AI Image
Generated by DELL-E 3

CSci Senior Seminar Spring 2024

## Unmasking Misinformation:

# The Potential of Natural Language Processing (NLP)



**Dongting Cai**
University of Minnesota, Morris
cai00151@morris.umn.edu


Ver 4.0 Fin_Fix 04/14/2024

# Work Based On…

**Missing Counter-Evidence Renders NLP Fact-Checking Unrealistic for Misinformation** ▸ Glockner et al, 2022

**A Survey on Automated Fact-Checking** ▸ Guo et al, 2022

## Missing Counter-Evidence Renders NLP Fact-Checking Unrealistic for Misinformation


**Max Glockner[1], Yufang Hou[2], Iryna Gurevych[1]**
[1] Ubiquitous Knowledge Processing Lab (UKP Lab),
Department of Computer Science and Hessian Center for AI (hessian.AI),
Technical University of Darmstadt
[2] IBM Research Europe, Ireland
www.ukp.tu-darmstadt.de, yhou@ie.ibm.com



### Abstract

Misinformation emerges in times of uncertainty when credible information is limited. This is challenging for NLP-based fact-checking as it relies on counter-evidence, which may not yet be available. Despite increasing interest in automatic fact-checking, it is still unclear if automated approaches can realistically refute harmful real-world misinformation. Here, we contrast and compare NLP fact-checking with how professional fact-checkers combat misinformation in the absence of counter-evidence. In our analysis, we show that, by design, existing NLP task definitions for fact-checking cannot refute misinformation as professional fact-checkers do for the majority of claims. We then define two requirements that the evidence in datasets must fulfill for realistic fact-checking: It must be (1) sufficient to refute the claim and (2) not leaked from existing fact-checking articles. We survey existing fact-checking datasets and find that all of them fail to satisfy both criteria. Finally, we perform experiments to demonstrate that models trained on a large-scale fact-checking dataset rely on leaked evidence, which makes them unsuitable in real-world scenarios. Taken together, we show that current NLP fact-checking cannot realistically combat real-world misinformation because it depends on unrealistic assumptions about counter-evidence in the data[1].

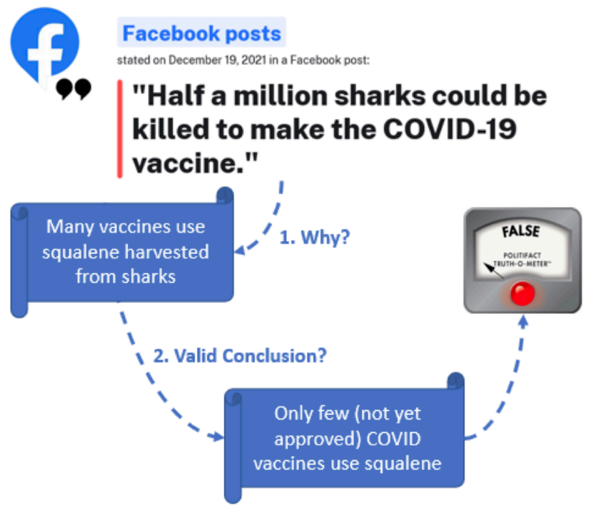




Figure 1: A false claim from PolitiFact. It is unlikely to find counter-evidence. Fact-checkers refute the claim by disproving why it was made.

### 1 Introduction

According to van der Linden (2022), misinformation is "*false or misleading information masquerading as legitimate news, regardless of intent*". Misinformation is dangerous as it can directly impact human behavior and have harmful real-world consequences such as the Pizzagate shooting (Fisher et al., 2016), interfering in the 2016 democratic US election (Bovet and Makse, 2019), or the promotion of false COVID-19 cures (Aghababaeian et al., 2020). Surging misinformation during the COVID-19 pandemic, coined "infodemic" by WHO (Zarocostas, 2020), exemplifies the danger coming from misinformation. To combat misinformation, journalists from fact-checking organizations (e.g., PolitiFact or Snopes) conduct a laborious manual effort to verify claims based on possible harms and their prominence (Arnold, 2020). However, manual fact-checking cannot keep pace with the rate at which misinformation is posted and circulated. Automatic fact-checking has gained significant attention within the NLP community in recent years, with the goal of developing tools to assist fact-checkers in combating misinformation. For the past few years, NLP researchers have created a wide range of fact-checking datasets with claims from fact-checking organization websites (Vlachos and Riedel, 2014; Wang, 2017; Augenstein et al., 2019; Hanselowski et al., 2019; Ostrowski et al., 2021; Gupta and Srikumar, 2021; Khan et al., 2022). The fundamental goal of fact-checking is, given a *claim* made by a *claimant*, to find a collection of *evidence* and provide a *verdict* about the claim's veracity based

[1]Code provided at https://github.com/UKPLab/emnlp2022-missing-counter-evidence

## A Survey on Automated Fact-Checking


**Zhijiang Guo*, Michael Schlichtkrull*, Andreas Vlachos**
Department of Computer Science and Technology
University of Cambridge, UK
{zg283,mss84,av308}@cam.ac.uk



### Abstract

Fact-checking has become increasingly important due to the speed with which both information and misinformation can spread in the modern media ecosystem. Therefore, researchers have been exploring how fact-checking can be automated, using techniques based on natural language processing, machine learning, knowledge representation, and databases to automatically predict the veracity of claims. In this paper, we survey automated fact-checking stemming from natural language processing, and discuss its connections to related tasks and disciplines. In this process, we present an overview of existing datasets and models, aiming to unify the various definitions given and identify common concepts. Finally, we highlight challenges for future research.


### 1 Introduction

Fact-checking is the task of assessing whether claims made in written or spoken language are true. This is an essential task in journalism, and is commonly conducted manually by dedicated organizations such as PolitiFact. In addition to *external* fact-checking, *internal* fact-checking is also performed by publishers of newspapers, magazines, and books prior to publishing in order to promote truthful reporting. Figure 1 shows an example from PolitiFact, together with the evidence (summarized) and the verdict.

Fact-checking is a time-consuming task. To assess the claim in Figure 1, a journalist would need to search through potentially many sources to find job gains under Trump and Obama, evaluate the reliability of each source, and make a comparison. This process can take professional fact-checkers several hours or days (Hassan et al., 2015; Adair et al., 2017). Compounding the problem, fact-checkers often work under strict and tight deadlines, especially in the case of internal processes (Borel, 2016; Godler and Reich, 2017), and some studies have shown that less than half of all published articles have been subject to verification (Lewis et al., 2008). Given the amount of new information that appears and the speed with which it spreads, manual validation is insufficient.

Automating the fact-checking process has been discussed in the context of computational journalism (Flew et al., 2010; Cohen et al., 2011; Graves, 2018), and has received significant attention in the artificial intelligence community. Vlachos and Riedel (2014) proposed structuring it as a sequence of components—identifying claims to be checked, finding appropriate evidence, producing verdicts—that can be modeled as natural language processing (NLP) tasks. This motivated the development of automated pipelines consisting of subtasks that can be mapped to tasks well-explored in the NLP community. Advances were made possible by the development of datasets, consisting of either claims collected from fact-checking websites, for example Liar (Wang, 2017), or purpose-made for research, for example, FEVER (Thorne et al., 2018a).

A growing body of research is exploring the various tasks and subtasks necessary for the automation of fact checking, and to meet the need for new methods to address emerging challenges. Early developments were surveyed in Thorne and Vlachos (2018), which remains the closest to an exhaustive overview of the subject. However, their proposed framework does not include work on determining *which* claims to verify (i.e., claim detection), nor does their survey include the recent work on producing explainable, convincing verdicts (i.e., justification production).

Several recent papers have surveyed research focusing on individual components of the task. Zubiaga et al. (2018) and Islam et al. (2020) focus on identifying rumors on social media, Küçük and Can (2020) and Hardalov et al. (2021)

*Equal contribution.

OUTLINE

PART 1

## Background

PART 2

## How Humans Fact-Check

PART 3

## NLP Fact-Checking: Techniques

PART 4

## Future Directions

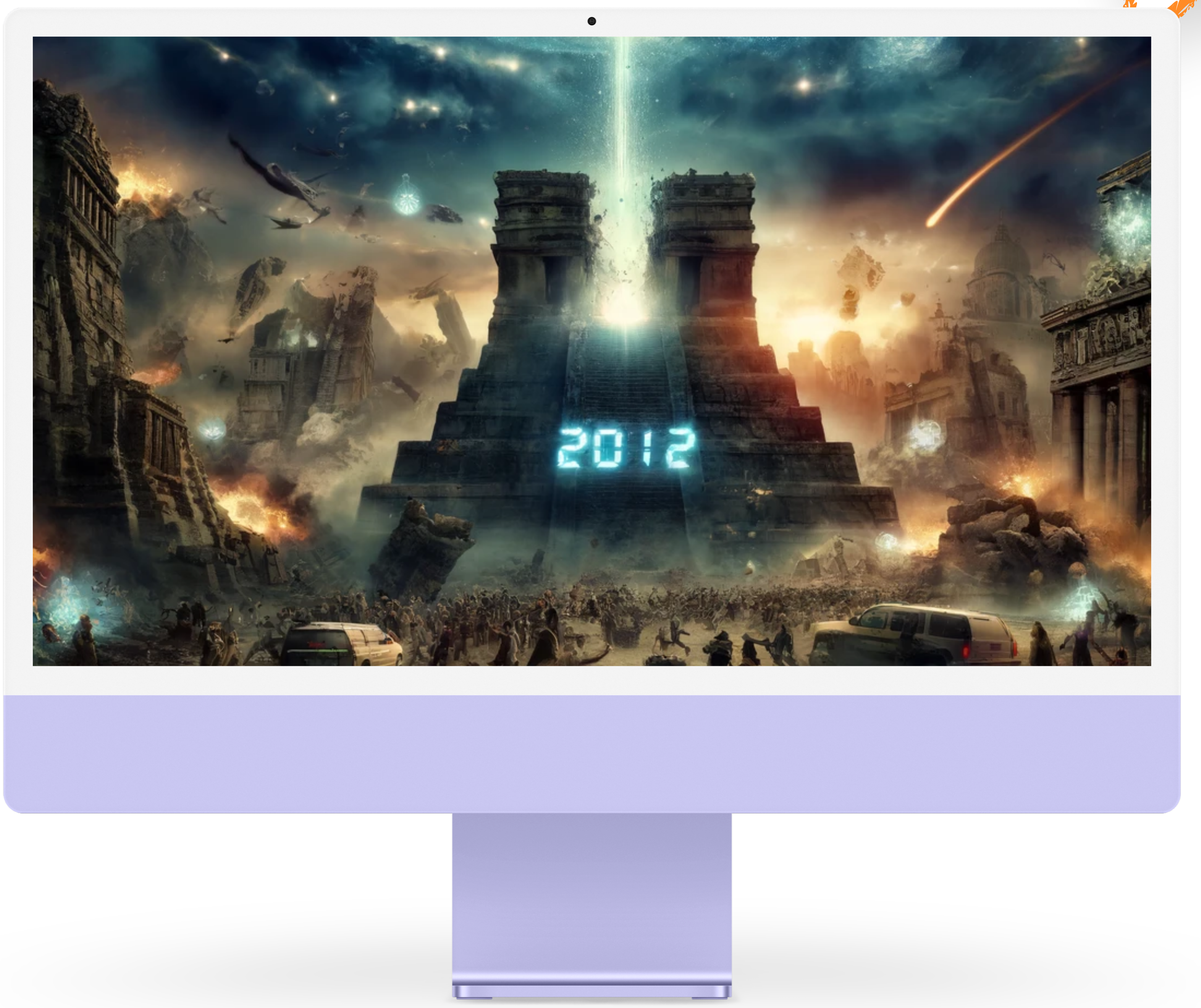
PART 1

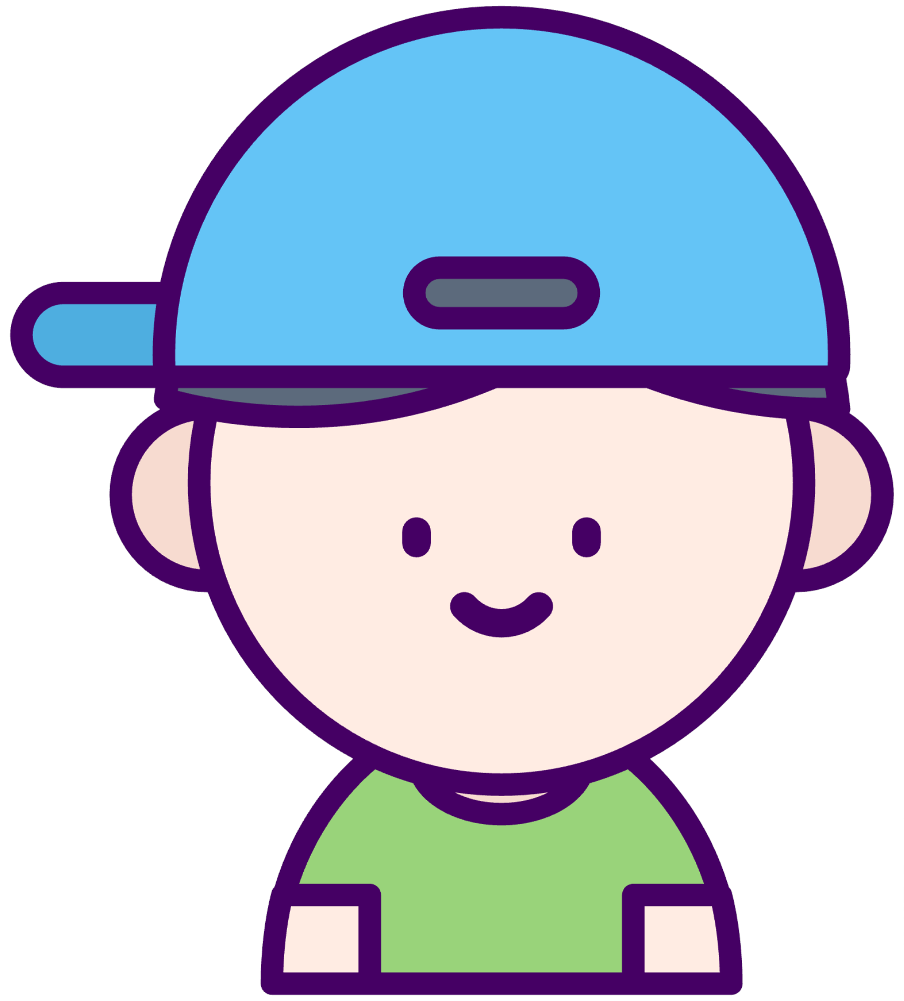
# Background

# Back to 2012...

Hello Mr. Wang, it's great to spend the 2012 prophecy days with you in your class. Let's look forward to seeing if we will ***die*** together.

Mayan rumors: Earth is about to become extinct



**10%**

People from
21 Countries
**Believe it's REAL**

AI Image
Generated by DELL·E 3

"You won't believe what this celebrity did next! FIND OUT NOW!"

"You won't believe what this cele

imple trick will change the way you … FOREVER!"

"You won't believe what happened when … or su

believe the secret person has been HIDING all these years!"

"You won't believe what this celebrity

# Misinformation

lieve me - the doctor will NEVER TELL YOU THIS!"

"You won't believe what this celebrity did nex

mind! See the SHOCKING video here!"

You WON'T BELIEVE what happened when [something unexpe

n't believe what this celebrity did next! Find out now!"

"Find out THE UNTOLD TRUTH about … tha

# Negative of Misinformation

**6x**

**False News/Stories** spread
**Faster** than the True one on Twitter

**Misinformation brings:**
- Influencing public opinion
- Hurt trust in institutions
- Threatening public health and safety



Vosoughi et al, 2018
Lazer et al, 2018

PART 2

# How Humans Fact-Check

# Human Fact-Check Approaches

1 **Global Counter-Evidence (GCE)** | 2 Local Counter-Evidence (LCE) | 3 Non-Credible Sources (NCS) | 4 No Evidence Assertion (NEA)

Finding **counter-evidence** that refutes the claim through arbitrarily complex reasoning, **without** requiring a specific source guarantee

DailyFakeNews 3h

**Everyone knows,**
**All birds can fly** 🐦

↩ 2w 👍 66,823 ❤️ 66,820

→

Fact-Checker Own Knowledge

We know penguins
and ostriches are birds,
but they **can't fly**

→

× **False**



▸ Glockner et al, 2022

# Human Fact-Check Approaches

1 Global Counter-Evidence (GCE) | **2 Local Counter-Evidence (LCE)** | 3 Non-Credible Sources (NCS) | 4 No Evidence Assertion (NEA)

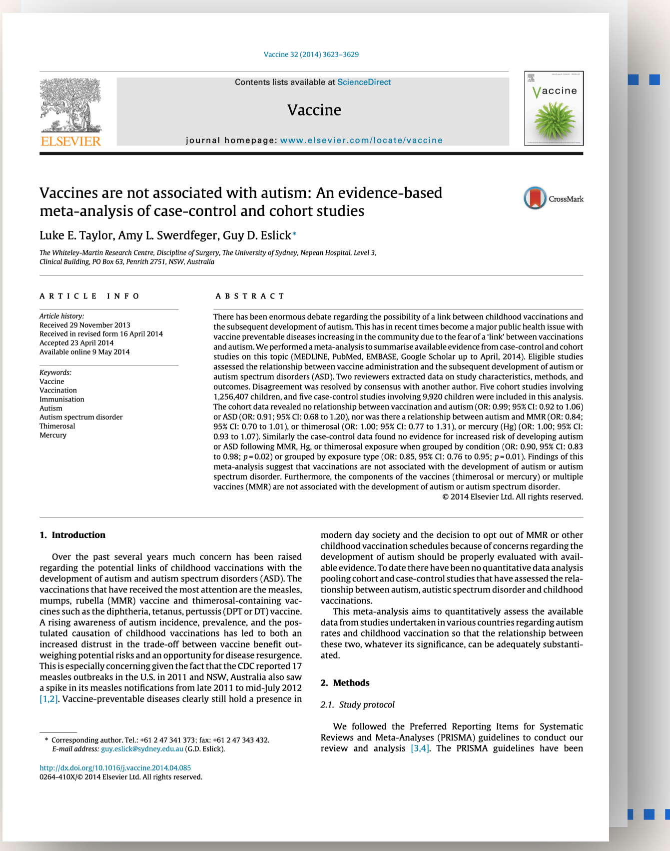
Finding evidence **from a trustworthy source** (source guarantee) to refute the reasoning behind the claim

FakeUser_3650 2d

⚠️ **You should know!**
New Study Claims Vaccines Linked to Autism!

1000 👍 568 ❤️ 5000

→


Vaccine

Vaccines are not associated with autism: An evidence-based meta-analysis of case-control and cohort studies

Luke E. Taylor, Amy L. Swerdfeger, Guy D. Eslick

Professional Evidence

“The cohort data revealed **no relationship** between vaccination and autism”

(Taylor et al., 2014)

→

× **False**



▸ Glockner et al, 2022

# Human Fact-Check Approaches

1 Global Counter-Evidence (GCE) | 2 Local Counter-Evidence (LCE) | 3 **Non-Credible Sources (NCS)** | 4 No Evidence Assertion (NEA)

Finding **evidence from a trustworthy source** (source guarantee) to refute the claim **based on the non-credibility of the sources** used to support the claim

FakeNewsToday.com 08/10/2023

🚨 **BREAKING NEWS** 🚨
Computer science is dead!
Grads Can't Find a Job!

👀 10,000,000

→

FactCheck.com

"**FakeNewsToday.com** is **not a Trustworthy** News site, they published fake news multiple times in..."

→

× **False**

# Human Fact-Check Approaches

1 Global Counter-Evidence (GCE) | 2 Local Counter-Evidence (LCE) | 3 Non-Credible Sources (NCS) | 4 **No Evidence Assertion (NEA)**

Refuting the claim by **asserting that no trusted evidence** supports it.

dcai.io 04/13/2024

**Dongting is the most handsome man in the world**

👀 200,823

→

**Can't find** any other sources to support that opinion

→

× **False**



Glockner et al, 2022

# Human Fact-Check Challenges

- **Time-consuming** process
- Dealing with **complex or ambiguous** claims
- Keeping up with the **rapid spread of information**
- Potential for **human biases and errors**
- Difficulty in finding **suitable counter-evidence for some claims**



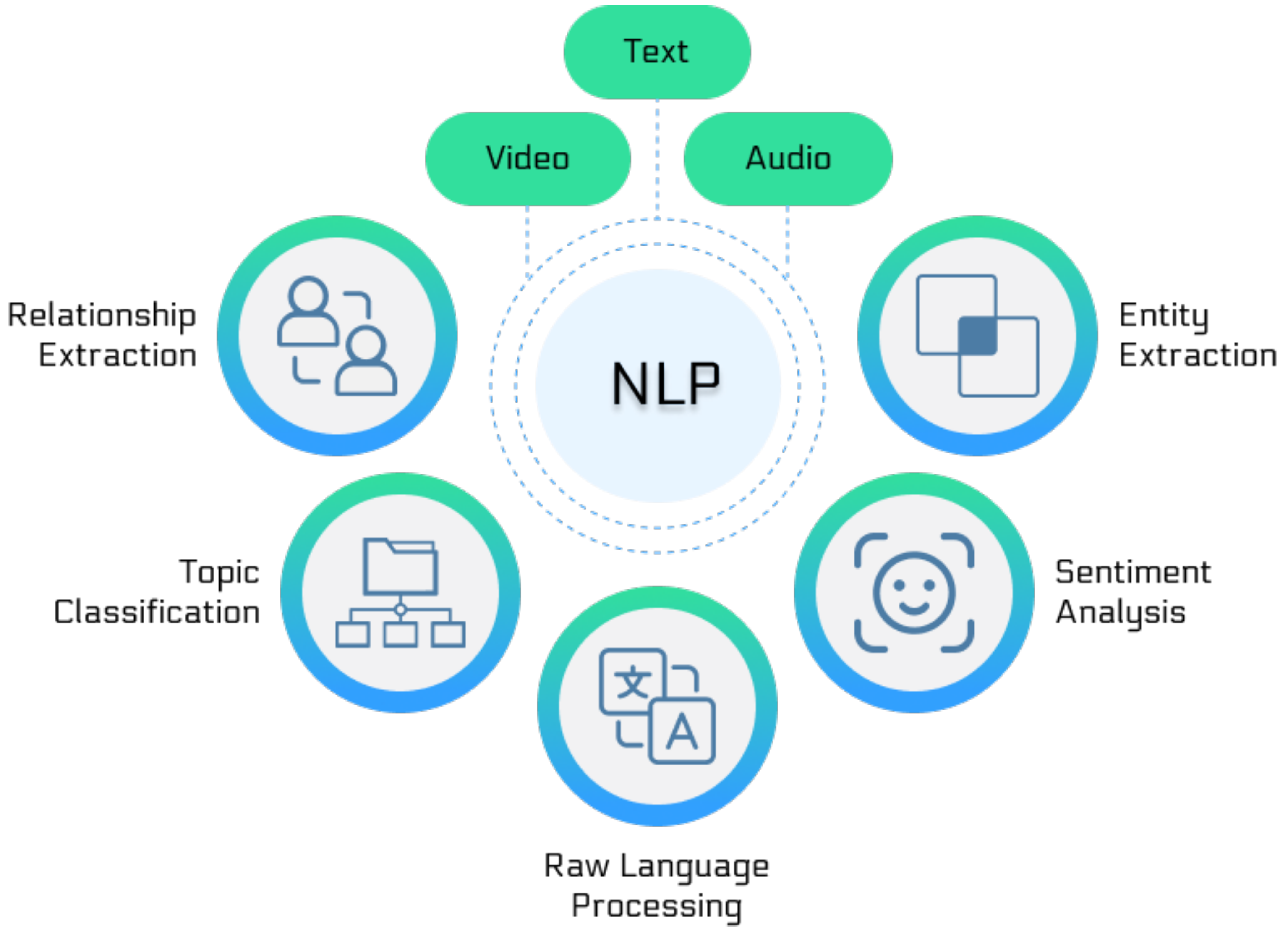
Glockner et al, 2022
Graves, 2018

PART 3

# NLP Fact-Checking: Techniques

# Natural Language Processing (NLP)

Focuses on **teaching computers to understand, interpret, and generate human language.**



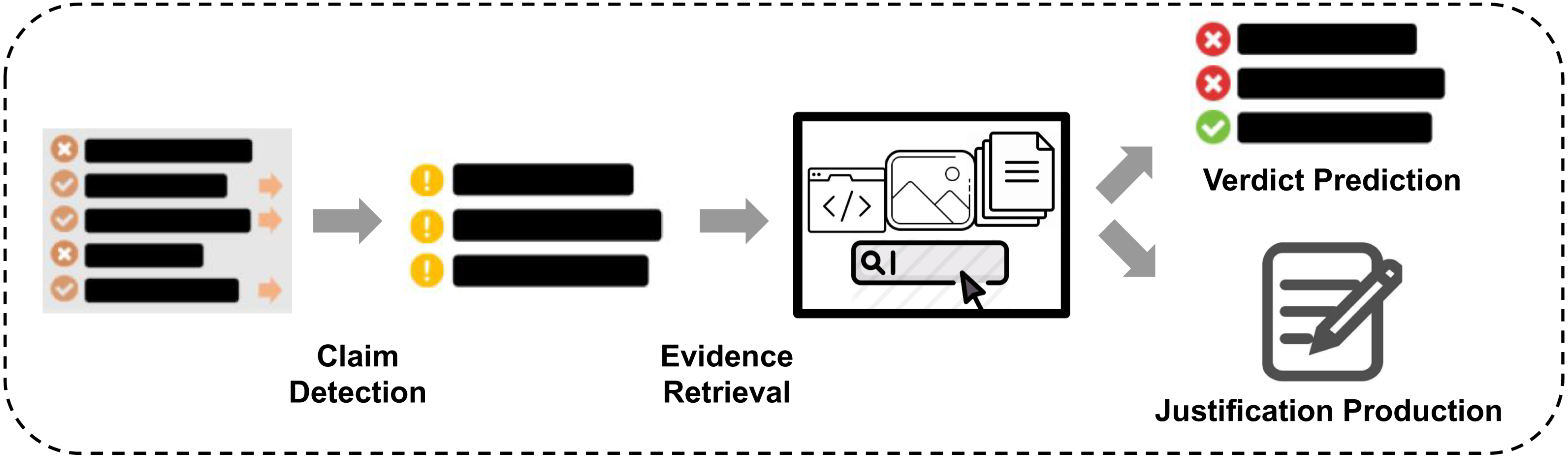

**The potential of NLP for fact-checking:**
- Automatically identifying claims
- Retrieving relevant evidence
- Verifying the truthfulness of claims



Guo et al, 2022
Jurafsky & Martin, 2020
Amazinum, 2024

# NLP Fact-Checking Pipeline

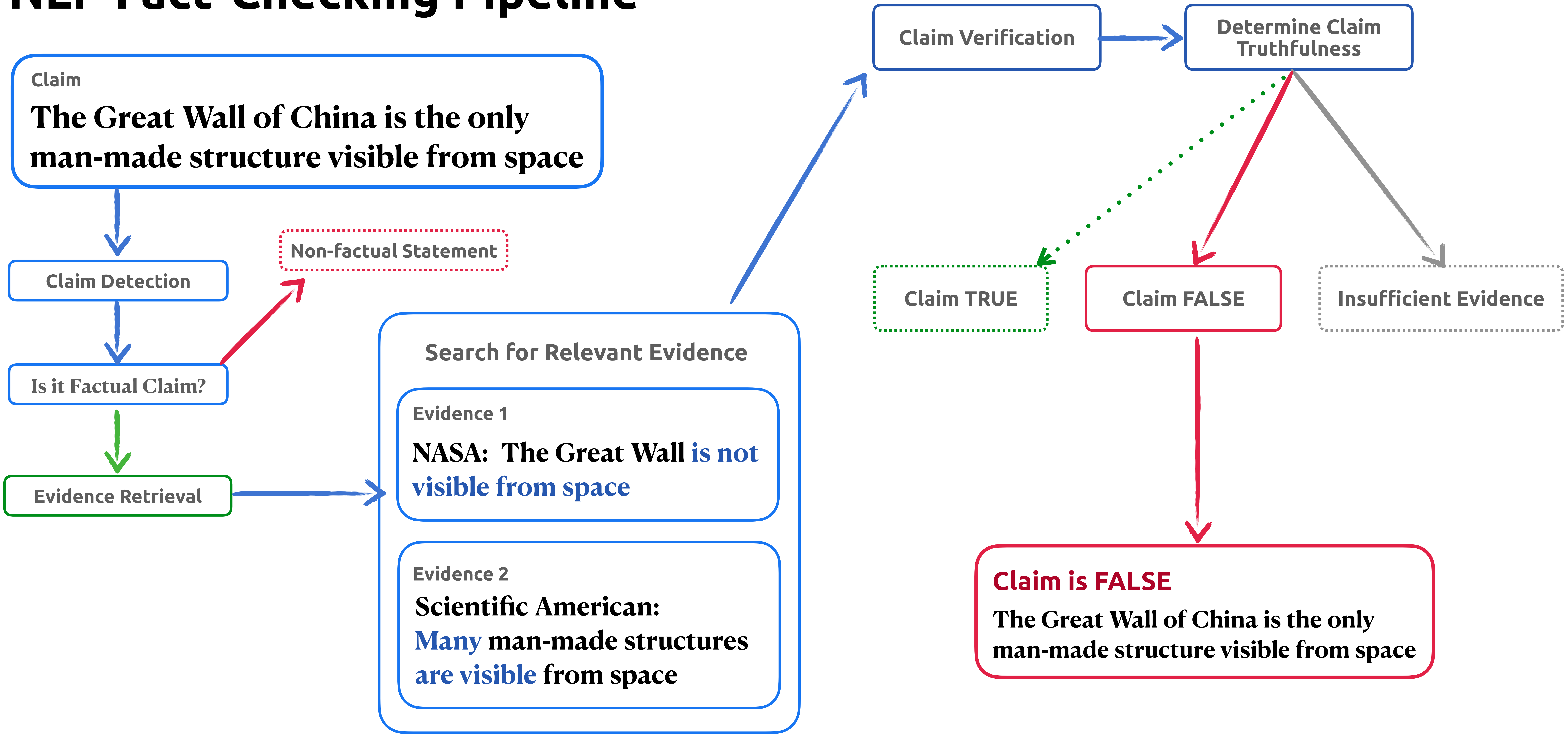
- **Claim Detection**: Identifying factual claims in text
- **Evidence Retrieval**: Gathering relevant evidence from reliable sources
- **Claim Verification**: Determining the truthfulness of the claim based on the evidence






Guo et al, 2022

# NLP Fact-Checking Pipeline






Guo et al, 2022

# Categories of NLP Models for Fact-Checking

## Single-task Models

Separate models for each stage of the fact-checking pipeline

## Multi-task models

Single models trained to perform multiple fact-checking tasks simultaneously

## Knowledge-based Models

Rely on external knowledge bases or fact-checking websites to verify the truthfulness of claims

## Hybrid Models

combine multiple approaches, such as single-task and multi-task models, to enhance the fact-checking process

**And MORE...**



Guo et al, 2022

# Single-Task Models

Separate models are trained for each stage of the fact-checking pipeline

### Example Models

## ClaimBuster

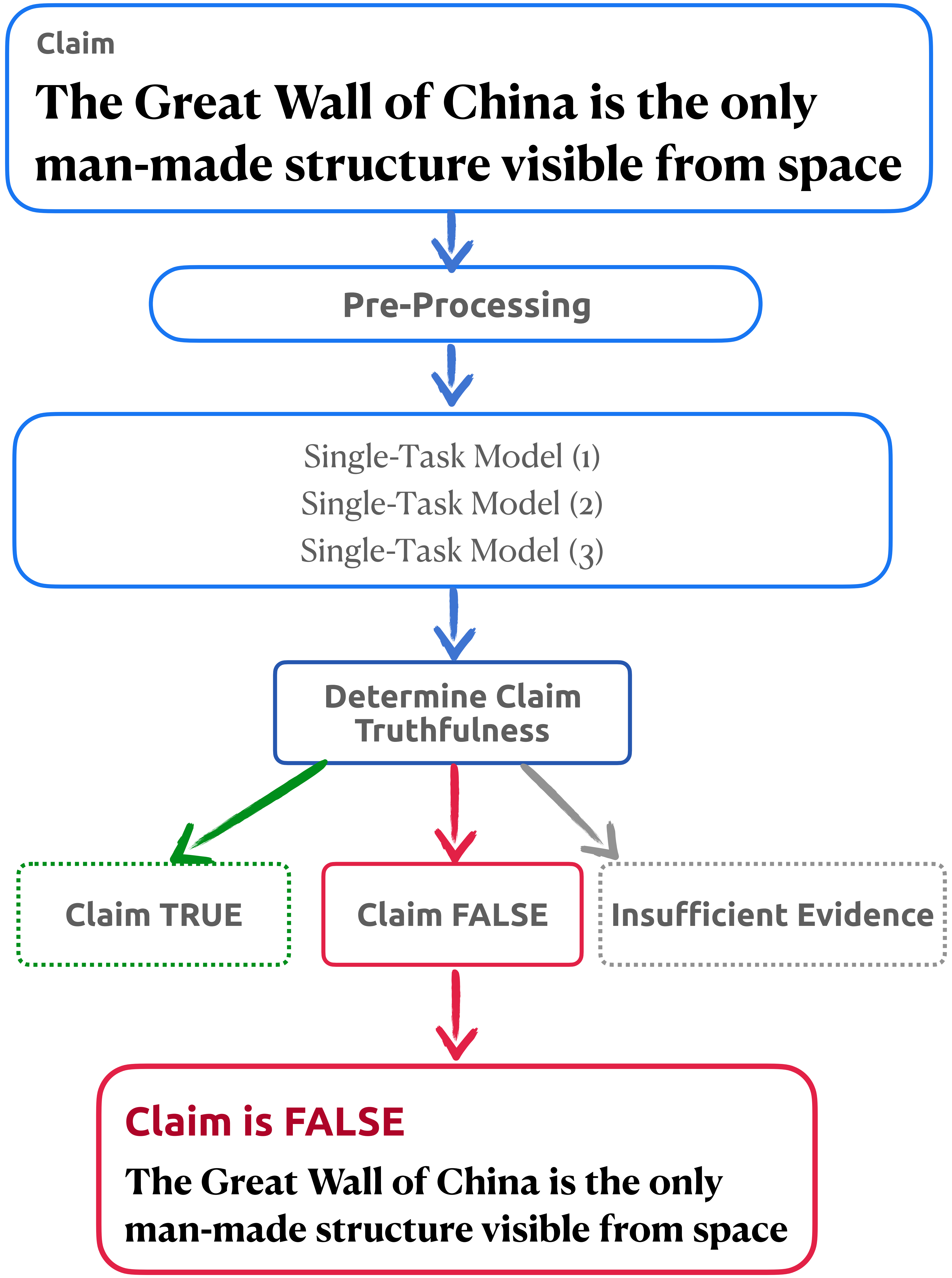
Claim Detection

## TF-IDF

Evidence Retrieval

## Textual entailment

Claim Verification






Hassan et al, 2017

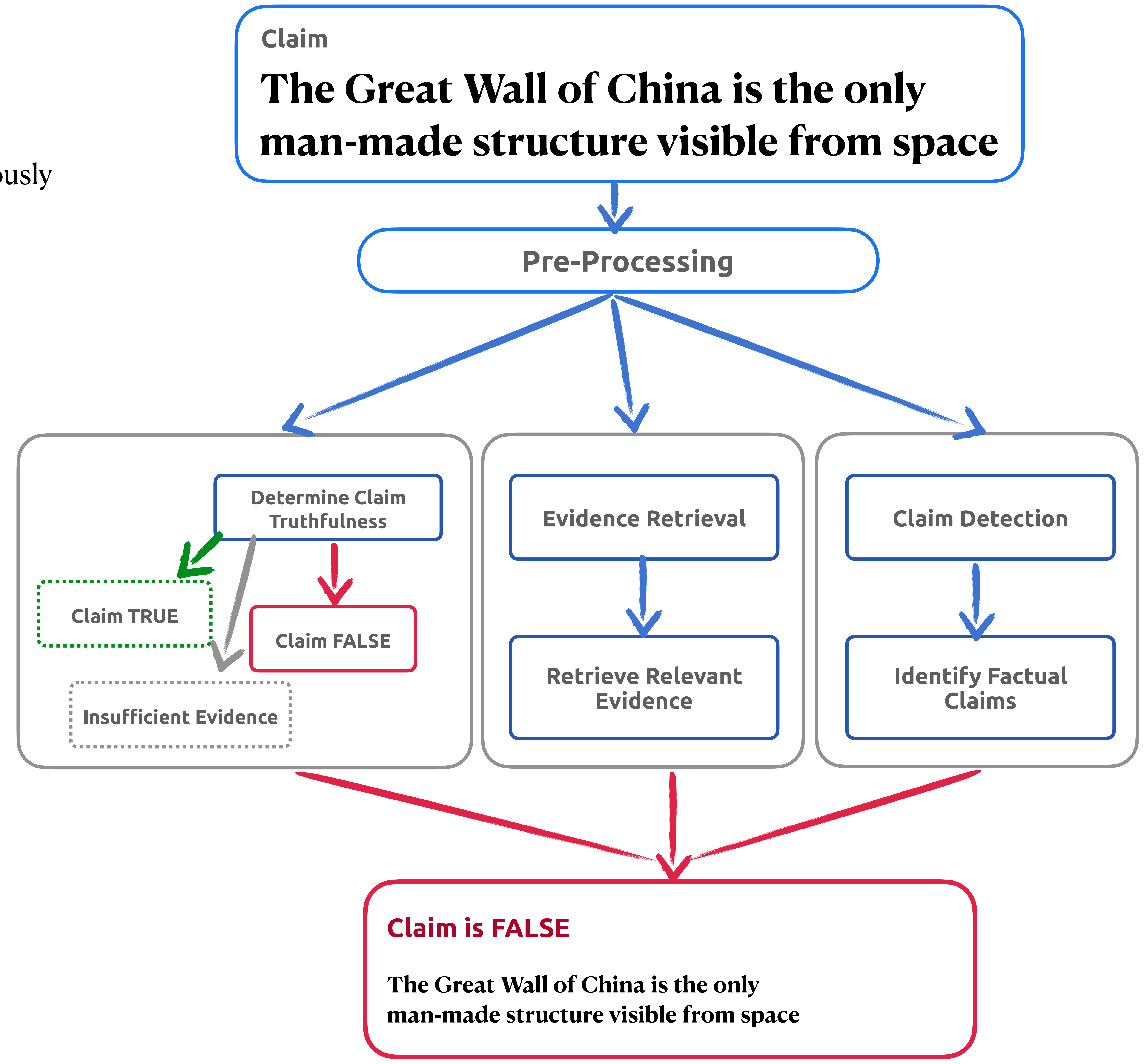
# Multi-Task Models

Single models are trained to perform multiple fact-checking tasks simultaneously

## Example Models

### UNC-NLP
Document Retrieval, Sentence Selection, Textual Entailment

### DREAM
Evidence retrieval & Claim Verification






Nie et al, 2019
Zhong et al, 2020

# Evaluation Metrics and Benchmarks: FEVER

A large-scale dataset **consisting of claims and their corresponding evidence sentences** from Wikipedia

**Evidence Retrieval** | **Claim Verification**

**FEVER SCORE**

**The percentage of claims** for which the system correctly retrieves all the required evidence sentences and assigns the correct label. The FEVER score is the primary metric used to rank the participating systems in the FEVER shared task.



Thorne et al, 2018

# FEVER Score

| Model | FEVER Score |
|---|---|
| **UNC-NLP**<br>Combine-FEVER-NSMN | 67.98% |
| **DREAM**<br>**D**ual **R**etrieval **E**vidence<br>Enh**a**nced **M**ulti-task Learning | 70.60% |

**Strengths**
Multi-task Learning
Claim Detection
Evidence Retrieval

**Limitations**
Limited Context Understanding
Handling Complex Claims
Bias and Fairness
Explainability



Nie et al, 2019
Zhong et al, 2020

PART 4

# Future Directions

# Limitations of Current NLP Fact-Checking Models

**Limited ability to handle complex claims**

- Current models struggle with claims that require reasoning, common sense, or world knowledge
- Example: "The Earth is flat because if it were round, people on the bottom would fall off"

**Dependence on high-quality, labeled data**

- NLP fact-checking models require large amounts of labeled data for training and evaluation
- Creating such datasets is time-consuming, expensive, and prone to human biases and errors

**Limited adaptability to new domains and types of misinformation**

- Models trained on one domain or type of misinformation may not generalize well to others
- Example: A model trained on political fact-checking may not perform well on scientific or medical misinformation



Thorne et al, 2021

# Improving NLP Models for Fact-Checking

~~**Time-consuming** process~~

Dealing with **complex or ambiguous** claims

~~Keeping up with the **rapid spread of information**~~

Potential for **human biases and errors**

Difficulty in finding **suitable counter-evidence for some claims**



Glockner et al, 2022
Graves, 2018

# Improving NLP Models for Fact-Checking

- **Real-time fact-checking and early detection**
  - Developing NLP models that can identify and flag potential misinformation in real-time, before it spreads widely
  - Integrating fact-checking systems with social media platforms and news aggregators to provide early warnings and corrections
- **Collaborative and decentralized fact-checking**
  - Encouraging collaboration between human fact-checkers and NLP models to improve accuracy and coverage
  - Exploring decentralized fact-checking approaches, such as blockchain-based systems, to increase transparency and trust
- **Proactive fact-checking and misinformation prevention**
  - Using NLP techniques to identify and address the root causes of misinformation, such as biased or misleading content
  - Developing educational tools and resources to improve media literacy and critical thinking skills among the public



Shao et al, 2016
Nakov et al, 2021

# Conclusion

- NLP techniques have **shown promise** in automating fact-checking and combating misinformation
- Current NLP fact-checking models **have limitations and face challenges in real-world applications**
- Future directions include improving model performance, scalability, and explainability, as well as addressing ethical and societal considerations

# What Could We Do?

- **Check IT** - Be a Fact-Checker
- **Think IT** - Think before share
- **Tag IT** - Report it to the platform or website where it appears
- Maybe... **Involve the NLP Fact-Check** Development Process

# Acknowledgement


**Advisor**
Dr. Wenkai Guan

**Faculty**
Dr. Elena Machkasova
Dr. Kristin Lamberty
Dr. Nic McPhee
Dr. Peter Dolan
Dr. Kristofer Schlieper

**Special Thanks To...**

| | |
|---|---|
| **My Mom** | Ms. Weiyun Zhang |
| **My Grandfather** | Mr. Tong Zhang |
| **My Grandmother** | Ms. Cuizhen Wang |
| **My Best Friend** | Alex Chen |

And, **All of You**

# Q & A Session

## Thanks for your Listening!

# References

**[1]** Lucas Graves. 2018. Understanding the promise and limits of automated fact-checking. (2018). https://doi.org/10.60625/RISJ-NQNX-BG89

**[2]** Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. Toward Automated Fact-Checking: Detecting Check-worthy Factual Claims by ClaimBuster. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, August 13, 2017. ACM, Halifax NS Canada, 1803–1812. https://doi.org/10.1145/3097983.3098131

**[3]** Dan Jurafsky and James H. Martin. 2024. Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition (3rd ed ed.). Pearson Prentice Hall, Upper Saddle River, N.J.

**[4]** David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. 2018. The science of fake news. *Science* 359, 6380 (March 2018), 1094–1096. https://doi.org/10.1126/science.aao2998

**[5]** Sophia Matsuk. 2022. What is NLP and how It is Implemented in Our Lives. *Amazinum*. Retrieved April 15, 2024 from https://amazinum.com/insights/what-is-nlp-and-how-it-is-implemented-in-our-lives/

**[6]** Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated Fact-Checking for Assisting Human Fact-Checkers. August 09, 2021. 4551–4558. https://doi.org/10.24963/ijcai.2021/619

**[7]** Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining Fact Extraction and Verification with Neural Semantic Matching Networks. *AAAI* 33, 01 (July 2019), 6859–6866. https://doi.org/10.1609/aaai.v33i01.33016859

**[8]** Chengcheng Shao, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. 2016. Hoaxy: A Platform for Tracking Online Misinformation. In *Proceedings of the 25th International Conference Companion on World Wide Web* (*WWW '16 Companion*), April 11, 2016. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 745–750. https://doi.org/10.1145/2872518.2890098

**[9]** James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2018. Association for Computational Linguistics, New Orleans, Louisiana, 809–819. https://doi.org/10.18653/v1/N18-1074

**[10]** Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (March 2018), 1146–1151. https://doi.org/10.1126/science.aap9559

**[11]** Wanjun Zhong, Jingjing Xu, Duyu Tang, Zenan Xu, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. 2020. Reasoning Over Semantic-Level Graph for Fact Checking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020. Association for Computational Linguistics, Online, 6170–6180. https://doi.org/10.18653/v1/2020.acl-main.549