

Reality Check of LLM- driven Fact Verification

Pepa Atanasova

Tenure Track Assistant Professor
University of Copenhagen, Denmark
pepa@di.ku.dk

UNIVERSITY OF COPENHAGEN



5.11.2025

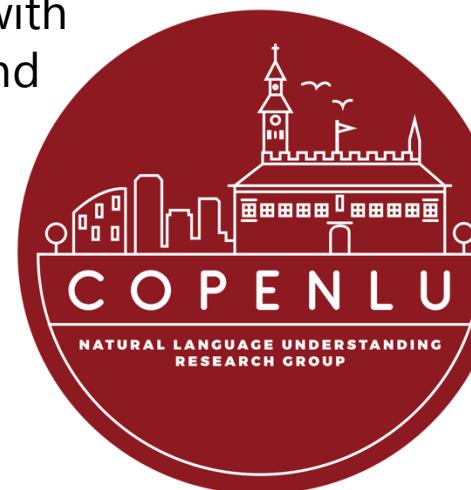
DIGITAL
TECH
SUMMIT



About Me

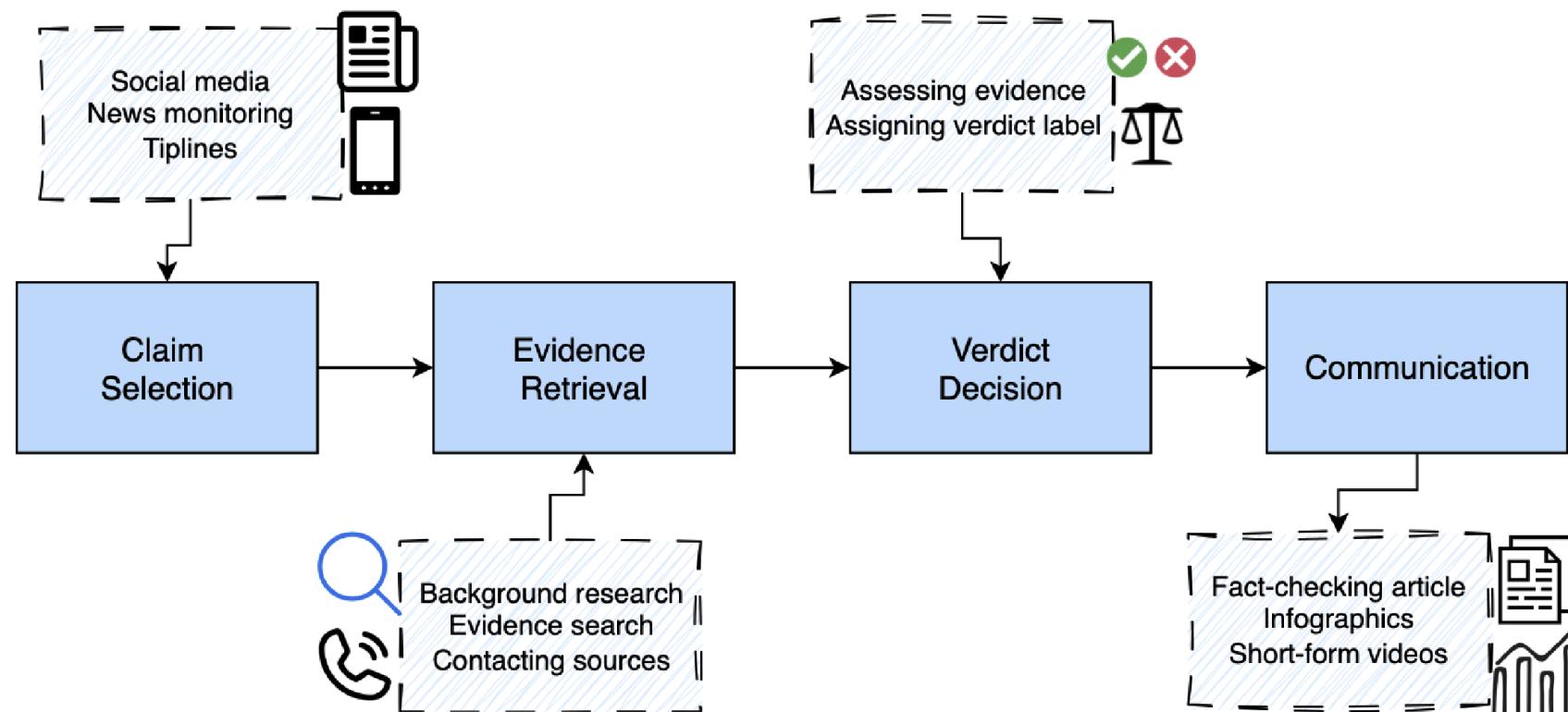
Research Interests:

- **Factuality in LMs:** Addressing the challenge of maintaining factual accuracy in language models.
- **Explainability Methods:** Designing Robust and user-aligned explainability techniques that enhance the understanding of complex models.
- **Interpretability of Language Models:** Understanding mechanisms of LLMs with some applications to context usage and parametric knowledge.



*Pioneer Center for AI
University of Copenhagen, Denmark*

Journalistic Fact Checking - How?



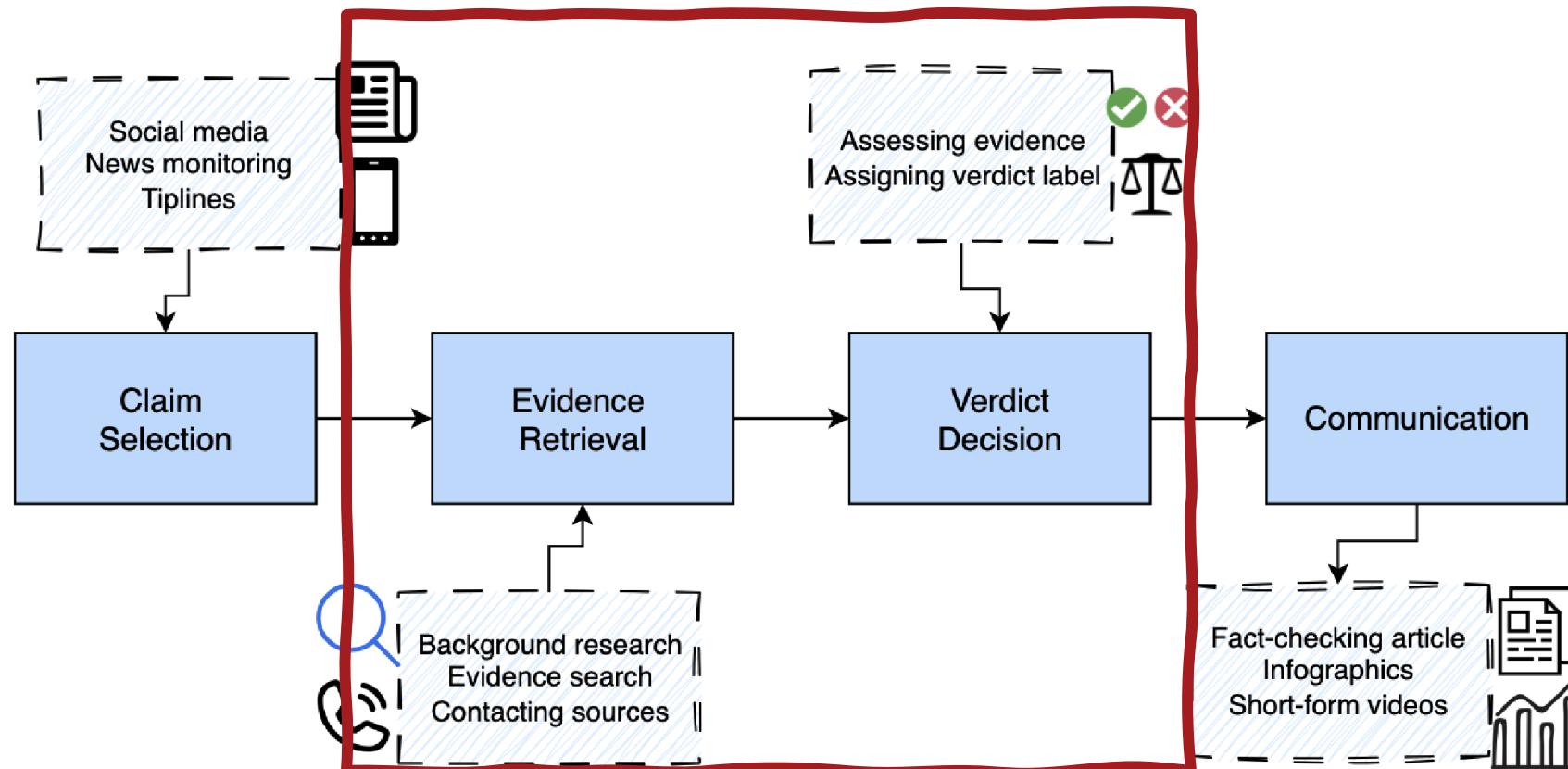
Lovisa Hagström, Sara Vera Marjanović, Haeun Yu, Arnav Arora, Christina Lioma, Maria Maistro, Pepa Atanasova, Isabelle Augenstein.

A Reality Check on Context Utilisation for Retrieval-Augmented Generation. ACL 2025

Image credits: Isabelle Augenstein

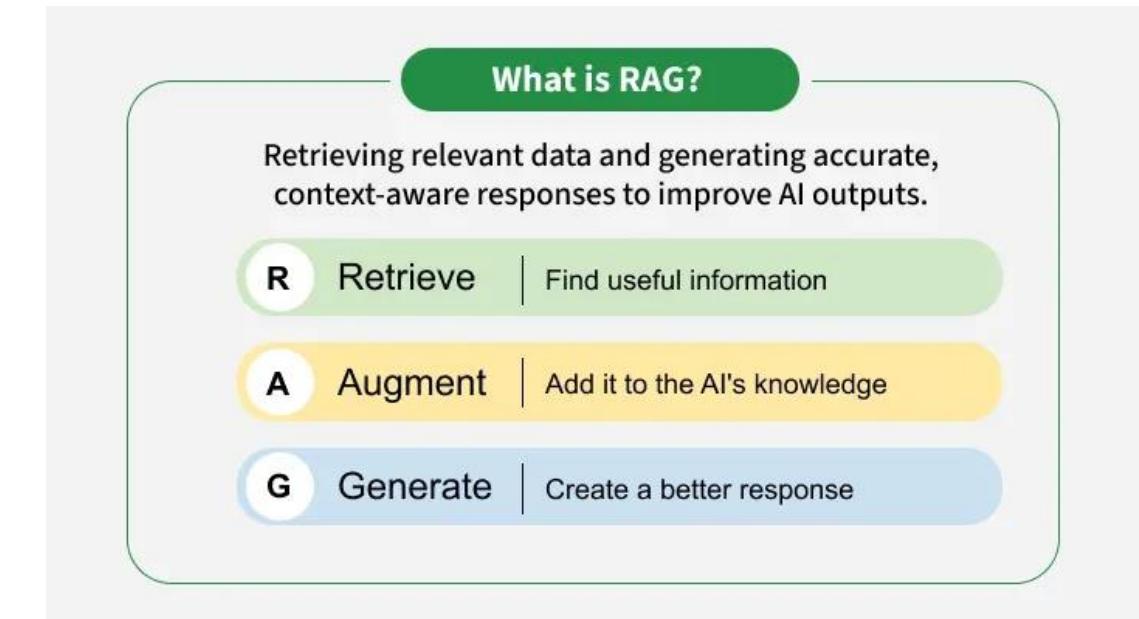
Journalistic Fact Checking - How?

Retrieval Augmented Generation (RAG)



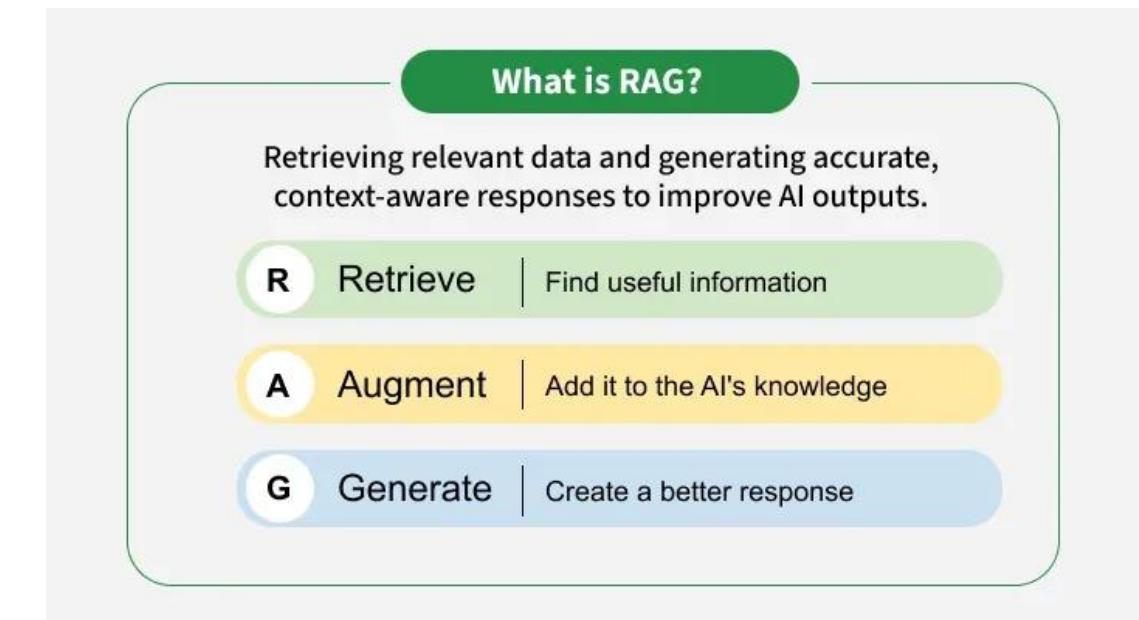
The RAG Revolution: Promise vs Reality

- RAG promises to solve LLM knowledge limitations
- Success depends on two critical factors:
 - Quality of retrieved information
 - Model's ability to utilize that information effectively
- Current research uses synthetic datasets that don't reflect real-world complexity



The RAG Revolution: Promise vs Reality

- RAG promises to solve LLM knowledge limitations
- Success depends on two critical factors:
 - Quality of retrieved information
 - Model's ability to utilize that information effectively
- Current research uses synthetic datasets that don't reflect real-world complexity
- Key question: How do LLMs actually perform with **messy, real-world evidence?**



Synthetic vs Reality: The Disconnect

- Most studies of how retrieved context is utilised use artificial datasets
 - Template-based, overly simplistic scenarios
 - Perfect evidence that always has clear stance
 - Unrealistic knowledge conflicts

CounterFact

Context #1

The capital of Japan is Stockholm.



Context #2

The capital of Japan is definitely ~~100~~ Stockholm.



Query

Q: What is the capital of Japan?

Controlled ✓
Realistic ✗
Real-world ✗

[Yu et al. \(2023\)](#)
[Du et al. \(2024\)](#)

ConflictQA

Context

George Rankin graduated from Harvard Law School in 2005 and has been practicing law for the past 15 years...



Query

What is George Rankin's occupation?

Controlled ✓
Realistic ✓
Real-world ✗

[Xie et al. \(2024\)](#)

Context characteristics

	knowledge conflict		unreliable
	assertive		hedging
	generated		insufficient

Synthetic vs Reality: The Disconnect

- Real-world evidence is messy:
 - Often insufficient or unclear (50% in our findings)
 - Contains hedging and uncertainty markers
 - Comes from unreliable sources
 - May contradict itself

The slide is titled "Synthetic vs Reality: The Disconnect". It features two examples of real-world evidence from CES 2019:

- Context #1:** CES 2019: Scientists have developed a blood pressure monitoring app to replace the 100-year-old cuff. [...] The Biospectal app, still in testing, could? essentially replace the traditional blood pressure cuff.
- Context #2:** FULL CLAIM: Blood pressure tracking apps can replace a cuff [...] Despite the way it was shown in the promotional Facebook post, there is no indication that the app is able to measure blood pressure. Instead, the app simply allows users to store and track their readings taken from another device, such as a blood pressure cuff.

A query is posed: "Is it true that ‘blood pressure tracking apps can replace a cuff’?" The answer is: Controlled ✓, Realistic ✓, Real-world ✓.

Synthetic vs Reality: The Disconnect

- Real-world evidence is messy:
 - Often insufficient or unclear (50% in our findings)
 - Contains hedging and uncertainty markers
 - Comes from unreliable sources
 - May contradict itself

This gap leads to unrealistic performance estimates.

DRUID

Context #1

CES 2019: Scientists have developed a blood pressure monitoring app to replace the 100-year-old cuff. [...] The Biospectal app, still in testing, could? essentially replace the traditional blood pressure cuff.

Query

Is it true that “blood pressure tracking apps can replace a cuff”?

Controlled ✓
Realistic ✓
Real-world ✓

Our work

Context #2

FULL CLAIM: Blood pressure tracking apps can replace a cuff [...] Despite the way it was shown in the promotional Facebook post, there is no indication that the app is able to measure blood pressure. Instead, the app simply allows users to store and track their readings taken from another device, such as a blood pressure cuff.

DRUID: A Reality Check Dataset

- Dataset of Retrieved Unreliable, Insufficient and Difficult-to-understand contexts for fact verification.
- 5,490 real-world claim-evidence pairs from 7 fact-checking sources.
- Manually annotated for relevance and stance.

The screenshot shows the DRUID interface with two contexts for fact-checking:

Context #1
CES 2019:
Scientists have developed a blood pressure monitoring app to replace the 100-year-old ! cuff. [...] The Biospectal app, still in testing, could? essentially replace the traditional blood pressure cuff. !

Query
Is it true that “blood pressure tracking apps can replace a cuff”?

**Controlled ✓
Realistic ✓
Real-world ✓**

Context #2
FULL CLAIM:
Blood pressure tracking apps can replace a ! cuff [...] Despite the way it was shown in the promotional Facebook post, there is no indication that the app is able to measure blood pressure. Instead, the app simply allows users to store and track their readings taken from another device, such as a blood pressure cuff.

Our work

DRUID: A Reality Check Dataset

- Key findings:
 - 50% of automatically retrieved contexts are insufficient
 - 34% of claims have conflicting evidence
 - Average evidence length: 3x longer than synthetic datasets
- Represents actual RAG retrieval scenarios.

The screenshot shows the DRUID interface with two contexts and a query section.

Context #1: CES 2019: Scientists have developed a blood pressure monitoring app to replace the 100-year-old cuff. [...] The Biospectal app, still in testing, could? essentially replace the traditional blood pressure cuff.

Context #2: FULL CLAIM: Blood pressure tracking apps can replace a cuff [...] Despite the way it was shown in the promotional Facebook post, there is no indication that the app is able to measure blood pressure. Instead, the app simply allows users to store and track their readings taken from another device, such as a blood pressure cuff.

Query: Is it true that “blood pressure tracking apps can replace a cuff”?

Controlled: ✓
Realistic: ✓
Real-world: ✓

Real Evidence Looks Nothing Like Synthetic Data

- Real-world retrieved evidence shows:
 - Higher reading difficulty (lower Flesch scores)
 - More uncertainty markers and hedging language
 - Greater implicitness in connections to claims
 - Varied source reliability
- Memory conflicts less prevalent in reality.

 DRUID

Our work

Context #1	Context #2
<p>CES 2019: Scientists have developed a <u>blood pressure monitoring app</u> to replace the <u>100-year-old</u>  cuff. [...] The Biospectal app, still in testing, <u>could?</u> essentially <u>replace the traditional blood pressure cuff.</u> </p> <p>Query</p> <p>Is it true that “blood pressure tracking apps can replace a cuff”?</p> <p>Controlled ✓ Realistic ✓ Real-world ✓</p>	<p><u>FULL CLAIM:</u> <u>Blood pressure tracking apps</u> can replace a  <u>cuff</u> [...] Despite the way it was shown in the promotional Facebook post, there is no indication that the app is able to measure blood pressure. Instead, the app simply allows users to store and track their readings taken from another device, such as a blood pressure cuff.</p>

Real Evidence Looks Nothing Like Synthetic Data

- Real-world retrieved evidence shows:
 - Higher reading difficulty (lower Flesch scores)
 - More uncertainty markers and hedging language
 - Greater implicitness in connections to claims
 - Varied source reliability
- Memory conflicts less prevalent in reality.

Implication: Models trained/tested on synthetic data may fail in production.

DRUID

Context #1

CES 2019: Scientists have developed a blood pressure monitoring app to replace the 100-year-old ⚠ cuff. [...] The Biospectal app, still in testing, could? essentially replace the traditional blood pressure cuff. ⚠

Query

Is it true that “blood pressure tracking apps can replace a cuff”?

Controlled ✓
Realistic ✓
Real-world ✓

Our work

Context #2

FULL CLAIM: Blood pressure tracking apps can replace a ⚠ cuff [...] Despite the way it was shown in the promotional Facebook post, there is no indication that the app is able to measure blood pressure. Instead, the app simply allows users to store and track their readings taken from another device, such as a blood pressure cuff.

Models Usually Utilize Real Evidence

- Synthetic datasets show "context repulsion"
- Real-world behavior is different:
 - Less preference for supporting evidence
 - Rarely see context repulsion
 - More balanced utilization patterns

Models Usually Utilize Real Evidence

- Synthetic datasets show "context repulsion"
- Real-world behavior is different:
 - Less preference for supporting evidence
 - Rarely see context repulsion
 - More balanced utilization patterns

Different models (e.g. Llama vs Pythia) show dramatically different behaviors

Context Usage performance on synthetic \neq performance on real data

Individual Features Don't Predict Success

- Traditional focus on single characteristics (length, similarity, perplexity) shows weak correlations
- What actually matters:
 - Source credibility (fact-checking sites: +0.6 correlation)
 - Aggregated feature combinations
 - Context published after claim made
- References to external sources? Nearly irrelevant
- Claim-evidence similarity? Low impact in real scenarios

Individual Features Don't Predict Success

- Traditional focus on single characteristics (length, similarity, perplexity) shows weak correlations
- What actually matters:
 - Source credibility (fact-checking sites: +0.6 correlation)
 - Aggregated feature combinations
 - Context published after claim made
- References to external sources? Nearly irrelevant
- Claim-evidence similarity? Low impact in real scenarios

RAG failure causes are more complex than previously thought