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

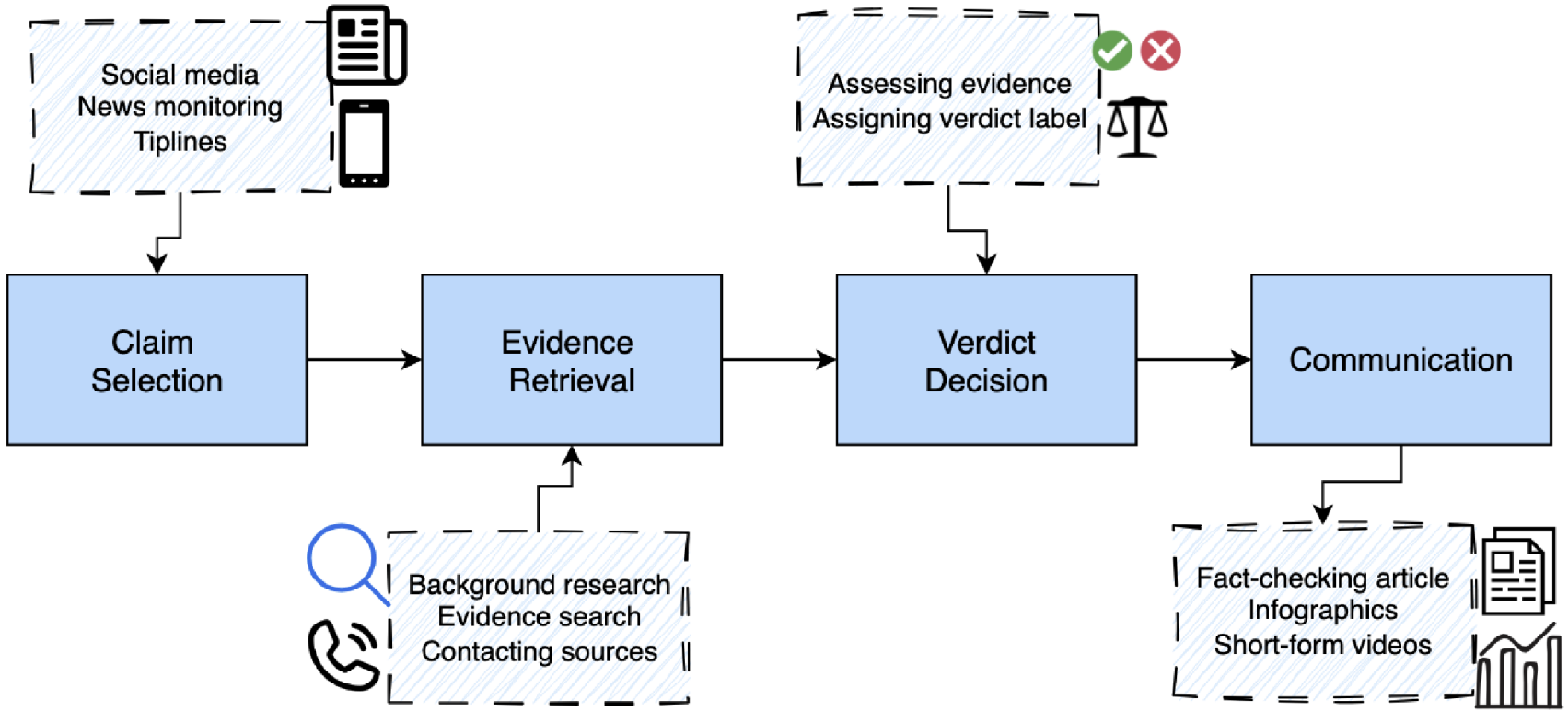
Research Interests:

- 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.

COPENLU
NATURAL LANGUAGE UNDERSTANDING RESEARCH GROUP

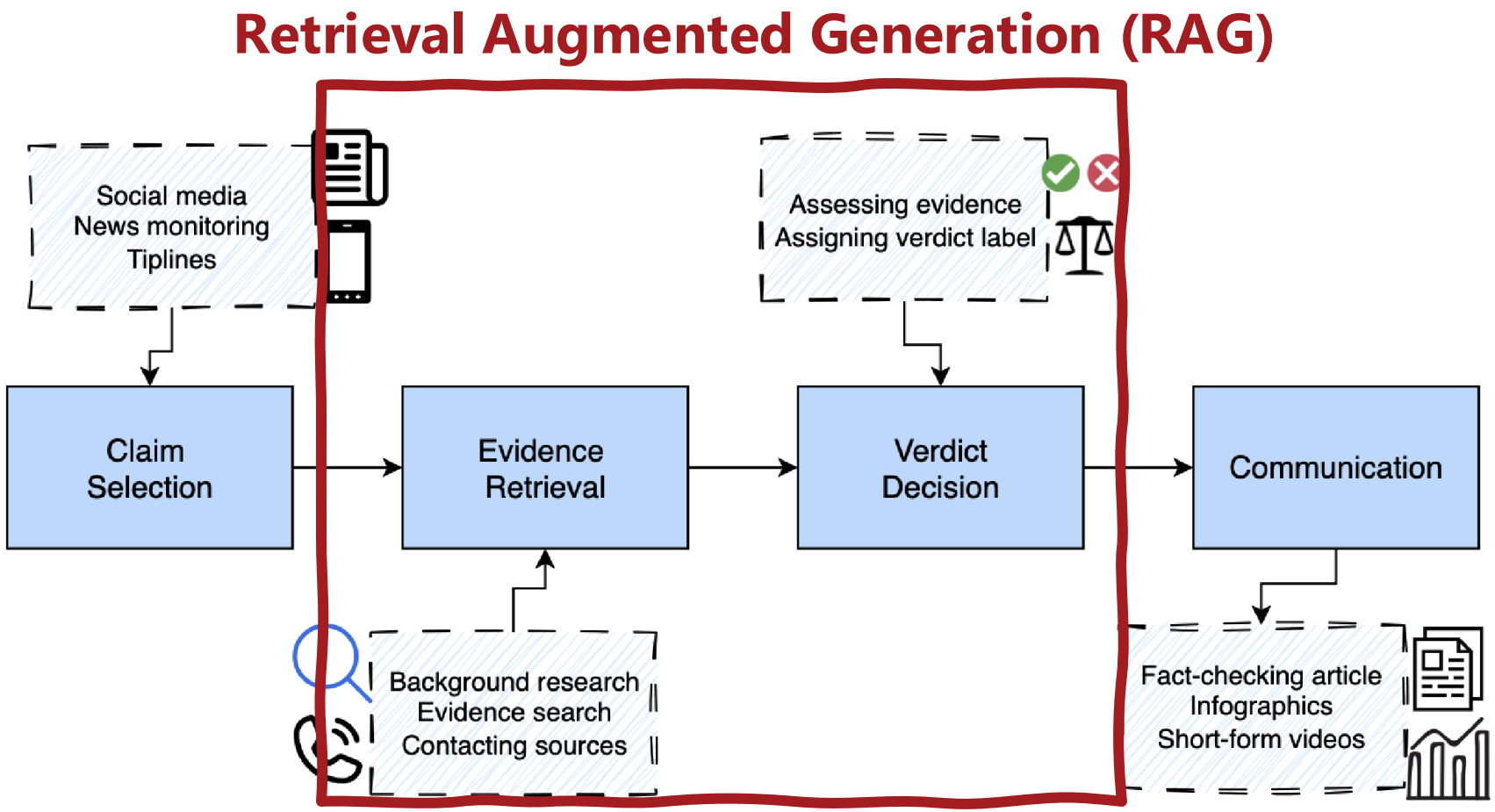
Pioneer Center for AI
University of Copenhagen, Denmark

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

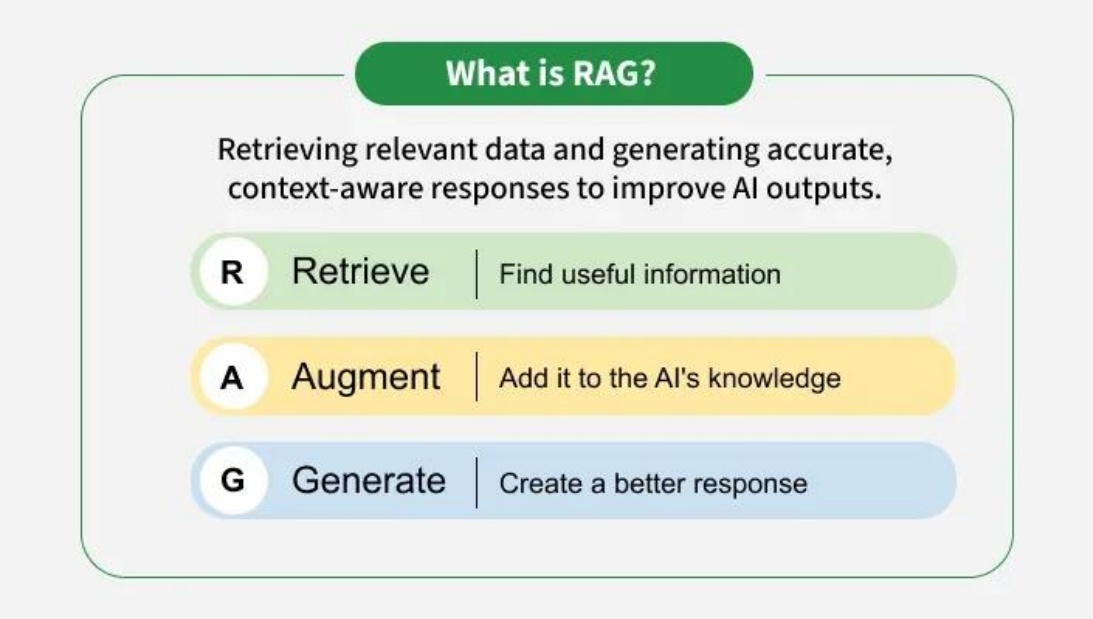
Journalistic Fact Checking - How?



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

The RAG Revolution: Promise vs Reality

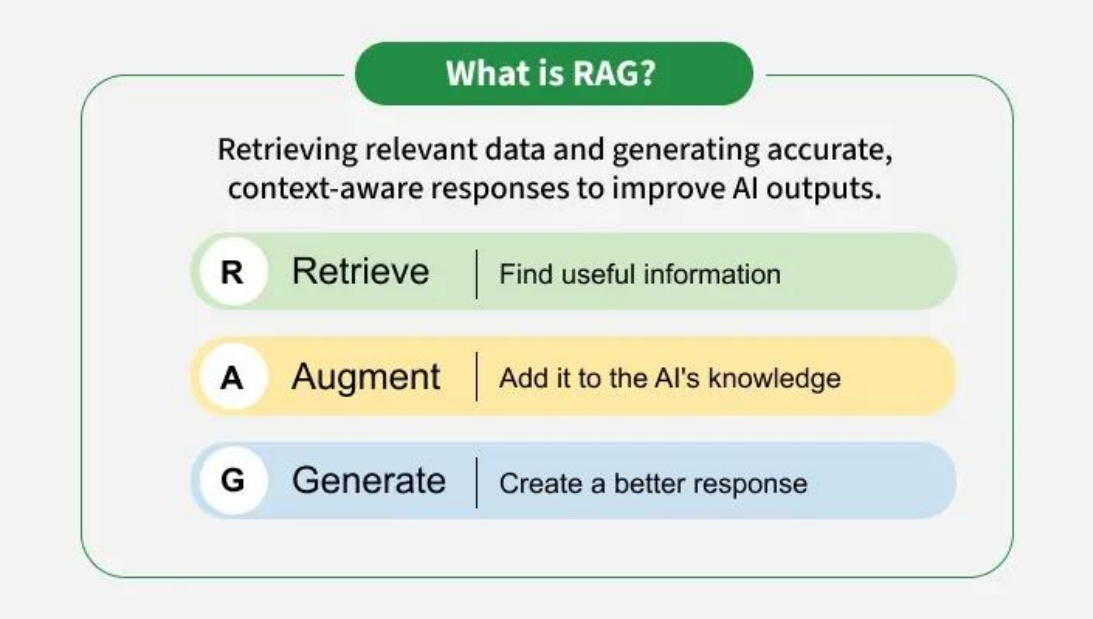
- RAG promises to solve LLM knowledge limitations
- Success depends on two critical factors:
  - Quality of retrieved information
  - Model's ability to utilize that information effectively
- Current research uses synthetic datasets that don't reflect real-world complexity



The RAG Revolution: Promise vs Reality

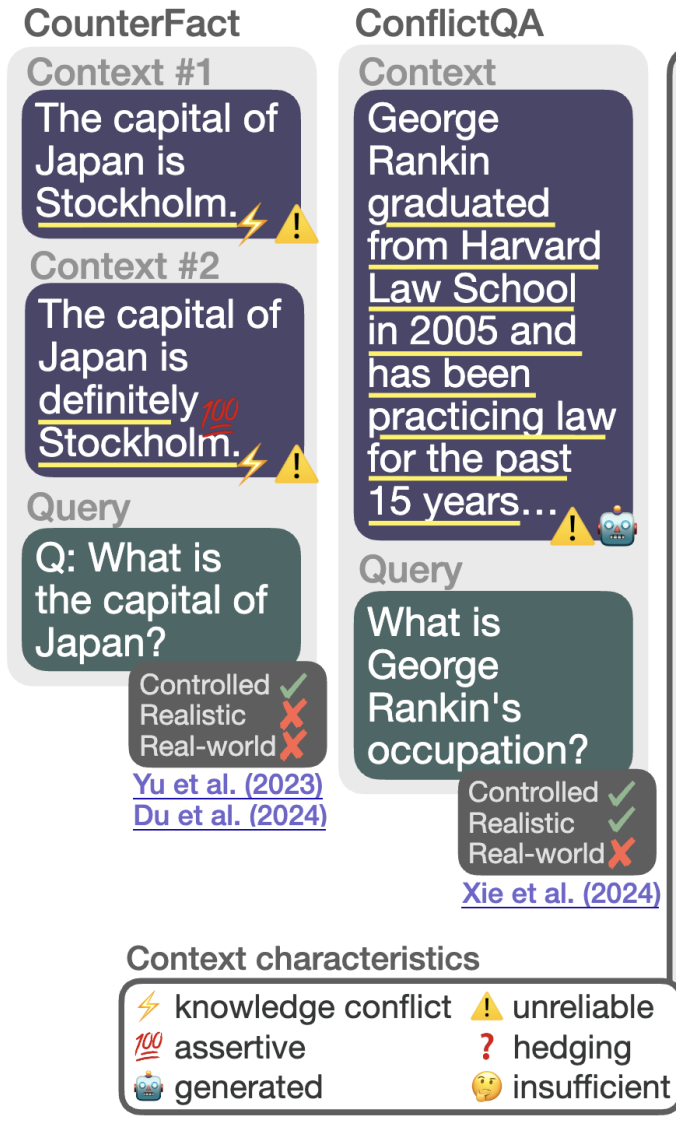
- RAG promises to solve LLM knowledge limitations
- Success depends on two critical factors:
  - Quality of retrieved information
  - Model's ability to utilize that information effectively
- Current research uses synthetic datasets that don't reflect real-world complexity

- Key question: How do LLMs actually perform with **messy, real-world evidence**?



Synthetic vs Reality: The Disconnect

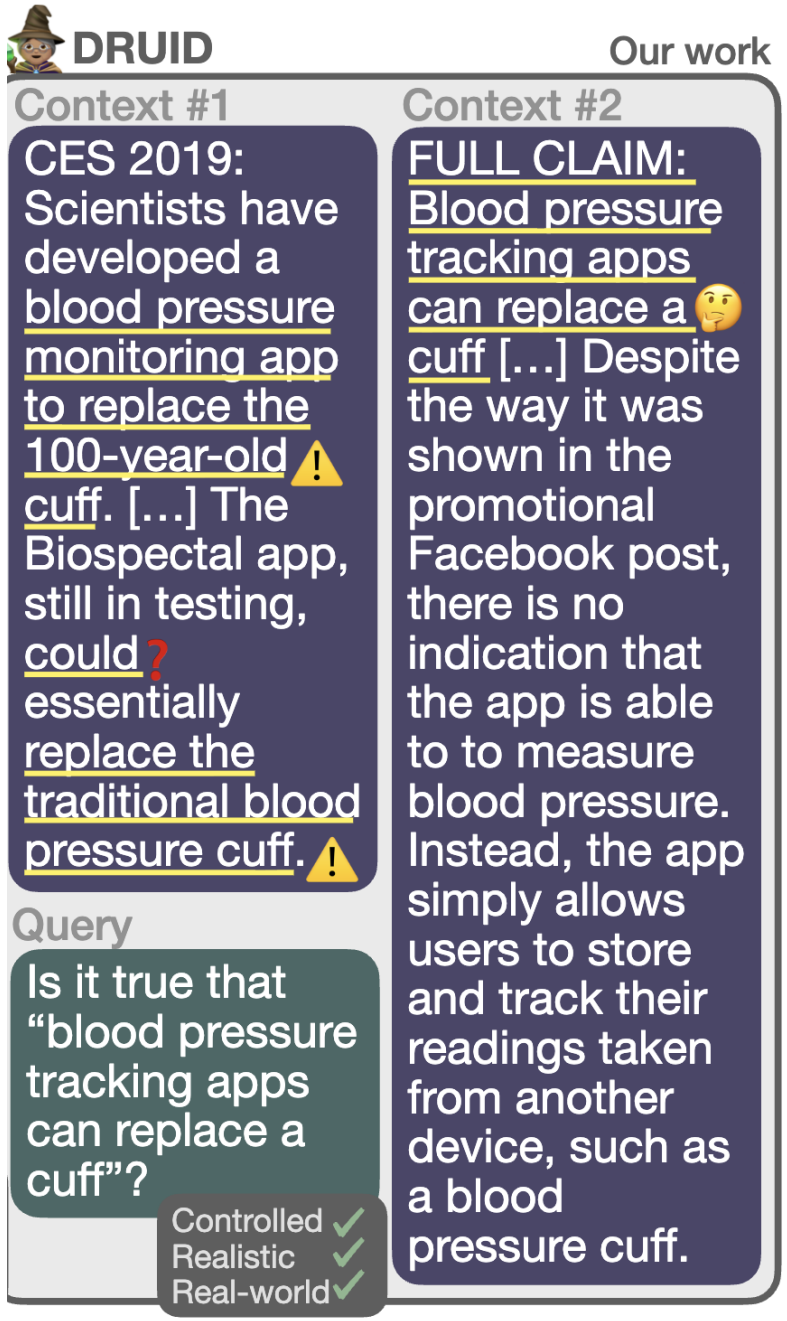
- Most studies of how retrieved context is utilised use artificial datasets
  - Template-based, overly simplistic scenarios
  - Perfect evidence that always has clear stance
  - Unrealistic knowledge conflicts



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

Synthetic vs Reality: The Disconnect

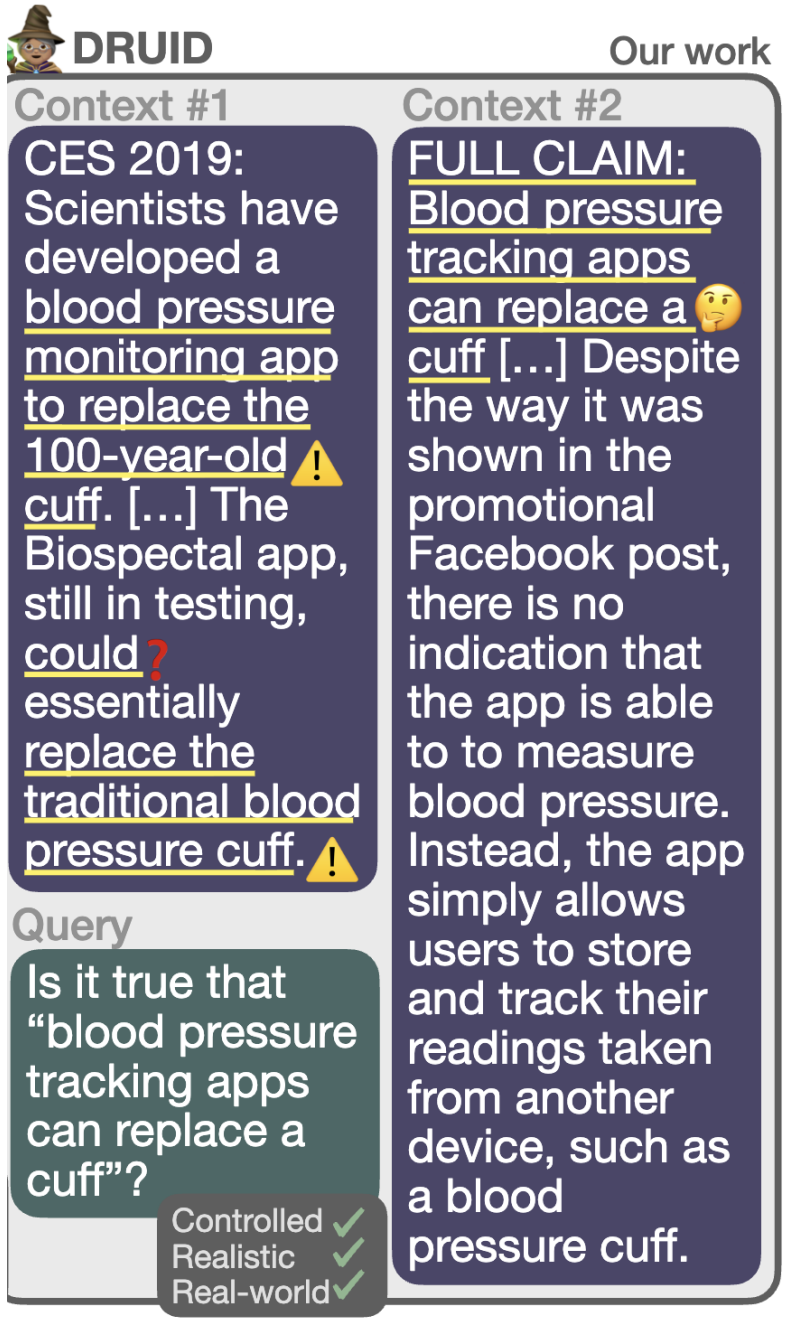
- Real-world evidence is messy:
  - Often insufficient or unclear (50% in our findings)
  - Contains hedging and uncertainty markers
  - Comes from unreliable sources
  - May contradict itself



Synthetic vs Reality: The Disconnect

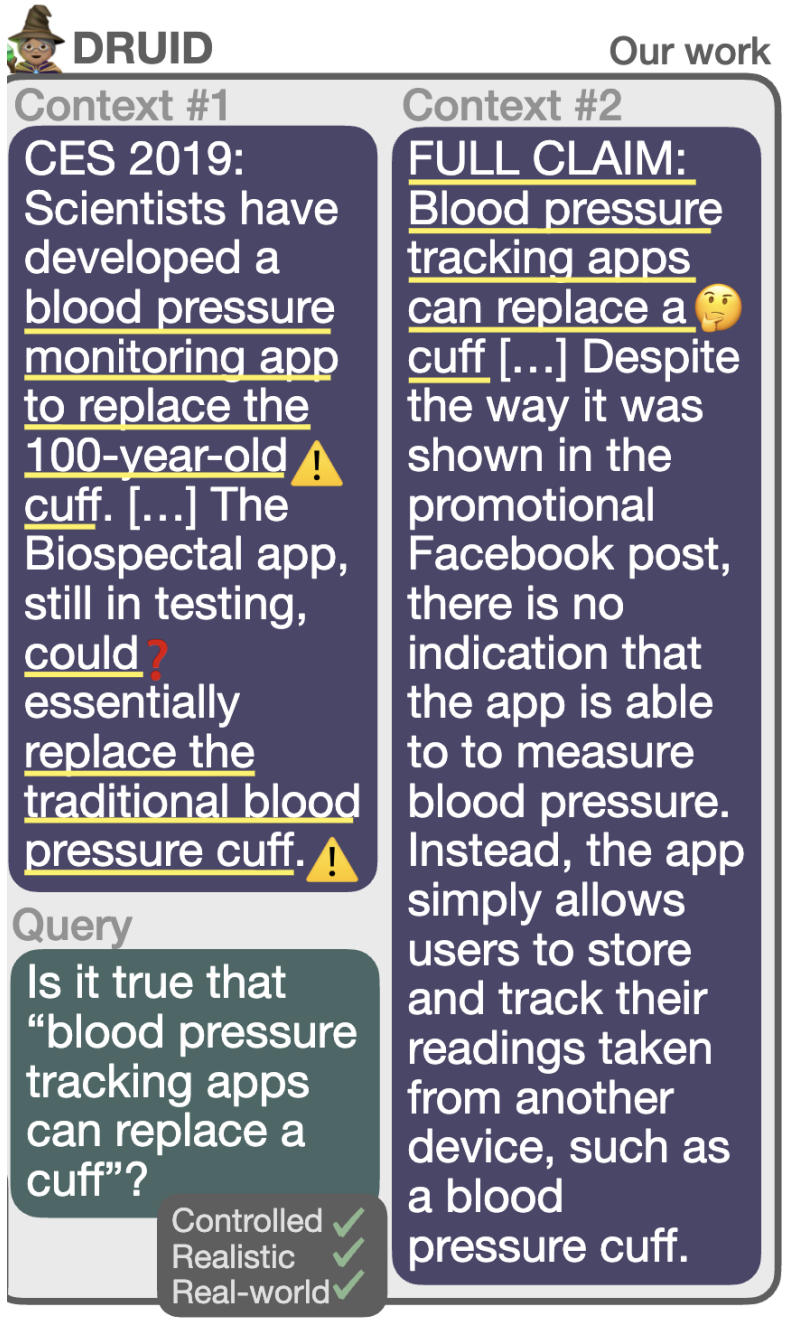
- Real-world evidence is messy:
  - Often insufficient or unclear (50% in our findings)
  - Contains hedging and uncertainty markers
  - Comes from unreliable sources
  - May contradict itself

This gap leads to unrealistic performance estimates.



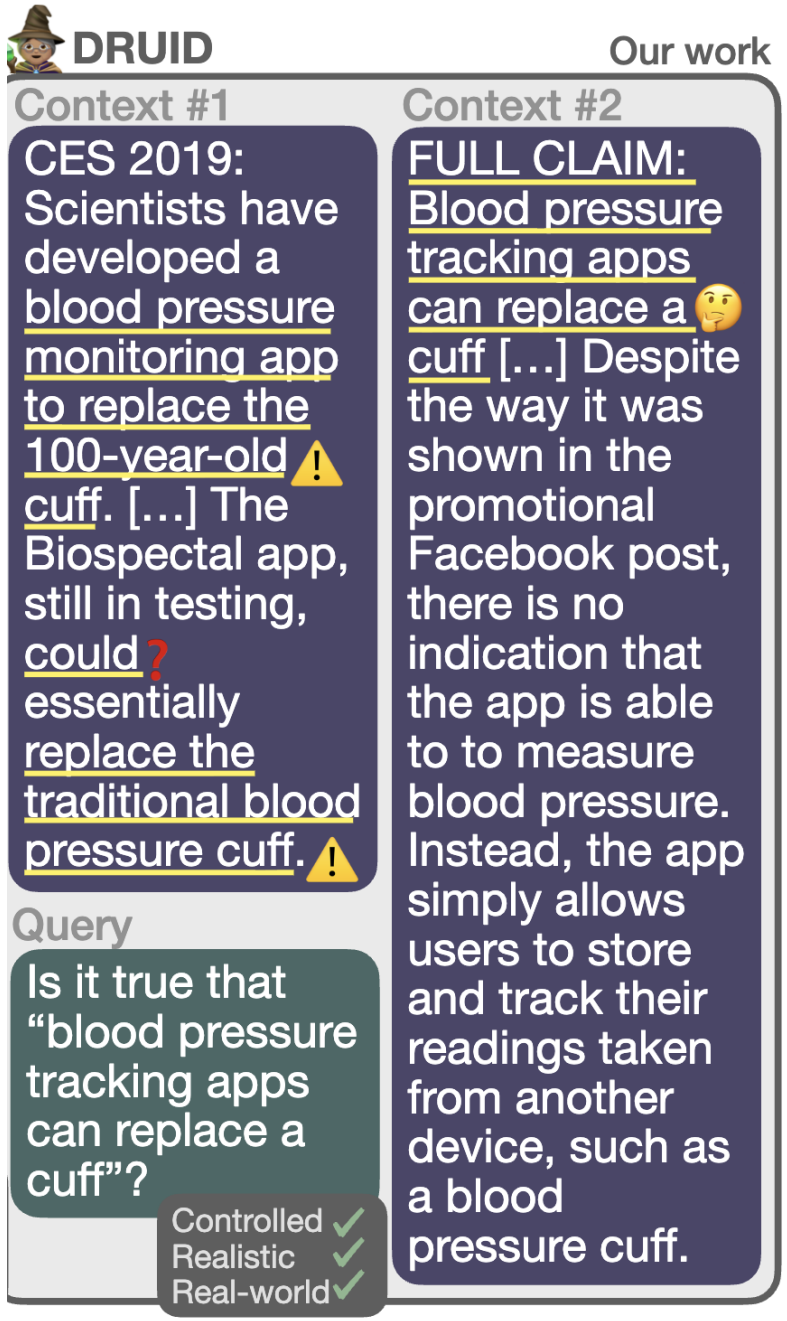
DRUID: A Reality Check Dataset

- 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.



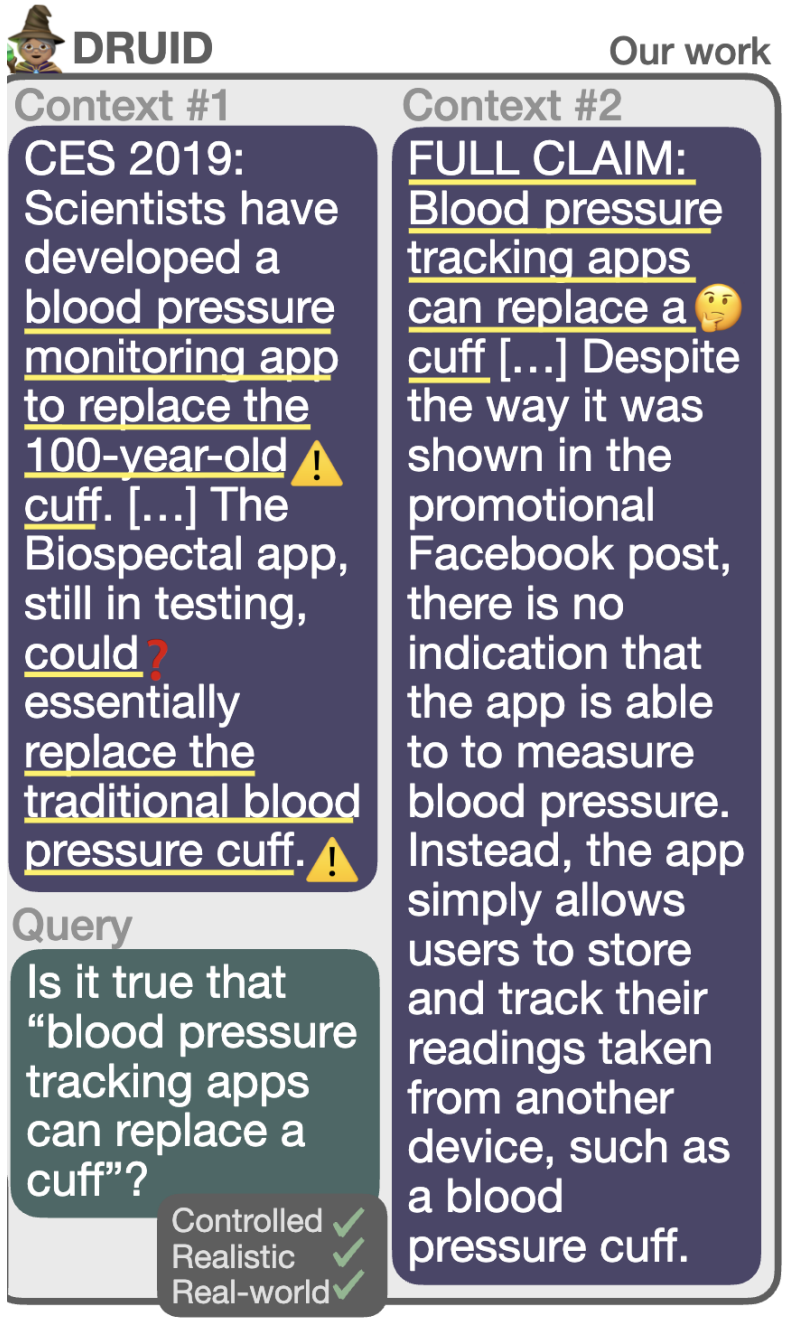
DRUID: A Reality Check Dataset

- 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.



Real Evidence Looks Nothing Like Synthetic Data

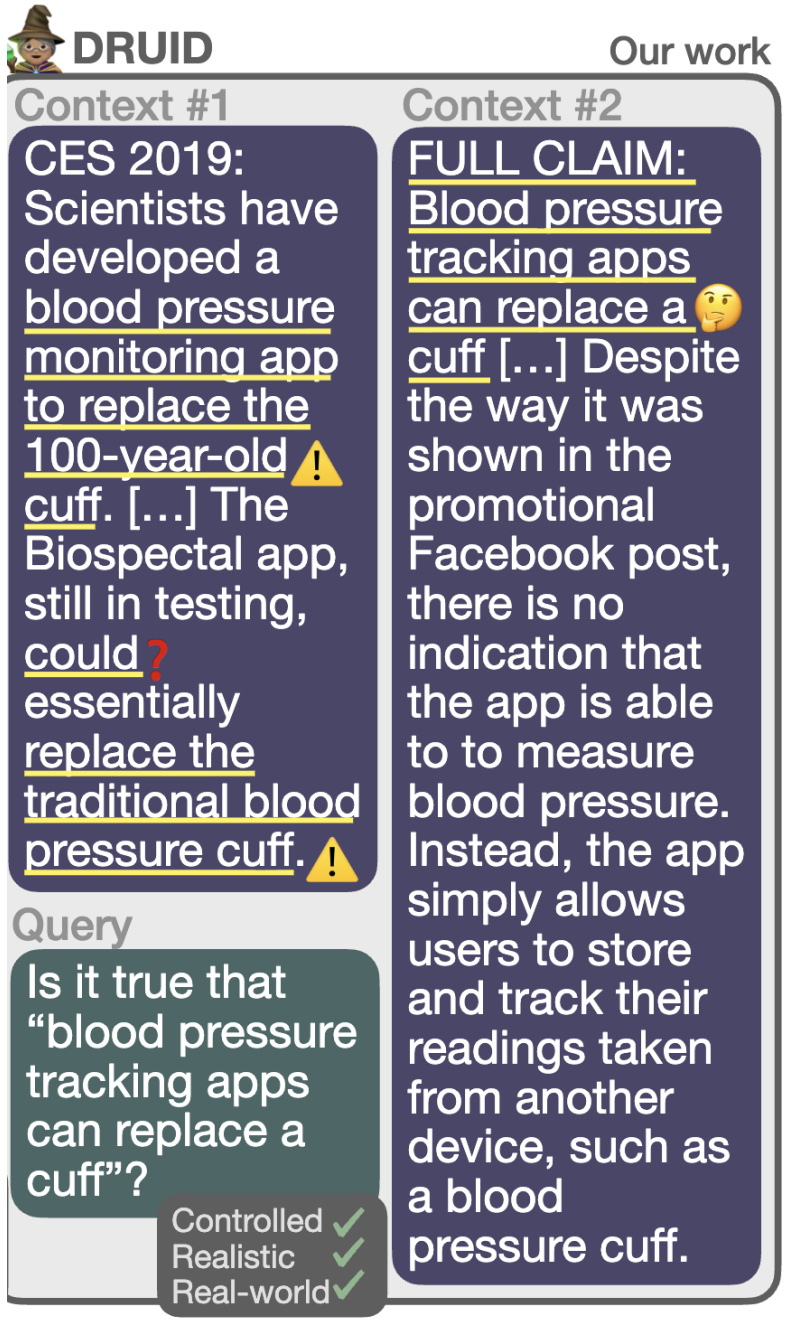
- Real-world retrieved evidence shows:
  - Higher reading difficulty (lower Flesch scores)
  - More uncertainty markers and hedging language
  - Greater implicitness in connections to claims
  - Varied source reliability
- Memory conflicts less prevalent in reality.



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.



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 ≠ 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