

# *SmartBook*: an AI Prophetess for Disaster Reporting and Forecasting


Revanth Gangi Reddy, Sha Li, Chi Han, Steeve Huang, Yi R. Fung, Ziqi Wang, Qi Zeng, Manling Li, Heng Ji (UIUC)

Based on Collaborations with Eduard Hovy (University of Melbourne, CMU) and Clare R. Voss (ARL)

[hengji@illinois.edu](mailto:hengji@illinois.edu)

# It is the Best of Times

- LLMs produce text sequences word-by-word by computing conditional probability based on context
- At a sufficiently large scale, they can answer questions, generate arguments, write poetry, impersonate characters, negotiate contracts.
- They achieve competitive results across a wide variety of standard NLP tasks including entity typing, sentiment analysis, and textual entailment, showcasing “emergent behavior” such as in-context learning (Wei et al., 2022)



Whether it's the best  
of times or the worst  
of times, it's the only  
time we've got.

Art Buchwald

BrainyQuote®



# It is the Worst of Times

- “One-fits-all” models cause self-identity crisis
  - Applied to hundreds or even thousands of tasks (Sanh et al., 2022)
  - Some of these tasks seem to be nearly solved (e.g., single-document news summarization, entity typing)
  - Many of the tasks are what our current PhD students are tackling
- To add insult to injury – NLP research becomes too luxury
  - Most academic institutions do not have enough GPUs to perform fine-tuning on LLMs, not to mention pre-train them from scratch (Ajao, 2023; Writer, 2023)
  - They must pay significant amounts to industry for access to computing clouds and LLM APIs (I’m paying \$40K/month to these services)
  - This access can be revoked at a whim due to usage caps, policy changes, or funding limits



# How Can We Stay Relevant?

- While LLMs seemingly close research avenues, they also open up new ones
- Current LLMs remain somewhat monolithic, expensive, amnesic, delusional, uncreative, static, assertive, stubborn, and biased black boxes
- LLMs lack of self-learning, self-updating, self-correction and knowledge generalization capabilities
- NLP!=NLG: There are many more exciting and challenging problems other than NLG out there
  - Structured Prediction
  - Cross-document reasoning
  - Online Learning
  - Theoretical Understanding of LLMs
  - Factual Error Correction
  - AI for scientific discovery
  - ....
- Case Study on Situation Reporting and Forecasting

# Some of us are not losing our jobs yet: State of LLMs for Information Extraction Today

LLMs have not caught up with SOTA in more complex IE tasks yet, but more and more people are applying LLMs for IE.

Task	Dataset	BERT	RoBERTa	SOTA	ChatGPT
<b>Entity Typing(ET)</b>	<b>BBN</b>	80.3	79.8	82.2 (Zuo et al., 2022)	85.6
	<b>OntoNotes 5.0</b>	69.1	68.8	72.1 (Zuo et al., 2022)	73.4
<b>Named Entity Recognition(NER)</b>	<b>CoNLL2003</b>	92.8	92.4	94.6 (Wang et al., 2021)	67.2
	<b>OntoNotes 5.0</b>	89.2	90.9	91.9 (Ye et al., 2022)	51.1
<b>Relation Classification(RC)</b>	<b>TACRED</b>	72.7	74.6	75.6 (Li et al., 2022a)	20.3
	<b>SemEval2010</b>	89.1	89.8	91.3 (Zhao et al., 2021)	42.5
<b>Relation Extraction(RE)</b>	<b>ACE05-R</b>	87.5   63.7	88.2   65.1	91.1   73.0 (Ye et al., 2022)	40.5   4.5
	<b>SciERC</b>	65.4   43.0	63.6   42.0	69.9   53.2 (Ye et al., 2022)	25.9   5.5
<b>Event Detection(ED)</b>	<b>ACE05-E</b>	71.8	72.9	75.8 (Liu et al., 2022a)	17.1
	<b>ACE05-E+</b>	72.4	72.1	72.8 (Lin et al., 2020)	15.5
<b>Event Argument Extraction(EAE)</b>	<b>ACE05-E</b>	65.3	68.0	73.5 (Hsu et al., 2022)	28.9
	<b>ACE05-E+</b>	64.0	66.5	73.0 (Hsu et al., 2022)	30.9

near SOTA

large gap

Table from Li et al. “Evaluating ChatGPT’s Information Extraction Capabilities: An Assessment of Performance, Explainability, Calibration, and Faithfulness”. Arxiv 2023.

# What are Situation Reports?

- The key to situation understanding for unfolding events is to extract the common truths across heterogeneous data for creating situation reports for guiding action planning and strategic development.
- Manual construction of situation reports by expert analysts
  - time-consuming, resource-intensive and unscalable
  - often limited to a few topics, scenarios or regions
  - can be biased, incomplete, difficult to keep up-to-date
  - incapable of consolidating information across foreign languages and multiple data modalities
- Situation reports are expected to have following characteristics:
  - **Salient information** about key events, trends, statistics relevant to subject of interest.
  - Clear **logical structure** to help understand, follow, and easily access information
  - **Organized as Timelines** to seamlessly update for new events and cover event progressions; can be covered over longer time periods.
  - **Grounded Factual Content** to build credibility and trust by allowing for cross-checking of information presented

***Our Goal:** Automatic generation of intelligence reports to facilitate timely, comprehensive, and organized presentation of information in support of prompt policy actions and tactical operations.*



# What Would the “SmartBook” Look Like? An Example Chapter

## Ukraine Crisis Smart Book

### Chapters >

#### Position of Third Parties >

Overview

What is NATO's position on the Ukraine–Russia conflict?

**What is the position of China on the Ukraine-Russia conflict?**

What is the position of the governments of Poland, Turkey and Moldova on the Ukraine–Russia conflict?

What is Venezuela's position on Ukraine–Russia Conflict?

Perspectives

## What is the position of China on the Ukraine–Russia conflict?

### Summary

China has publicly stated that it does not support the idea of ejecting Russia from the G20, and has [instead called for dialogue between all parties involved in the conflict](#). [China has also said that it is opposed to the spread of disinformation about China in an attempt to pressure the country](#).

### Claims

Claim Sentence	Context
Since the war began, China has tried to project a neutral stance. <a href="#">Source</a>	Biden–Xi call: A call between Chinese President Xi Jinping and US President Joe Biden was underway Friday. <a href="#">Since the war began, China has tried to project a neutral stance</a> . It has not condemned Russian actions, and has refused to label the attack an invasion. <a href="#">See More</a>
Cooperation with China was ambivalent, he said, adding the country was "struggling with the consequences of its own Covid strategy, which also has consequences for world trade."	While the G7 leaders reemphasized their condemnation of "Russia's unjustifiable, unprovoked and illegal war against Ukraine," they called on China to press Russia to withdraw its troops from Ukraine. Scholz said he also expected China not to undermine sanctions against Russia. Cooperation with China was ambivalent, he said, adding the country was "struggling with the consequences of its own Covid strategy, which also has consequences for world trade." <a href="#">See More</a>
China is firmly opposed to this and will never accept it. <a href="#">Source</a>	The US has indicated that China would pay an economic price if its support for Russia goes beyond rhetoric. Speaking Friday, Zhao repeated China's public rebuke, saying "some people in the US have been spreading disinformation to smear and <a href="#">put pressure on China</a> , which is extremely irresponsible and will not help solve the issue. China is firmly opposed to this and will never accept it." <a href="#">See More</a>
That is considered unlikely, as China has said it would not back kicking Russia out.	While US President Joe Biden has said Russia should no longer be in the G20, ejecting Moscow would require the support of all members. That is considered unlikely, as China has said it would not back kicking Russia out. <a href="#">See More</a>
The White House is realistic the G20 will not collectively remove Russia from its ranks, since the decision would likely require consensus and China has been clear it doesn't support such a move.	The White House is realistic the G20 will not collectively remove Russia from its ranks, since the decision would likely require consensus <a href="#">and China has been clear it doesn't support such a move</a> . <a href="#">See More</a>

### Multi-Lingual Claims

#### China's position:

[Source](#)

#### Original Post

中国常驻联合国副代表耿爽：各方都应摒弃政治私利，为化解乌克兰危机做切实努力 中国日报联合国9月7日电（记者张旻璐）当地时间7日，联合国安理会召开会议审议乌克兰人道问题。中国常驻联合国副代表耿爽出席会议并发言表示，乌克兰危机中各方都应摒弃政治私利，停止激化矛盾、停止制造分裂、停止鼓吹对抗，为化解危机做出切实努力。耿爽说，历史反复证明，外交努力、谈判沟通是化解冲突、结束危机的唯一正确出路。中方再次呼吁有关当事方保持对话接触，探索政治解决的可能性，为早日停火止战、恢复和平稳定积累条件。耿爽指出，扎波罗热核电站的安全和安保问题事关乌克兰和整个地区上亿人的健康福祉，决不容许发生意外。中方呼吁有关当事方认真落实格罗西总干事提出的七项支柱原则，避免采取危及核安全和安保的行动。中方欢迎国际原子能机构访问团上周对扎波罗热核电站进行实地考察，赞赏机构专家驻留在核电站继续开展工作，鼓励机构就相关问题同各方保持沟通。

#### Translation (Google)

Geng Shuang, Deputy Permanent Representative of China to the United Nations: All parties should abandon political self-interest and make practical efforts to resolve the Ukrainian crisis China Daily, United Nations, September 7 (Reporter Zhang Minlu) On the 7th local time, the UN Security Council held a meeting to discuss the humanitarian issue in Ukraine. China's Deputy Permanent Representative to the United Nations, Geng Shuang, attended the meeting and said that all parties in the Ukraine crisis should

### What is the position of China on the Ukraine–Russia conflict?

Summary

Claims

Multi-Lingual Claims

China's position:

[Original Post](#)

[Translation \(Google\)](#)



# What Would the "SmartBook" Look Like? An Example Chapter

## Ukraine Crisis Smart Book

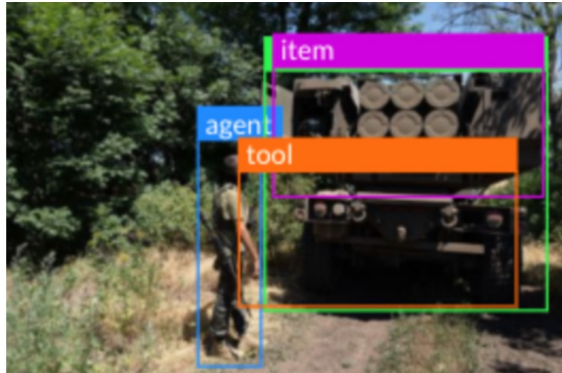
- Chapters >
- Position of Third Parties >
  - Overview
  - What is NATO's position on the Ukraine-Russia conflict?
  - What is the position of China on the Ukraine-Russia conflict?
  - What is the position of the governments of Poland, Turkey and Moldova on the Ukraine-Russia conflict?
  - What is Venezuela's position on Ukraine-Russia Conflict?
  - Perspectives

### What is NATO's position on the Ukraine-Russia conflict?

#### Summary

NATO has not imposed a no-fly zone in the Ukraine-Russia conflict because neither Ukraine nor Russia are members of NATO. NATO's focus has been on the "tactical approach" to the war in Ukraine, responding to the developments on the ground. However, the foreign minister of Lithuania, a NATO member state, has said that NATO needs to think strategically about the longer-term and be prepared for the possibility that Russia "might war again, and not excluding NATO countries".

#### Claims

Claim Sentence	Context
The US and NATO oppose creating a no-fly zone in Ukraine; such a move, they have warned, could lead to "full-fledged war in Europe." <a href="#">Source</a>	<p>His message was delivered amid ongoing conversations between the US and European allies about the possibility of providing Ukraine with fighter jet aircrafts from Eastern European countries, five sources familiar with the discussions told CNN. Officials say there is a divide among countries about whether or not individual nations should provide Ukraine aircraft, given the associated risks. The US and NATO oppose creating a no-fly zone in Ukraine; such a move, they have warned, could lead to "full-fledged war in Europe." <a href="#">See More</a></p> 
	<p>Ukraine is not a member of NATO, but has long hoped to join the alliance. This is a sore point for Russia, which sees NATO as a threat and vehemently opposes the move. <a href="#">See More</a></p> <p><b>agent</b></p>

What is NATO's position on the Ukraine-Russia conflict?

- Summary
- Claims

# Wait ... Can GPT do this already?

HE

Is Russian positioning forces inside of Russian for possible follow-on actions into Ukraine after initial invasion?



I don't have the current information regarding any Russian military actions or deployments. It's important to rely on credible news sources for the most up-to-date and accurate information on political and military developments.

HE

What is the reporting size of Russian and Ukrainian forces currently involved in Ukraine?



I don't have the current information on the exact size of Russian and Ukrainian forces involved in any potential conflict in Ukraine. Information on military deployments and strengths can change rapidly and may not be immediately available or publicly disclosed. It's important to rely on credible news sources for the most up-to-date and accurate information.

- Fluent, but short, generic and boring
- No concrete or up-to-date information
- No evidence
- No guarantee of truthful information
- English text only

# Asking Our *SmartBook* the Same Questions

- Is Russian positioning forces inside of Russian for possible follow-on actions into Ukraine after initial invasion?
  - The arrival of Russian forces in Slavutych came after several days of shelling against the city, which is strategically located close to the Dnieper River, about 95 miles (150 kilometers) north of Kyiv and close to the border with Belarus.
  - Russia has been using Belarus as a springboard for many of its air operations in Ukraine, according to intelligence collected by NATO surveillance planes flying over the Polish-Ukrainian border and radar seen by CNN.
  - So far, however, US officials have not seen Belarusian troops "being readied to move into Ukraine" or "that they are moving or are in Ukraine," a senior US defense official said Monday, adding that the forces inside Ukraine are Russian.
- What is the size of Russian and Ukrainian forces currently involved in Ukraine?
  - It said the Ukrainians had created a force of 20,000 close to the Belarus border, which "requires a response from us."
  - It also said 400 Russian soldiers are "present full time" at the site.

## Compared to Bing Search + GPT3.5

HE

What is the status of efforts to secure the release of Paul Whelan and Brittney Griner?



I'm sorry, but as a language model AI, I do not have the most up-to-date information on the current status of individual cases like Paul Whelan and Brittney Griner. However, as of my knowledge cutoff in 2021, Paul Whelan, a former US Marine, was being held in Russia on espionage charges and has been advocating for his release through his legal team and government representatives. Similarly, Brittney Griner is a basketball player who was playing in Russia and has been involved in legal issues there, but I don't have any more specific details on her case. It's best to check reliable news sources for the most recent updates on their cases.



# Asking Our SmartBook the Same Question

← → ↻ Not Secure | https://blenderdemo.com/smartbook

## Ukraine Crisis Smart Book

Sep 16th - 30th >

Paul Whelan and Brittney Griner >

[What is the status of efforts to secure the release of Paul Whelan and Brittney Griner?](#)

What are the respective situations of Paul Whelan and Brittney Griner?

How can the United States best pressure Russia to engage in good faith negotiations for the release of Paul Whelan and Brittney Griner?

What is the US government's position on the case of Brittney Griner?

What are the risks and potential benefits of a prisoner swap between the United States and Russia?

Zaporizhzhia nuclear power plant >

What is the significance of the Zaporizhzhia nuclear power plant?

What are the implications of the Russia-Ukraine conflict for nuclear safety and security in Europe?

Why is the Zaporizhzhia nuclear power plant a target for Russia?

What are the risks associated with the Zaporizhzhia nuclear power plant?

How could a Chernobyl-style disaster be prevented at the Zaporizhzhia nuclear power plant?

American prisoners of war released >

What motivated the US government to get involved in the prisoner exchange between Russia and Ukraine?

What challenges do the American

## What is the status of efforts to secure the release of Paul Whelan and Brittney Griner?

### Summary

The United States has made a "serious offer" to negotiate the return of WNBA star Brittney Griner and former US Marine Paul Whelan, but Russia has not yet responded. Griner is currently serving a nine-year sentence in Russia after pleading guilty to drug charges, and Whelan is serving a 16-year sentence on espionage-related charges. The US government regards both as wrongfully detained, and Secretary of State Antony Blinken has said that the administration has made a substantial proposal to Russia.

### Claims

Claim Sentence	Context
National Security Council coordinator John Kirby says Russia has not responded to the "substantial proposal" the U.S. offered back in July regarding the release of WNBA star Brittney Griner and former U.S. Marine Paul Whelan. <a href="#">Source</a>	National Security Council coordinator John Kirby says Russia has not responded to the "substantial proposal" the U.S. offered back in July regarding the release of WNBA star Brittney Griner and former U.S. Marine Paul Whelan. Kirby told the media during Fridays White House press briefing that the U.S. has made a "serious offer" to negotiate the return of Griner and Whelan, but added that those discussions are "still ongoing."
Where the negotiations stand: A senior administration official told CNN Thursday that there has been "movement but not breakthrough" on the efforts to free Whelan and WNBA star Brittney Griner. <a href="#">Source</a>	Paul Whelan has been imprisoned in Russia for more than three years after being convicted on espionage charges that he vehemently denies. He was sentenced in June 2020 to 16 years in prison in a trial US officials denounced as unfair. Where the negotiations stand: A senior administration official told CNN Thursday that there has been "movement but not breakthrough" on the efforts to free Whelan and WNBA star Brittney Griner. The official said that the United States has urged Russia to put forward "a serious counteroffer" to the proposal on the table to secure the release of the two, but "we've not gotten a serious response back."
Her legal team in Russia has appealed the nine-year sentence. <a href="#">Source</a>	The tournament start comes a week after Griners wife, Cherelle Griner, met with US President Joe Biden to plead her wifes case as the US negotiates the release of the WNBA star and former US Marine Paul Whelan. Griner was detained in February for carrying vape cartridges containing cannabis oil in her luggage and pleaded guilty to drug charges, saying she accidentally packed the drugs in a hurry. Her legal team in Russia has appealed the nine-year sentence. Griners teammates and other WNBA players have also shown solidarity with the eight-time All-Star by boycotting playing in Russia during the offseason.
The U.S. government regards both as wrongfully detained, placing their cases with the office of its top hostage negotiator. <a href="#">Source</a>	She was sentenced last month to nine years in prison after pleading guilty and has appealed the punishment. Whelan is serving a 16-year sentence on espionage-related charges that he and his family say are false. The U.S. government regards both as wrongfully detained, placing their cases with the office of its top hostage negotiator. Secretary of State Antony Blinken took the unusual step of announcing two months ago that the administration had made a substantial proposal to Russia.
He was sentenced in June 2020 to 16 years in prison in a trial US officials denounced as unfair. <a href="#">Source</a>	Her legal team in Russia has appealed the sentence. Whelan has been imprisoned in Russia for more than three years after being convicted on espionage charges that he vehemently denies. He was sentenced in June 2020 to 16 years in prison in a trial US officials denounced as unfair.



## Compared to Bing Search + GPT3.5

HE

Why is the Zaporizhzhia nuclear power plant a target for Russia?



The Zaporizhzhia Nuclear Power Plant is not a specific target for Russia. It is important to note that there is no evidence to suggest that Russia has any intention of targeting the plant. Nuclear power plants are generally considered critical infrastructure and are protected by international laws and regulations. Any deliberate attack on a nuclear power plant could have severe consequences for the environment and human health.

# Asking Our SmartBook the Same Question

← → ↻ Not Secure | https://blenderdemo.com/smartbook

## Ukraine Crisis Smart Book

pressure Russia to engage in good faith negotiations for the release of Paul Whelan and Brittney Griner?

What is the US government's position on the case of Brittney Griner?

What are the risks and potential benefits of a prisoner swap between the United States and Russia?

Zaporizhzhia nuclear power plant >

What is the significance of the Zaporizhzhia nuclear power plant?

What are the implications of the Russia-Ukraine conflict for nuclear safety and security in Europe?

[Why is the Zaporizhzhia nuclear power plant a target for Russia?](#)

What are the risks associated with the Zaporizhzhia nuclear power plant?

How could a Chernobyl-style disaster be prevented at the Zaporizhzhia nuclear power plant?

American prisoners of war released >

What motivated the US government to get involved in the prisoner exchange between Russia and Ukraine?

What challenges do the American prisoners of war face as they return home?

What role did Saudi Arabia play in the prisoner exchange between Russia and Ukraine?

How do the American prisoners of war feel about being released?

What are the conditions that the American prisoners of war were released under?

Russia detains anti-war protesters >

## Why is the Zaporizhzhia nuclear power plant a target for Russia?

### Summary

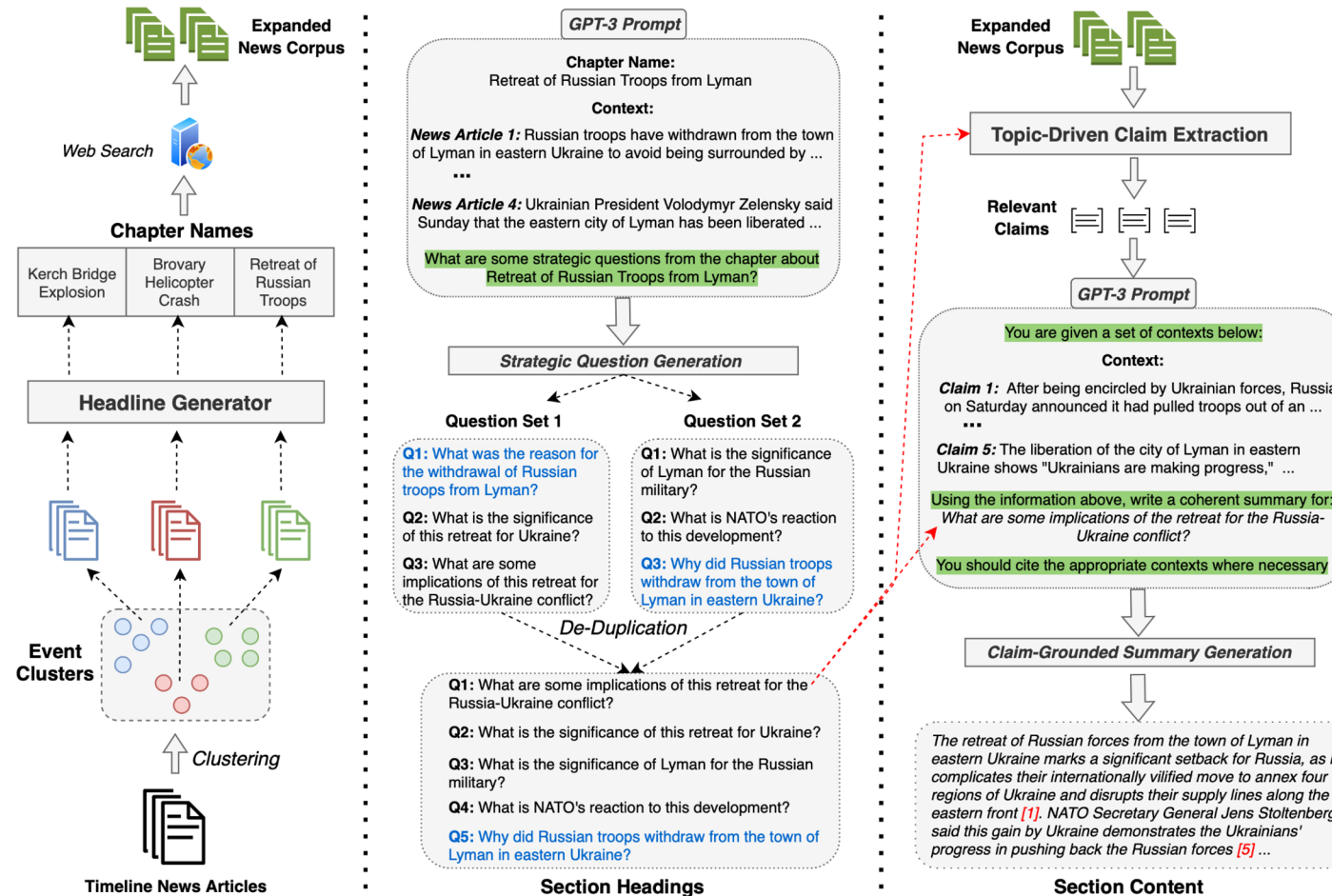
The Zaporizhzhia nuclear power plant is a target for Russia because it is a large source of electricity for Ukraine. If Russia were to capture the plant, they would be able to control a large portion of the country's electricity supply. Additionally, the plant is located in a war zone, making it a more difficult target to protect.

### Claims

Claim Sentence	Context
Each of Zaporizhzhia's reactors would cost \$7 billion to replace, making the plant a target for Russians to capture undamaged, with hopes of serving its own electricity market, according to analysis by defense and security intelligence firm Janes. <a href="#">Source</a>	The Zaporizhzhia plant is located on the eastern bank of the Dnipro river in Ukraine. The area, and the nuclear complex, has been under Russian control since the beginning of the war, but the plant is still mostly operated by Ukrainian workers. Each of Zaporizhzhia's reactors would cost \$7 billion to replace, making the plant a target for Russians to capture undamaged, with hopes of serving its own electricity market, according to analysis by defense and security intelligence firm Janes. Should Russia keep it, Ukraine would lose 20% of its domestic electricity generating capacity.
Reason: Russia hopes to use the plant to serve its own electricity market. <a href="#">Source</a>	The plant, located on the eastern bank of the Dnipro river, has been under Russian control since the beginning of the war but is still being operated by Ukrainian workers. Though the plant has been witnessing intense shelling and explosion, experts don't think anything disastrous would happen, at least for now. Reason: Russia hopes to use the plant to serve its own electricity market. Analysts from defence and security intelligence firm Janes believe Russia would likely keep the plant, thereby depriving Ukraine of 20 per cent of its domestic electricity-generating capacity.
Fears that Russias war in Ukraine could cause a nuclear catastrophe had focused so far on the countrys largest atomic energy station, in Zaporizhzhia, where all six reactors have now been shut down. <a href="#">Source</a>	"We have to stop it before its too late." Russia did not immediately comment on the allegation. Fears that Russias war in Ukraine could cause a nuclear catastrophe had focused so far on the countrys largest atomic energy station, in Zaporizhzhia, where all six reactors have now been shut down. Repeated shelling and fires had disconnected the Zaporizhzhia plant from the national electricity grid, requiring emergency measures to prevent essential cooling procedures from being interrupted.
It remains close to the frontline, and has come under fire repeatedly in recent months, raising fears of a nuclear catastrophe. <a href="#">Source</a>	MOSCOW, Sept 21 (Reuters) - Russia's defence ministry on Wednesday said that a large-caliber shell had damaged a technical water pipe at the Zaporizhzhia nuclear plant in Ukraine, Russian state-owned news agency RIA reported. The Zaporizhzhia nuclear plant was captured by Russian forces in March. It remains close to the frontline, and has come under fire repeatedly in recent months, raising fears of a nuclear catastrophe. Both Ukraine and Russia blame each other for shelling the facility.
Russia and Ukraine have blamed each other for shelling at the site of the Zaporizhzhia nuclear power plant (ZNPP) that has damaged buildings and caused the disconnection of power lines. <a href="#">Source</a>	"With the main line's reconnection yesterday afternoon, the three back-up power lines are again being held in reserve," the International Atomic Energy Agency watchdog said in a statement. "The three other main external 750 kV (kilovolt) power lines that were lost earlier during the conflict remain down," it added. Russia and Ukraine have blamed each other for shelling at the site of the Zaporizhzhia nuclear power plant (ZNPP) that has damaged buildings and caused the disconnection of power lines. "While the ZNPP's power status has improved over the past week in sharp contrast to earlier this month when all power lines at one stage were down and it depended on its last operating reactor for vital electricity supplies the general situation for the plant located in the middle of a war zone remains precarious," the IAEA said.

# SmartBook Overview

- We propose **SmartBook**, an automated framework that generates comprehensive, real-time situation reports from various sources, by discovering critical strategic questions and grounding the report with trustworthy factual evidence.



Overall workflow for constructing **SmartBook**. Given the articles corresponding to a specific timeline, the figure shows the process for obtaining the chapters, their section headings, and the corresponding section content.

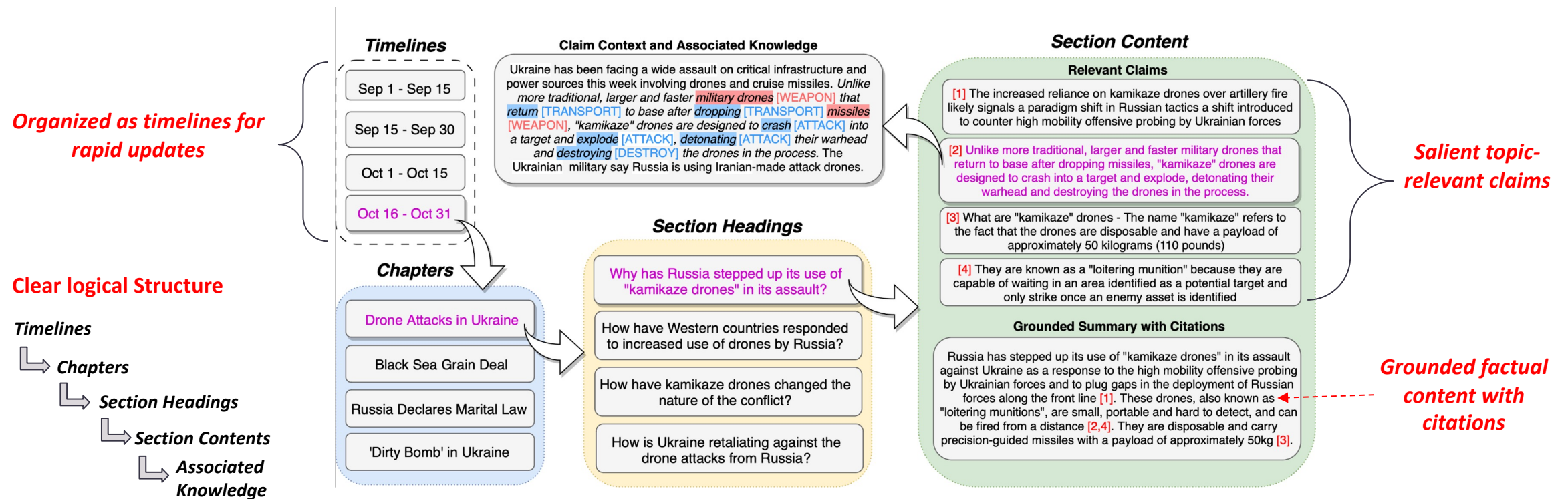
# Limitations of Human-Generated Questions

- Questions from human analyst are static and become outdated as the conflict has been evolving too rapidly
  - e.g., *What are the reports on Naval activity in the Black Sea?*
  - The above question doesn't have much relevant content from the past 3 weeks.
- Some human-generated questions are too vague or complex to parse
  - e.g., *Is Russian positioning forces inside of Russian for possible follow-on actions into Ukraine after initial invasion?*
- Human-generated questions often require multi-hop reasoning and aggregation
  - e.g., *What is the reporting size of Russian and Ukrainian forces currently involved in Ukraine?*
- Can we automatically generate questions from news articles?
  - *What was the reason for the withdrawal of Russian troops from Lyman?*
  - *What are the possible motives for sabotage of the Nord Stream gas pipelines?*
- We aim to generate and update the smartbook for regular intervals, based on the events in that time period (e.g., bi-weekly).



# SmartBook Overview

- Major events for each timeline are identified by clustering the news articles within the given time period.
- Derive a corresponding short chapter name for each cluster based on news story headline generation
- For generating chapter content, we automatically discover strategic questions that cover various strategic aspects of the event by prompting GPT-3 with the chapter name and a grounded context, and filtering duplicated questions:
  - *What was the reason for the withdrawal of Russian troops from Lyman?*
  - *What are the possible motives for sabotage of the Nord Stream gas pipelines?*

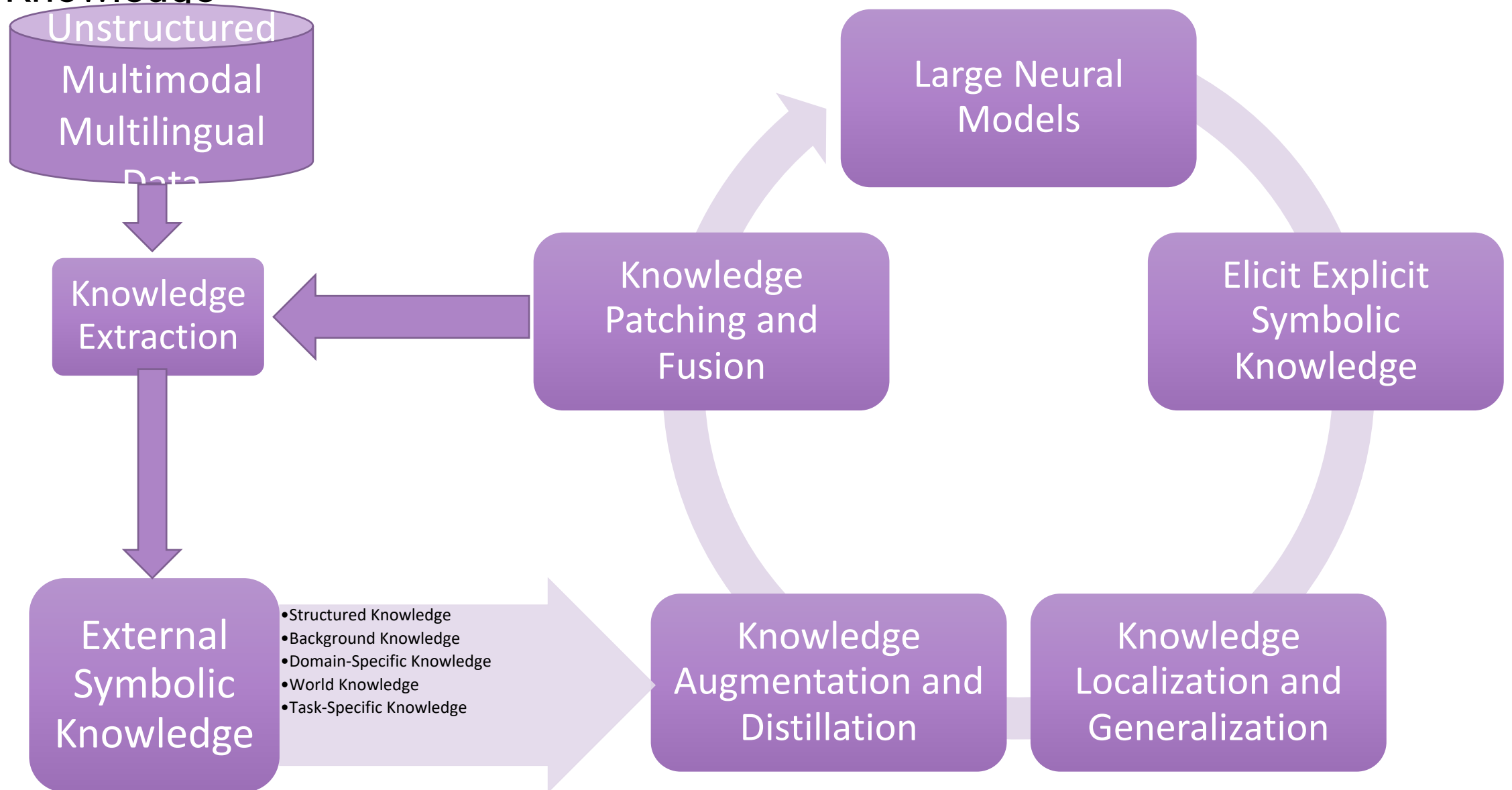




# Challenges and Our Solutions

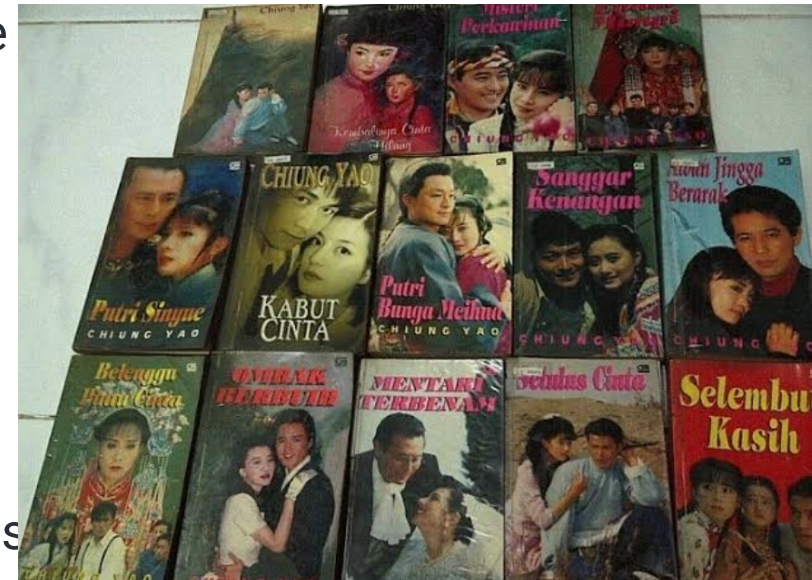
- Challenge 1: Open-domain Information Extraction techniques lack of global prior knowledge and cannot perform prediction
  - Event Schema Induction for better Information Extraction, Event Prediction and Situation Forecasting
- Challenge 2: Existing Vision+ Language neural models fail to understand complex situations
  - Explicitly probing large V+L neural models to force them to pay attention to event and action related knowledge
- Challenge 3: Conflicting or complementary claims may come from a wide range of multilingual multimedia sources
  - Cross-lingual Cross-media Claim Verification and Factual Error Correction

# Completing Knowledge Lifecycle: Neural Models Self-Learning Symbolic Knowledge



# Story Repeats Itself

- TV shows about “CEO boss guy”
  - He always has this really slicked back hair-style, always wears a suit
  - He has this really non.cha.lant look
  - He speaks at least 5 languages
  - He spent 5 years somewhere in France just for the sake of learning everything pretentious about the wine
  - And he spent his recent 5 years somewhere in NYC, of course the upper east side
  - At age 15 he got into Harvard, accepted by 10 majors
  - Ever since then he started to run this big multi-billion dollar company
  - He met this girl who graduated from an unknown college somewhere applied for internship in the company run by this guy
- Love stories by Qiong Yao
  - Pretty women are good at crying in a pretty way
  - At least two people get their hearts broken
  - Main characters are very poetical
  - The original wife is evil
  - The main character is always a jerk but very popular, jobless, and yells



# • What does ChatGPT know about these schemas?



Qiong Yao is a prolific Taiwanese author and screenwriter known for her romantic novels and melodramatic television dramas. While her stories vary in plot and characters, there are several recurring patterns and themes that can be found in many of her works:

1. **Romantic Love:** Qiong Yao's stories often revolve around romantic love and the complexities of romantic relationships. Her protagonists typically experience intense and passionate love affairs that are central to the plot.
2. **Forbidden Love:** Many of her stories involve forbidden or star-crossed love, where the main characters face significant obstacles to being together. These obstacles can include family opposition, societal norms, class differences, or prior commitments.
3. **Suffering and Sacrifice:** Qiong Yao's characters often endure great suffering and make significant sacrifices for the sake of love. These sacrifices can include giving up personal ambitions, enduring hardship, or even facing tragic consequences.
4. **Strong Female Protagonists:** Qiong Yao's female protagonists are often portrayed as strong, independent, and resilient individuals who are not afraid to pursue their own desires and dreams, even in the face of adversity.
5. **Dramatic Plot Twists:** Her stories are known for their dramatic and unexpected plot twists that keep the audience engaged. These twists can involve the revelation of hidden secrets, unexpected reunions, or sudden changes in fortune.
6. **Beautiful Scenery:** Qiong Yao's works often feature descriptions of beautiful natural scenery and romantic settings, which add to the overall romantic atmosphere of her stories.

7. **Tearful Moments:** Many of her stories evoke strong emotions and are known for their tearful and emotional scenes. These emotional moments often come from the challenges and hardships faced by the characters.
8. **Music and Poetry:** Music and poetry are integral elements in Qiong Yao's stories. She often incorporates song lyrics and poetic elements into her works to enhance the emotional depth and impact of the narrative.
9. **Timeless Love:** Qiong Yao's stories often explore the idea of timeless love, where the love between the main characters transcends time and obstacles, enduring through the years.
10. **Family and Relationships:** Family dynamics and relationships are important in her stories. Conflicts and resolutions within families often play a significant role in the plot.
11. **Cultural and Social Context:** Qiong Yao's works often reflect the cultural and social norms of the time and place in which they are set. This can include exploring traditional values, societal expectations, and gender roles.



# • What does ChatGPT know about these schemas?



Stories featuring CEO or boss characters often follow specific patterns and tropes, particularly in the context of romance novels or dramas. These patterns can create a dynamic and engaging narrative. Here are some common patterns associated with CEO or boss characters in stories:

1. **The Arrogant CEO:** The CEO or boss is initially portrayed as arrogant, cold, and distant. They often have a reputation for being ruthless and uncompromising in business.
2. **The Opposites Attract:** The love interest, often an employee or subordinate, is the complete opposite of the CEO. They may be more down-to-earth, compassionate, and relatable, creating tension and conflict.
3. **Power Dynamics:** The power imbalance between the CEO and the love interest is a central theme. The challenges and obstacles they face are often related to this dynamic, as workplace relationships can be complicated.
4. **Office Romance:** The workplace serves as the primary setting for the story, and the characters' interactions occur within the professional context. This can lead to secrecy and complications, as they try to keep their relationship hidden from colleagues and superiors.
5. **The Transformation:** Over the course of the story, the CEO undergoes a significant transformation. They may become more compassionate, understanding, and willing to compromise as they fall in love.
6. **Conflict and Misunderstandings:** Misunderstandings and conflicts are common plot devices, often arising from the CEO's initial arrogance or from external factors trying

5. **The Transformation:** Over the course of the story, the CEO undergoes a significant transformation. They may become more compassionate, understanding, and willing to compromise as they fall in love.
6. **Conflict and Misunderstandings:** Misunderstandings and conflicts are common plot devices, often arising from the CEO's initial arrogance or from external factors trying to keep the couple apart.
7. **The Redemption Arc:** The CEO often goes through a redemption arc, where their character evolves, and they learn important life lessons about love and relationships.
8. **Family and Social Expectations:** Family and societal expectations can add another layer of conflict, especially if the CEO comes from a wealthy or prestigious family.
9. **Wealth and Luxury:** The CEO's wealth and extravagant lifestyle are often featured prominently, creating a sense of opulence and glamour in the story.
10. **The Happy Ending:** Despite the challenges and obstacles, these stories typically conclude with a happy ending where the CEO and love interest overcome their differences and find lasting love together.
11. **Business Intrigue:** Subplots involving business deals, corporate rivalries, or hostile takeovers may be woven into the narrative, adding complexity to the story.
12. **Personal Growth:** Both the CEO and the love interest experience personal growth and self-discovery as they navigate their relationship and face challenges together.

Reg



# Quiz Time! Order these episodes in a CEO Boss Guy TV show

1



2



3



4



5





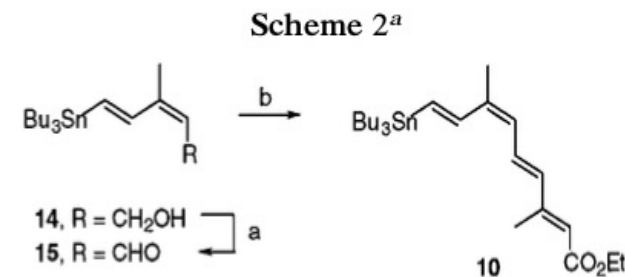
# Procedure Repeats Itself





# Procedure Repeats Itself

1. MnO<sub>2</sub> (8.31 g, 0.10 mol) and K<sub>2</sub>CO<sub>3</sub> (13.2 g, 0.10 mol) were added to a stirred solution of alcohol (2.06 g, 5.32 mmol) in CH<sub>2</sub>Cl<sub>2</sub> (38 mL).
2. Stir (the mixture) for 30 min at 25 °C
3. The suspension was filtered through Celite
4. The filtrate (things that has passed through the filter) was evaporated
5. Purify the residue by column chromatography (SiO<sub>2</sub>, 90:8:2 hexane/EtOAc/Et<sub>3</sub>N)
6. Afforded 1.88 g (92%) of 15 (a chemical shown as its structure in the paper)



<sup>a</sup> (a) MnO<sub>2</sub>, K<sub>2</sub>CO<sub>3</sub>, CH<sub>2</sub>Cl<sub>2</sub>, 0 to 25 °C, 2 h, 92%. (b) diethyl 3-(ethoxycarbonyl)-3-methylprop-2-enylphosphonate, *n*-BuLi, DMPU, THF, -115 to -40 °C, 94%.

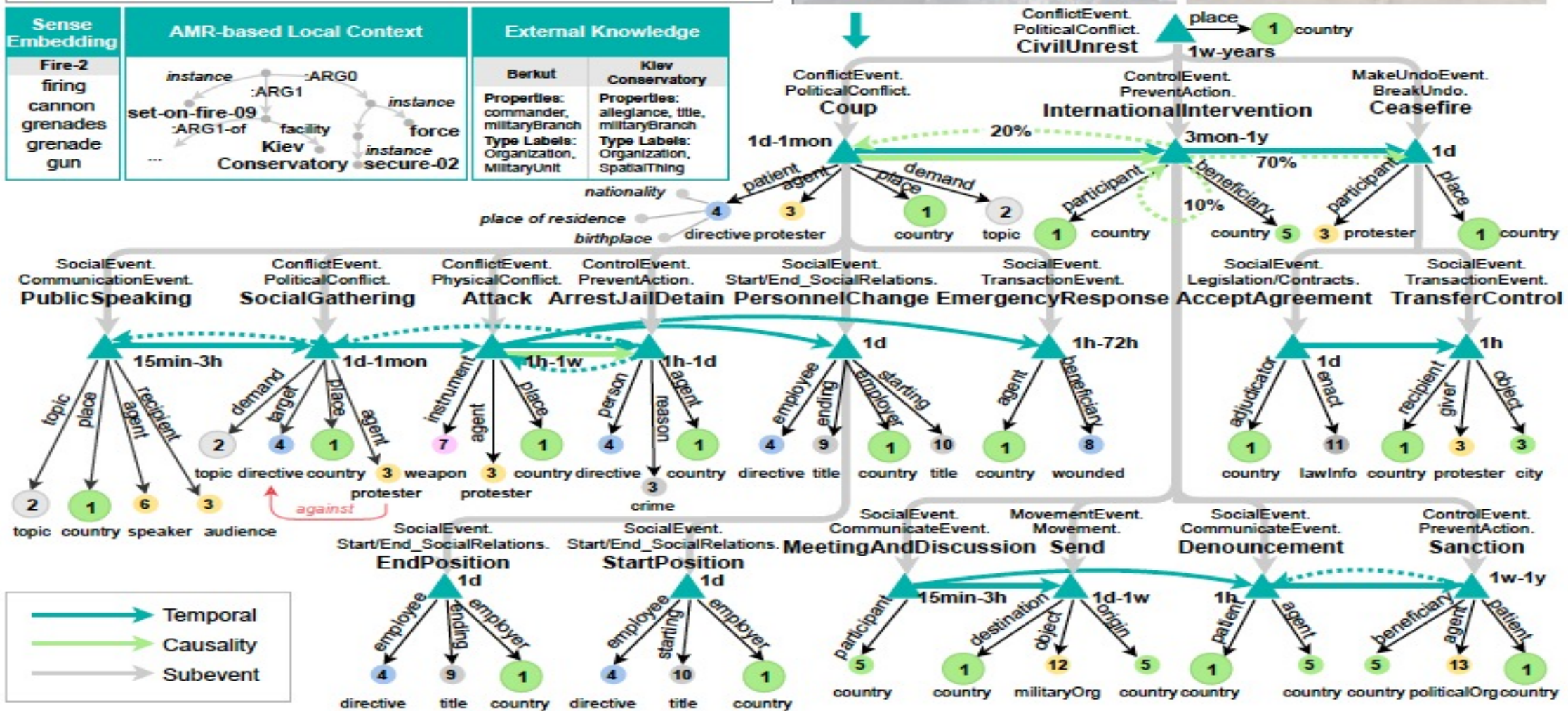
## Quiz Time!

- John purchased a high-pressure cooker and materials for bomb making. \_\_\_\_\_ . \_\_\_\_\_ . There was an explosion during Boston Marathon. \_\_\_\_\_ . Police found the bomb was inside a high-pressure cooker. \_\_\_\_\_ . \_\_\_\_\_ . John was sentenced to death.

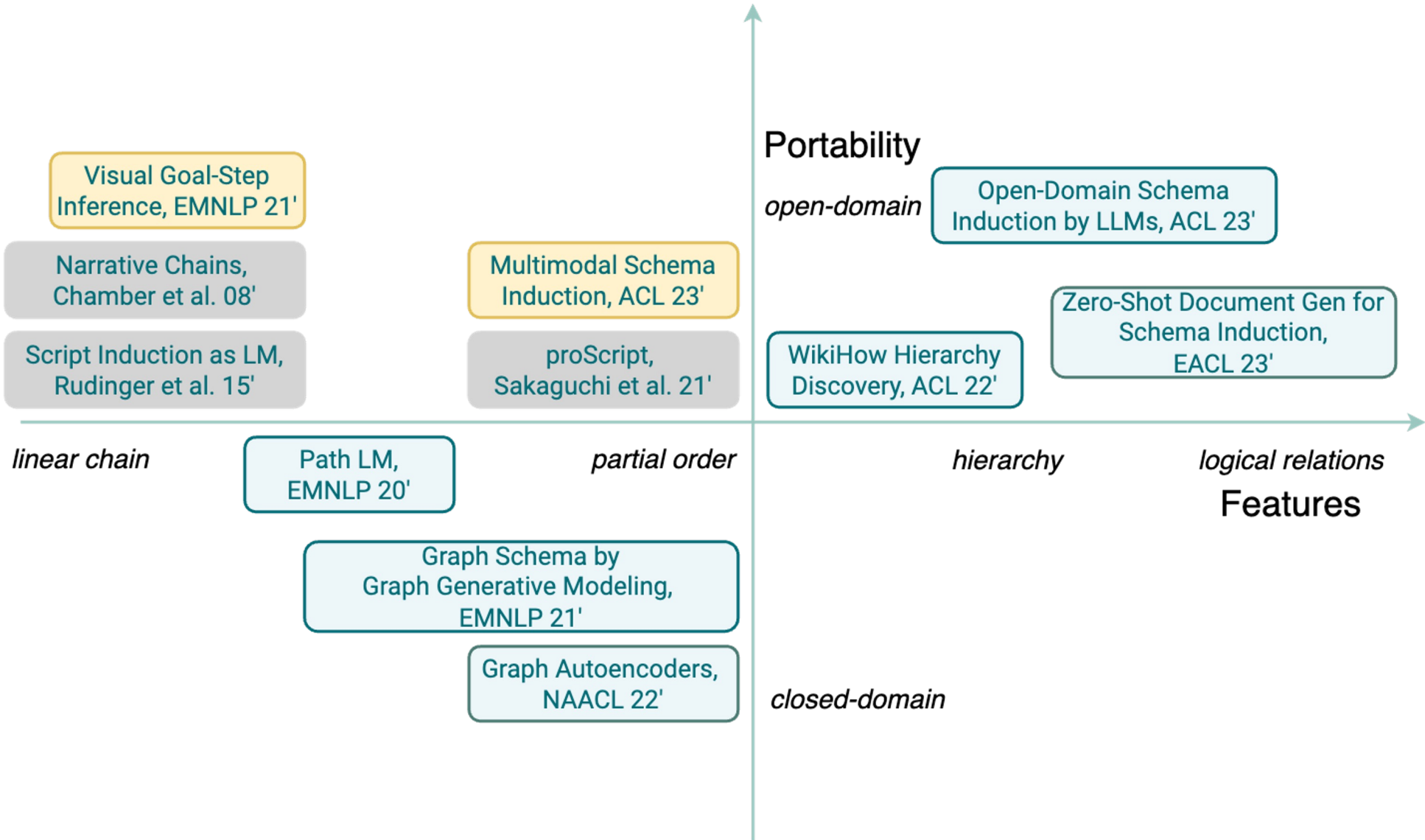


- Schemas, or scripts, were originally defined by (Schank and Abelson, 1975) as “ a predetermined, stereotyped sequence of actions that defines a well-known situation”

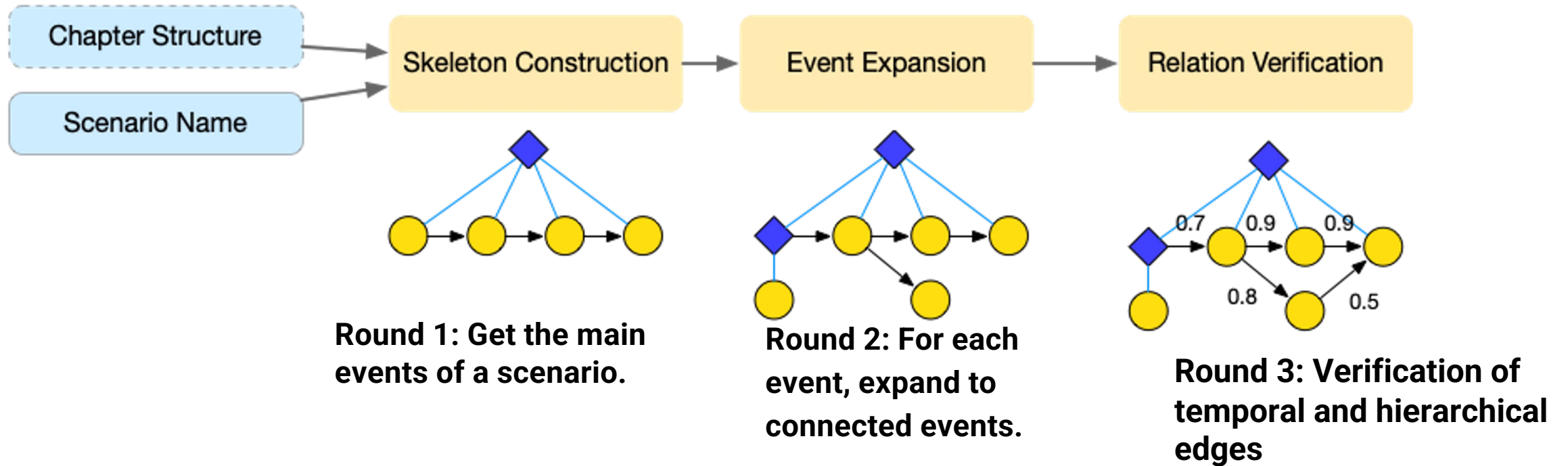
2014 Thai coup d'état: Однако протесты и блокада длятся уже почти 3 месяца, а военные так и не перешли к действиям  
 2013 Egyptian coup d'état: ... General Abdel Fattah el-Sisi announced that he there would be calling new presidential and Shura Council elections.  
 Ukrainian crisis: At 09:25, protesters pushed the Berkut back to the October Palace after security forces tried to set fire to Kiev Conservatory, which was being used as a field hospital for wounded protesters.



# Open-Domain Hierarchical Schema Induction



# Open Domain Hierarchical Schema Induction via LLMs [Li et al., ACL2023]



- Advantages Compared to Previous Methods

- **Open-Domain:** our model can induce schemas for any scenario given the scenario name
- **Extensible:** our paradigm can support new event-event relations by adding new prompt templates. We showcase this by including hierarchical relations.
- **Interpretable:** by representing events with sentences, human assessors consider our schemas to be more readable than prior approaches



# Retrieval-Augmented Prompting

- When humans curate schemas, they often refer to related news articles or Wikipedia
- Whenever our prompt is related to an event, we simulate this process by using the GPT3-generated event description to retrieve related passages to serve as extra context to the language model
  - To encourage the model to output a general answer, we retrieve 3 passages per prompt using a pretrained TCT-ColBERT model.
  - Retrieving multiple passage (ideally about different instances) is important for guiding the model to produce a **generalized answer**.

## Retrieval-Augmented Prompt

Based on the following passages  
{retrieved passages},  
{prompt}



# Event Skeleton Construction and Event Expansion

## Event Skeleton Prompt

{evt.name} is defined as "{evt.description}".  
List the major events that happen in the  
{evt.name} of a {scenario}:

## Event Expansion Prompt

What happened during "{evt.description}"?  
List the answers:

## Question Decomposition for Event-Event Relations

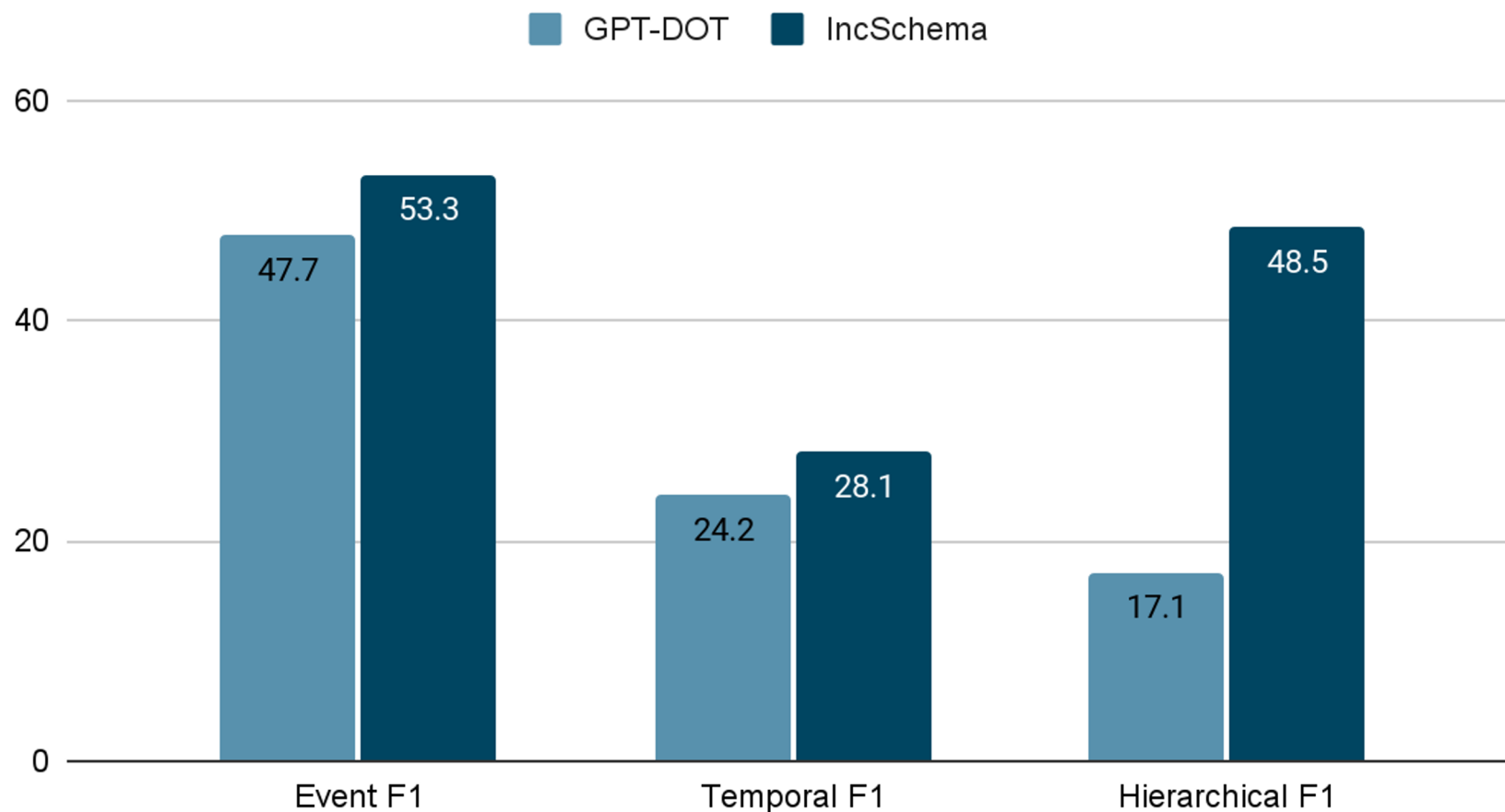
- Instead of directly asking for “Does event A happen before event B?”, we ask three questions about start time, end time and duration.
- This allows us to avoid conflicting hierarchical and temporal relations since the hierarchical relation can be defined as spatial-temporal containment.

Relation	Allen's base relations	$e_1$ starts before $e_2$ ?	$e_1$ ends before $e_2$ ?	Duration
$e_1 \prec e_2$	$e_1$ precedes $e_2$ , $e_1$ meets $e_2$	Yes	Yes	-
$e_1 \succ e_2$	$e_1$ is preceded by $e_2$ , $e_1$ is met by $e_2$	No	No	-
$e_1 \subset e_2$	$e_1$ starts $e_2$ , $e_1$ during $e_2$ , $e_1$ finishes $e_2$	No	Yes	$d(e_1) < d(e_2)$
$e_1 \supset e_2$	$e_1$ is started by $e_2$ , $e_1$ contains $e_2$ , $e_1$ is finished by $e_2$	Yes	No	$d(e_1) > d(e_2)$
$e_1 \parallel e_2$	$e_1$ overlaps with $e_2$ , $e_1$ is equal to $e_2$	Yes	No	$d(e_1) \leq d(e_2)$
$e_1 \parallel e_2$	$e_1$ is overlapped by $e_2$	No	Yes	$d(e_1) > d(e_2)$

# Hierarchical Schema Quality: Findings

Dataset	# Scenarios	# Event	# Temp.	# Hier.
RESIN-11	11	579	381	603
ODIN	18	593	398	569
ProScript	2077	14997	13946	0

## Schema Quality Evaluation

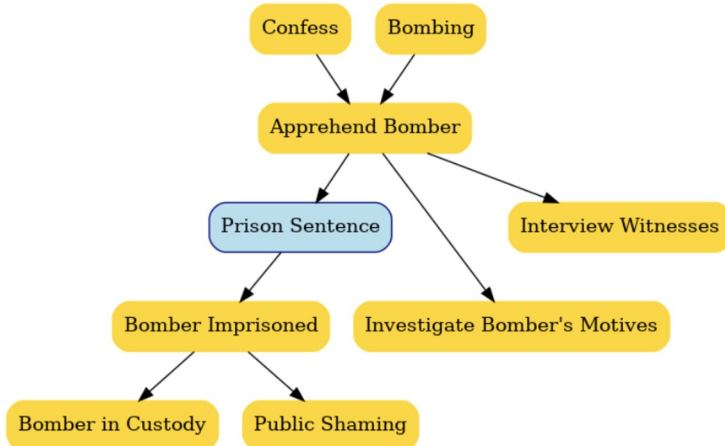


Our baseline GPT-DOT, directly asks GPT3 to output a linearized graph format of the final schema given some in-context examples.

Compared to our model, GPT-DOT generates much fewer events (10.11 events for GPT-DOT VS 52.6 events on ODIN) which leads to high precision but low recall.

GPT-DOT struggles with hierarchical relations.

# Interpretability Evaluation Results

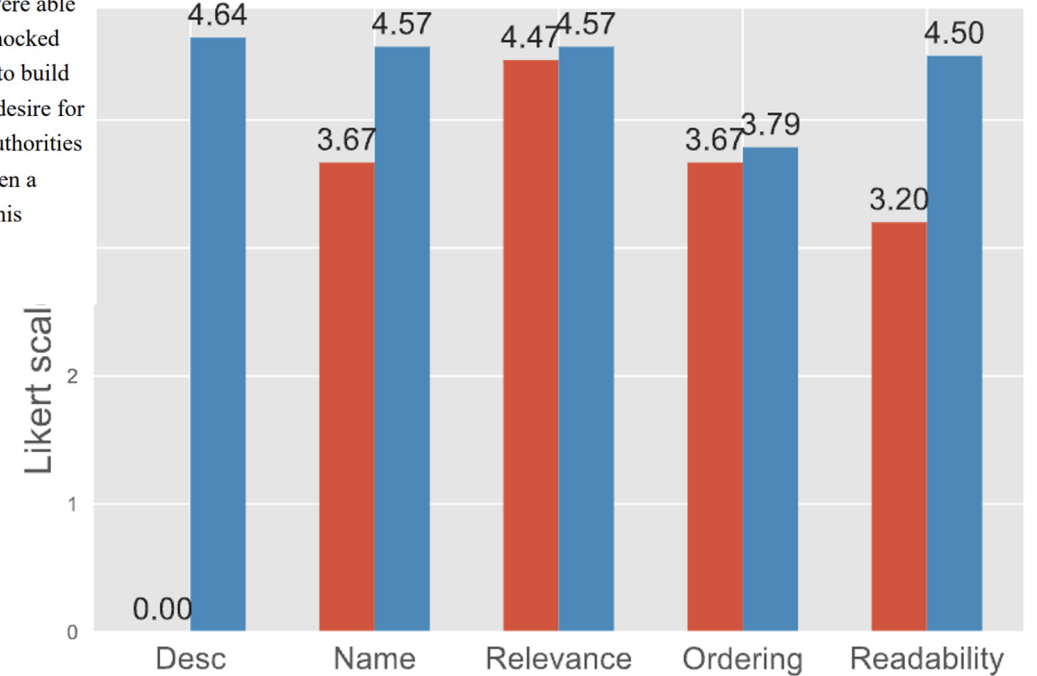


This is part of a schema about the Criminal Investigation of a car bombing event. Please write a story describing the figure:

The authorities were on high alert after reports of a car **bombing** in the city. They immediately began their **investigation**, determined to apprehend the bomber and bring them to justice. They **interviewed witnesses** and **gathered evidence**, piecing together the details of the heinous act. Their hard work paid off when they were able to **track down** and **apprehend** the bomber. During **questioning**, the bomber shocked investigators by **confessing to the crime**, sparing them the trouble of having to build a case against him. The bomber cited his **motives for the attack** as a twisted desire for **public shaming**. With the **bomber in custody** and a confession in hand, the authorities were able to swiftly bring him to **trial**. The bomber was found guilty and given a harsh **prison sentence**, ensuring he would spend a long time behind bars for his terrible crime. ...

- Human assessors are able to compose a longer story with better event coverage using our schema while taking roughly the same amount of time.
- Human assessors rate our event descriptions and event names to be very helpful (>4.5 score) and our schemas are easier to understand compared to the baseline.

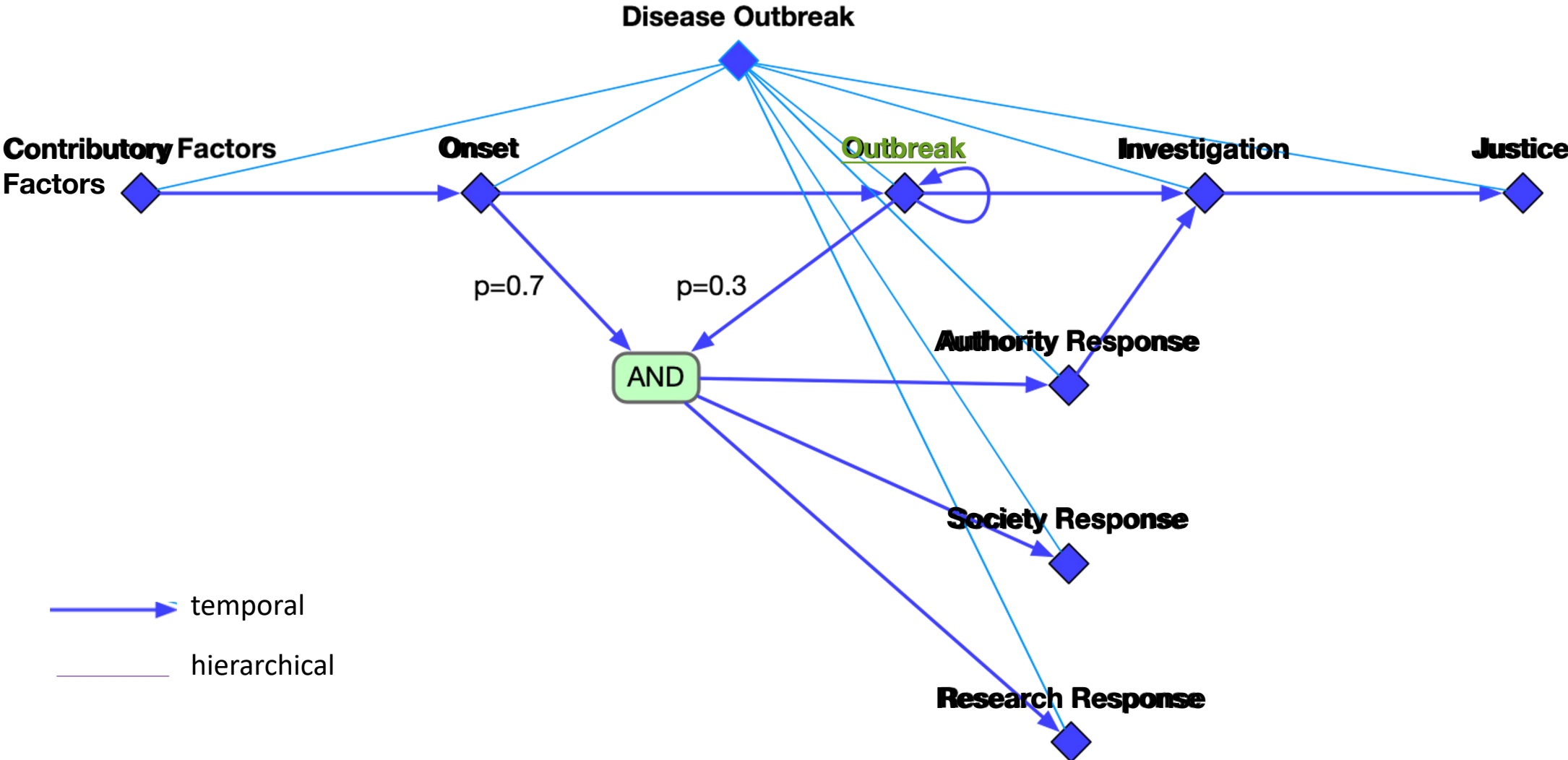
Model	Coverage↑	Len(words)↑	Time(mins)↓
Double-GAE	79.8	9.62	0.998
INCPROMPT	89.7	15.53	1.137



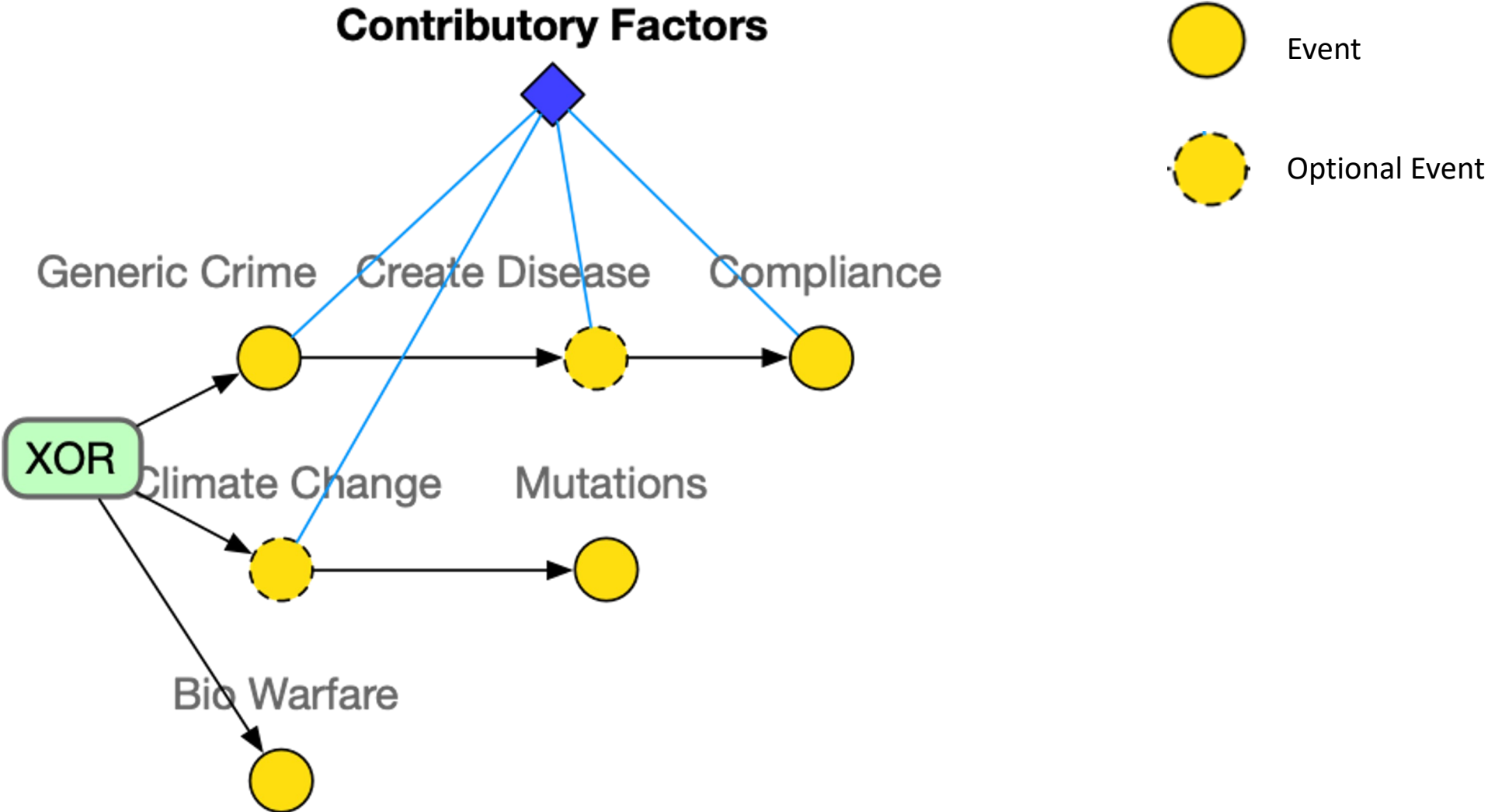
Blue: Our method IncPrompt  
Red: Double-GAE



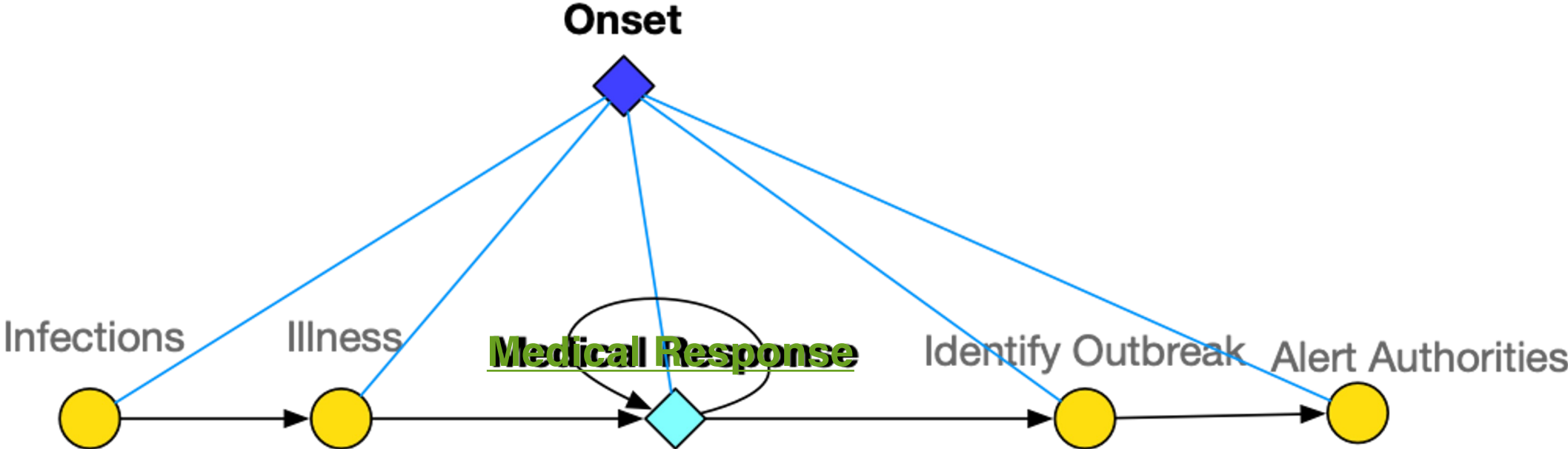
# Epidemic Schema Example



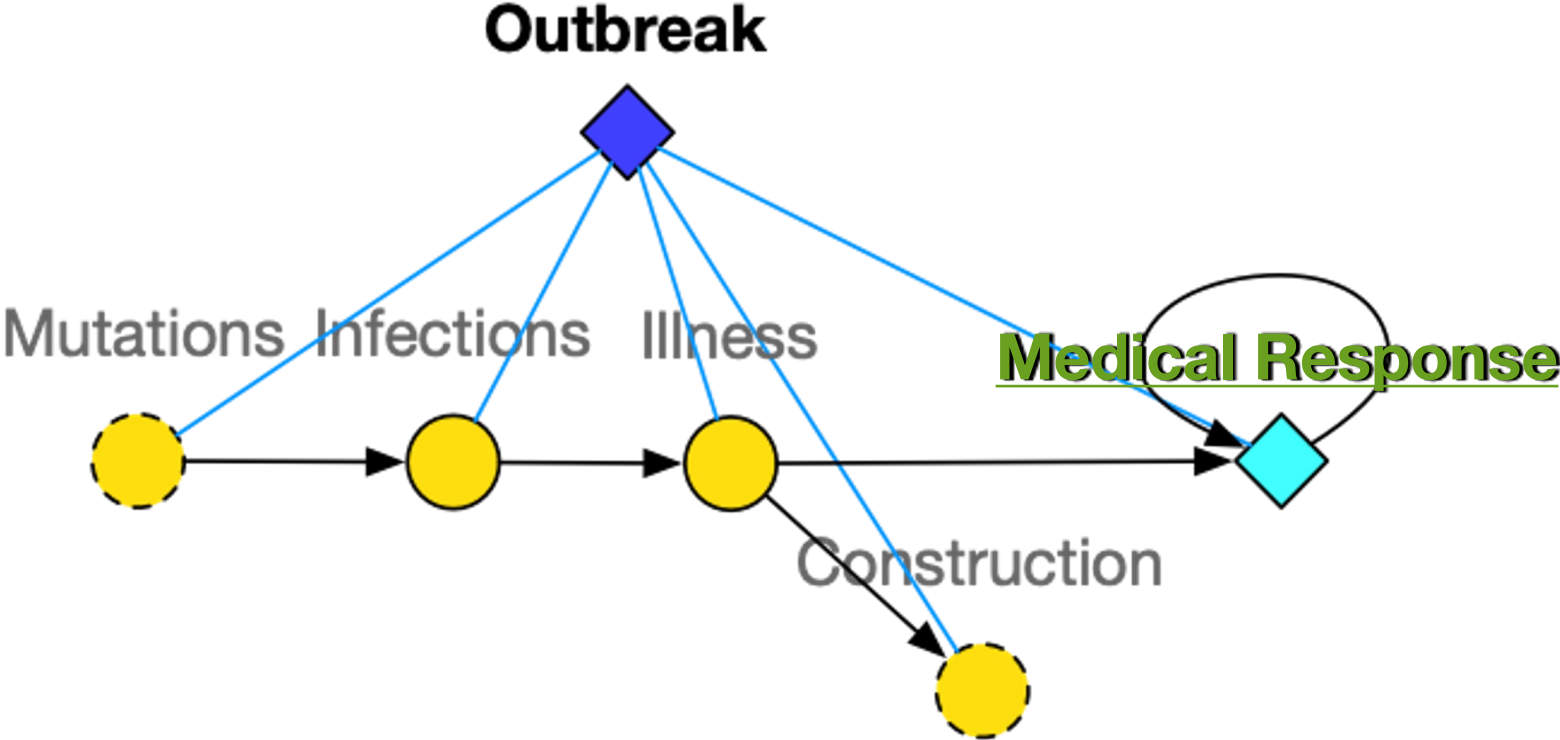
# Contributory Factors



# Onset

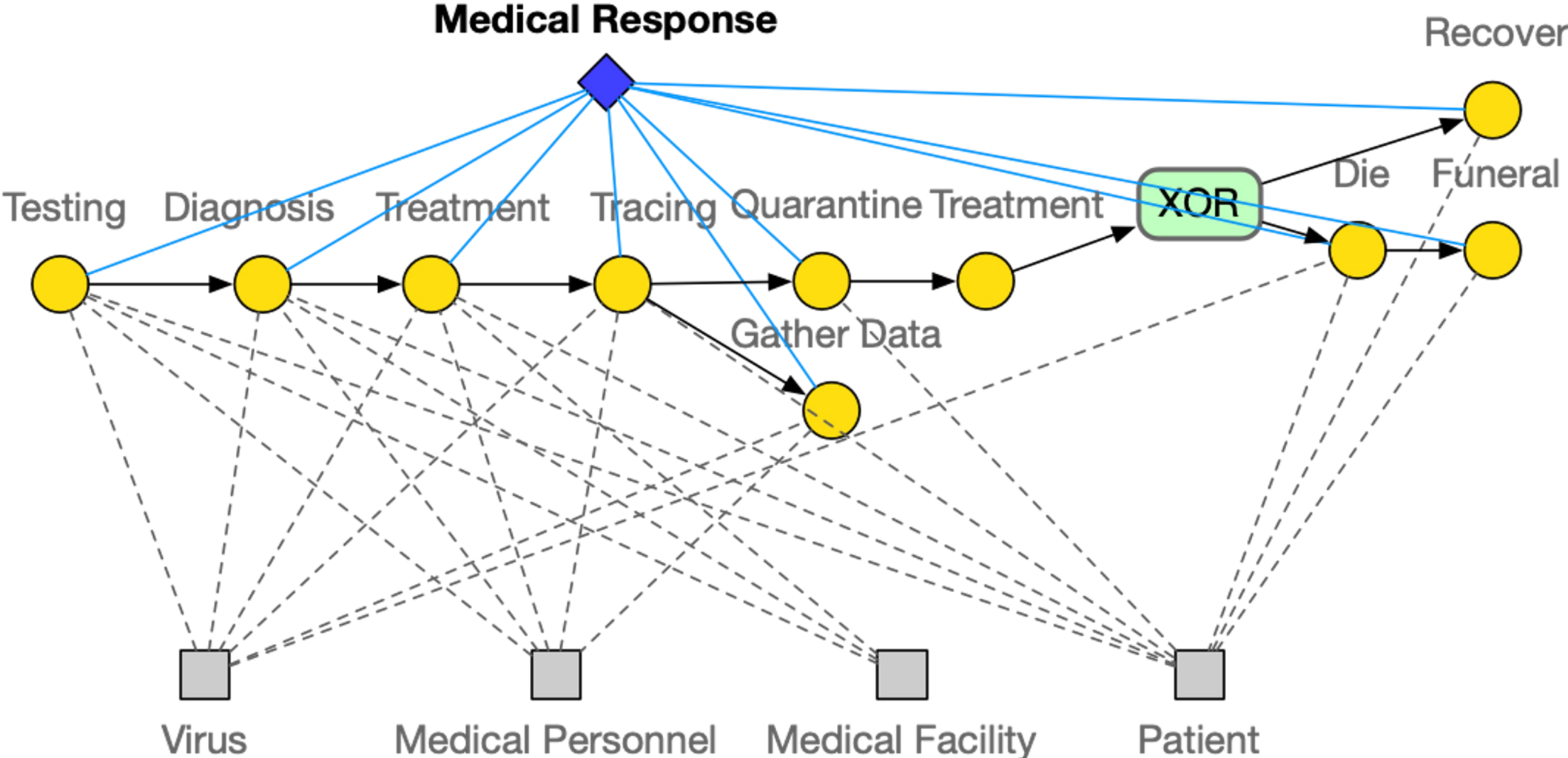


# Outbreak

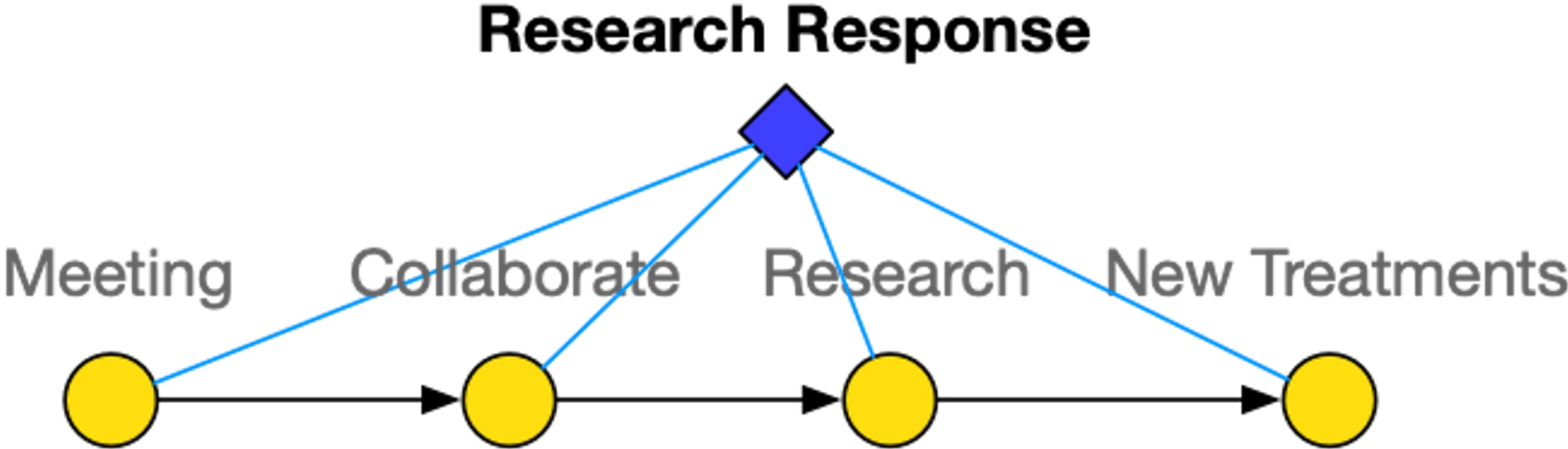




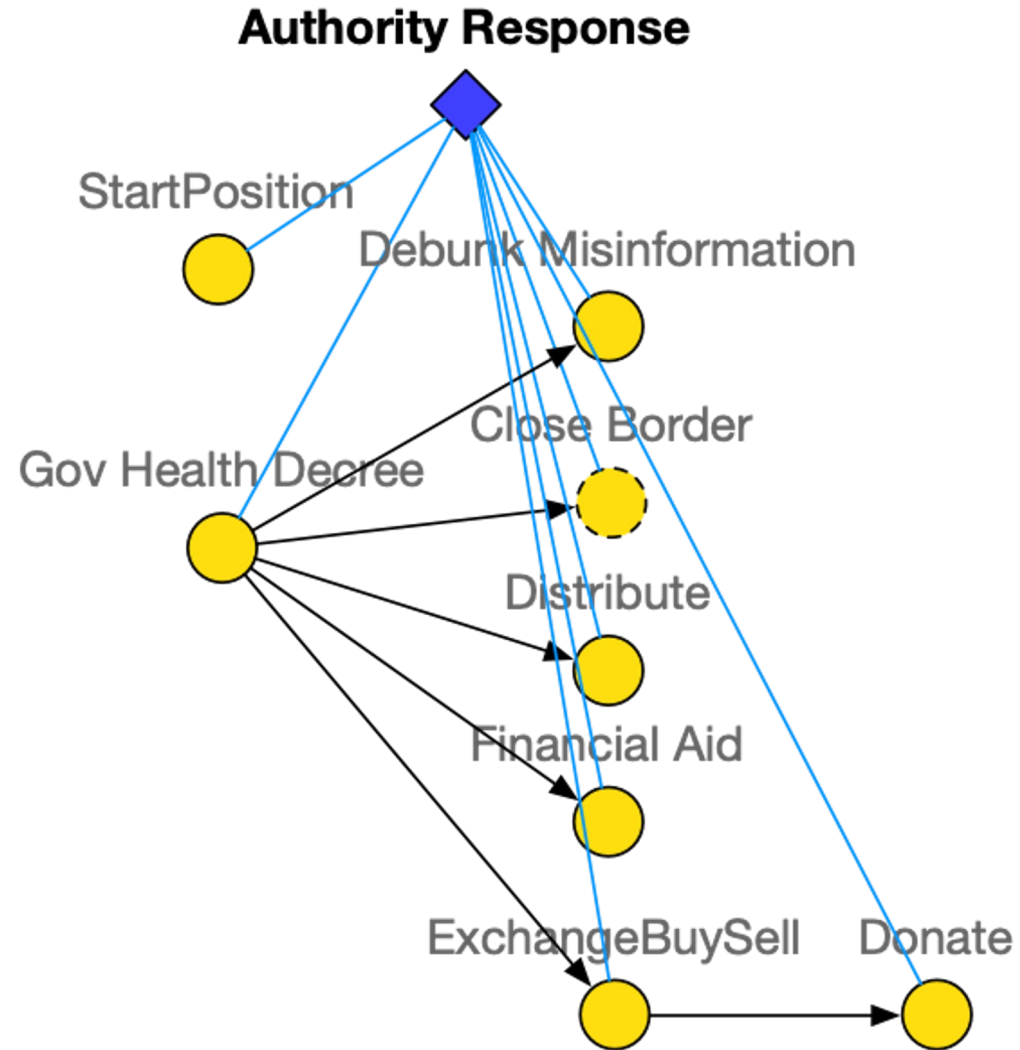
# Medical Response



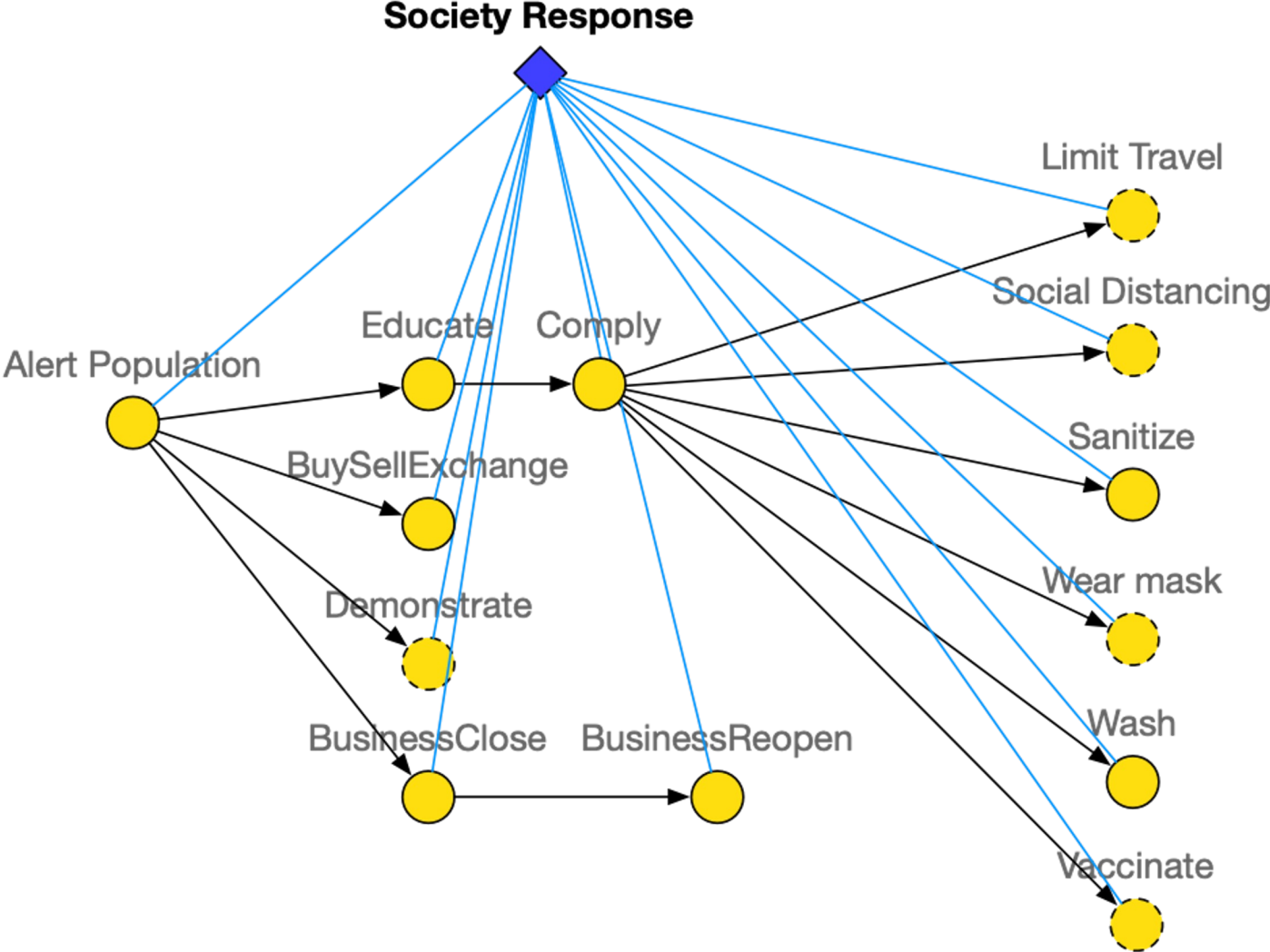
# Research Response



# Authority Response

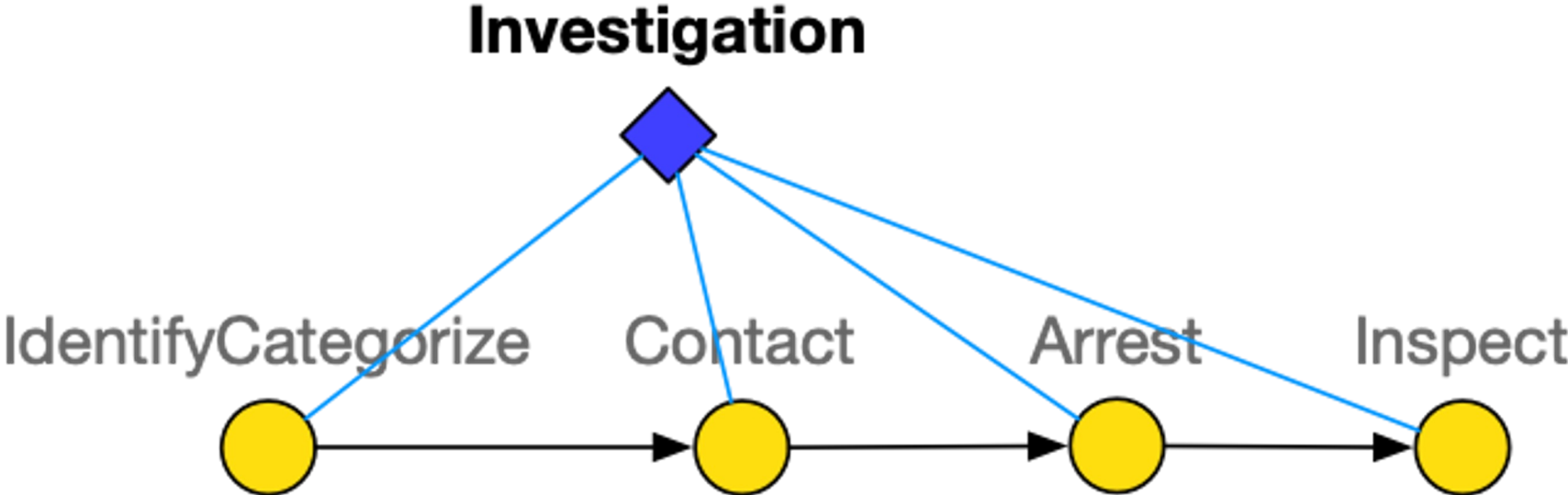


# Society Response

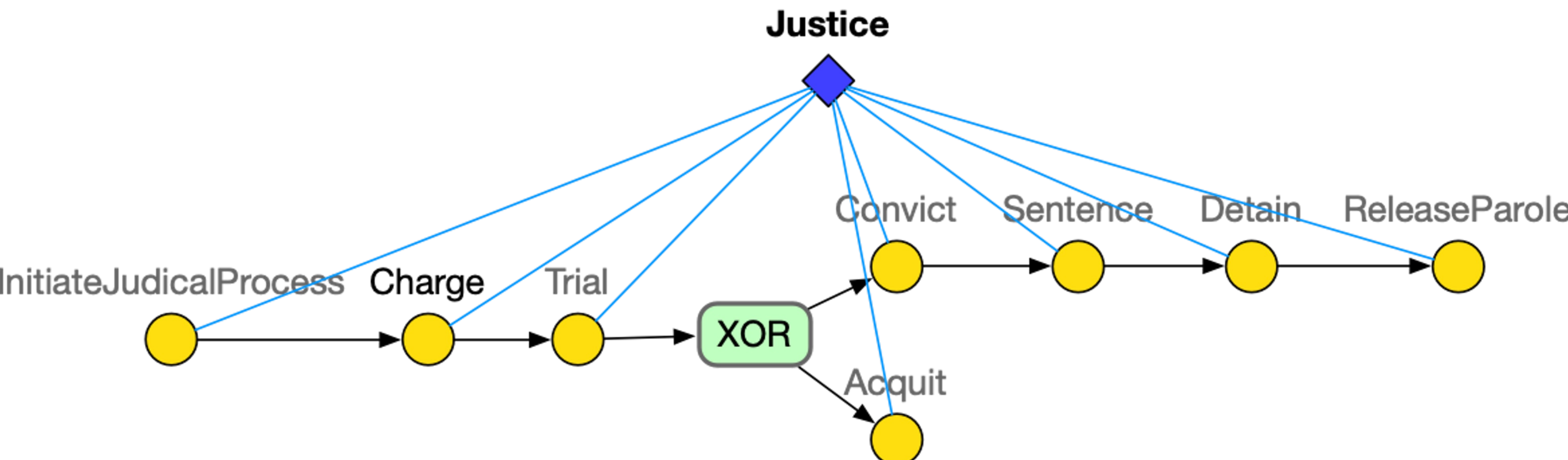




# Investigation

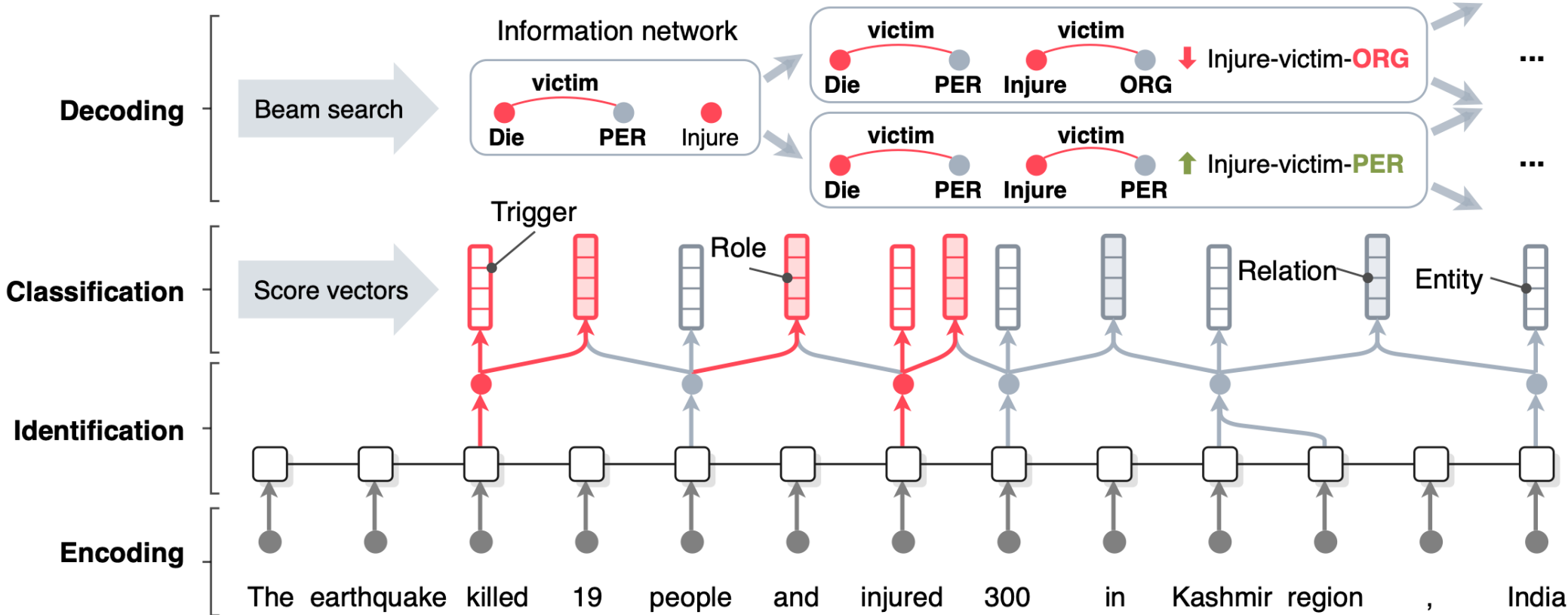


# Justice



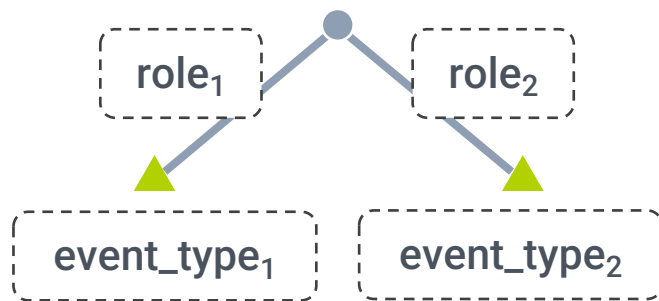
# Schema Enhanced Information Extraction

- A typical sequence-to-graph information extraction model [Lin et al., ACL2020]

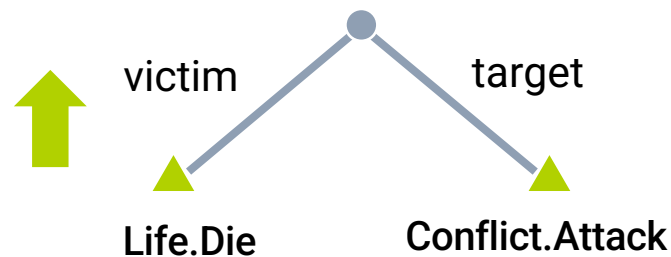


# Schema-Enhanced Information Extraction

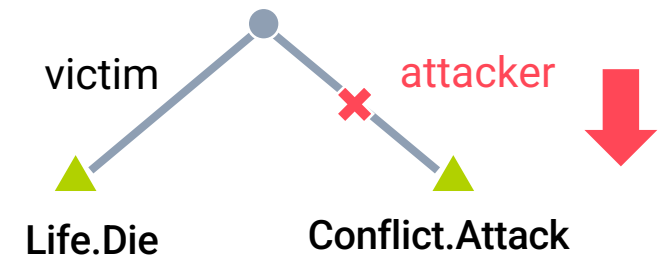
- We design a set of *global feature templates* (e.g.,  $\text{event\_type}_1 - \text{role}_1 - \text{role}_2 - \text{event\_type}_2$ : an entity acts a  $\text{role}_1$  argument for an  $\text{event\_type}_1$  event and a  $\text{role}_2$  argument for an  $\text{event\_type}_2$  event in the same sentence). A more comprehensive event schema library is inducted following (Li et al, 2020).
- The model learns the *weight* of each feature during training



Template



Positive weight



Negative weight

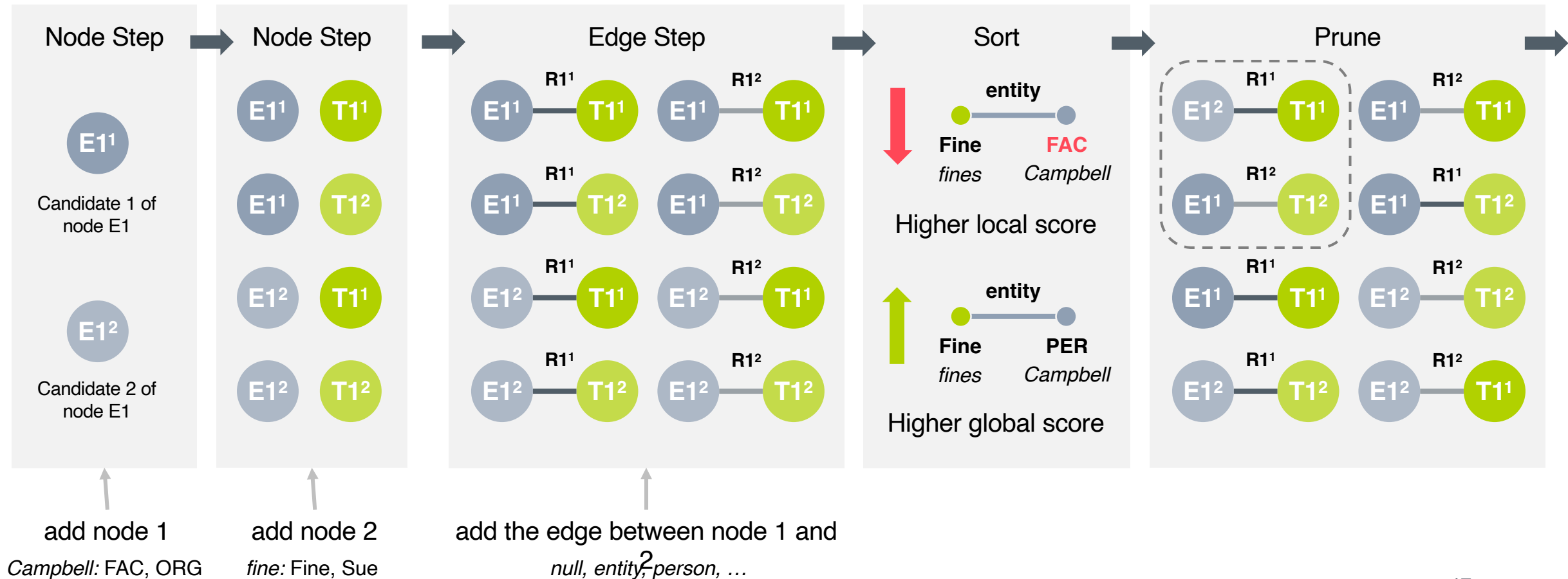
- Global score of a graph: local graph score + global feature score

$$s(G) = s'(G) + \mathbf{u}f_G$$



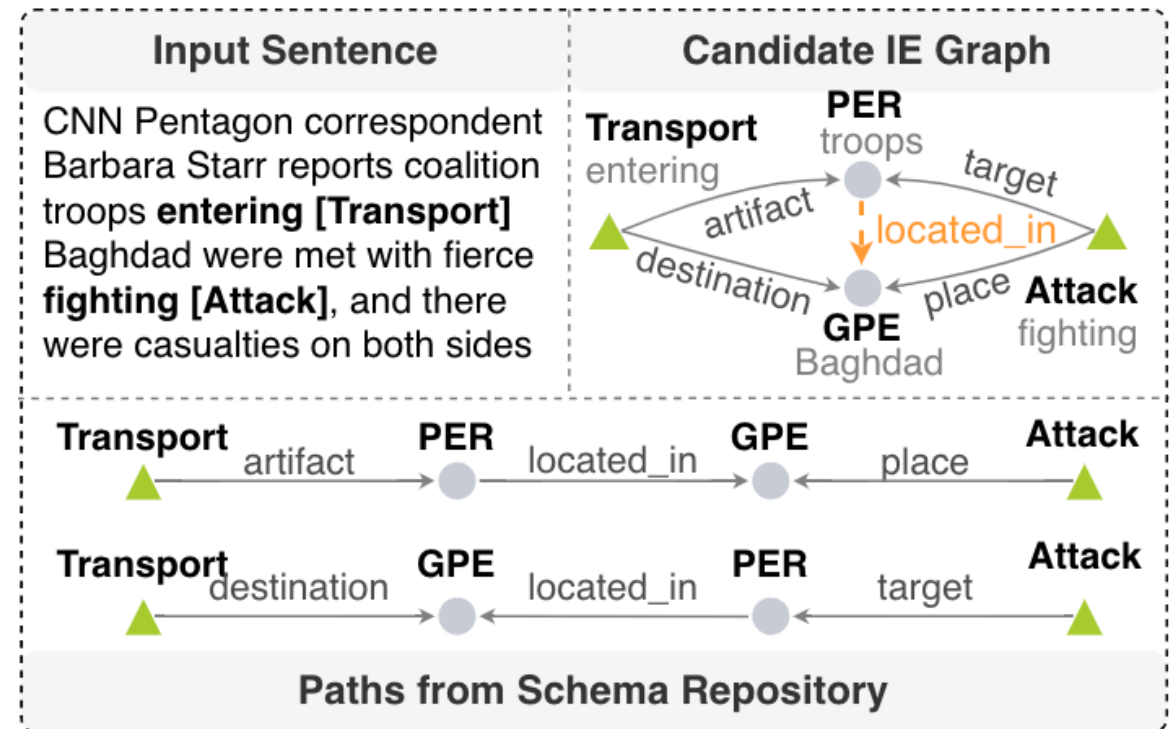
# Schema-Enhanced Information Extraction

- We use beam search to decode the information graph (Lin, et al, 2020)
- Example: *He also brought a check from **Campbell** to pay the **fin**es and fees.*



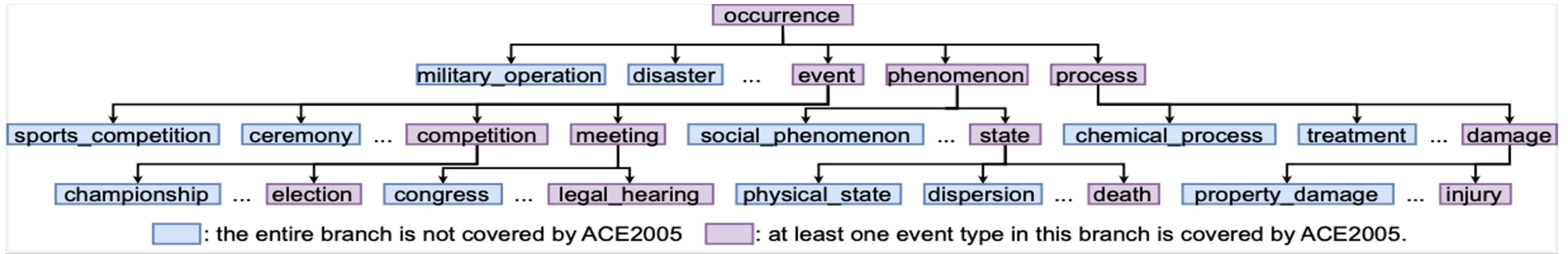
# Schema-Enhanced Information Extraction (Li et al., EMNLP2020)

- Use the state-of-the-art IE system OneIE (Lin et al, 2020) to decode converts each input document into an IE graph
- Each path in the graph schema is encoded as a single global feature for scoring candidate IE graphs
- OneIE promotes candidate IE graphs containing paths matching schema graphs
- <http://blender.cs.illinois.edu/software/oneie>
- F-scores (%) on ACE2005 data [Lin et al., ACL2020]:



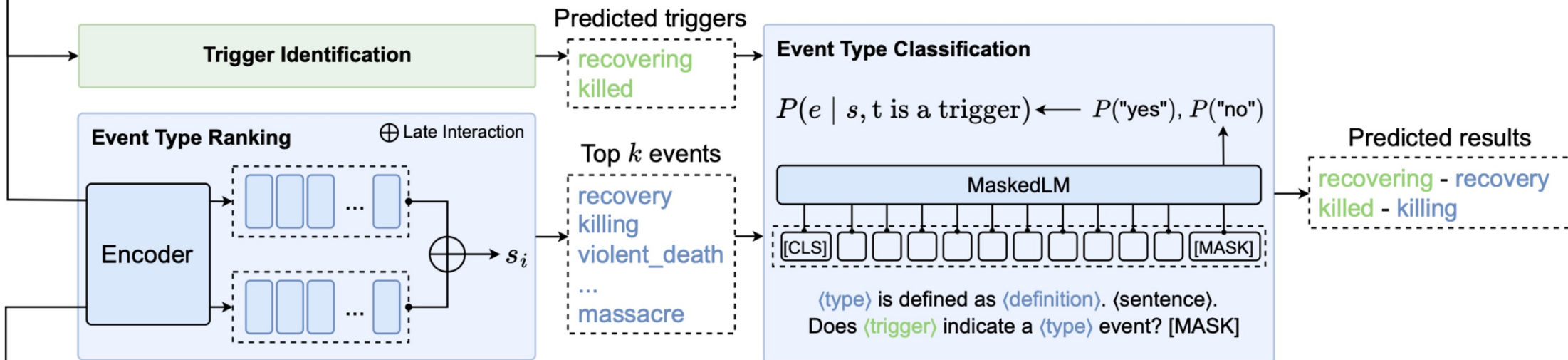
Dataset	Entity	Event Trigger Identification	Event Trigger Classification	Event Argument Identification	Event Argument Classification	Relation
Baseline	90.3	75.8	72.7	57.8	55.5	44.7
+PathLM	90.2	<b>76.0</b>	<b>73.4</b>	<b>59.0</b>	<b>56.6</b>	<b>60.9</b>

# General-Purpose Event Extraction for 3,465 types [Li et al., EMNLP2023]



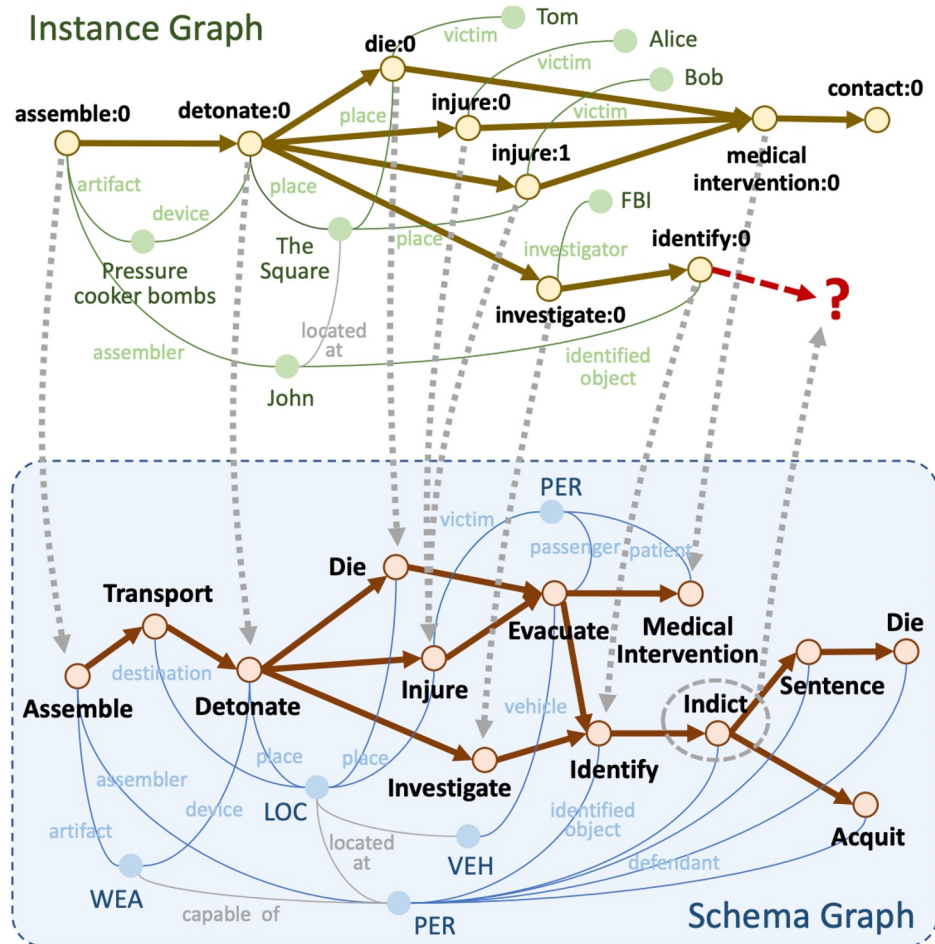
**Sentence  $s$**  : A police officer taking part in recovering the bodies said about 17 people were killed.

  Trained with clean annotation  
  Trained with noisy annotation



**Event Node  $e_i$** : recovery - return of a lost or stolen item to its owners  
 ...  
 } Thousands of events

# Hierarchical Schema Matching and Event Prediction [Wang et al., AKBC2022]

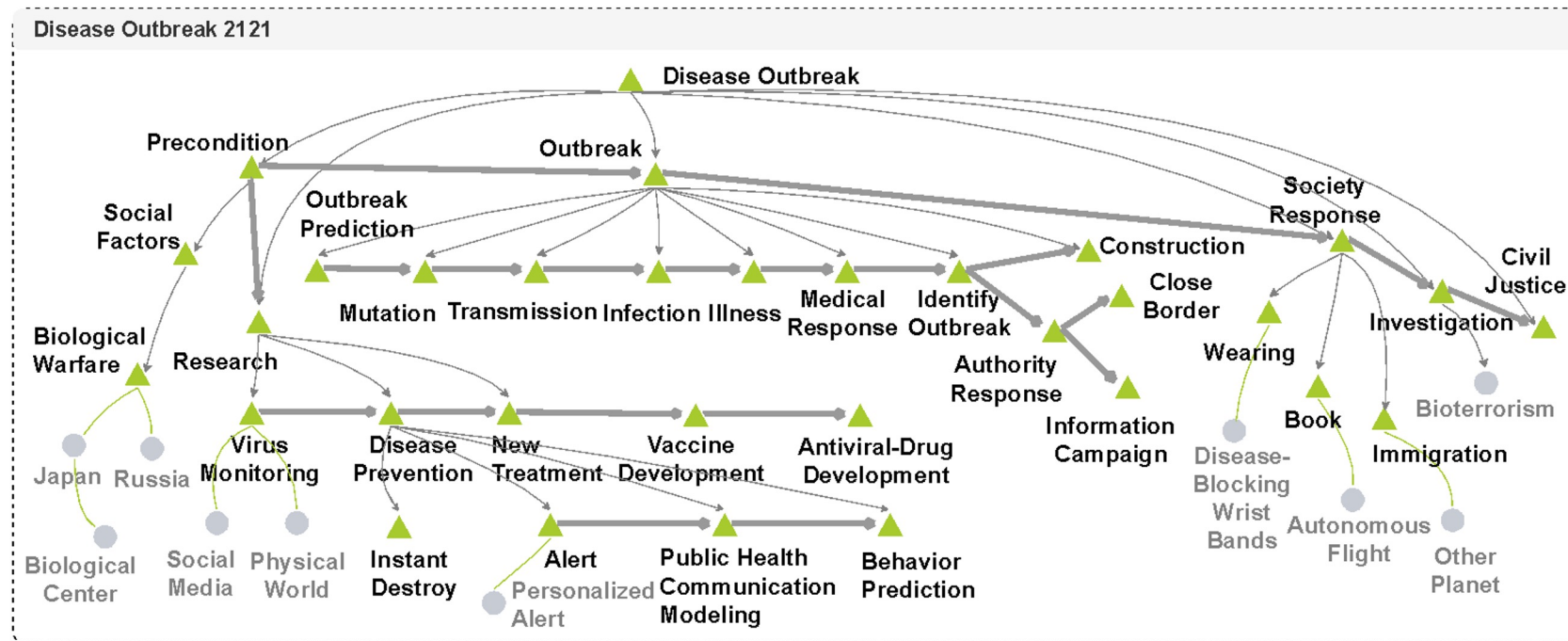


- Compute the probability for each schema event being missing conditioned on the current instance graph.
  - Neighbor Information (GNN)
  - Path Information (Bag of paths representations)
- Select the schema events with high probabilities as our predictions.
- Instantiate arguments from the matched events.



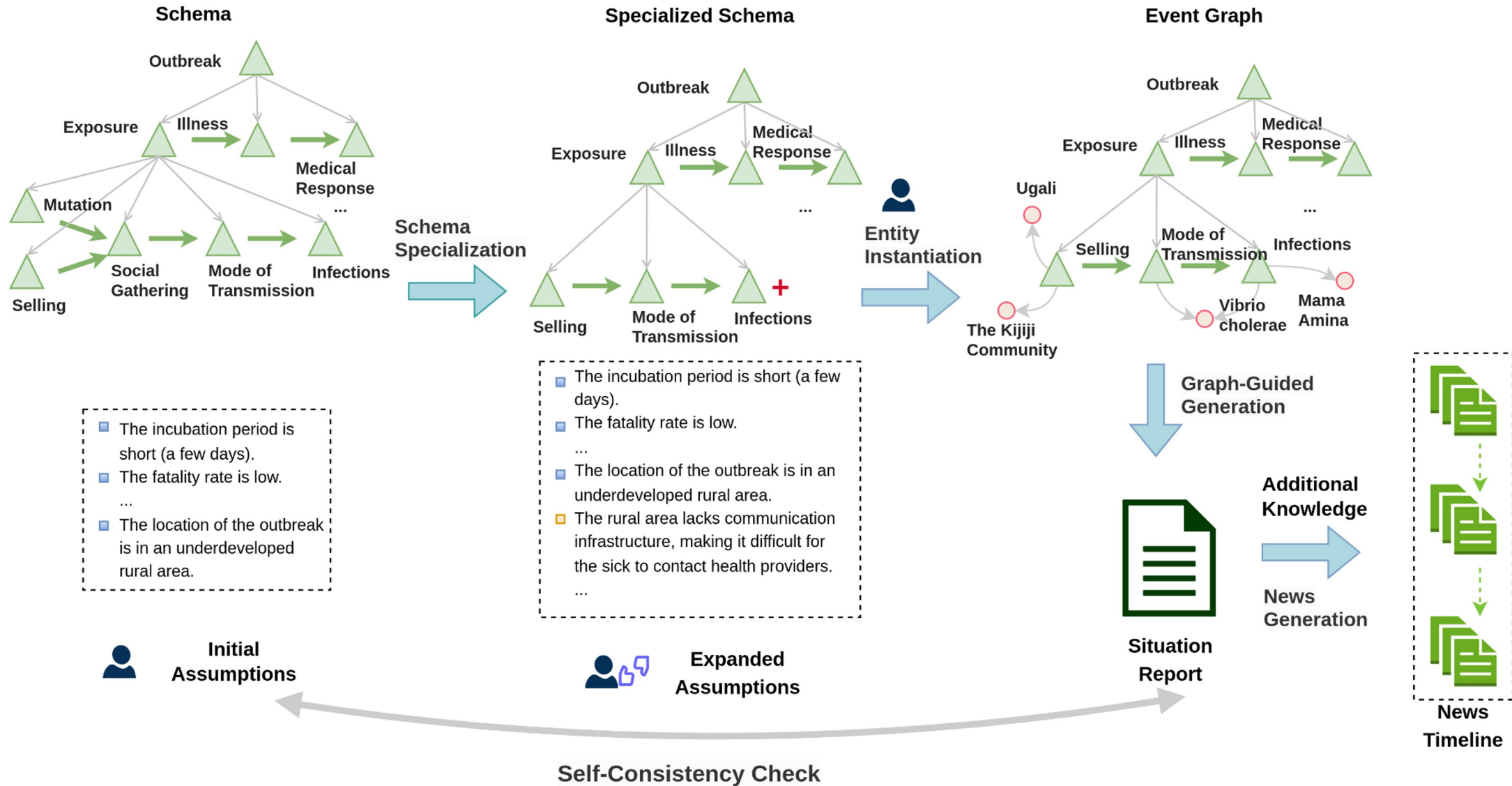
# News Event Simulator for Disaster Forecasting and Prevention

- **New Complex Event Graph Skeleton for a user Imagined Disease Outbreak in 2121**



- **Then generate a timeline of news articles about this imagined scenario**
- **Conduct factual error correction through inference and neuron subset pattern mining and model editing**

# “What-If” Situation Simulation for Disaster Forecasting

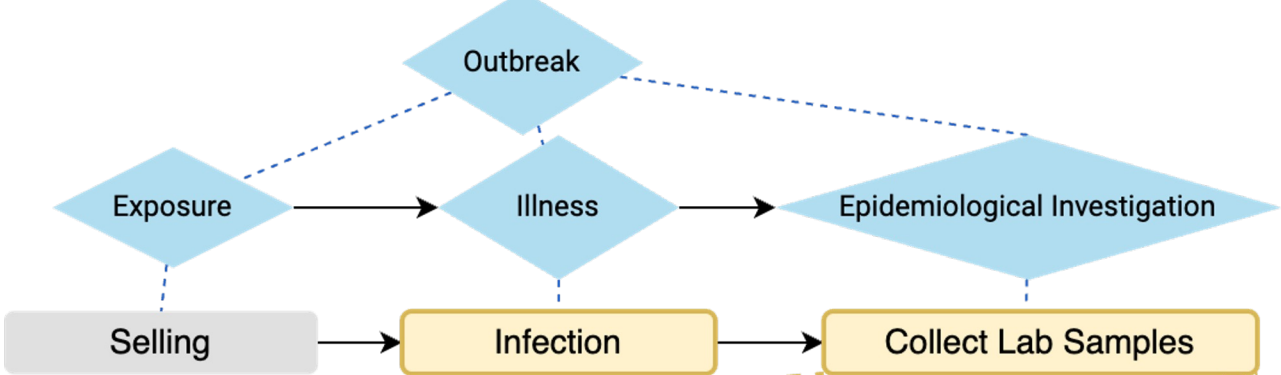


# Schema Specialization


## Assumptions

- The infectivity of the disease is high.
- The fatality rate of the disease is low.
- The location of the outbreak is in an underdeveloped rural area.

## Schema



Collect Lab Samples: Collecting and processing samples with the pathogen.



*You are an expert in public health whose task is to explain how disease outbreaks events unfold under different conditions.*

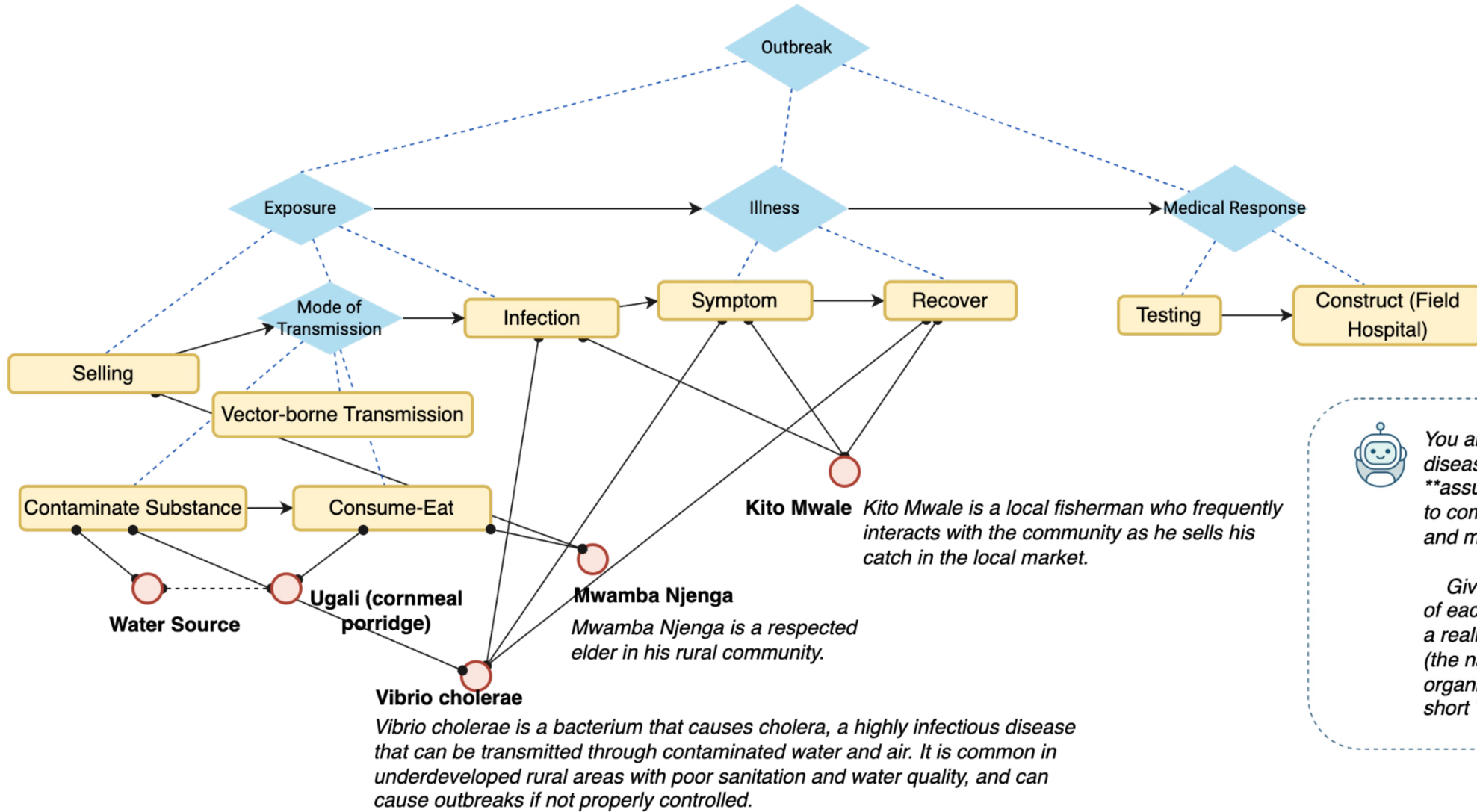
*Given the {assumptions}, how likely would the following event occur?*

Somewhat Likely

New Assumption:  
There is a well-equipped mobile lab team that can reach the rural area and collect samples despite the lack of communication infrastructure

Very Likely

# Entity Instantiation



You are writing a story on a disease outbreak given some **assumptions**. You are starting to come up with the background and main characters in the story.

Given an abstract description of each **entity**, come up with a realistic **name** of the entity (the name of the person, organization, disease etc.) and a short **backstory** of the entity.



# Disaster Forecasting

In the rural areas of Nandi County, a local vendor named Joe Mwangi unknowingly spreads the H3N2 Influenza to others in the community. The distinctive cough of one of the first infected individuals, Sam Odera, becomes a telltale sign of the disease's presence. The Nandi County Health Department swiftly quarantines Sam, recognizing his distinctive cough as a symptom of H3N2 Influenza. Rural Connect Telecommunications (RCT) initiates a comprehensive health training program for its employees, including information on how to recognize the distinctive symptoms of the disease, measures to prevent airborne transmission, and protocols to follow if they or someone in the community falls ill. The government issues a decree enforcing strict quarantine measures, mandatory mask-wearing, and social distancing, and allocates additional resources to the Nandi County Health Department to enhance their capacity to manage the outbreak. The Nandi County Health Department imposes travel restrictions within the county, advising residents to stay within their settlements and avoid unnecessary travel to neighboring areas.

## **1. The distinctive cough of the H3N2 Influenza virus:**

The distinctive cough associated with the H3N2 Influenza virus is a key factor in identifying the outbreak. The cough is characterized by a dry, hacking sound, often accompanied by a sore throat and difficulty breathing. This distinctive cough is a telltale sign of the virus's presence and can help healthcare professionals and the public identify and respond to the outbreak.

## **2. The role of the Nandi County Health Department:**

The Nandi County Health Department plays a crucial role in managing the H3N2 Influenza outbreak. They are responsible for identifying the outbreak, implementing quarantine measures, and providing healthcare services to those affected. Their swift response and effective communication with the public are essential in containing the outbreak and minimizing its impact on the community.

## **3. The impact of the outbreak on the rural areas of Nandi County:**

The H3N2 Influenza outbreak has a significant impact on the rural areas of Nandi County. The lack of healthcare facilities and communication infrastructure in these areas makes it challenging to manage the outbreak effectively. The travel restrictions imposed by the government further exacerbate the situation, limiting access to essential services and resources.

## **4. The role of Rural Connect Telecommunications (RCT) in managing the outbreak:**

Rural Connect Telecommunications (RCT) plays a crucial role in managing the H3N2 Influenza outbreak. Their comprehensive health training program for employees and proactive measures to prevent the spread of the disease are commendable. Their efforts to improve communication infrastructure and allocate additional resources to the Nandi County Health Department are also essential in managing the outbreak effectively.

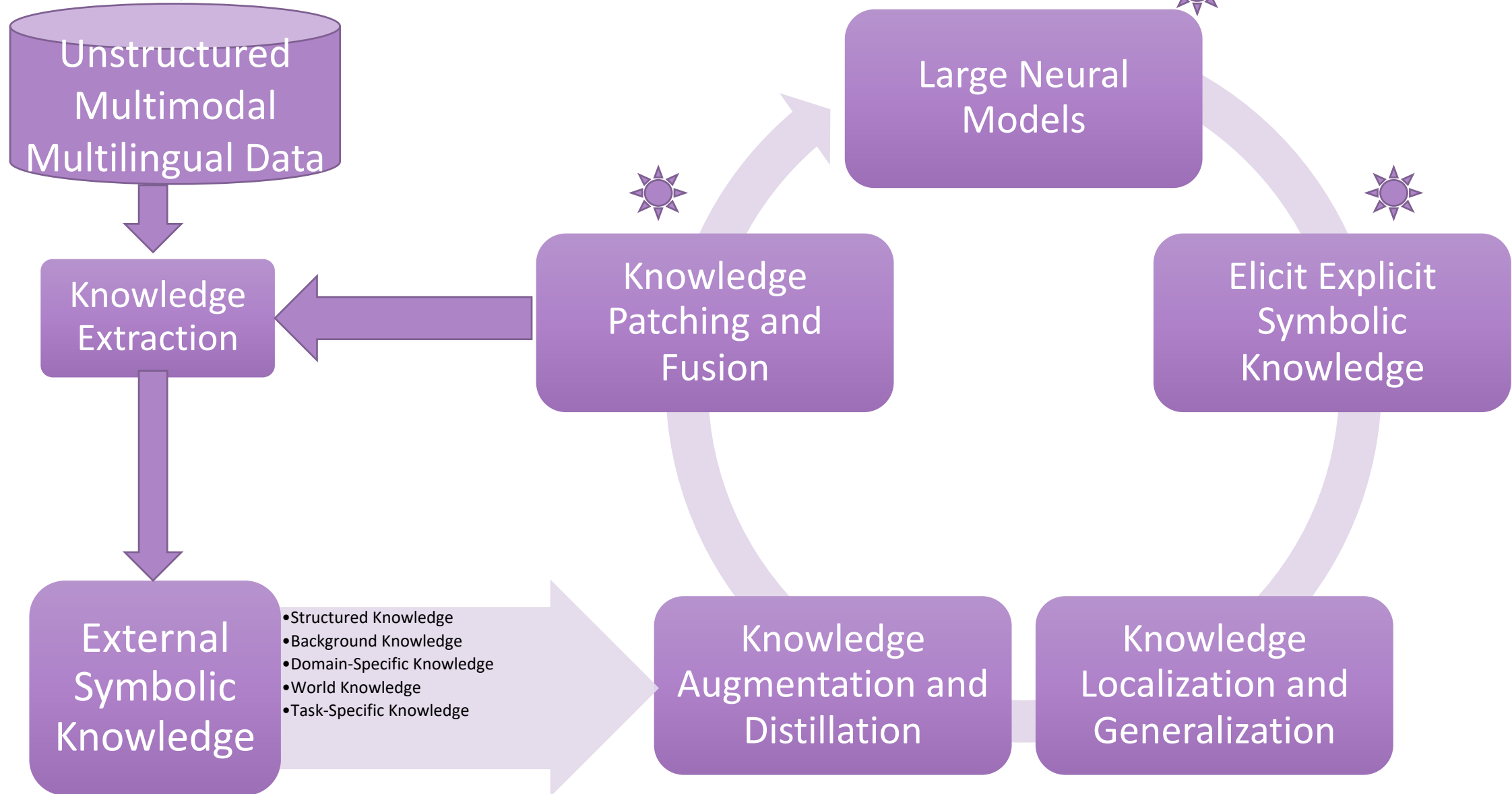
# Impact of Schema Induction on Event Understanding

Task	Before Using Schema	After Using Schema	Compared to LLMs Only Methods
Ontology Induction	Manual	Automatic open-domain entity/relation/event/argument role induction	Cannot do
Schema Induction	Linear chain, closed-domain, from text only	Hierarchical, logical, temporal, open-domain (43 newsworthy scenarios+general domain WikiHow/Youtube), from multi-modal	5.6% higher event F1, 31.4% higher hierarchical F1 than GPT
Text Event Extraction	33 types, supervised, require 5000 event mentions to train	3,465 types, ontology and schema guided extraction, low cost/distant supervision	44.3% higher F1 on identification; 42.1% higher F1 on classification than InstructGPT
Multimodal Event Extraction	Single data modality, supervised, closed domain, surface and entity-centric representation	Multimodal (text, image, video), zero-shot cross-media transfer, open-domain, event-centric representation	Cannot do
Temporal Ordering	Pairwise classification	Global temporal consistency, reduce exposure bias	5% higher accuracy than T5
Event Organization	Isolated primitive events	Hierarchically, logically, temporally well-organized event structures	Cannot do
Event Prediction	None	Schema-guided multimedia event prediction	Cannot do

# Challenges and Our Solutions

- Challenge 1: Open-domain Information Extraction techniques lack of global prior knowledge and cannot perform prediction
  - Event Schema Induction for better Information Extraction, Event Prediction and Situation Forecasting
- Challenge 2: Existing Vision+ Language neural models fail to understand complex situations
  - Explicitly probing large V+L neural models to force them to pay attention to event and action related knowledge
- Challenge 3: Conflicting or complementary claims may come from a wide range of multilingual multimedia sources
  - Cross-lingual Cross-media Claim Verification and Factual Error Correction

# Case Study 2: Patching Video-Language Foundation Models with Action Knowledge



# Visual Activity Examples Where SOTA Vision-Language Models Failed

- Generated caption: “A soldier feeding a baby to a dog” [Kamath et al., 2022]





# Visual Activity Examples Where SOTA Vision-Language Models Failed

## Visual Genome Relation

Assessing relational understanding (23,937 test cases)



- ✓ the horse is eating the grass
- ✗ the grass is eating the horse

## Visual Genome Attribution

Assessing attributive understanding (28,748 test cases)



- ✓ the paved road and the white house
- ✗ the white road and the paved house

<https://arxiv.org/abs/2210.01936>



BLIP

the grass is eating the horse 81%

the horse is eating the grass 78%

Recent VLMs face challenges in understanding visual language concepts beyond object nouns (e.g., recognizing attributes, relations, states)

# SOTA Foundation Models are Clueless about Certain Types of Knowledge

- ▶ Example: Design Knowledge-Specific Probing Tasks for Knowledge Patching
- ▶ Design Action Dynamics Benchmark (ActionBench) based on two VL datasets: SSv2, Ego4d
  - Probing tasks: Action Antonym (AA), Video Reversal (VR)
  - Baseline task: Object Replacement (OR)

## Probing Task: Action Antonym (AA)



"Book **falling** like a rock"  
*Original Action Text*



GT VidLM Result  
23.2%

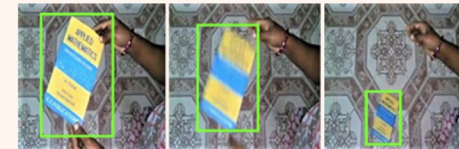
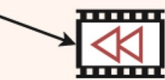
"Book **rising** like a rock"  
*Action Antonym Text*



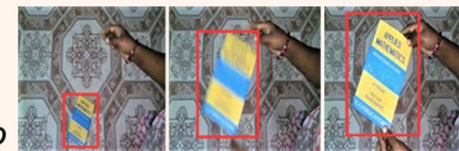
76.8%

## Probing Task: Video Reversal (VR)

"Book **falling** like a rock"  
*Original Action Text*



GT VidLM Result  
49.9%



50.1%

## Baseline Task: Object Replacement (OR)



"**Book** falling like a rock"  
*Original Action Text*



GT VidLM Result  
77.9%

"**Cellphone** falling like a rock"  
*Object Replaced Text*

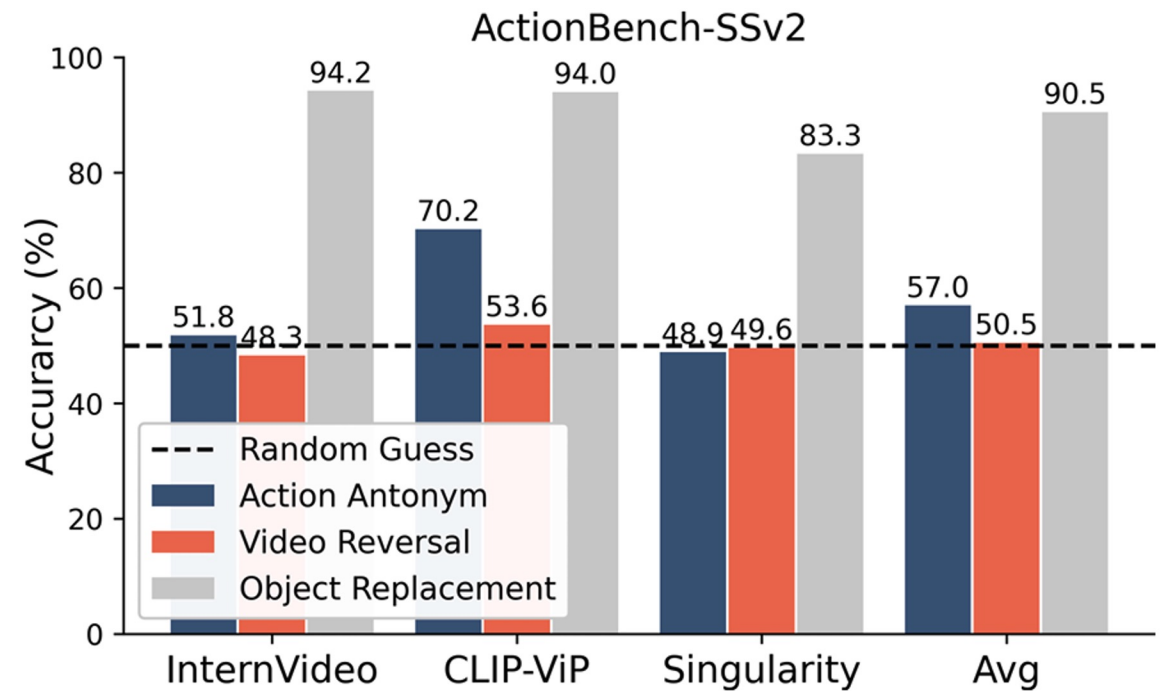
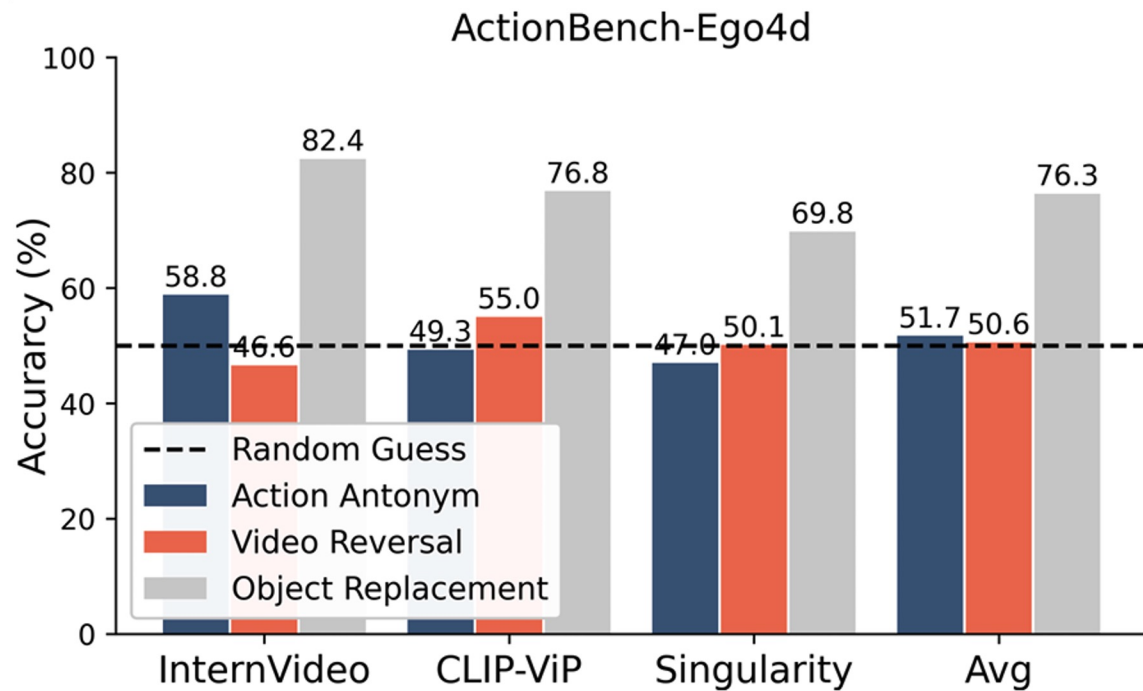


22.1%

# Do SOTA VLMs really understand actions?

Near random performance on **Action Antonym (AA)** and **Video Reversal (VR)**

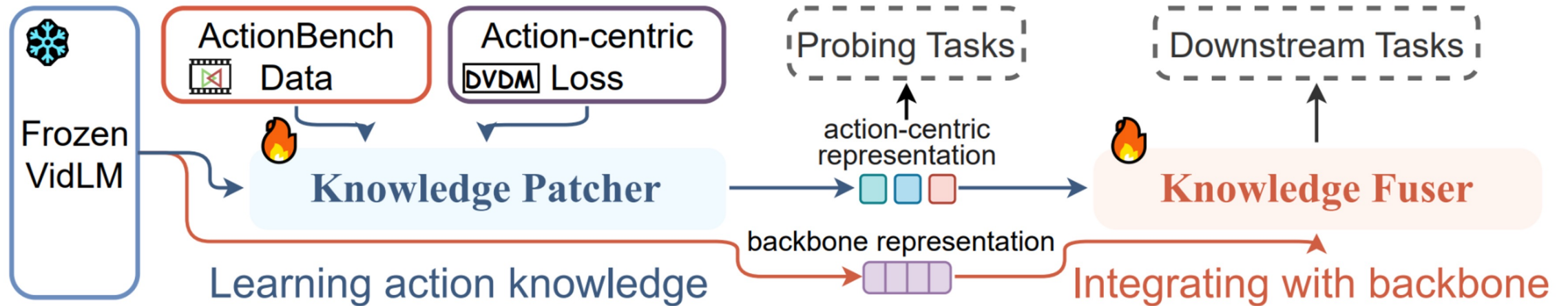
Clear **biases towards object nouns** compared to actions



# PAXION Framework Overview [Wang et al., NeurIPS2023 Spotlight]

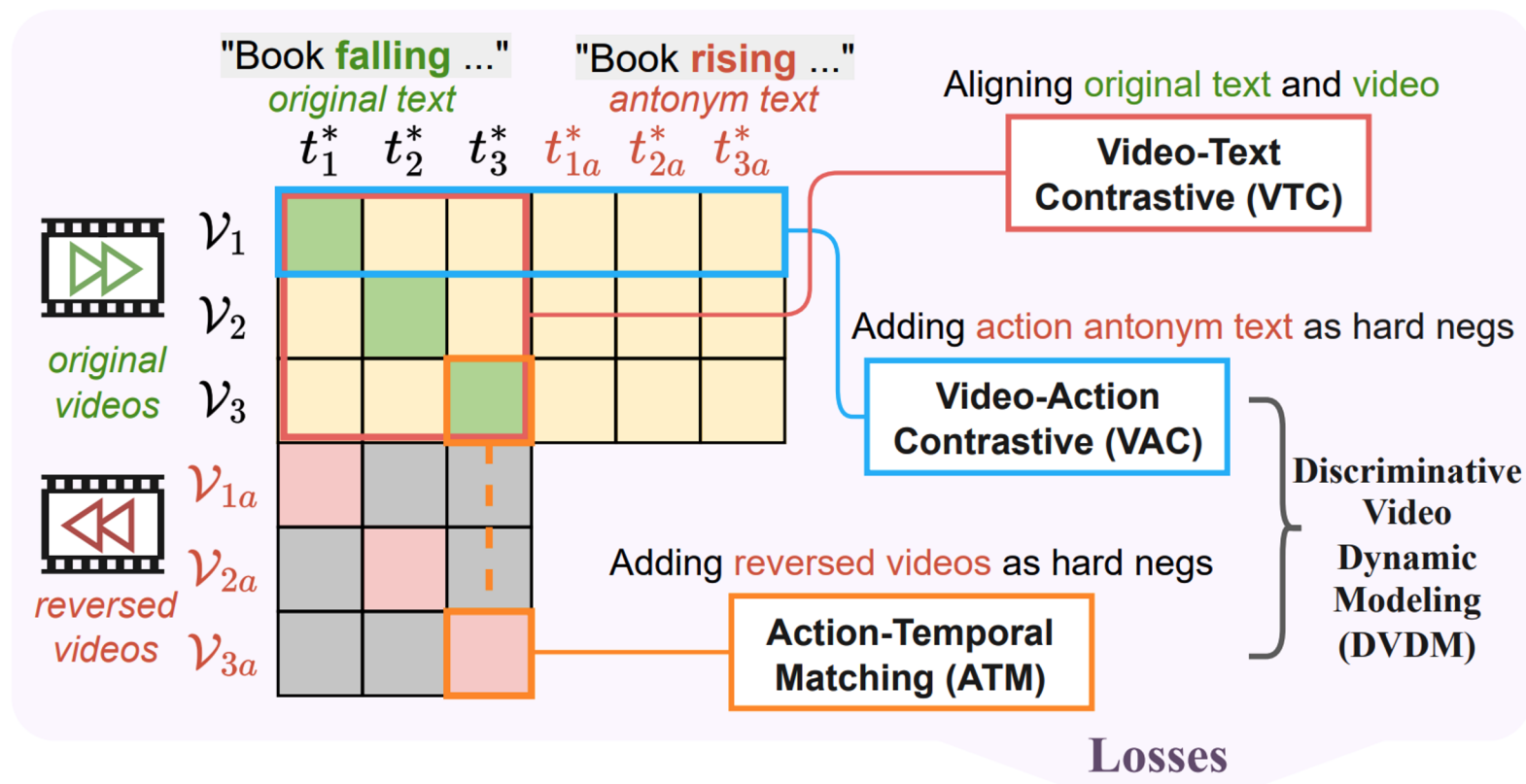
## ► Patch → Fuse

How can we **patch action knowledge** into existing VidLMs **without compromising their general VL capabilities?**



# PAXION

# Knowledge Patcher: Patching frozen VLMs with Action Knowledge



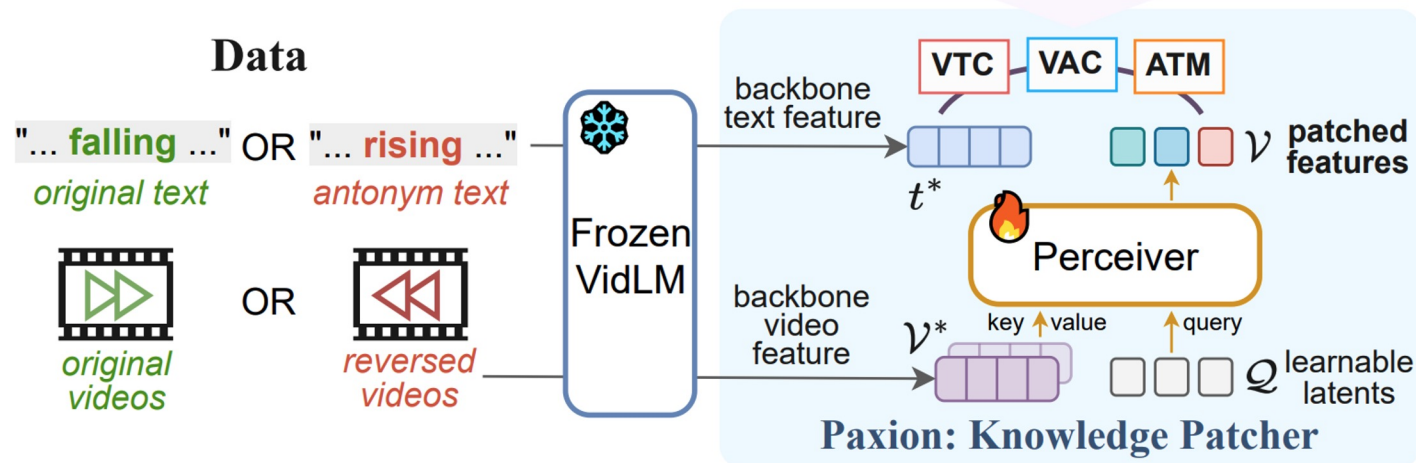
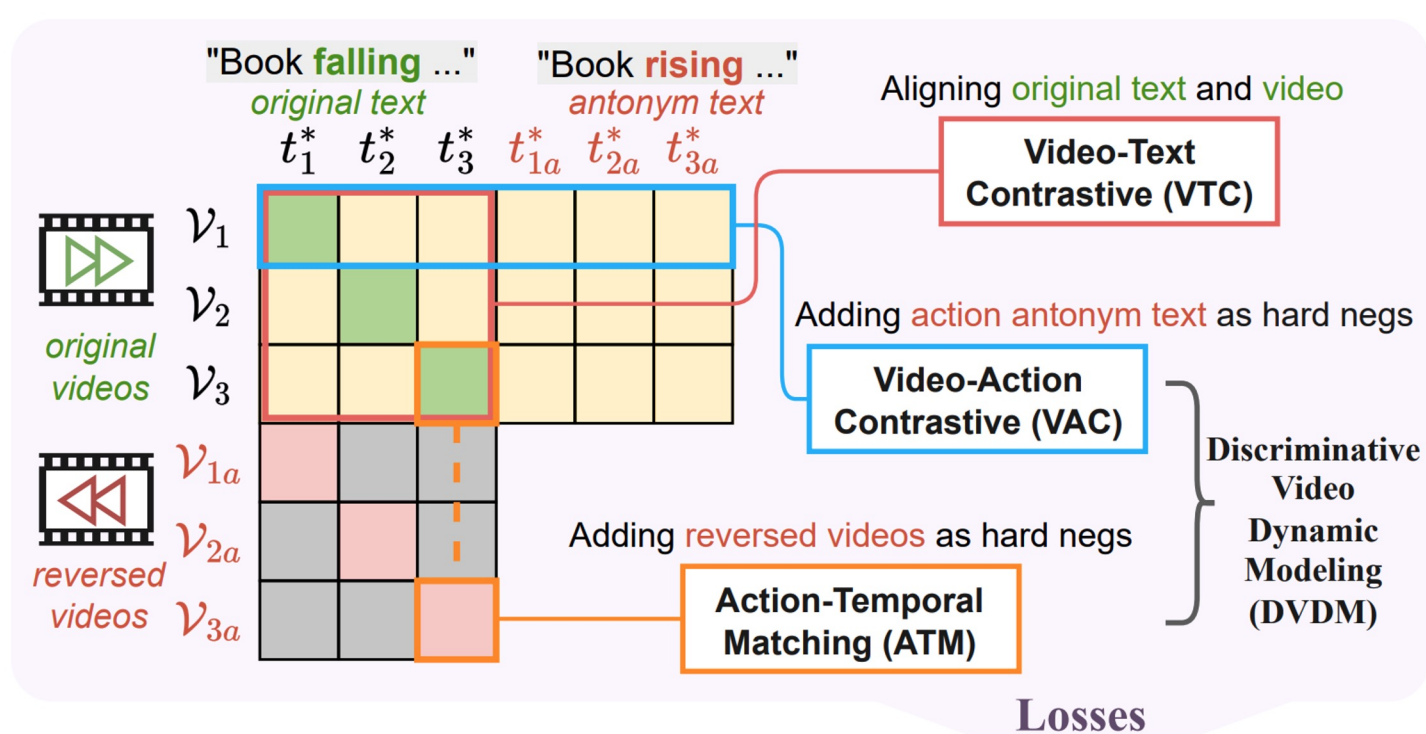
**New training objects DVDM** (VAC, ATM losses) to force the model to encode action dynamics

**Video-Action Contrastive (VAC):** encourages learning the **alignment** between the **video** and the **action verbs**

**Action-Temporal Matching:** encourages learning the correct **temporal ordering** implied by the **action text**

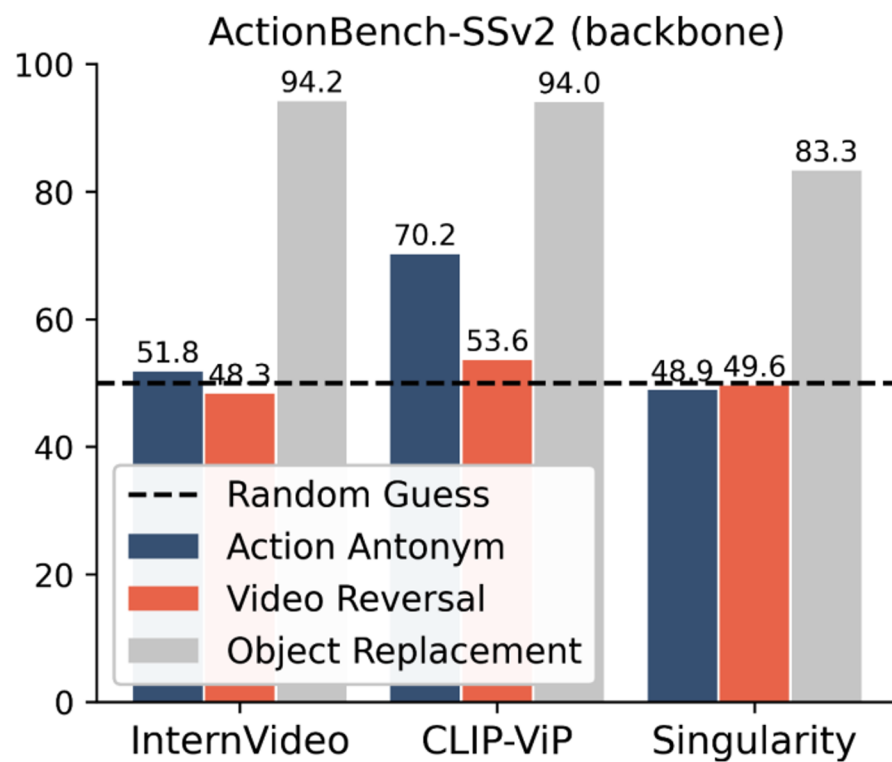


# Knowledge Patcher: Patching frozen VLMs with Action Knowledge

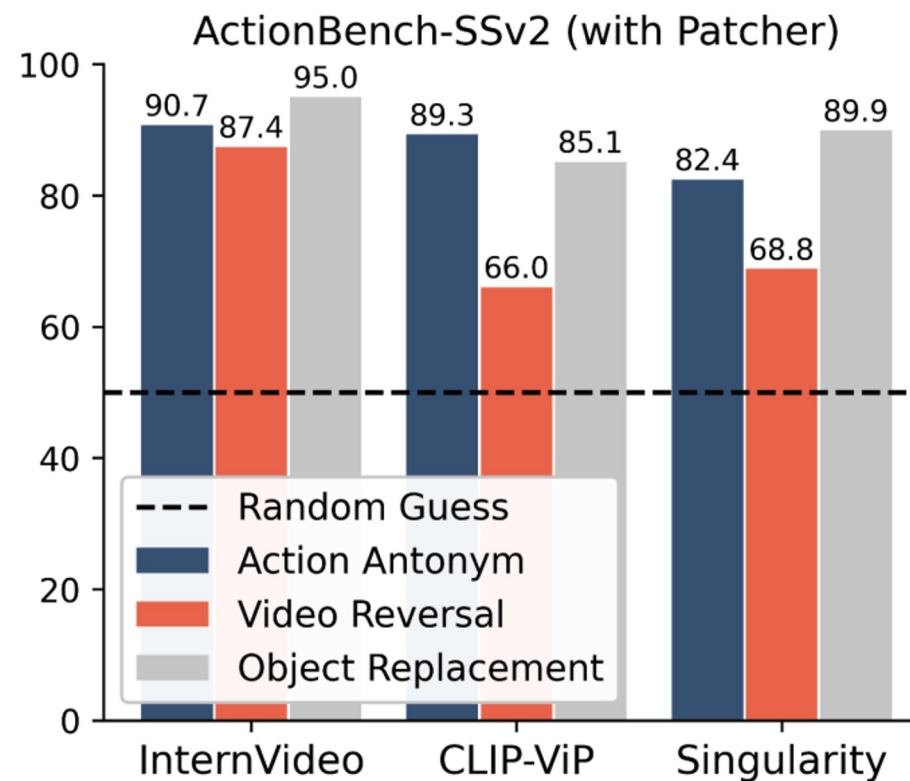


A light-weight **perceiver-based module** attached to a frozen VidLM

# Knowledge Patcher: Patching frozen VLMs with Action Knowledge

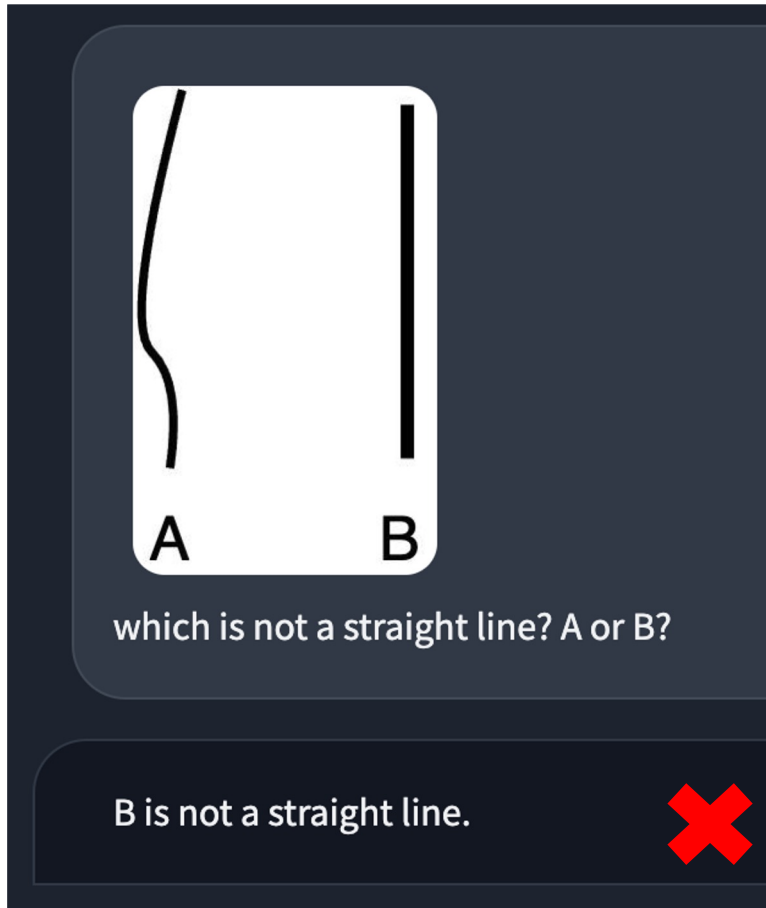


**Before** knowledge patching

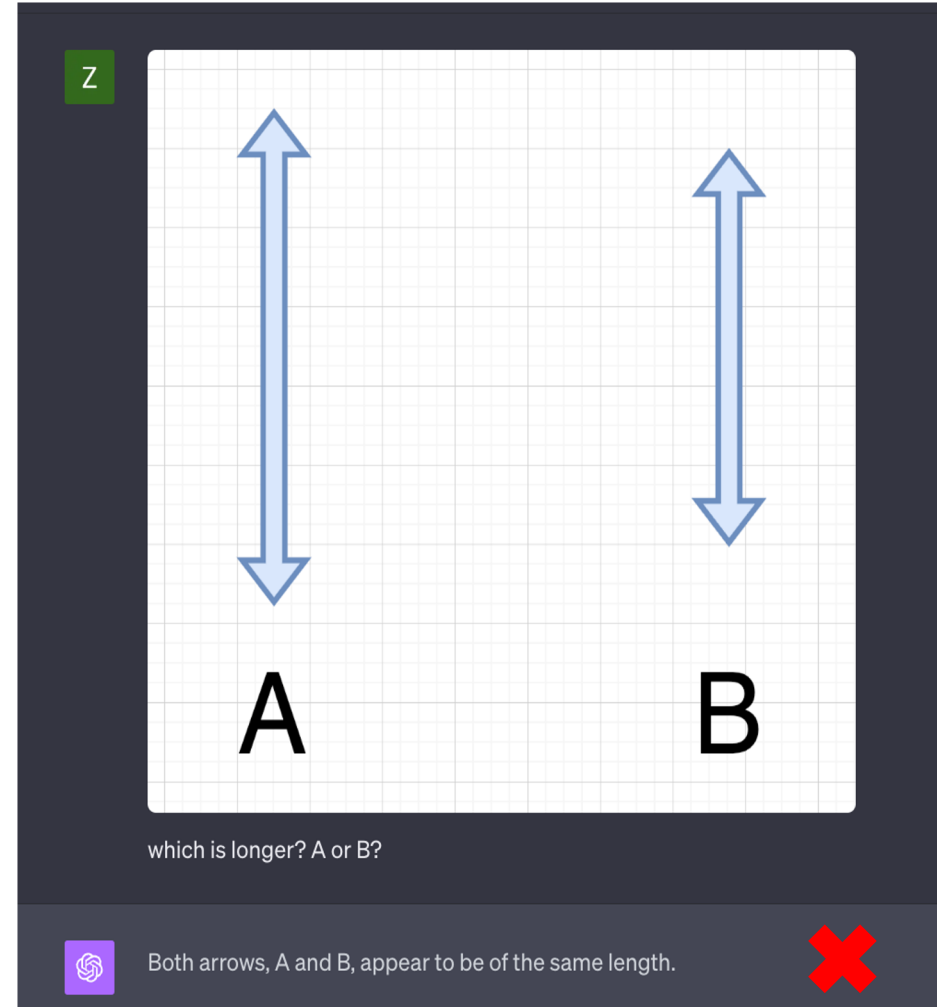


**After** knowledge patching

# Similar Problem on Geometry perception



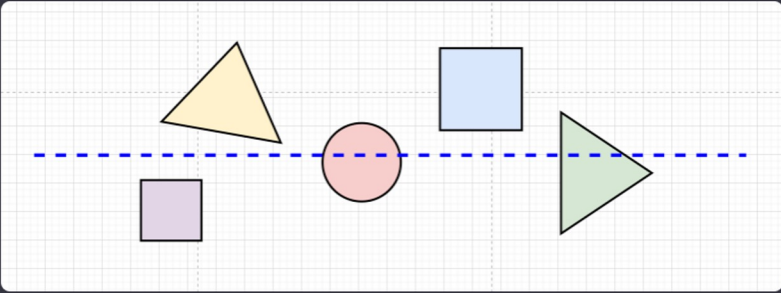
Llava v1.5



GPT4-V

# Similar Problem on Geometry perception

Z



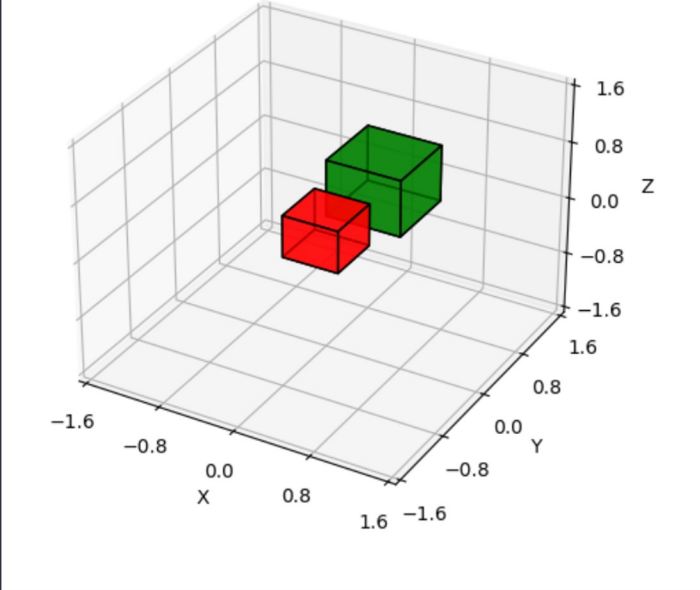
what shapes are intersecting with the blue dashed line?

The shapes intersecting with the blue dashed line are:

- Triangle (beige color)
- Square (purple color)
- Circle (pink color)
- Square (blue color)
- Triangle (green color)

**X**

GPT4-V



Is the green block located at a higher Y axis value than the red block?

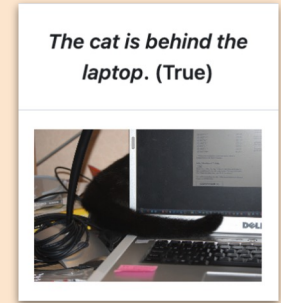
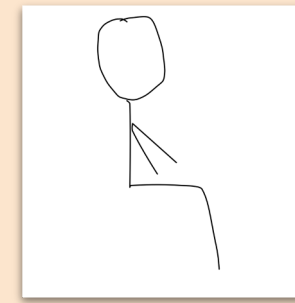
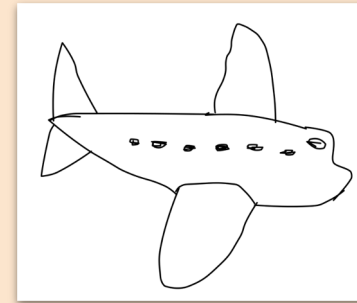
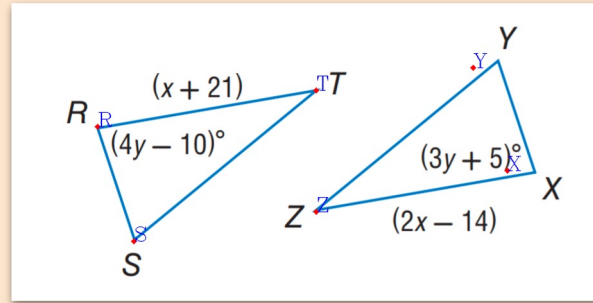
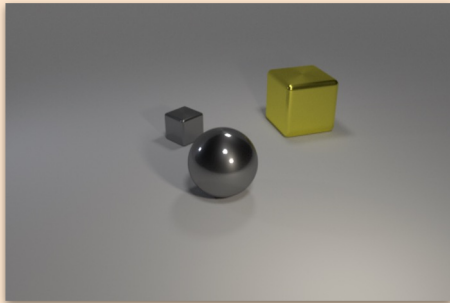
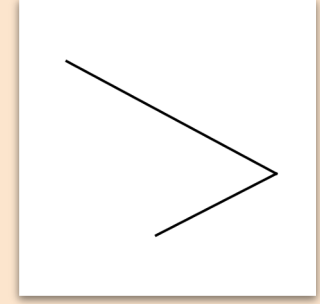
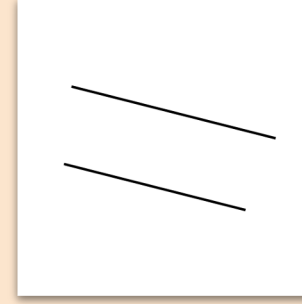
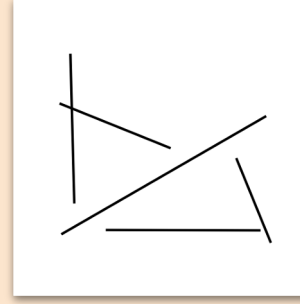
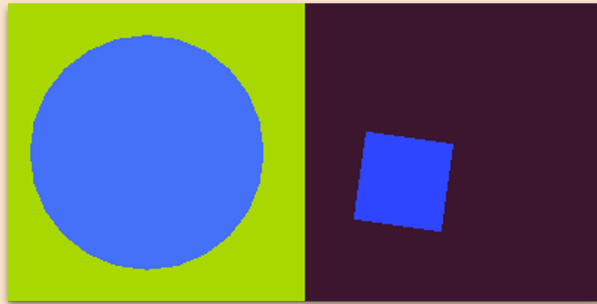
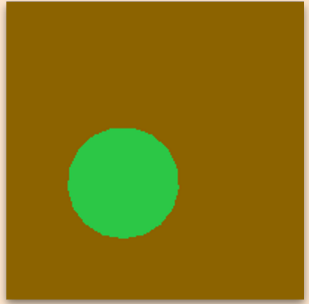
**ChatGPT**

No, based on the orientation of the axes in the image, the green block is not located at a higher Y-axis value than the red block. Both blocks appear to have the same Y-axis value, meaning they are aligned front to back. Instead, the green block is located at a higher Z-axis value, meaning it is above the red block.

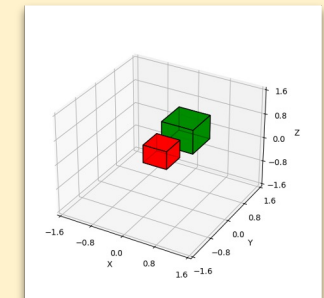
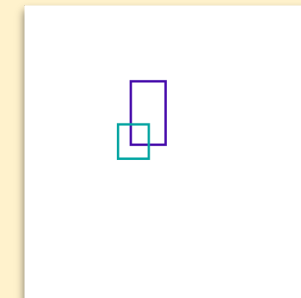
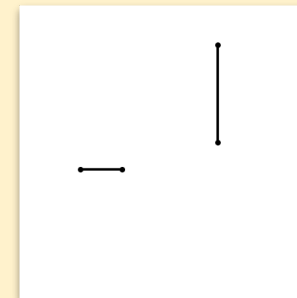
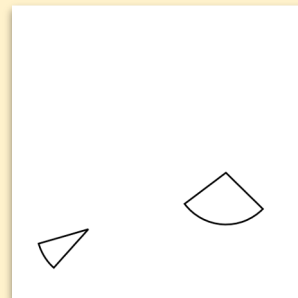
GPT4-V

# Create a new benchmark to force the model to explicitly learn geometry

**Geo-Mix:** a new geometric-centric instruction-tuning mixture containing seven tasks; we repurpose and generate perception focused QA pairs based on six existing datasets



**Geo-Eval:** Four newly proposed tasks aiming at evaluating very basic geometric perception ability, with unseen shapes, styles or question types not overlapping with Geo-Mix.





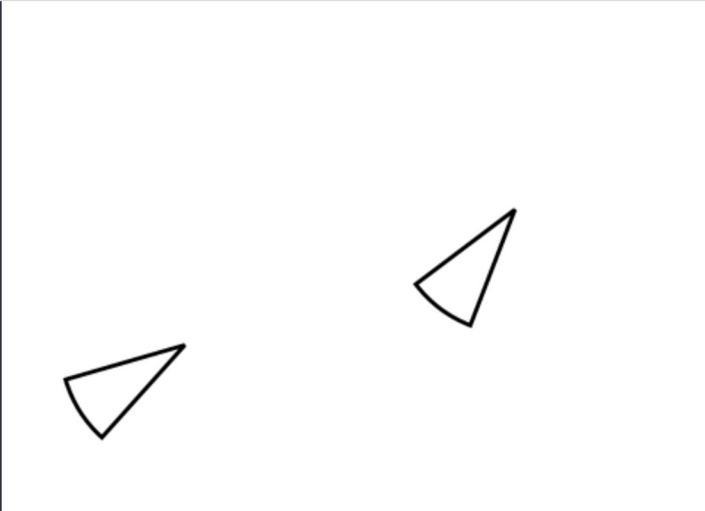
# Geometric Perception Results

- To prevent catastrophic forgetting, we include a small subset of original llava training data (llava-mix) as replay data


Models	Finetuning Mixture	Unseen Evaluation Tasks			
		Angles	Lines	Intersect	Position3D
Random Guess		0.5	0.5	0.5	0.5
Llava-v1.5-7b	-	0.51	0.608	0.518	0.523
Llava-v1.5-7b	geo-mix + llava-mix 5%	0.786	0.613	0.564	0.608
<b>LlavaGeo-7b</b>	geo-mix + llava-mix 5%	<b>0.81</b>	<b>0.644</b>	<b>0.643</b>	<b>0.684</b>


Ablation

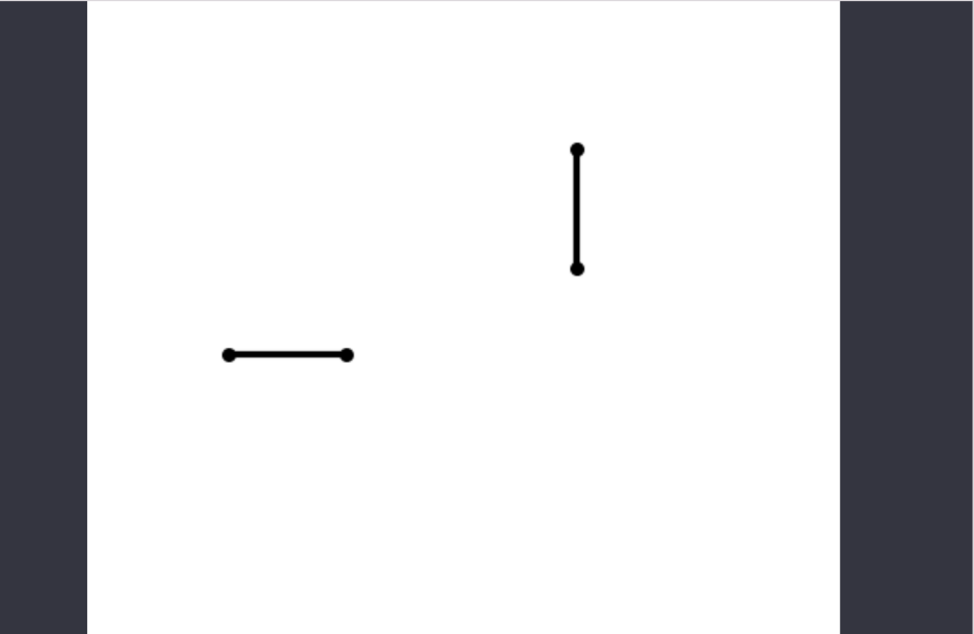
# Qualitative Examples (compare with Latest GPT4-V)




Does the two central angles have the same measure? Choose from "Yes" or "No".


 **ChatGPT**  
No, the two central angles do not have the same measure. The central angle on the left appears larger than the one on the right. Central angles are measured by the arc that they intercept on the circumference of the circle, and the arc corresponding to the left angle looks longer than the one on the right, indicating a larger angle measure. ❌

**LLavaGeo: Yes.** 



Are the length of the two line segments the same? Choose from "Yes" or "No".

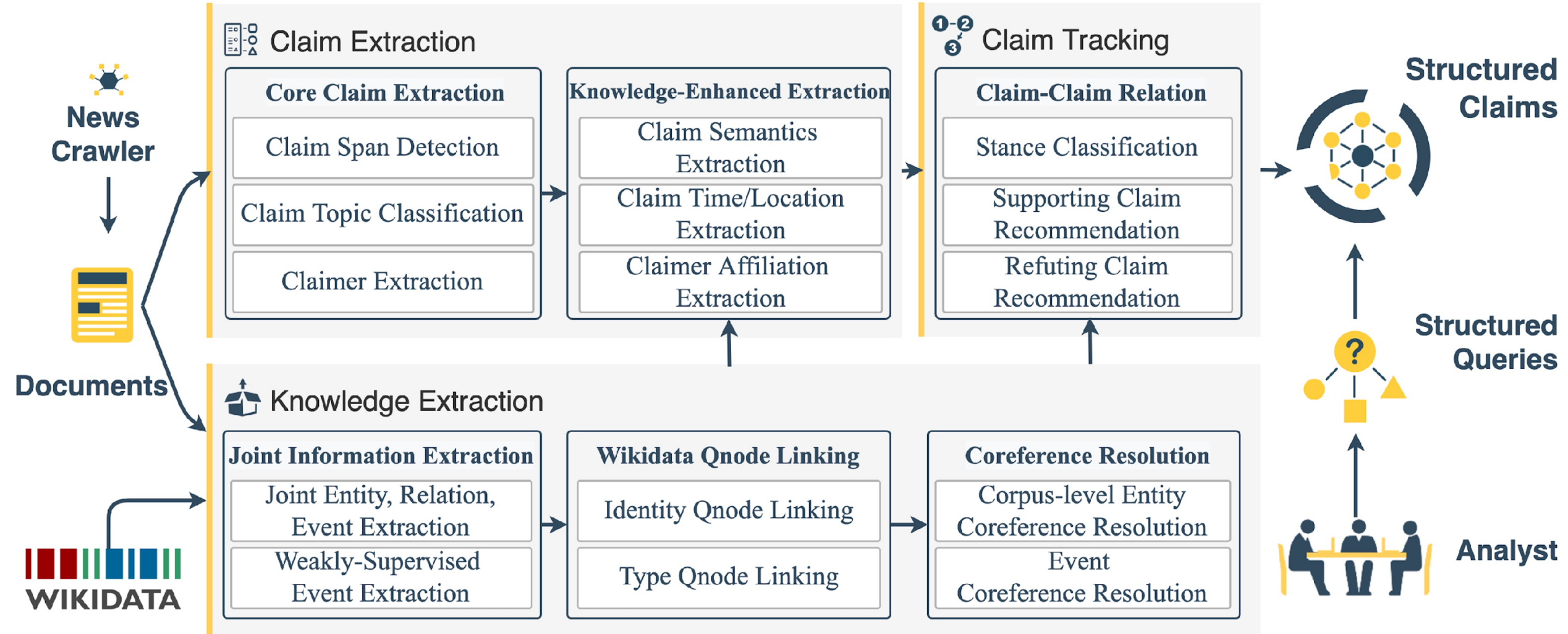
 **ChatGPT**  
No, the lengths of the two line segments are not the same. ❌

**LLavaGeo: Yes.** 

# Challenges and Our Solutions

- Challenge 1: Open-domain Information Extraction techniques lack of global prior knowledge and cannot perform prediction
  - Event Schema Induction for better Information Extraction, Event Prediction and Situation Forecasting
- Challenge 2: Existing Vision+ Language neural models fail to understand complex situations
  - Explicitly probing large V+L neural models to force them to pay attention to event and action related knowledge
- Challenge 3: Conflicting or complementary claims may come from a wide range of multilingual multimedia sources
  - Cross-lingual Cross-media Claim Verification and Factual Error Correction

# Zero-shot QA-based Claim Detection Framework [Reddy et al., COLING2022; AAI2023]




# Cross-media Information Consistency Checking [Fung et al., ACL2021]

- ❖ Combine *local* and *global* features
- ❖ Leverage external knowledge to help pinpoint misinformation

bbc.com

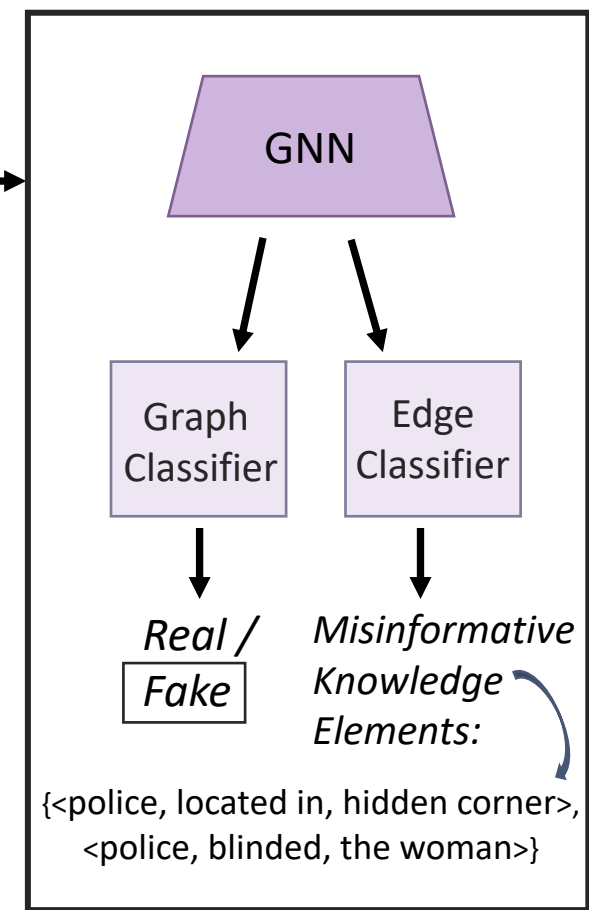
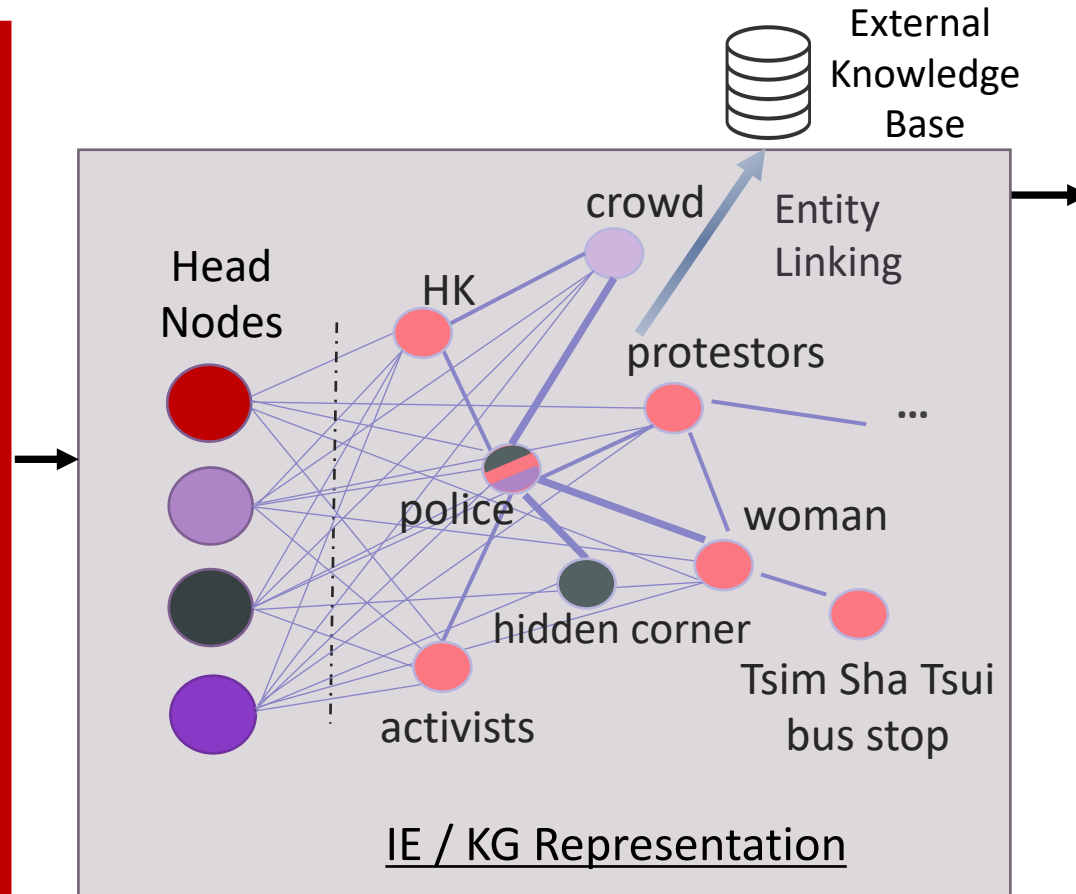
## Police Brutality in HK at new Extreme Levels

Aug 11, 2019 Lisa Lu



Police brutality has risen to a new, extreme level in HK this past weekend. HK police started shooting at protestors on the streets, including the unarmed, peaceful protestors. One notable incidence involved a woman at the Tsim Sha Tsui bus stop being shot in the eye by a policeman hiding behind corners. No warning was issued beforehand, and the woman was permanently blinded. Local activists are avidly calling for international attention on the HK police brutality.

HK police shoot cold bullets at protestors from hidden corners.

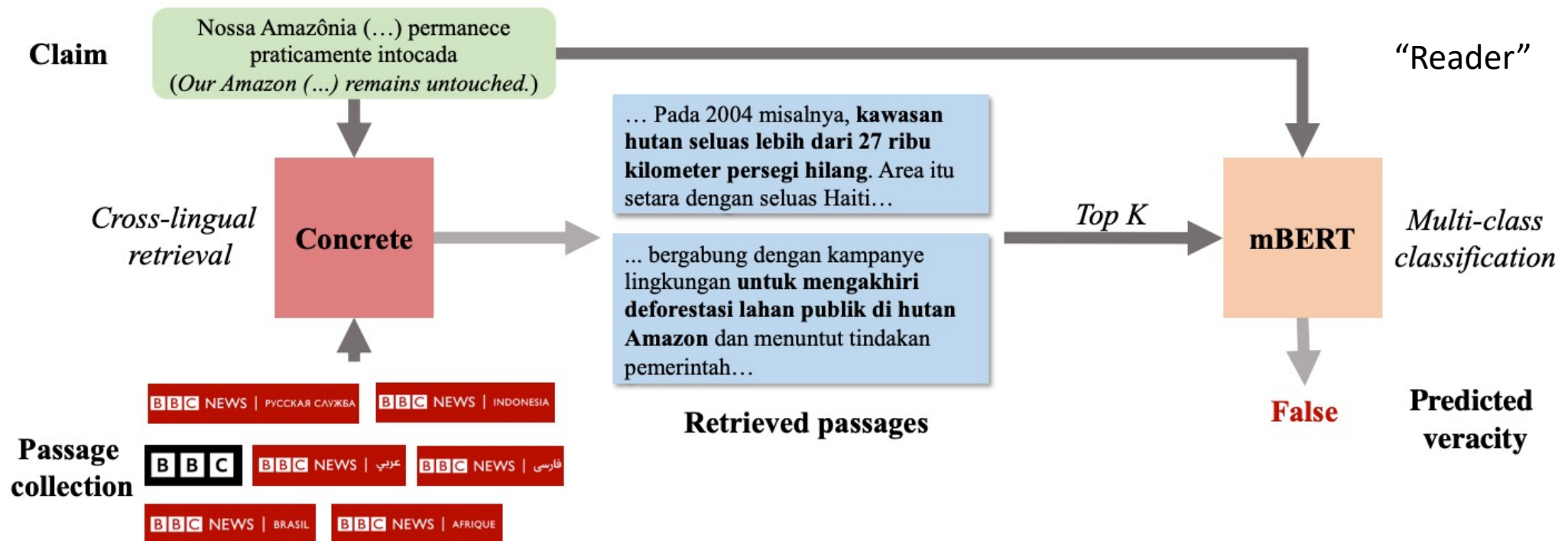


**InfoSurgeon**



# Cross-lingual Fact-checking [Huang et al., COLING2022]

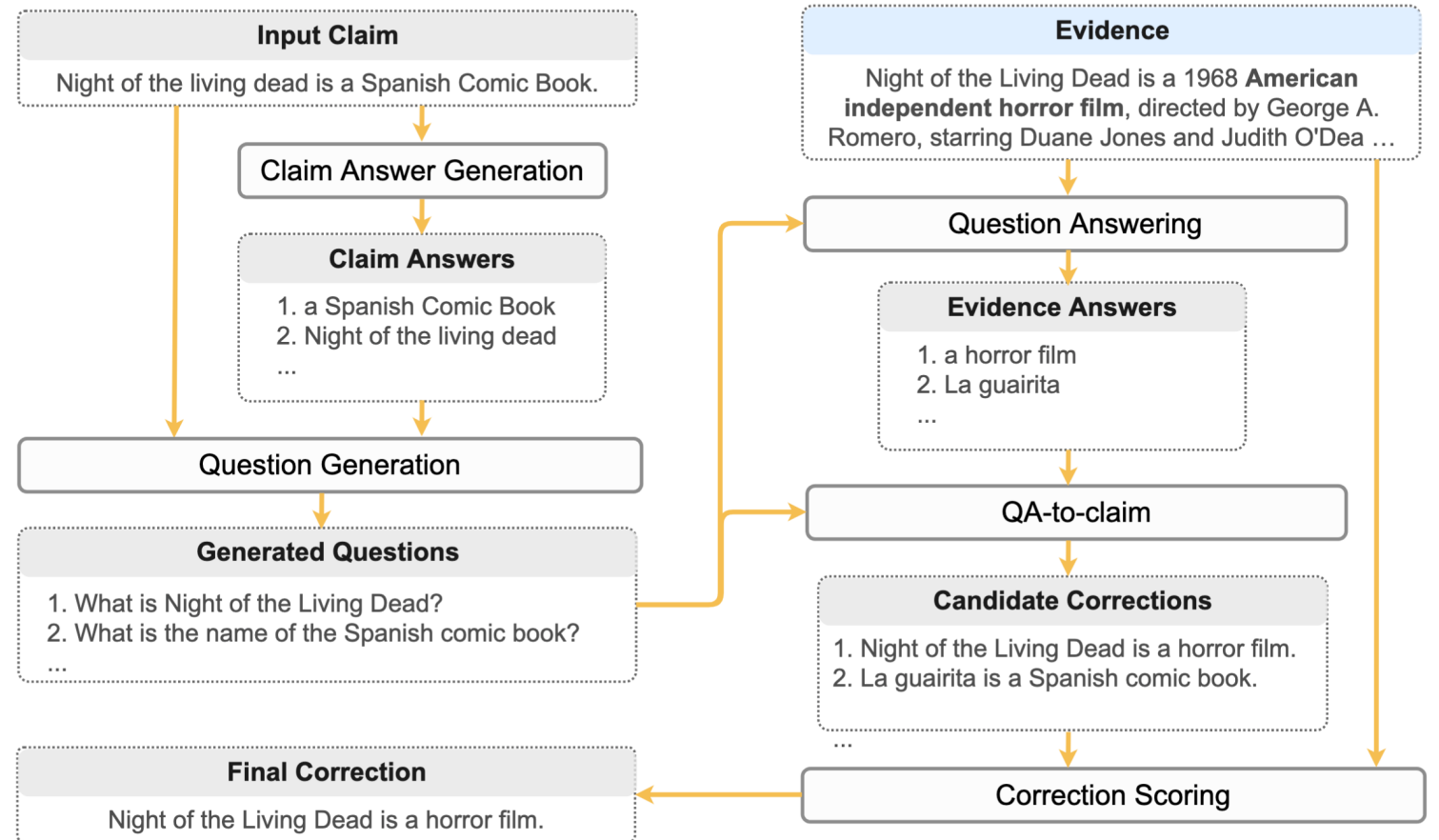
- We constructed a trustworthy multilingual passage collection by collecting news articles from bbc.com.
- We propose a cross-lingual retriever, CONCRETE, that can retrieve evidence from any languages, extended from a cross-lingual retriever, mDPR [Asai et al., 2021].
- We propose a new Cross-lingual Inverse Cloze Task (X-ICT) to identify the passage corresponding to a translated title; X-ICT warm-start the retriever by predicting the context given a randomly sampled sentence.
- We further fine-tune mDPR on the proposed X-ICT task.



# Zero-shot Faithful Fact Error Correction [Huang et al., 2023 submission]

Motivated by how humans correct factual errors, we break this task into five sub-tasks:

- Claim answer generation
- Question generation
- Question answering
- QA-to-claim
- Correction scoring

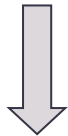


# Automatic Citation Generation: Grounded Abstractive Summarization with News Claim Linking

- In-context learning to generate query-focused abstractive summaries + human expert analyst editing
- Citations are directly added into the generated summary using the instruction of the form:  
*“summarize the above, regarding <strategic question X>, with citations”*

**Question:** Who are the likely suspects in the Nord Stream pipeline sabotage?

**Summary:** The most likely suspects in the Nord Stream pipeline sabotage are either the U.S. or a foreign power.



**Summary:** The most likely suspects in the Nord Stream pipeline sabotage are either the U.S. [2] or a foreign power [5].

Claims
1) Experts and also Russia, which built the network of gas pipelines, said the possibility of sabotage could not be ruled out
2) Russian-state-run media has hinted the U.S. is behind this.
3) European and NATO officials are blaming sabotage for three leaks in the Nord Stream 1 and 2 undersea pipelines running from Russia to Germany under the Baltic Sea.
4) Russian energy giant Gazprom, a state entity, owns a controlling interest in both pipelines.
5) Sweden's security service has opened a "gross sabotage" investigation regarding the incident at the Nord Stream pipelines, the agency said in a statement Wednesday, adding that it cannot be ruled out "that a foreign power is behind it."

# Challenges and Our Solutions

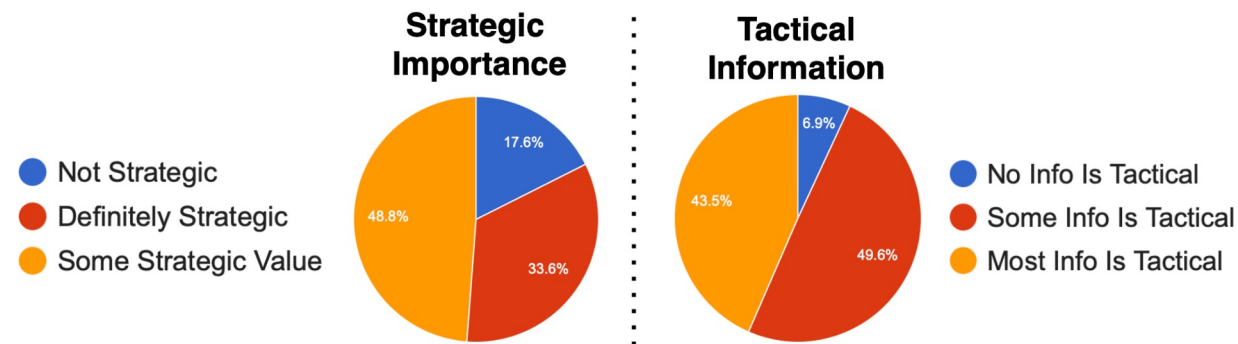
- Challenge 1: Open-domain Information Extraction techniques lack of global prior knowledge and cannot perform prediction
  - Event Schema Induction for better Information Extraction, Event Prediction and Situation Forecasting
- Challenge 2: Conflicting or complementary claims may come from a wide range of multilingual multimedia sources
  - Cross-lingual Cross-media Claim Verification and Factual Error Correction
- Challenge 3: Existing large language models fail to generate meaningful long texts
  - A novel lambda-shaped attention mask to generate situation report of infinite length

# Experiments: **Strategic Importance of Questions**

We performed a human evaluation of the questions within SmartBook by measuring for the following aspects:

- **Strategic Importance:** Introspection on whether the question provided within SmartBook would have been asked as a strategic question by a senior analyst. One of three categories, namely, *Not Strategic*, *Some Strategic Value*, and *Definitely Strategic*.
- **Relevant Tactical Information:** Question is evaluated based on its corresponding SmartBook summary, as a measure of how much relevant tactical information useful to an analyst can be gathered using the question. Categorized into *No information is tactical*, *Some information is tactical*, and *Most information is tactical*.

Results from evaluation of 125 randomly selected questions by 3 annotators shows that most questions are strategic, with at least **82%** of the questions having some strategic value. These questions can help gather relevant tactical information in roughly **93%** of the cases.



Results from the human evaluation of the questions (section headings) in SmartBook.



# Experiments: Quality of Generated Section Content

- We evaluated our claim-drive summarization approach in SmartBook against two alternative approaches:
  - **Query-focused Summarization**: Directly pass the news articles as input to an LLM (GPT-3.5) to generate the summary, as opposed to using question-driven claim extraction and validation for identifying relevant content.
  - **Web Search + LLM**<sup>1</sup>: Analogous to Bing Search + ChatGPT by using the web as the background corpora for obtaining relevant web pages and summarizing them using an LLM (GPT-3.5).
- Three annotators are asked to assess the summaries on a scale of 1 to 5 for each of the following aspects:
  - Coherence: measures the quality of all sentences collectively regardless of the query
  - Relevance: quantifies whether the summary contains key points that answer the given question.
  - Strategicness: indication of whether the summary provides non-trivial and insightful information for analysts
- Results show that:
  - Our question-driven claim extraction step (3 vs 1) leads to considerably more relevant summaries.
  - Directly obtaining information from the web (2 vs 3) can give less strategic content compared to using focused news articles.

<i>Summary-Type</i>	<i>Background</i>	<i>Coherence</i>	<i>Relevance</i>	<i>Strategicness</i>
(1) Query-Focused	News	4.39	3.85	3.77
(2) Web Search + LLM	Web	4.44	4.07	3.83
(3) SmartBook	News	<b>4.56</b>	<b>4.23</b>	<b>4.17</b>

Results of human evaluation comparing the summaries within SmartBook against direct query-focused summarization and a *Bing Search+ChatGPT* analogous system. Results are on a 1-5 scale (worst to best).

<sup>1</sup>For this, we used [perplexity.ai](https://perplexity.ai), which combines web-search with LLM for summarization.

# Experiments: Quality of Generated Section Content

**Question:** How have kamikaze drones changed the nature of the Russia-Ukraine conflict?

Query-Focused	Web Search + LLM	SmartBook
<p>Kamikaze drones, or suicide drones, have been used by Russia in its attacks on Ukrainian cities such as Kyiv, Vinnytsia and Odesa. They are known as a “loitering munition” because they are capable of waiting in an area identified as a potential target and only strike once an enemy asset is identified. Russia is using Iranian-made Shahed-136 drones, which have a payload of approximately 50 kilograms (110 pounds) and are hard to detect due to their small size. Ukraine has responded by developing an app that civilians can use to report sightings of incoming drones and missiles, and increasing the proportion shot down before they hit the ground.</p>	<p>Russia and Ukraine have been using Iranian-made Shahed-136 drones in the conflict. These drones are known as ‘kamikaze’ drones due to their ability to self-destruct on impact. The use of these drones has been highlighted by Russia’s growing reliance on them and their use as both military weapons and psychological weapons. On October 13, 2022, kamikaze drones and missiles hit north and south Ukraine, though the Kremlin denied any involvement from President Putin in talks between Russia and the West to end the war in Ukraine.</p>	<p>The Russia-Ukraine conflict has seen a paradigm shift in the nature of Russian tactics, with the increased reliance of kamikaze drones over artillery fire. These drones, which are designed to crash into a target and explode, can spray potentially deadly debris and are being used to plug gaps in the deployment of Russian forces along the front line. Russia is reportedly using Iranian-made Shahed series drones, which are showcased at Kashan Airfield south of Tehran. This shift in tactics has seen an escalation in the aerial bombardment of densely populated urban areas and an attack on critical infrastructure and power sources.</p>

Web Search + LLM output contains hallucinated information suggesting both Russia and Ukraine have been using the ‘kamikaze’ drones.

SmartBook summary contains considerably more relevant information

An example showing the summaries generated by different approaches. Question-relevant information in the summaries is highlighted in blue.

# Experiments: Editing by Expert Analyst

- We observe that token overlap between the generated and post-edited summaries is high, with a BLEU score of 59.0% and a Rouge-L score of 74.1%.
- Percentage of tokens deleted is only 2.3% meaning SmartBook can act as a good starting point for analysts to expand upon for the generation of situation reports.

---

**SmartBook Summary:** Russia has reportedly stepped up its use of kamikaze drones in its assault against Ukraine. The increased reliance on kamikaze drones over artillery fire likely signals a paradigm shift in Russian tactics a shift introduced to counter high mobility offensive probing by Ukrainian forces.

---

**Analyst-edited Summary:** Russia has reportedly stepped up its use of kamikaze drones in its assault against Ukraine. **The aircraft are called kamikaze drones because they attack once and don't come back.** The increased reliance on kamikaze drones over artillery fire likely signals a paradigm shift in Russian tactics - a shift introduced to counter high mobility offensive probing by Ukrainian forces. **Their low price means the drones can be deployed in large numbers and they hover before they strike, so they have a psychological effect on civilians as they watch and wait for them to strike. These drones allow Russia to target Ukrainians far away from the front line, away from the primary battle space. The emergence of swarms of drones in Ukraine is part of a shift in the nature of the Russian offensive, which some speculate indicates that Moscow may be running low on long-range missiles.**

---

Analyst adds an interesting insight (in green) at the end based on the information in the summary

Human analyst mainly tends to add additional tactical information (in blue) in order to elaborate on certain aspects (e.g. what is special about the 'kamikaze' type of drone)

Example showing intelligence analyst edits (in color) for an automatic machine generated SmartBook summary on the use of 'kamikaze' drones in the Ukraine-Russia crisis.

# Conclusions & What We Need

- Our overall aim is to combat both the stultification of NLP as a mere evaluation optimization endeavor and to dispel fears that LLMs and generative AI will shut down the field
- Embrace the new challenges brought by LLMs and improve our research standard (e.g., massively multilingual NLP, NLP for science)
- More open-minded to collaborate with researchers from other fields, including social science, natural science, computer vision, knowledge representation and reasoning, and human-computer interaction
- What Humans are Good at & Machines are still Bad at
  - Information consistency checking with background knowledge and commonsense knowledge
  - Capture other aspects of human fake news generation, such as novelty and elaboration, engaging linguistic style, consistency with the remaining content
- What We Need
  - Deeper thinking about the foundational conceptual models driving our field
  - Global infrastructures to dramatically scale up computing resources to develop academia owned real open-source LLMs



# Publicly Available Demos, Systems and Resources

- SmartBook for Natural Disaster Monitoring: <https://blenderdemo.com/smartbook-turkey-syria-earthquake>
- SmartBook for Ukraine War: <https://blenderdemo.com/smartbook>
- SmartBook for DoS secretary's visit to East Asia: <https://blenderdemo.com/smartbook-july>
- ClaimRadar for COVID19: <https://blenderdemo.com/covid-list>
- ClaimRadar for Ukraine Crisis (Old data): <https://blenderdemo.com/ukraine-list>
- ClaimRadar for Ukraine Crisis (New data): <https://blenderdemo.com/ukraine-list-new>
- Real Time Claim Extraction for COVID19: <https://blenderdemo.com/covid-extract>
- Real Time Claim Extraction for Ukraine Crisis: <https://blenderdemo.com/ukraine-extract>
- Stance Controlled Generation: [https://incas.csl.illinois.edu/blender/INCAS Framing \(LM Stance Re-switching\)](https://incas.csl.illinois.edu/blender/INCAS_Framing_(LM_Stance_Re-switching))
- Stance Detection: [https://incas.csl.illinois.edu/blender/INCAS Political Stance Classification](https://incas.csl.illinois.edu/blender/INCAS_Political_Stance_Classification)
- InfoSurgeon: <https://github.com/yrf1/InfoSurgeon>
- ClaimRadar source code: <https://github.com/uiucnlp/covid-claim-radar>
- ClaimRadar docker: <https://hub.docker.com/repository/docker/blendernlp/covid-claim-radar>
- ClaimRadar demo video: [http://blender.cs.illinois.edu/aida/covid\\_claim\\_radar.mp4](http://blender.cs.illinois.edu/aida/covid_claim_radar.mp4)
- KAIROS RESIN Cross-document Cross-lingual Cross-media Information Extraction system: <https://github.com/RESIN-KAIROS/RESIN-pipeline-public>
- Joint Neural Information Extraction system (Lin et al., ACL2020)
  - <http://blender.cs.illinois.edu/software/oneie/>
- GAIA Multimedia Event Extraction system (Li et al., ACL2020 best demo award)
  - GitHub: [https://github.com/GAIA-AIDA/uiuc\\_ie\\_pipeline\\_fine\\_grained](https://github.com/GAIA-AIDA/uiuc_ie_pipeline_fine_grained)
  - Text IE DockerHub: <https://hub.docker.com/orgs/blendernlp/>
  - Visual IE repositories: <https://hub.docker.com/u/dannapierskitopta>



# Representative Publications

- [1] Revanth Gangi Reddy, Yi R. Fung, Qi Zeng, Manling Li, Ziqi Wang, Paul Sullivan, Clare R. Voss and Heng Ji. 2023. SmartBook: AI Assisted Situation Report Generation. *Arxiv*.
- [2] Sha Li, Ruining Zhao, Manling Li, Heng Ji, Chris Callison-Burch and Jiawei Han. 2023. Open-Domain Hierarchical Event Schema Induction by Incremental Prompting and Verification. *Proc. The 61st Annual Meeting of the Association for Computational Linguistics (ACL2023)*.
- [3] Chi Han, Qifan Wang, Wenhan Xiong, Yu Chen, Heng Ji, Sinong Wang. 2023. LM-Infinite: Simple On-the-Fly Length Generalization for Large Language Models. *Arxiv*.
- [4] Sha Li, Qiusi Zhan, Kathryn Conger, Martha Palmer, Heng Ji and Jiawei Han. 2023. GLEN: General-Purpose Event Detection for Thousands of Types. *Proc. The 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP2023)*
- [5] Revanth Gangi Reddy, Sai Chinthakindi, Zhenhailong Wang, Yi R. Fung, Kathryn Summerville Conger, Ahmed S. Elsayed, Martha Palmer, Preslav Nakov, Eduard Hovy, Kevin Small and Heng Ji. 2022. NewsClaims: A New Benchmark for Claim Detection from News with Attribute Knowledge. *Proc. The 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP2022)*.
- [6] Revanth Gangi Reddy, Sai Chetan Chinthakindi, Yi Fung, Kevin Small and Heng Ji. 2022. A Zero-Shot Claim Detection Framework using Question Answering. *Proc. The 29th International Conference on Computational Linguistics (COLING2022)*.
- [7] Manling Li, Revanth Gangi Reddy, Ziqi Wang, Yi-Shyuan Chiang, Tuan M. Lai, Pengfei Yu, Zixuan Zhang and Heng Ji. 2022. COVID-19 Claim Radar: A Structured Claim Extraction and Tracking System. *Proc. The 60th Annual Meeting of the Association for Computational Linguistics (ACL2022) Demo Track*.
- [8] Revanth Gangi Reddy, Heba Elfardy, Hou Pong Chan, Kevin Small and Heng Ji. 2023. SumREN: Summarizing Reported Speech about Events in News. *Proc. Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI2023)*.
- [9] Kung-Hsiang Huang, Kathleen McKeown, Preslav Nakov, Yejin Choi and Heng Ji. 2022. Faking Fake News for Real Fake News Detection: Propaganda-loaded Training Data Generation. *Proc. The 61st Annual Meeting of the Association for Computational Linguistics (ACL2023)*
- [10] Yi R. Fung, Christopher Thomas, Revanth Gangi Reddy, Sandeep Polisetty, Heng Ji, Shih-Fu Chang, Kathleen McKeown, Mohit Bansal and Avi Sil. 2021. InfoSurgeon: Cross-Media Fine-grained Information Consistency Checking for Fake News Detection. *Proc. The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*.

# Representative Publications (Cont.)

- [11] Kung-Hsiang Huang, ChengXiang Zhai and Heng Ji. 2022. CONCRETE: Improving Cross-lingual Fact Checking with Cross-lingual Retrieval. Proc. The 29th International Conference on Computational Linguistics (COLING2022).
- [12] Xueqing Wu, Kung-Hsiang Huang, Yi Fung and Heng Ji. 2022. Cross-document Misinformation Detection based on Event Graph Reasoning. Proc. The 2022 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT2022).
- [13] Chenkai Sun, Jinning Li, Yi Fung, Hou Pong Chan, Tarek Abdelzaher, ChengXiang Zhai, Heng Ji. 2023. Decoding the Silent Majority: Inducing Belief Augmented Social Graph with Large Language Model for Response Forecasting. *Proc. The 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP2023)*.
- [14] Chenkai Sun, Jinning Li, Hou Pong Chan, ChengXiang Zhai and Heng Ji. 2023. Measuring the Effect of Influential Messages on Varying Personas. Proc. The 61st Annual Meeting of the Association for Computational Linguistics (ACL2023).
- [15] Charles Yu, Sullam Jeoung, Anish Kasi, Pengfei Yu and Heng Ji. 2023. Unlearning Bias in Language Models by Partitioning Gradients. Proc. The 61st Annual Meeting of the Association for Computational Linguistics (ACL2023) Findings.
- [16] Ali Omrani, Alireza Salkhordeh Ziabari, Charles Yu, Preni Golazizian, Brendan Kennedy, Mohammad Atari, Heng Ji and Morteza Dehghani. 2023. Social-Group-Agnostic Bias Mitigation via the Stereotype Content Model. Proc. The 61st Annual Meeting of the Association for Computational Linguistics (ACL2023).
- [17] Chi Han, Jialiang Xu, Manling Li, Yi Fung, Chenkai Sun, Nan Jiang, Tarek Abdelzaher and Heng Ji. 2023. LM-Switch: Lightweight Language Model Conditioning in Word Embedding Space. Arxiv.
- [18] Xingyao Wang, Sha Li and Heng Ji. 2023. Code4Struct: Code Generation for Few-Shot Event Structure Prediction. Proc. The 61st Annual Meeting of the Association for Computational Linguistics (ACL2023).
- [19] Manling Li, Ruochen Xu, Shuohang Wang, Luwei Zhou, Xudong Lin, Chenguang Zhu, Michael Zeng, Heng Ji and Shih-Fu Chang. 2022. CLIP-Event: Connecting Text and Images with Event Structures. Proc. Conference on Computer Vision and Pattern Recognition (CVPR2022).
- [20] Manling Li, Ruochen Xu, Shuohang Wang, Luwei Zhou, Xudong Lin, Chenguang Zhu, Michael Zeng, Heng Ji and Shih-Fu Chang. 2022. CLIP-Event: Connecting Text and Images with Event Structures. Proc. Conference on Computer Vision and Pattern Recognition (CVPR2022).
- [21] Sha Li, Heng Ji and Jiawei Han. 2021. Document-Level Event Argument Extraction by Conditional Generation. Proc. The 2021 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT2021).