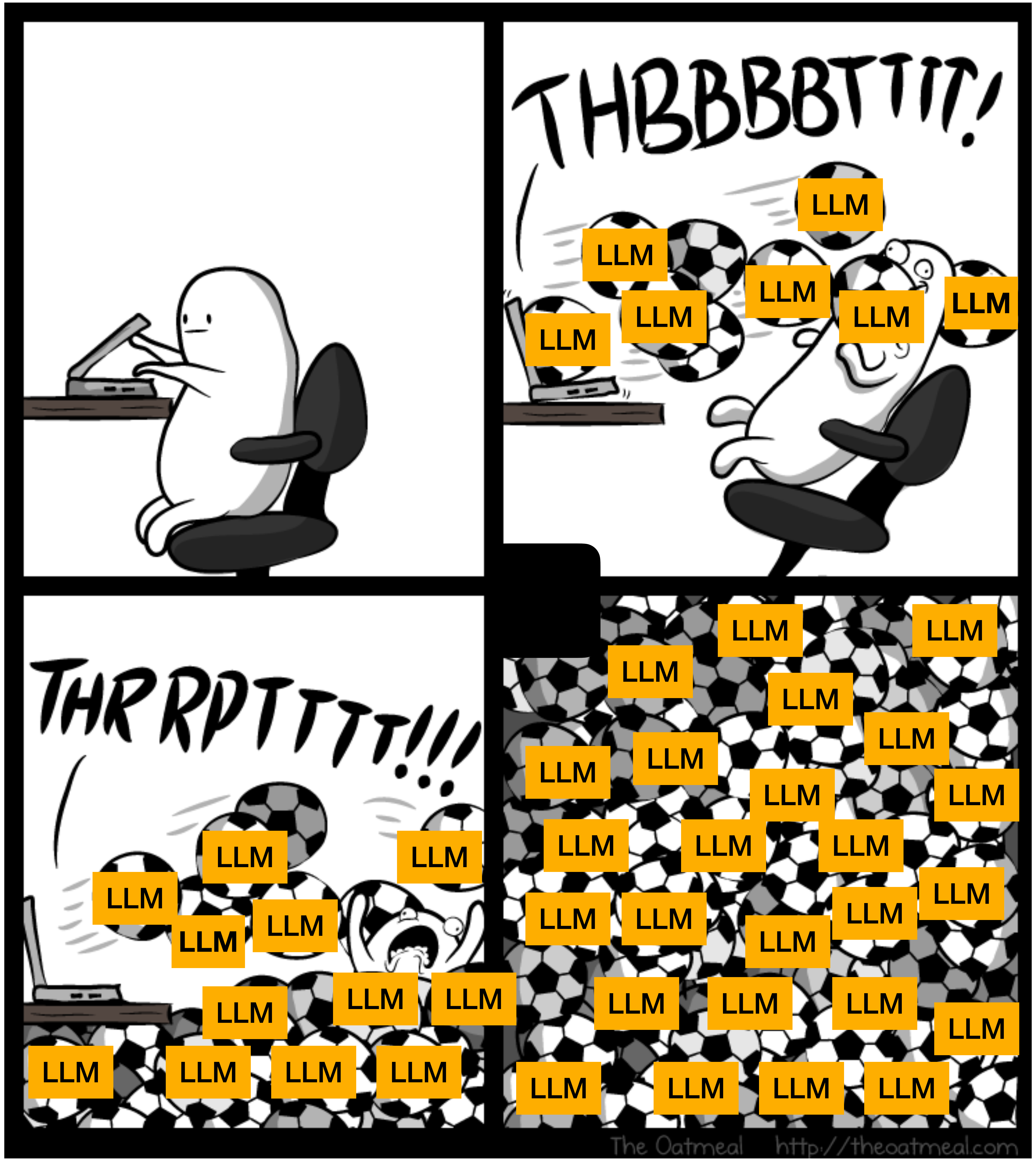
Position paper

Anamaria Crisan
University of Waterloo
ana.crisan@uwaterloo.ca

I volunteered to serve on 4 program committees

CHI’24/ FAccT’24/ CSCW ’24/ VIS'24

What is saw was concerning


THBBBBTTTT!
LLM
LLM
LLM
LLM
LLM
LLM
LLM
LLM
THRRPTTTTT!!!
LLM
LLM
LLM
LLM
LLM
LLM
LLM
LLM
LLM
LLM
LLM
LLM
LLM
LLM
LLM
LLM
LLM
LLM
LLM
LLM
LLM
LLM
LLM
LLM
LLM
LLM
LLM
LLM
LLM
LLM
LLM
LLM
LLM
LLM
LLM
LLM
LLM
The Oatmeal http://theoatmeal.com

here were some common reviewer critiques

1. LLMS have non-deterministic behaviour

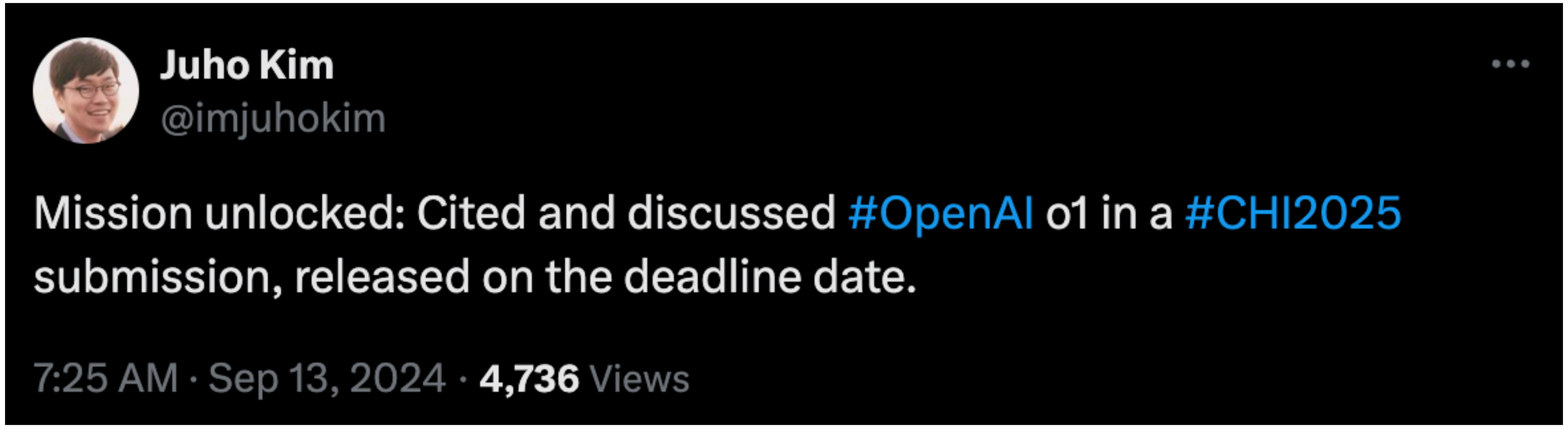
The **same** LLM provides a **different** response to exactly the **same** prompt

2. You didn’t evaluate against enough LLMs

Different LLMs provide a **different** response to exactly the **same** prompt

3. You didn't evaluate against the latest LLM

New LLMs are constantly emerging with **substantive** performance gains



Juho Kim
@imjuhokim

Mission unlocked: Cited and discussed #OpenAI o1 in a #CHI2025 submission, released on the deadline date.

7:25 AM · Sep 13, 2024 · **4,736** Views

1. LLMS have non-deterministic behaviour

2. You didn’t evaluate against enough LLMs

3. You didn’t evaluate against the latest LLM

You got lucky and didn’t assess the full range of performance

You got lucky and didn’t assess the full range of performance

- Its expensive to run multiple LLMs multiple times
- There is unequal access to LLMs
- Privileges specific labs/groups

4.You didn’t discuss the latest LLM paper

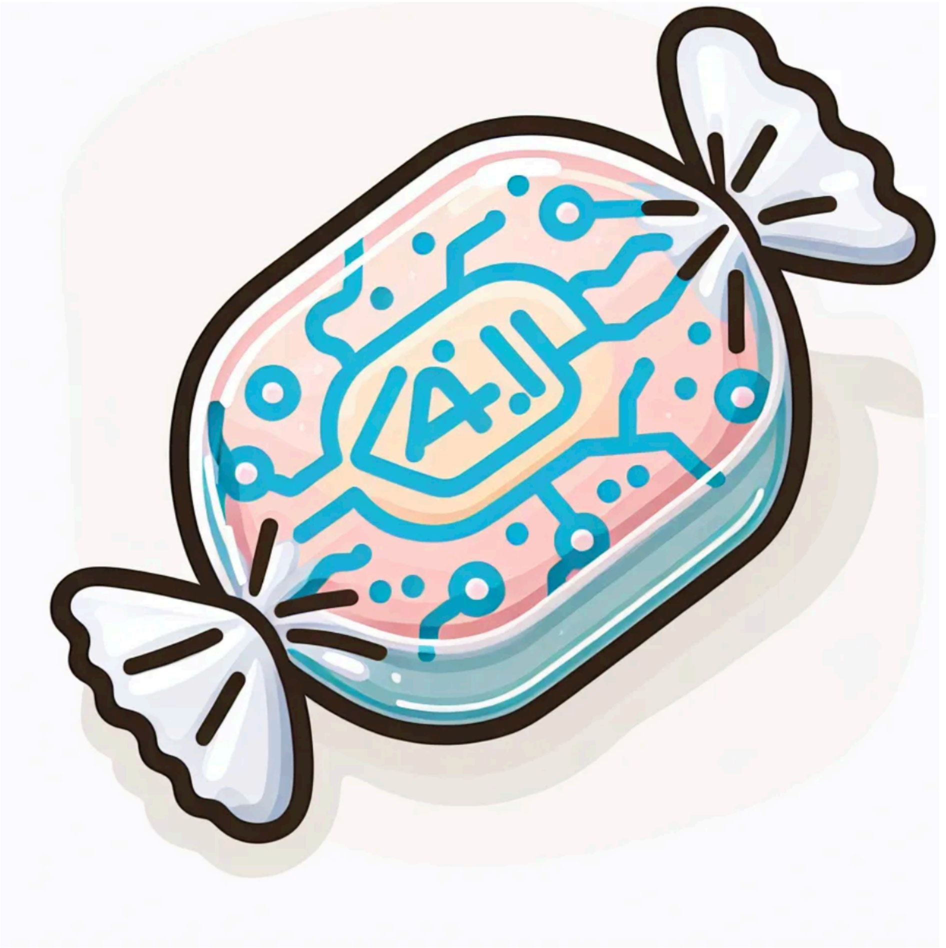
…. and it obviates your results

5.You didn’t cite ANY relevant research.

No VIS/CHI/CSCW/FAcct etc. papers were referenced

6.The LLM Wrapper Paper

Your novel research is just low-hanging fruit of LLM capabilities



An "LLM wrapper," according to Dall-E-3.

4. You didn’t discuss the latest LLM paper

5. You didn’t cite ANY relevant research.

6. The LLM Wrapper Paper

Your novel research is just low-hanging fruit of LLM capabilities and timing

Your novel research is just low-hanging fruit of LLM capabilities and timing

- There is too much flag planting on arXiv
- LLM (“AI for X”) applications are worth exploring
- Research communities as a counterpoint to industry claims

7.I don’t like LLMs / I am so over them

….yup, it happens

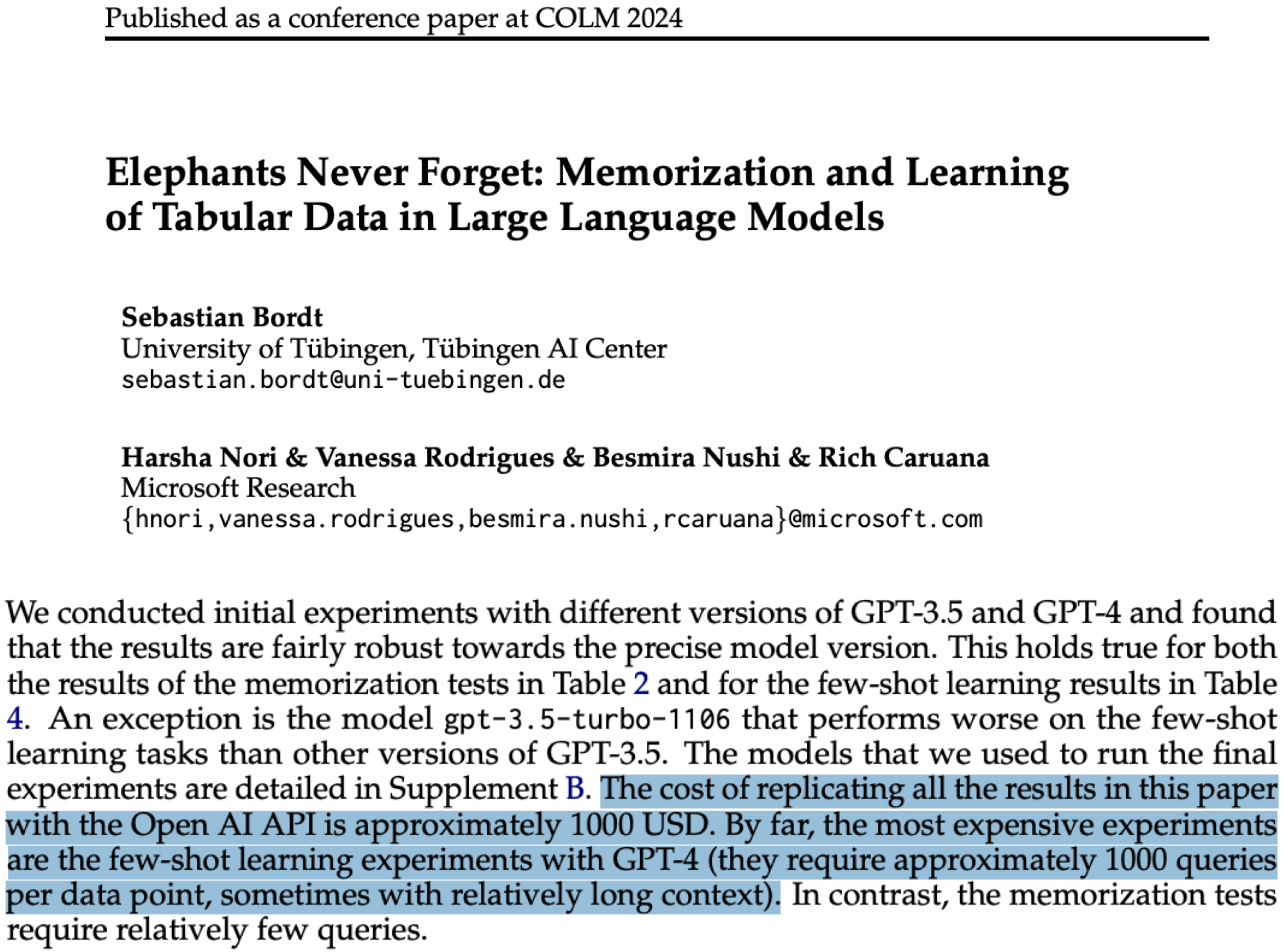
what should we do about it?

1. LLMS have non-deterministic behaviour
2. You didn’t evaluate against enough LLMs
3. You didn’t evaluate against the latest LLM

1. Make a reasonable attempt to understand the variability of LLM outputs
2. Consider requiring a budget and access statement
3. Make a reasonable attempt to justify the use of an LLM

Example of a Budget Statement

Shows people how expensive it is to reproduce the results



Published as a conference paper at COLM 2024

Elephants Never Forget: Memorization and Learning of Tabular Data in Large Language Models

Sebastian Bordt
University of Tübingen, Tübingen AI Center
sebastian.bordt@uni-tuebingen.de

Harsha Nori & Vanessa Rodrigues & Besmira Nushi & Rich Caruana
Microsoft Research
{hnori,vanessa.rodrigues,besmira.nushi,rcaruana}@microsoft.com

We conducted initial experiments with different versions of GPT-3.5 and GPT-4 and found that the results are fairly robust towards the precise model version. This holds true for both the results of the memorization tests in Table 2 and for the few-shot learning results in Table 4. An exception is the model `gpt-3.5-turbo-1106` that performs worse on the few-shot learning tasks than other versions of GPT-3.5. The models that we used to run the final experiments are detailed in Supplement B. The cost of replicating all the results in this paper with the Open AI API is approximately 1000 USD. By far, the most expensive experiments are the few-shot learning experiments with GPT-4 (they require approximately 1000 queries per data point, sometimes with relatively long context). In contrast, the memorization tests require relatively few queries.

4.You didn’t discuss the latest LLM paper

5.You didn’t cite ANY relevant research.

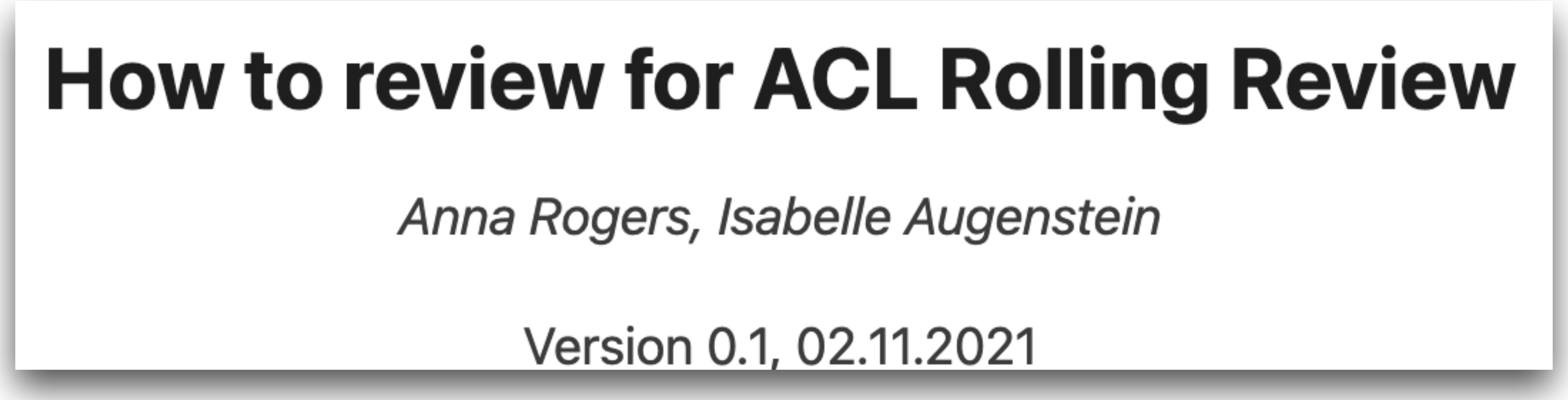
6.The LLM Wrapper Paper

4. Limit citation requirements for new research

5. Make Better Use of the Desk Reject

6. Actively Engage in the Discussion Period

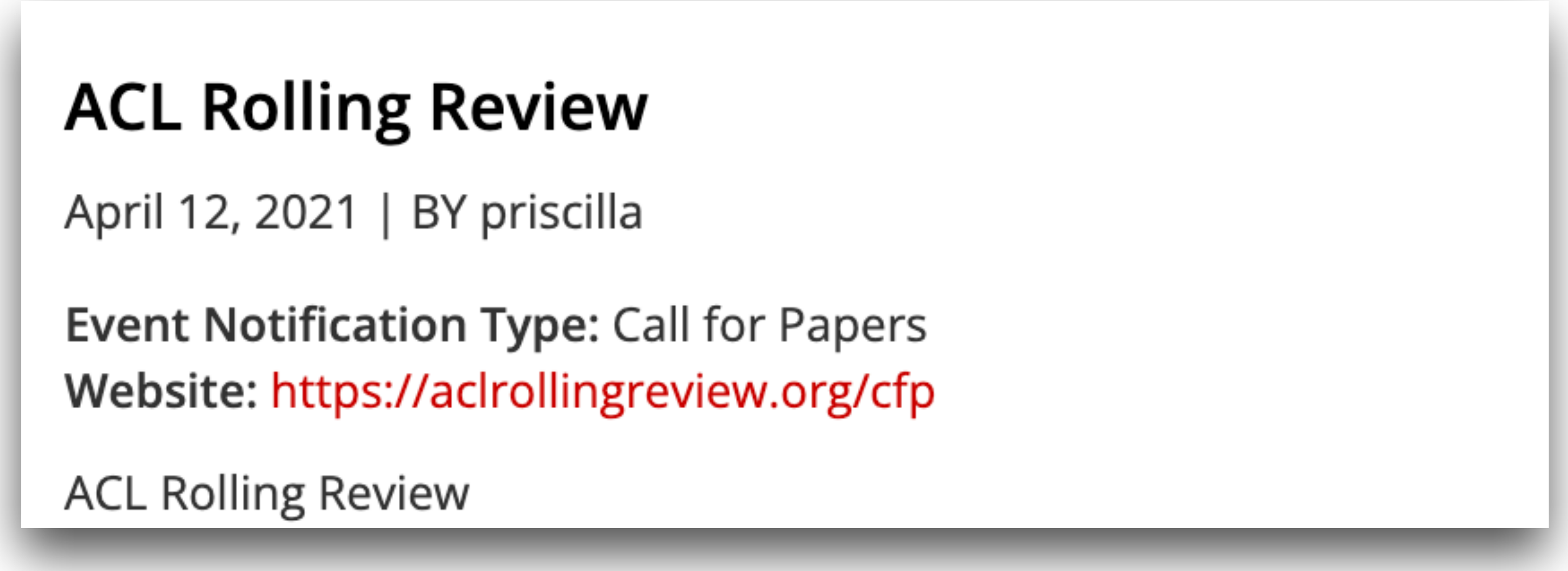
Example of Limiting Required Work To Cite



How to review for ACL Rolling Review

Anna Rogers, Isabelle Augenstein

Version 0.1, 02.11.2021

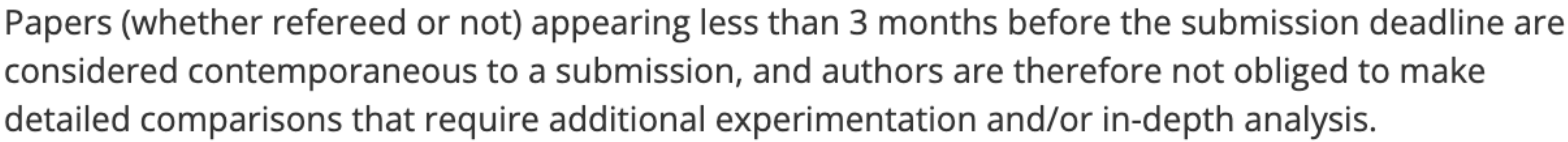


ACL Rolling Review

April 12, 2021 | BY priscilla

Event Notification Type: Call for Papers
Website: https://aclrollingreview.org/cfp

ACL Rolling Review



Papers (whether refereed or not) appearing less than 3 months before the submission deadline are considered contemporaneous to a submission, and authors are therefore not obliged to make detailed comparisons that require additional experimentation and/or in-depth analysis.

1. Make a reasonable attempt to understand the variability of LLM outputs
2. Consider requiring a budget and access statement
3. Make a reasonable attempt to justify the use of an LLM
4. Limit citation requirements for new research
5. Make Better Use of the Desk Reject
6. Actively Engage in the Discussion Period

We don’t Know how to Assess LLM Contributions in VIS/HCI

Position paper

Anamaria Crisan

University of Waterloo

ana.crisan@uwaterloo.ca