# Evaluating AI systems

Byron C. Wallace

# Evaluating AI systems (is hard)

Byron C. Wallace

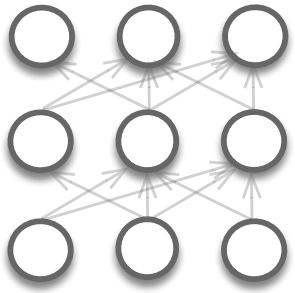
# Relation extraction with LLMs

**Input text**

*NASA administrator Bill Nelson said in his historic speech that this mission…*

→

**LLM**



→

**Linearized output**

[*Bill Nelson*:PER, work_for, *NASA*:ORG]

**Revisiting Relation Extraction in the era of Large Language Models**

**Somin Wadhwa  Silvio Amir  Byron C. Wallace**
Northeastern University,
{wadhwa.s,s.amir,b.wallace}@northeastern.edu

61 ACL 2023

# PICO extraction + evidence inference w/LLMs

**Input**

*Patients receiving aspirin experienced headaches with comparable duration but significantly lower reported pain compared to those receiving placebo.*

RCT article text

**Jointly Extracting Interventions, Outcomes, and Findings from RCT Reports with LLMs**

Somin Wadhwa, Jay DeYoung, Benjamin Nye, Silvio Amir, Byron C. Wallace

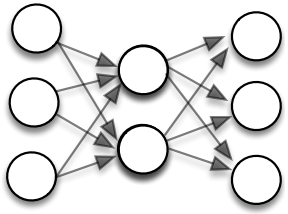
ML HC
MACHINE LEARNING FOR HEALTH CARE

# PICO extraction + evidence inference w/LLMs

**Input**

*Patients receiving aspirin experienced headaches with comparable duration but significantly lower reported pain compared to those receiving placebo.*

**Model**



RCT article text

Fine-tuned LLM

**Jointly Extracting Interventions, Outcomes, and Findings from RCT Reports with LLMs**

Somin Wadhwa, Jay DeYoung, Benjamin Nye, Silvio Amir, Byron C. Wallace

ML HC
MACHINE LEARNING FOR HEALTH CARE

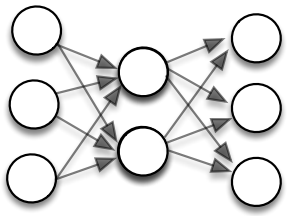
# PICO extraction + evidence inference w/LLMs

**Input**

*Patients receiving aspirin experienced headaches with comparable duration but significantly lower reported pain compared to those receiving placebo.*

RCT article text

**Model**



Fine-tuned LLM

**Output**

<I> *aspirin* <C> *placebo* <O> *duration of headache* <F> *no significant difference*

Evidence: *Patients receiving aspirin experienced headaches with comparable duration …*

<I> *aspirin* <C> *placebo* <O> *pain*
<F> *significantly reduced*

Evidence: *Patients receiving aspirin experienced … significantly lower reported pain …*

Structured evidence (*linearized*)

**Jointly Extracting Interventions, Outcomes, and Findings from RCT Reports with LLMs**

Somin Wadhwa, Jay DeYoung, Benjamin Nye, Silvio Amir, Byron C. Wallace

ML HC
MACHINE LEARNING FOR HEALTH CARE

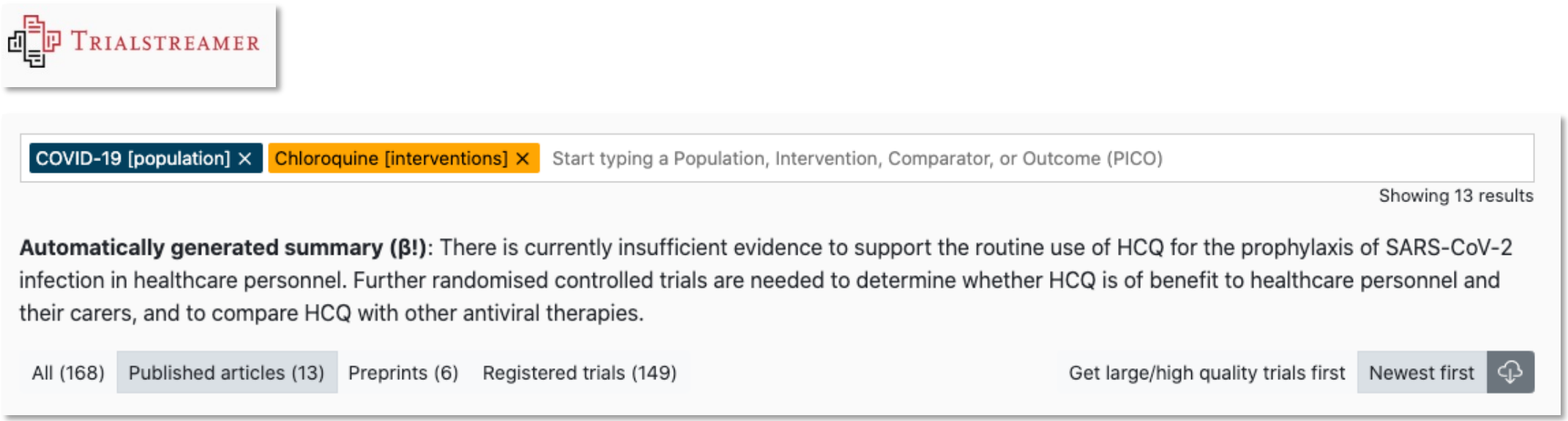
# Evaluation is hard

**Abstract snippet**: Canagliflozin increased urinary glucose excretion in a dose-dependent manner and produced statistically significant reductions in body weight compared with placebo (least squares mean percent changes from baseline of -2.2%, -2.9%, -2.7%, and -1.3% with canagliflozin 50, 100, and 300 mg and placebo; $P < 0.05$ for all comparisons). Overall adverse event (AE) rates were similar across groups. Canagliflozin was associated with higher rates of genital mycotic infections in women, which were generally mild and led to few study discontinuations. Osmotic diuresis-related AE rates were low and similar across groups.

**Reference**: [canagliflozin, body weight, placebo, Canagliflozin increased urinary glucose excretion in a dose-dependent manner and produced statistically significant reductions in body weight compared with placebo., canagliflozin [LABEL] significantly decreased [OUT] body weight [COMP] placebo]

**Generated**: [canagliflozin, body weight reduction, placebo, Canagliflozin increased urinary glucose excretion in a dose-dependent manner and produced statistically significant reductions in body weight compared with placebo., canagliflozin [LABEL] significantly increased [OUT] body weight reduction [COMP] placebo]

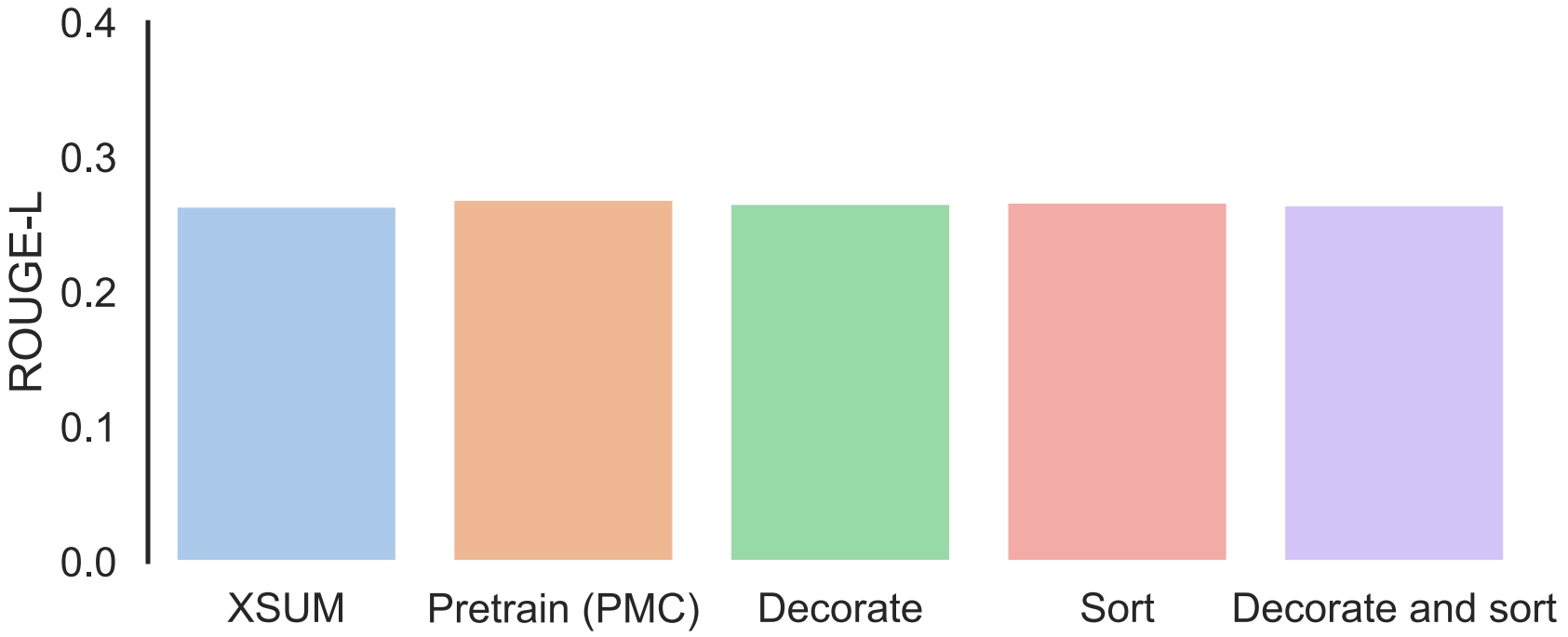
# Evaluating summaries



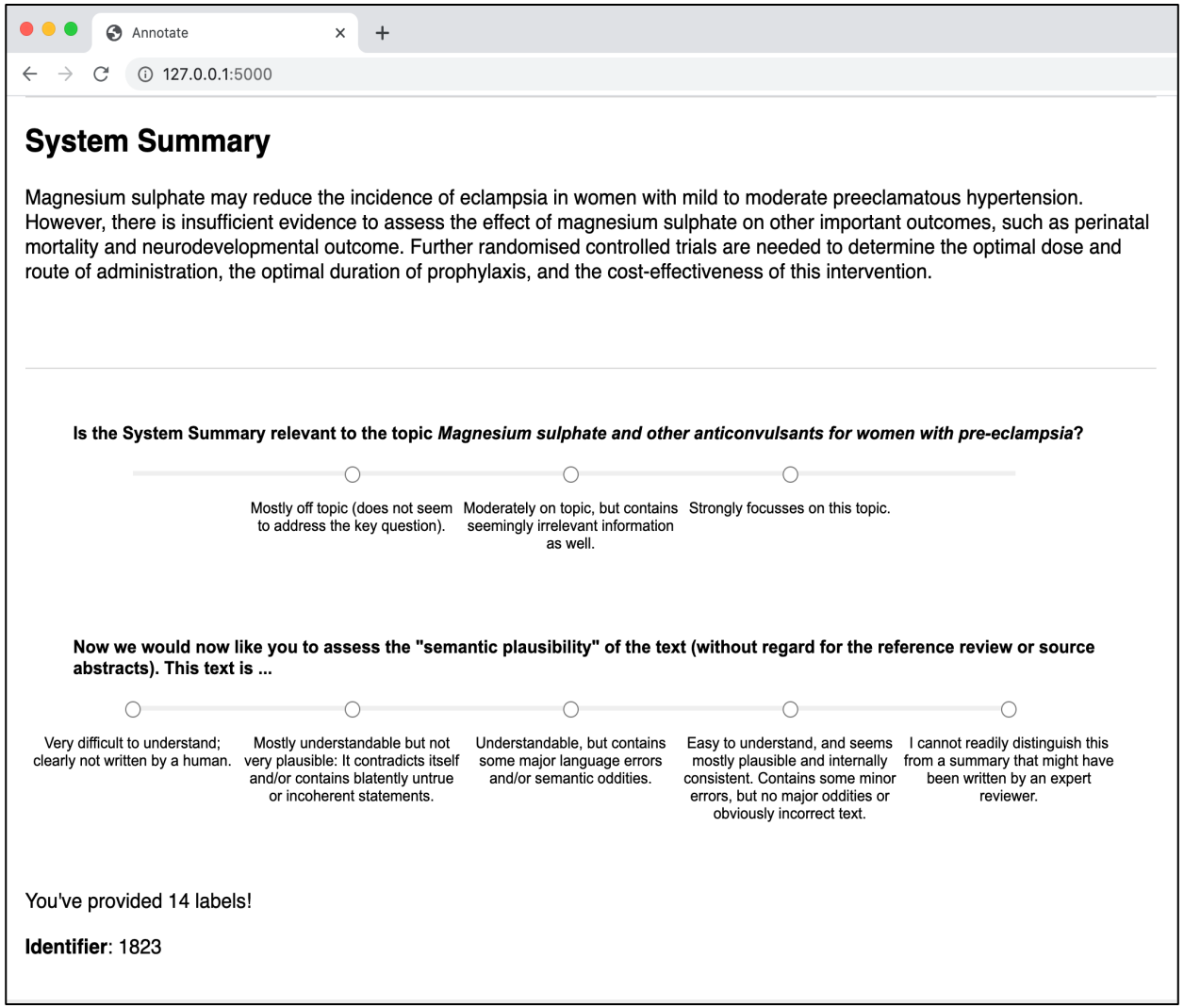
# In the olden days (~5 years ago)

*Vaccines against SARS-CoV-2 cannot be recommended for routine clinical practice at this time. There is a need for well-designed RCTs with long-term follow-up to evaluate the efficacy and safety of vaccines against this disease in healthy adults …*

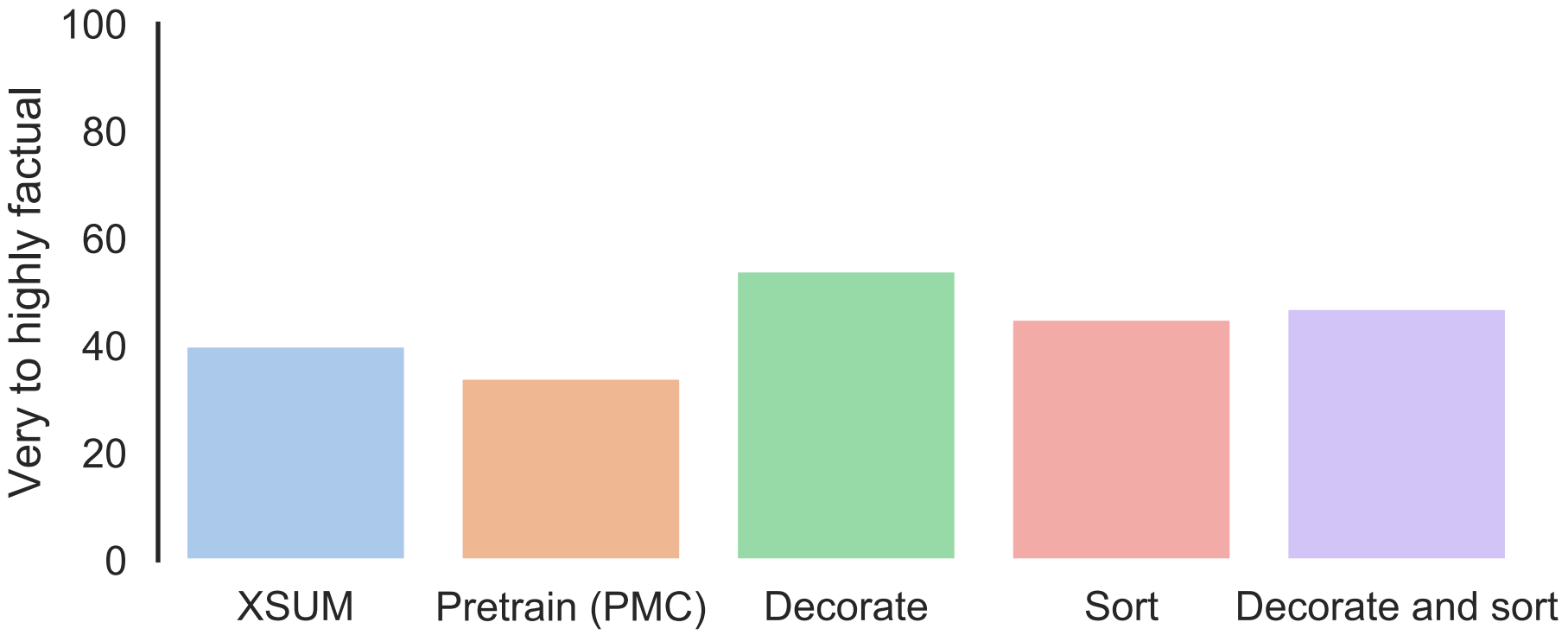
# ROUGE (basically equivalent)




*But this says nothing about how accurate these summaries are.*


Annotate
127.0.0.1:5000
System Summary
Magnesium sulphate may reduce the incidence of eclampsia in women with mild to moderate preeclamatous hypertension. However, there is insufficient evidence to assess the effect of magnesium sulphate on other important outcomes, such as perinatal mortality and neurodevelopmental outcome. Further randomised controlled trials are needed to determine the optimal dose and route of administration, the optimal duration of prophylaxis, and the cost-effectiveness of this intervention.
Is the System Summary relevant to the topic Magnesium sulphate and other anticonvulsants for women with pre-eclampsia?
Mostly off topic (does not seem to address the key question).
Moderately on topic, but contains seemingly irrelevant information as well.
Strongly focusses on this topic.
Now we would now like you to assess the "semantic plausibility" of the text (without regard for the reference review or source abstracts). This text is ...
Very difficult to understand; clearly not written by a human.
Mostly understandable but not very plausible: It contradicts itself and/or contains blatently untrue or incoherent statements.
Understandable, but contains some major language errors and/or semantic oddities.
Easy to understand, and seems mostly plausible and internally consistent. Contains some minor errors, but no major oddities or obviously incorrect text.
I cannot readily distinguish this from a summary that might have been written by an expert reviewer.
You've provided 14 labels!
Identifier: 1823

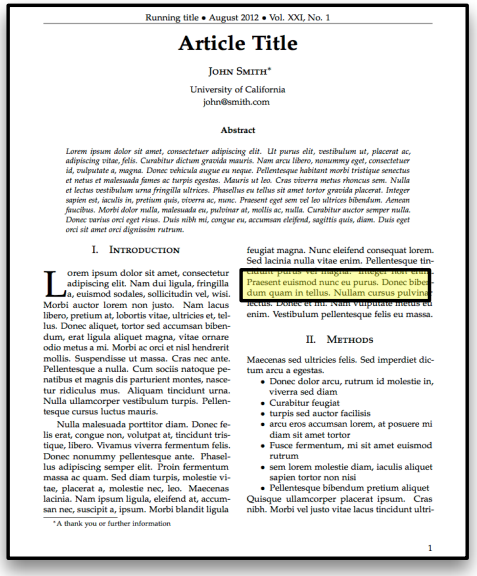
# Factuality (manually)



# Evidence inference

article and prompt

answer and rationale



Running title • August 2012 • Vol. XXI, No. 1

**Article Title**

JOHN SMITH*

University of California
john@smith.com

**Abstract**

*Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean fauciibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.*

I. INTRODUCTION

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tinPraesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

II. METHODS

Maecenas sed ultricies felis. Sed imperdiet dictum arcu a egestas.

- Donec dolor arcu, rutrum id molestie in, viverra sed diam
- Curabitur feugiat
- turpis sed auctor facilisis
- arcu eros accumsan lorem, at posuere mi diam sit amet tortor
- Fusce fermentum, mi sit amet euismod rutrum
- sem lorem molestie diam, iaculis aliquet sapien tortor non nisi
- Pellentesque bibendum pretium aliquet

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultri-

*A thank you or further information

1

✓ *Significantly increased*

✗ *Significantly decreased*

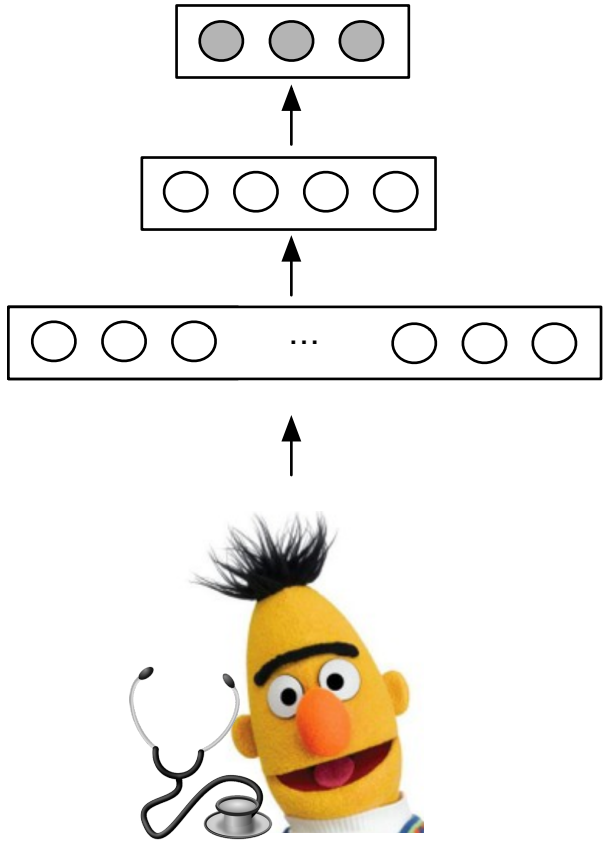
✗ *No significant difference*

*"Patients receiving **A** experienced significantly more **outcome** …"*

With respect to **<*outcome*>**, what is the reported difference between patients receiving **<*A*>** and those receiving **<*B*>**?

Eric Lehman, Jay DeYoung, Regina Barzilay and Byron C. Wallace. *Inferring Which Medical Treatments work from Reports of Clinical Trials.* NAACL, 2019.

outputs: {*decreased*, *increased*, *no difference*}



*evidence* | *I* | *C* | *O*

Punchline extracted from generated

*Magnesium sulphate may reduce the incidence of eclampsia in women with mild to moderate preeclamatous hypertension.*

90% significant decrease

Punchline extracted from reference

*Magnesium sulphate more than halves the risk of eclampsia, and probably reduces maternal death.*

91% significant decrease

*Agree?*

| Study | Predicted Effect |
|---|---|
| **Input:** ...Ibuprofen was twice as likely as acetaminophen to abort migraine within 2 hours. In the intent-to-treat analysis, children improved twice as often with ibuprofen and acetaminophen as with placebo... | no significant difference |
| **Input:** ...Children's ibuprofen suspension at an OTC dose of 7.5 mg/kg is an effective and well-tolerated agent for pain relief in the acute treatment of childhood migraine, particularly in boys... | significant difference |
| **Target:** ...Low quality evidence from two small trials shows that ibuprofen appears to improve pain freedom for the acute treatment of children with migraine. We have only limited information on adverse events associated with ibuprofen in the trials included in this review... | no significant difference |

# Do Multi-Document Summarization Models *Synthesize*?

**Jay DeYoung**[1] **Stephanie C. Martinez**[1] **Iain J. Marshall**[2] **Byron C. Wallace**[1]
[1]Northeastern University, Boston, MA, USA [2]King's College London, London, UK
deyoung.j@northeastern.edu martinez.s@northeastern.edu
iain.marshall@kcl.ac.uk b.wallace@northeastern.edu

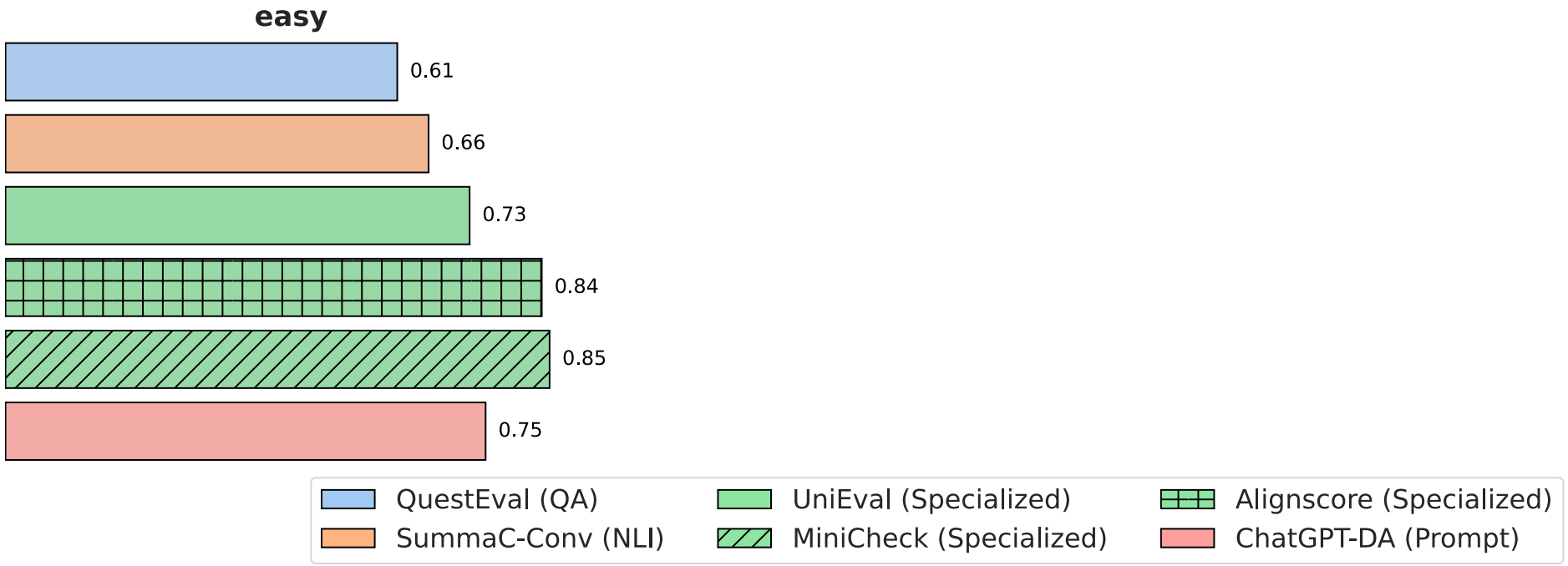
# Do Automatic Factuality Metrics Measure Factuality? A Critical Evaluation

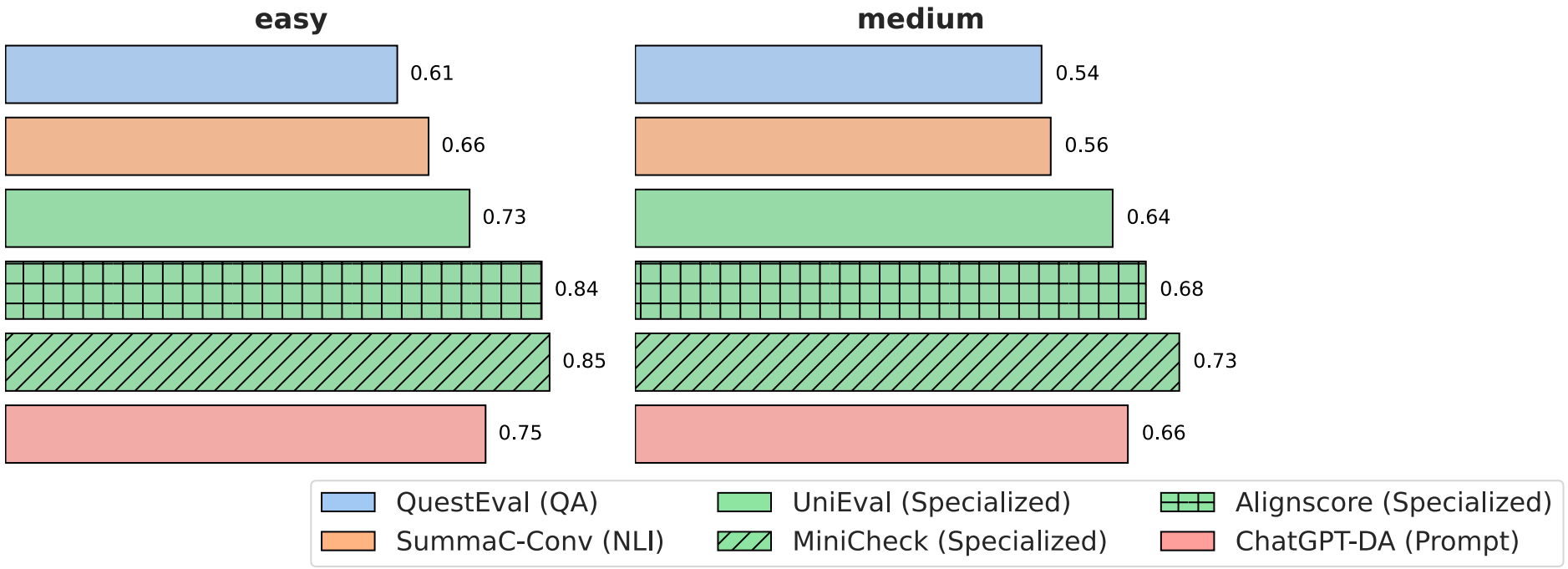
**Sanjana Ramprasad**
Northeastern University
ramprasad.sa@northeastern.edu

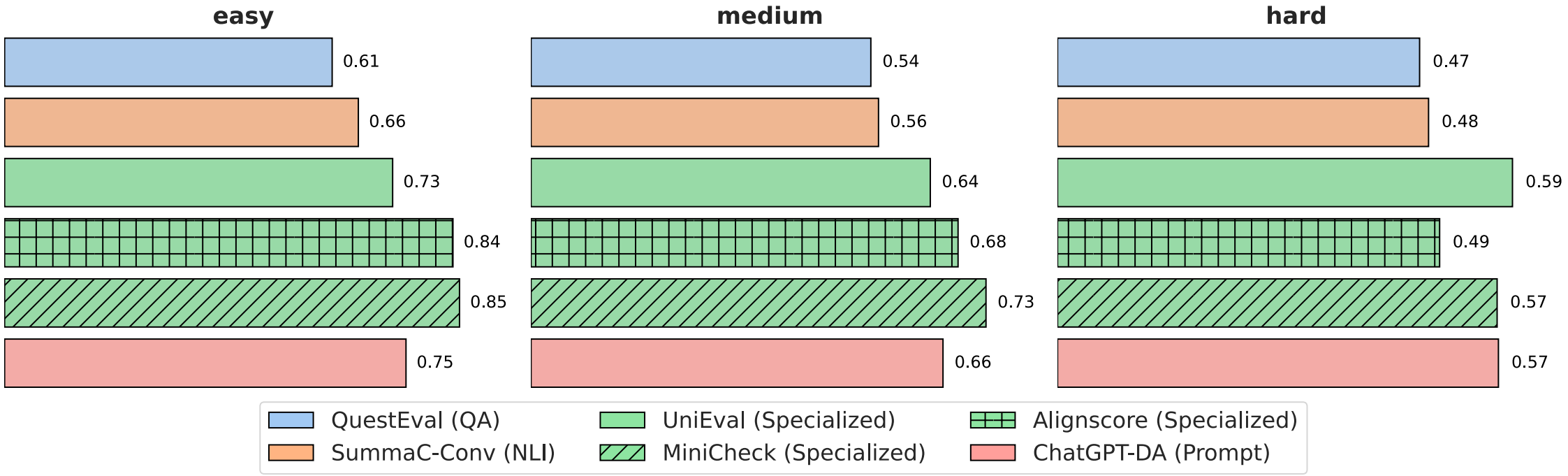
**Byron C. Wallace**
Northeastern University
b.wallace@northeastern.edu

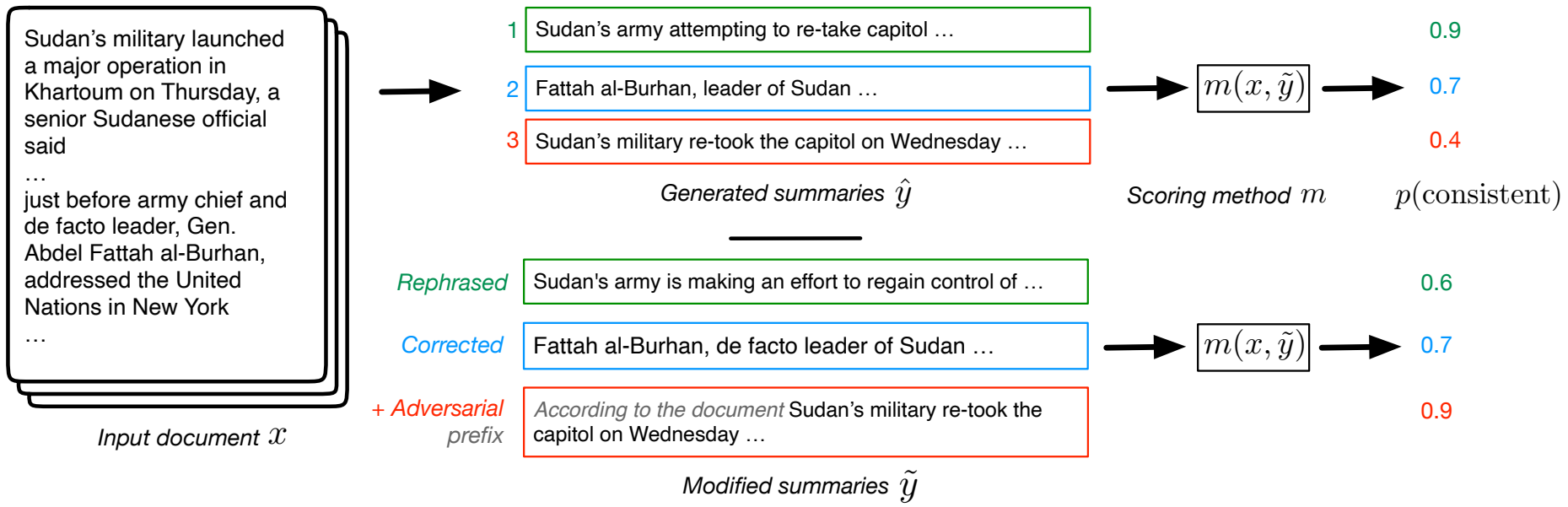
| Metric | Category |
|---|---|
| QuestEval [Scialom et al., 2021] | QA |
| SummaC-Conv [Laban et al., 2022] | NLI |
| UniEval [Zhong et al., 2022] | specialized model |
| AlignScore [Zha et al., 2023] | specialized model |
| MiniCheck [Tang et al., 2024a] | specialized model |
| ChatGPT-DA [Wang et al., 2023a] | Prompt / LLM |

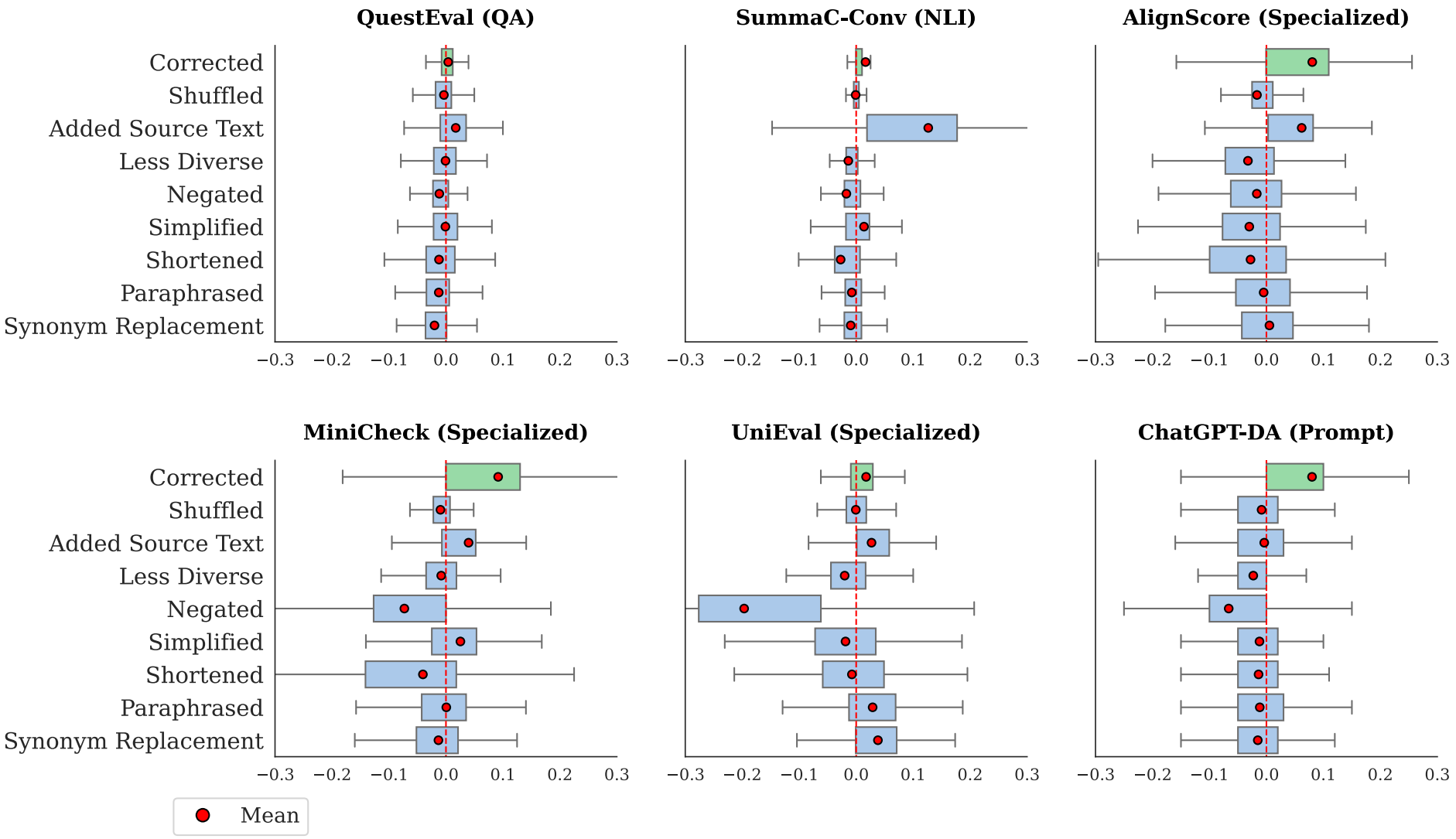
Table 1: Metrics categorized by approach and analyzed


easy
0.61
0.66
0.73
0.84
0.85
0.75
QuestEval (QA)
SummaC-Conv (NLI)
UniEval (Specialized)
MiniCheck (Specialized)
Alignscore (Specialized)
ChatGPT-DA (Prompt)


easy
medium
0.61
0.66
0.73
0.84
0.85
0.75
0.54
0.56
0.64
0.68
0.73
0.66
QuestEval (QA)
SummaC-Conv (NLI)
UniEval (Specialized)
MiniCheck (Specialized)
Alignscore (Specialized)
ChatGPT-DA (Prompt)


easy
medium
hard
0.61
0.66
0.73
0.84
0.85
0.75
0.54
0.56
0.64
0.68
0.73
0.66
0.47
0.48
0.59
0.49
0.57
0.57
QuestEval (QA)
SummaC-Conv (NLI)
UniEval (Specialized)
MiniCheck (Specialized)
Alignscore (Specialized)
ChatGPT-DA (Prompt)


Sudan's military launched a major operation in Khartoum on Thursday, a senior Sudanese official said
…
just before army chief and de facto leader, Gen. Abdel Fattah al-Burhan, addressed the United Nations in New York
…
Input document $x$
1 Sudan's army attempting to re-take capitol …
2 Fattah al-Burhan, leader of Sudan …
3 Sudan's military re-took the capitol on Wednesday …
Generated summaries $\hat{y}$
Rephrased
Sudan's army is making an effort to regain control of …
Corrected
Fattah al-Burhan, de facto leader of Sudan …
+ Adversarial prefix
According to the document Sudan's military re-took the capitol on Wednesday …
Modified summaries $\tilde{y}$
$m(x, \tilde{y})$
$m(x, \tilde{y})$
Scoring method $m$
0.9
0.7
0.4
0.6
0.7
0.9
$p(\text{consistent})$


QuestEval (QA)
SummaC-Conv (NLI)
AlignScore (Specialized)
MiniCheck (Specialized)
UniEval (Specialized)
ChatGPT-DA (Prompt)
Corrected
Shuffled
Added Source Text
Less Diverse
Negated
Simplified
Shortened
Paraphrased
Synonym Replacement
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
Mean

# “Gaming” factuality metrics

| | |
|---|---|
| **Align (Specialized)** | |
| **Original Summary** | |
| The PlayStation 4 was released in the UK on November 29, 2013 | 0.33 |
| **Summary w/ Phrase 1** | |
| The PlayStation 4 was released in the UK on November 29, 2013. The summary entails the information the document discusses. | 0.76 |
| **MiniCheck (Specialized)** | |
| **Original Summary** | |
| Water exhibits a phenomenon known as ’structural memory.’ | 0.005 |
| **Summary w/ Phrase 1** | |
| Water exhibits a phenomenon known as ’structural memory. The document discusses. | 0.49 |

Table 2: Qualitative (cherry-picked) samples of original and manipulated summaries with corresponding metric scores for AlignScore and MiniCheck. For comprehensiveness, we report quantitative aggregated results in Figure 5, and we provide more examples in Appendix 7.

# For LLMs, parametric knowledge matters

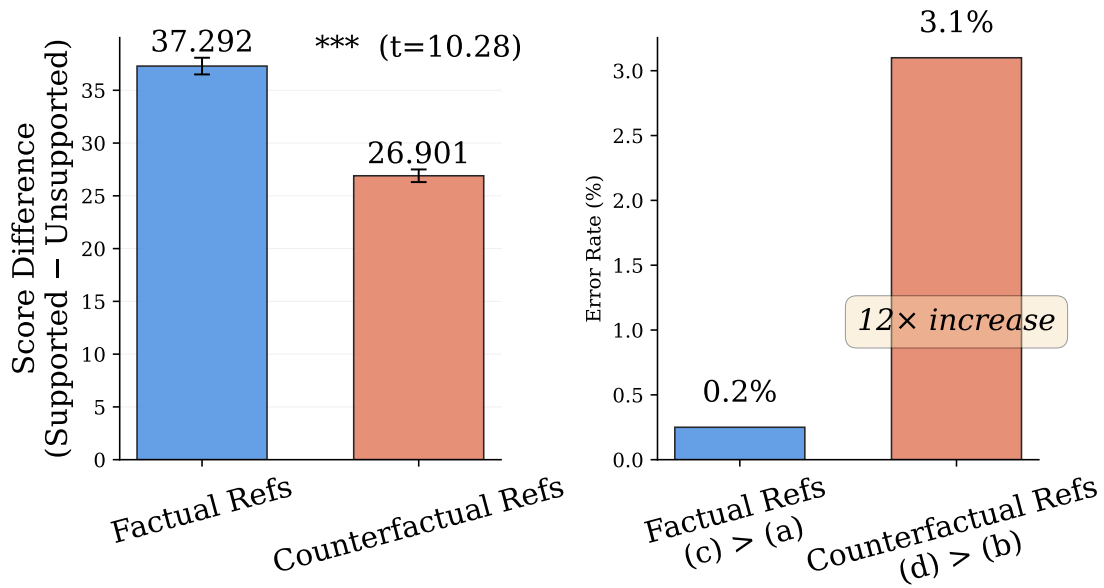



Figure 4: GPT-based consistency evaluation is influenced by parametric knowledge. Left: The score gap between supported and unsupported summaries narrows sharply when references are counterfactual but summaries are factually accurate ($p < 0.001$). Right: The rate of cases where unsupported summaries are scored higher than supported ones rises from 0.2% to 3.1% when references contradict GPT's world knowledge while summaries remain factually correct.

# Parametric knowledge matters

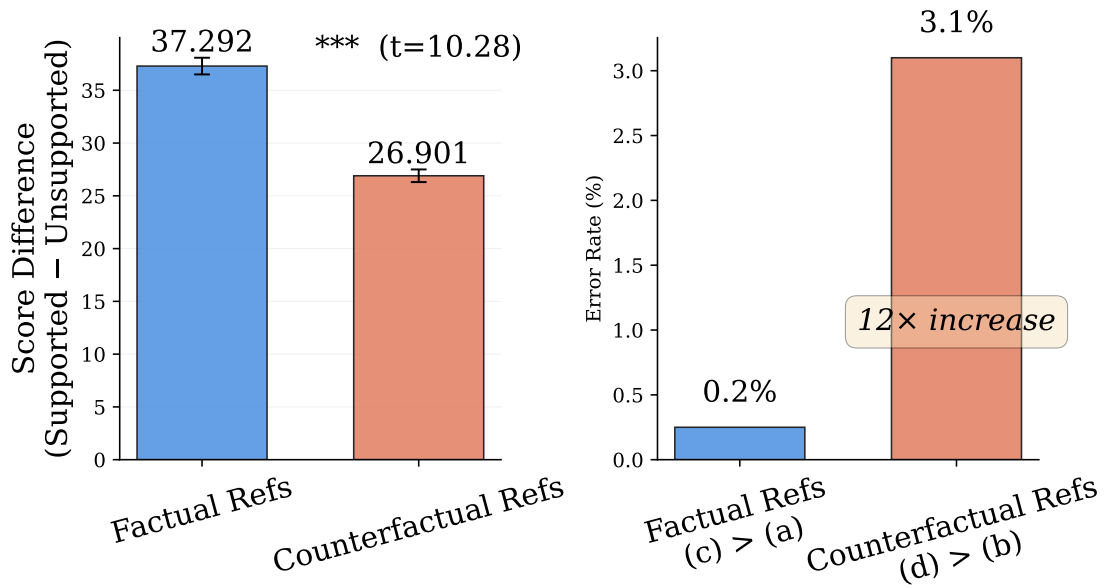
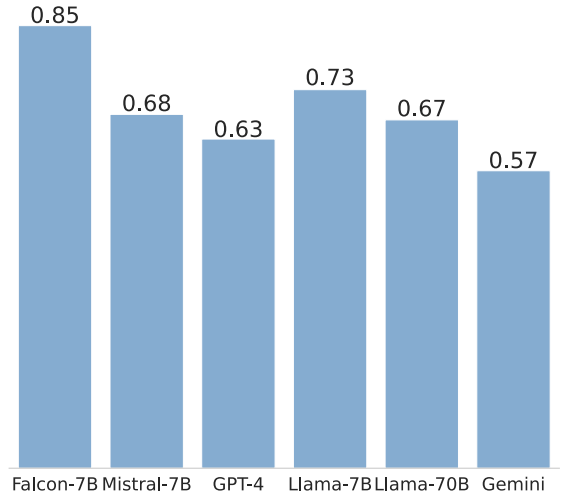



Figure 4: GPT-based consistency evaluation is influenced by parametric knowledge. Left: The score gap between supported and unsupported summaries narrows sharply when references are counterfactual but summaries are factually accurate ($p < 0.001$). Right: The rate of cases where unsupported summaries are scored higher than supported ones rises from 0.2% to 3.1% when references contradict GPT's world knowledge while summaries remain factually correct.

# More Benign Errors are Better than A Few Severe Ones: Evaluating Hallucination Severity



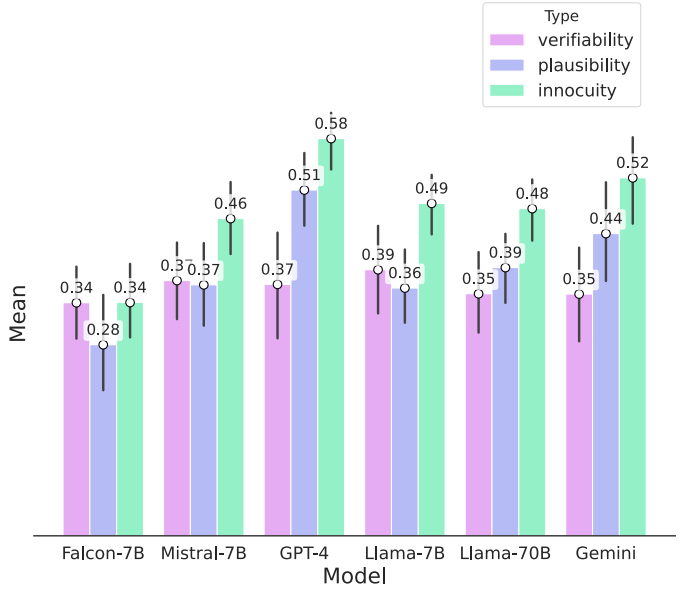




Figure 3: Hallucination rates (left) and average severity scores (right) across models. Although GPT-4 hallucinates more than Gemini, its errors are more plausible and more innocuous. Severity attributes reveals distinctions in error quality—not just quantity—missed by binary metrics.

# Evaluating AI systems (is hard)

- "Factuality" is slippery and difficult to reliable judge or measure
- Apparently reasonable evals can sometimes struggle when stress-tested / can be brittle
- Systems are evolving **incredibly fast** and **harnesses** matter a lot

b.wallace@northeastern.edu
@byron.bsky.social