



Toward Mitigating Misinformation and Social Media Manipulation in LLM Era

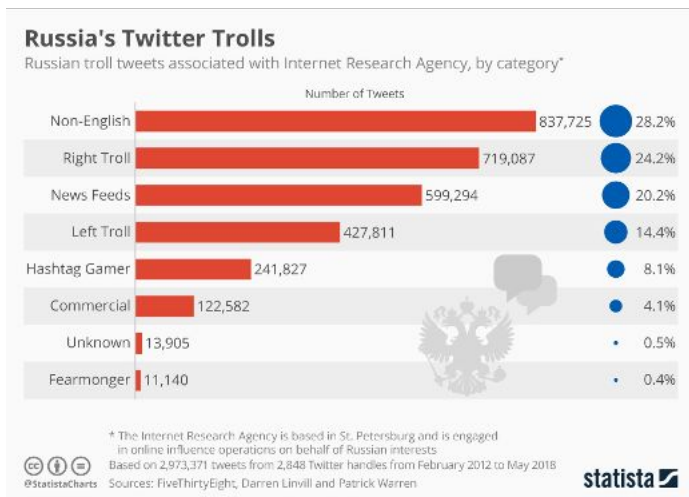


Yizhou Zhang¹, Lun Du², Karishma Sharma³, Yan Liu¹

¹University of Southern California, ²Coupang. Inc, ³Amazon. Inc

Background: Social Manipulation

- Social Manipulators within misinformation campaigns strategically manipulate public opinion on targeted topics by amplifying specific narratives and conspiracies.



Account 1

2020-03-06 19:22:40 RT @Jali_Cat:
🇷🇺: #COVID-19: a deep state attempt?! ...

2020-07-04 03:55:37 !!!Where are the law agencies and @realDonaldTrump administration regarding @CDCgov @CDCDirector DELIBERATELY exaggerating & running up COVID19 #s!?! WHERE!?! ...

2020-07-09 15:28:40 RT @Jali_Cat:
Where is @TheJusticeDept?? Where is @HHSGov?? Where is @DHS_Wolf??

Account 2

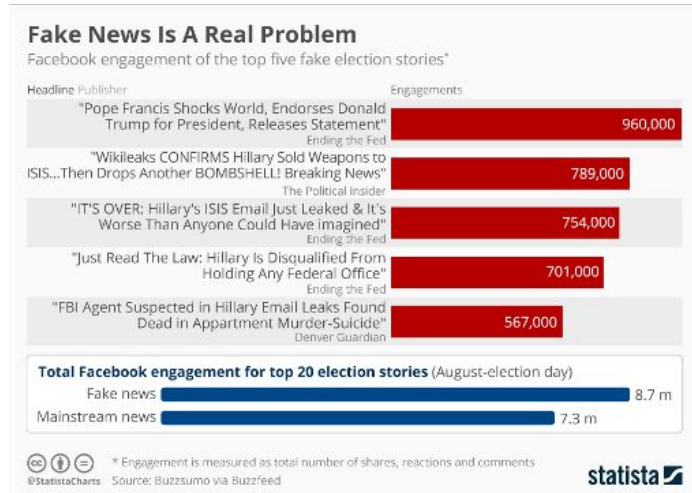
2020-03-07 01:58:57 RT @Jali_Cat:
🇷🇺: #COVID-19: a deep state attempt?! ...

2020-07-08 04:05:33 !!!Where are the law agencies and @realDonaldTrump administration regarding @CDCgov @CDCDirector DELIBERATELY exaggerating & running up COVID19 #s!?! WHERE!?! ...

2020-07-14 18:20:07 RT @Jali_Cat:
Where is @TheJusticeDept?? Where is @HHSGov?? Where is @DHS_Wolf??

Background: Social Manipulation and Misinformation Campaign




- Misinformation campaigns exploit online platforms to spread fake news, particularly during times of uncertainty, shaping public opinion, e.g.
 - Presidential Election.



Background: Social Manipulation and Misinformation Campaign

- Misinformation campaigns exploit online platforms to spread fake news, particularly during times of uncertainty, shaping public opinion, e.g.
 - Pandemic of COVID-19.

COVID-19 VACCINE

| MYTH ❌ | FACT ✅ |
|---|--|
| THEY AREN'T SAFE BECAUSE OF HOW FAST THEY WERE CREATED.  | NON-CLINICAL, CLINICAL, AND MANUFACTURING INFORMATION FOR BOTH VACCINES WAS REVIEWED BY A DATA SAFETY MONITORING BOARD BEFORE EMERGENCY APPROVAL. |
| THEY CAN CAUSE PEOPLE TO HAVE SERIOUS SIDE EFFECTS LIKE BELL'S PALS. Y. |  SERIOUS SIDE EFFECTS FROM BOTH THE PFIZER AND MODERNA VACCINES WERE RARE IN CLINICAL TRIALS. |
| THEY WILL MAKE PEOPLE SICK.  | THE MOST REPORTED SIDE EFFECT FOR EITHER VACCINE WAS INJECTION SITE REACTION/INJECTION SITE PAIN. VACCINES ALSO DO NOT CONTAIN A LIVE VIRUS. |


USC Melady Lab



Coronavirus Tweets Misinformation Analysis Trend Analysis Geo Analysis Contact

Coronavirus on Social Media: Misinformation Analysis

The 'infodemic' on Social Media impacts health and welfare of society, both at a local and global scale. We identify unreliable, misleading and clickbait information shared on Twitter to combat misinformation on Coronavirus from Mar 1- Apr 11.

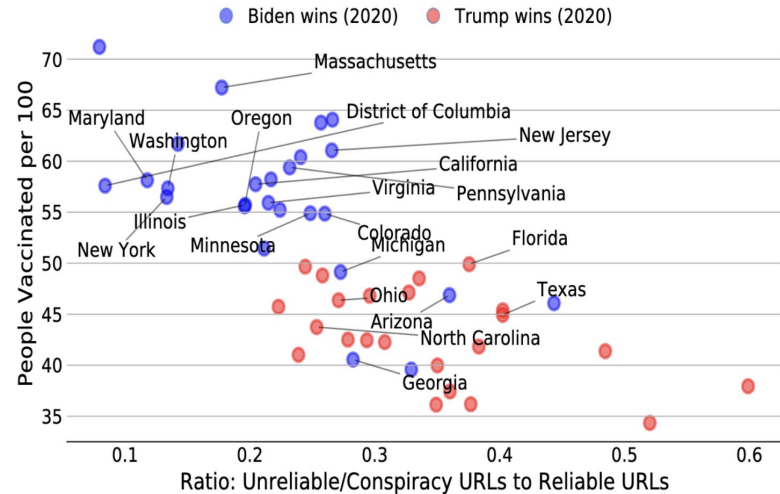
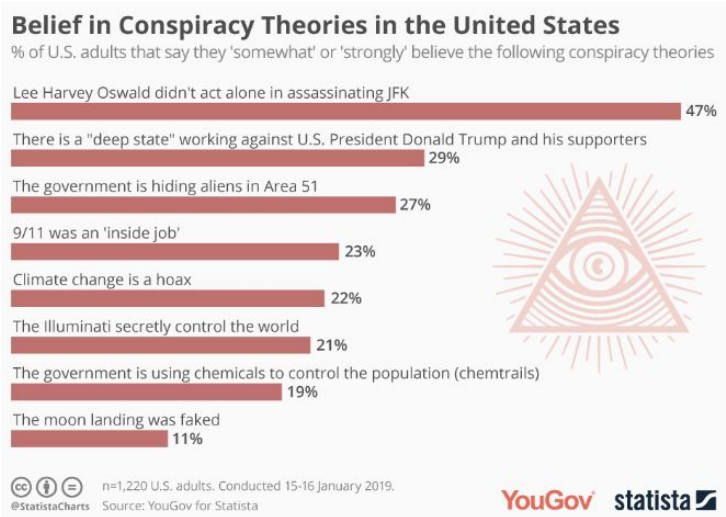
Distribution % (2020-03-01 - 2020-04-11)



| Website Type | Tweet | Twitter Link |
|----------------------------|---|----------------------|
| conspiracy-unreliable-bias | Convicted Murderer Re-Arrested After Being Freed from Prison Over Coronavirus - https://t.co/hL6svBAQeT  Via @NeonNettle #NeonNettle [2020-04-11T00:04:27] | link |
| conspiracy-unreliable-bias | 150 Members of Saudi Arabia Royal Family are Infected with Coronavirus - https://t.co/Yw4cvSYi5X  Via @NeonNettle #NeonNettle [2020-04-11T00:05:03] | link |
| unreliable | MORE FRAUD EXPOSED: Washington State Coronavirus Field Hospital Will Be Dismantled Without Treating a Single Patient https://t.co/QQmLjbo5KR [2020-04-11T00:05:46] | link |

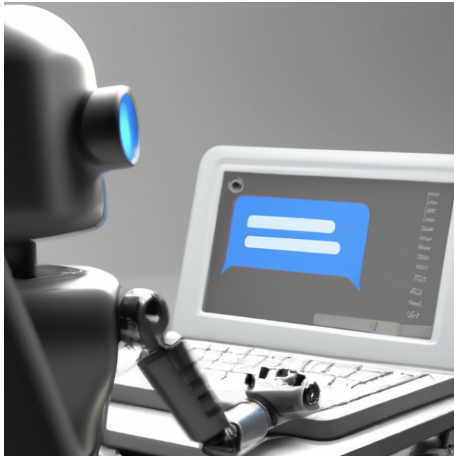
Background: Societal Impact of Misinformation and Manipulation

- The widespread fake news and social-manipulation contents have made substantial impacts on social-media users' opinions, beliefs and behaviors:



Background: Large Models are Challenging the Online Credibility

More Powerful
Social Bots for
Manipulation



More Efficiency
in Documenting
Fake News



Easier Creation of
Highly Deceptive
Misinformation



Eliot Higgins @EliotHiggins · Mar 20

Making pictures of Trump getting arrested while waiting for Trump's arrest. ...



1.8K

7.6K

39K

6.7M



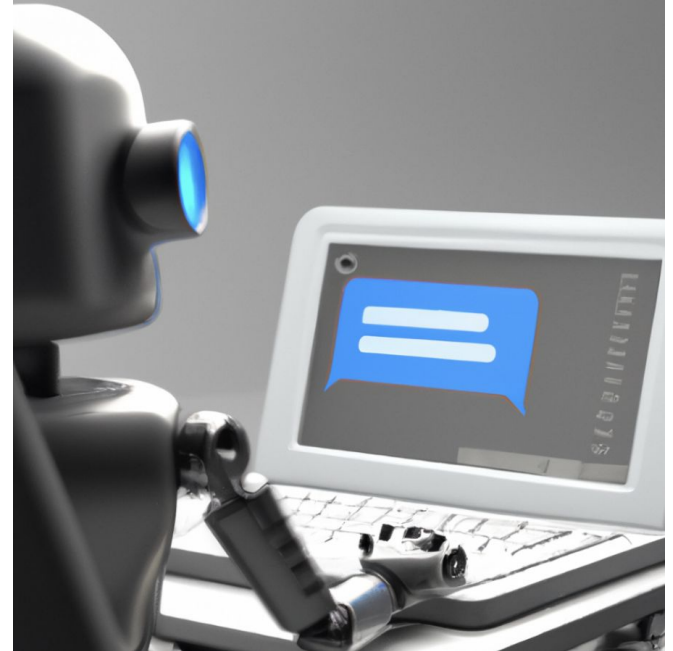
Tutorial Topics



- Manipulator Detection on Social Media
 - Detection of Social Manipulators (Social Bots)
 - LLM-based Detection of Social Manipulators
 - LLM-Inspired Collective Detection of Manipulator Campaigns
- Understanding the Causal Impact of Misinformation
 - Basic Introduction to Causal Inference and its Applications on Social Media
 - Causal Understanding on Misinformation Campaign Behavior
- LLM-based Misinformation Detection
 - Prompting and Reasoning Strategy for Misinformation Detection
 - LLM-based Multimodal Misinformation Detection

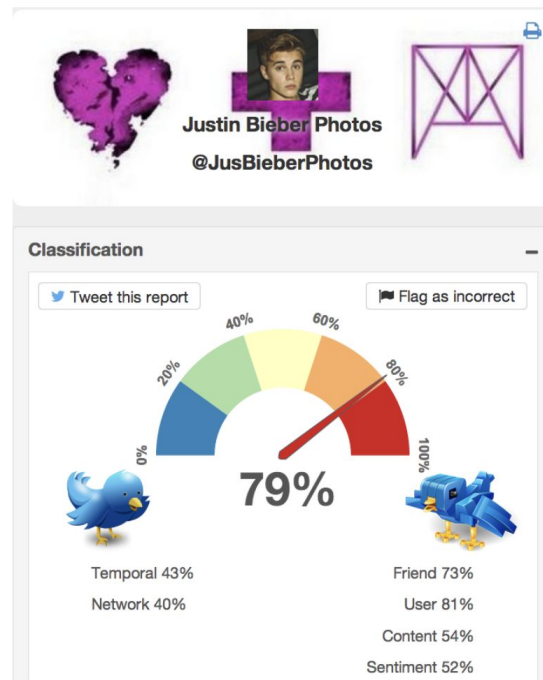
Manipulator Detection on Social Media

- Detection of Social Manipulators
- LLM-based Social Manipulator Detection
- LLM-Inspired Collective Detection of Manipulator Campaigns



Detection of Social Manipulators

In detection of social manipulators, we formulate this task as a supervised classification task, where each individual account is labeled as “Normal” and “Manipulator”. Our target is to learn a classifier to predict the labels for all accounts given its features.

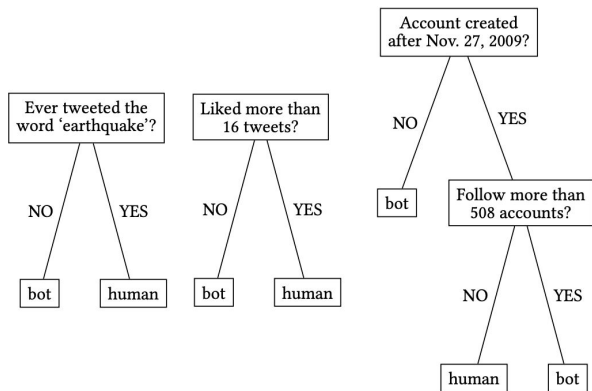


Interface of Botornot, the most famous social bot detection API

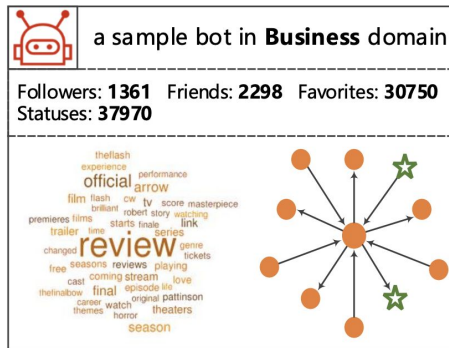
Features for Manipulator Detection

To classify social manipulators, we need to construct account features to capture useful information. The useful types of information include:

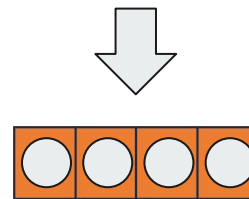
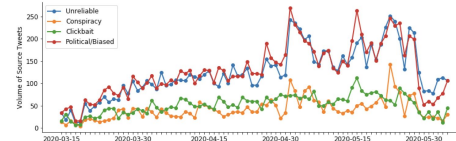
Classical Features: Metadata and Linguistic Cues



Structural Features: Features from Friends and Followers



Activity Traces: Patterns of Activity Time Series



Classical Features: Metadata and Language

To classify individual accounts, we need to construct account features to capture useful information. The useful types of information include:

- Metadata: e.g., statistics of activities, interests and IP address

| user metadata | | derived features | | |
|------------------------------|--------|---------------------------|-------------|---|
| feature name | type | feature name | type | calculation |
| statuses_count | count | tweet_freq | real-valued | $\text{statuses_count} / \text{user_age}$ |
| followers_count | count | followers_growth_rate | real-valued | $\text{followers_count} / \text{user_age}$ |
| friends_count | count | friends_growth_rate | real-valued | $\text{friends_count} / \text{user_age}$ |
| favourites_count | count | favourites_growth_rate | real-valued | $\text{favourites_count} / \text{user_age}$ |
| listed_count | count | listed_growth_rate | real-valued | $\text{listed_count} / \text{user_age}$ |
| default_profile | binary | followers_friends_ratio | real-valued | $\text{followers_count} / \text{friends_count}$ |
| profile_use_background_image | binary | screen_name_length | count | length of screen_name string |
| verified | binary | num_digits_in_screen_name | count | no. digits in screen_name string |
| | | name_length | count | length of name string |
| | | num_digits_in_name | count | no. digits in name string |
| | | description_length | count | length of description string |
| | | screen_name_likelihood | real-valued | likelihood of the screen_name |

Classical Features: Metadata and Language



To classify individual accounts, we need to construct account features to capture useful information. The useful types of information include:

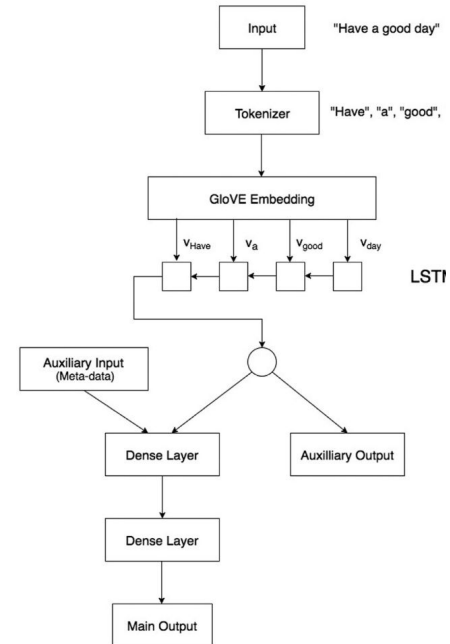
- Metadata: e.g., statistics of activities, interests and IP address
- Linguistic Cues: e.g., statistics of word usage and grammar errors

| Feature | Importance |
|----------------------------------|------------|
| out-degree (interaction network) | 100.00 |
| verb | 98.01 |
| conversational variety | 96.93 |
| conversational coverage | 96.65 |
| present | 94.66 |
| affect | 90.15 |
| personal pronoun | 89.71 |
| first person singular | 89.27 |
| conversational balance | 87.28 |
| motion | 87.28 |
| past | 86.56 |
| adverb | 86.20 |
| pronoun | 84.41 |
| negate | 84.33 |
| positive emotions | 83.25 |
| third person singular | 82.38 |
| social | 82.02 |
| exclusive | 81.86 |
| auxiliary verb | 81.70 |
| in-degree (interaction network) | 81.66 |

Classical Features: Metadata and Language

To classify individual accounts, we need to construct account features to capture useful information. The useful types of information include:

- Metadata: e.g., statistics of activities, interests and IP address
- Linguistic Cues: e.g., statistics of word usage and grammar errors
- Content Representation: e.g., hidden state extracted by neural networks (e.g., LSTM)



Structural Features: Features of Neighbors



Show me who your friends are and I'll tell you who you are
——Socrates

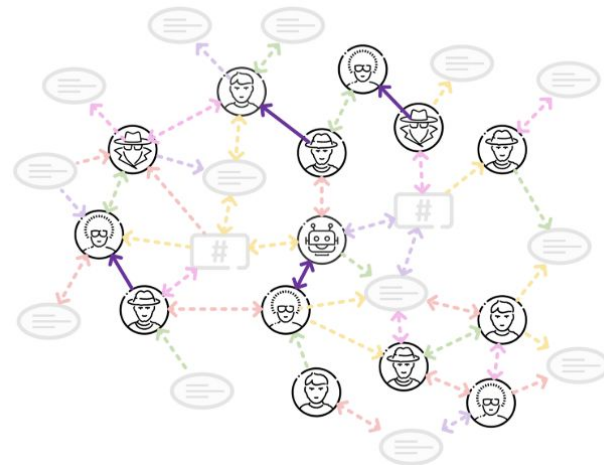
In addition to the accounts' own features, it is also important to model their neighbor's features based on the graph structure of social networks:

Structural Features: Features of Neighbors

Show me who your friends are and I'll tell you who you are
—Socrates

In addition to the accounts' own features, it is also important to model their neighbor's features based on the graph structure of social networks:

- Relationship Graph: The graph defined on the metadata, e.g., following and friends

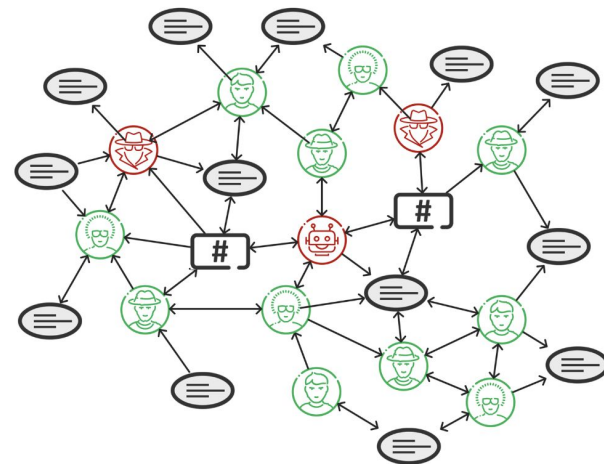


Structural Features: Features of Neighbors

Show me who your friends are and I'll tell you who you are
——Socrates

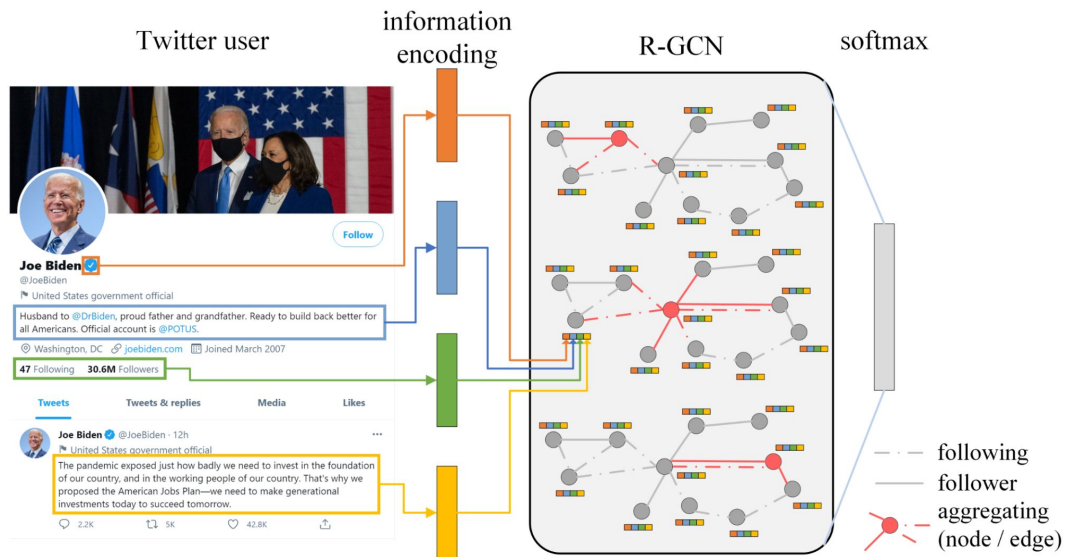
In addition to the accounts' own features, it is also important to model their neighbor's features based on the graph structure of social networks:

- Relationship Graph: The graph defined on the metadata, e.g., following and friends
- Interaction Graph: The graph defined on activity interactions, such as retweet.



Structural Features: Features of Neighbors

A typical strategy to use structural features for social manipulator detection is to first extract classical features for all accounts on the graph and then use graph neural networks (GNN) to propagate the features on the graph:



Structural Features boosts Classical Features

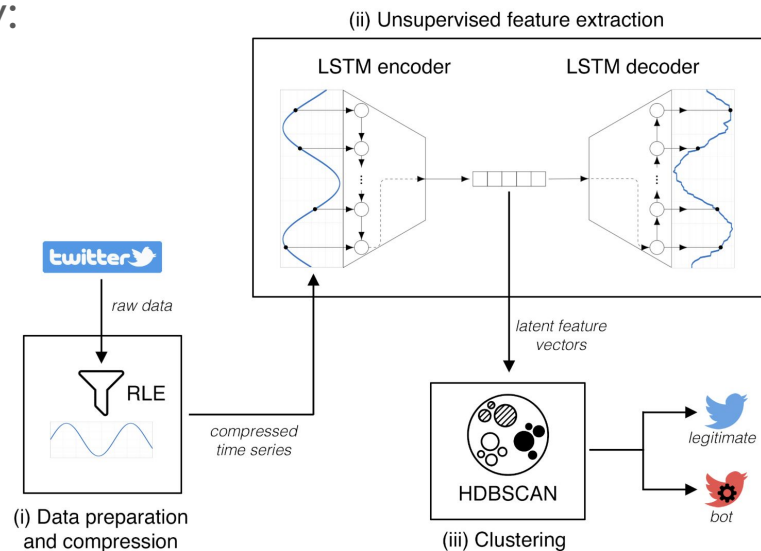
Existing benchmarking works have shown that structural features significantly boost performance of classical features (i.e., metadata+text)

| Method | Type | C-15 | G-17 | C-17 | M-18 | C-S-18 | C-R-19 | B-F-19 | Twibot-20 | Twibot-22 |
|------------------------------|------|---------------------------|--------------------|---------------------------|--------------------|---------------------------|---------------------------|---------------------------|---------------------------|---------------------------|
| Efthimiou <i>et al.</i> | FT | 92.5 (± 0.0) | 55.5 (± 0.0) | 88.0 (± 0.0) | 93.4 (± 0.0) | 70.8 (± 0.0) | 67.6 (± 0.0) | 69.8 (± 0.0) | 62.8 (± 0.0) | 74.1 (± 0.0) |
| Kantepe <i>et al.</i> | FT | 97.5 (± 1.3) | / | 98.2 (± 1.5) | / | / | / | / | 80.3 (± 4.3) | 76.4 (± 2.4) |
| Miller <i>et al.</i> | FT | 75.5 (± 0.0) | 51.0 (± 0.0) | 77.1 (± 0.2) | 83.7 (± 0.0) | 52.5 (± 0.0) | 54.4 (± 0.0) | 77.4 (± 0.0) | 64.5 (± 0.4) | 30.4 (± 0.1) |
| Varol <i>et al.</i> | FT | 93.2 (± 0.5) | / | / | / | / | / | / | 78.7 (± 0.6) | 73.9 (± 0.0) |
| Kouvela <i>et al.</i> | FT | 97.8 (± 0.5) | 74.7 (± 0.9) | <u>98.4</u> (± 0.1) | 97.0 (± 0.1) | 79.3 (± 0.3) | 79.7 (± 1.2) | 71.3 (± 0.9) | 84.0 (± 0.4) | 76.4 (± 0.0) |
| Santos <i>et al.</i> | FT | 70.8 (± 0.0) | 51.4 (± 0.0) | 73.8 (± 0.0) | 86.6 (± 0.0) | 62.5 (± 0.0) | 73.5 (± 0.0) | 71.7 (± 0.0) | 58.7 (± 0.0) | - |
| Lee <i>et al.</i> | FT | <u>98.2</u> (± 0.1) | 74.8 (± 1.2) | 98.8 (± 0.1) | 96.4 (± 0.1) | <u>81.5</u> (± 0.4) | 83.5 (± 1.9) | 75.5 (± 1.3) | 77.4 (± 0.5) | 76.3 (± 0.1) |
| LOBO | FT | 98.4 (± 0.3) | / | 96.6 (± 0.3) | / | / | / | / | 77.4 (± 0.2) | 75.7 (± 0.1) |
| Knauth <i>et al.</i> | FTG | 85.9 (± 0.0) | 49.6 (± 0.0) | 90.2 (± 0.0) | 83.9 (± 0.0) | 88.7 (± 0.0) | 50.0 (± 0.0) | 76.0 (± 0.0) | 81.9 (± 0.0) | 71.3 (± 0.0) |
| FriendBot | FTG | 96.9 (± 1.1) | / | 78.0 (± 1.0) | / | / | / | / | 75.9 (± 0.5) | - |
| SATAR | FTG | 93.4 (± 0.5) | / | / | / | / | / | / | 84.0 (± 0.8) | - |
| Botometer | FTG | 57.9 | 71.6 | 94.2 | 89.5 | 72.6 | 69.2 | 50.0 | 53.1 | 49.9 |
| Rodriguez-Ruiz <i>et al.</i> | FTG | 82.4 (± 0.0) | / | 76.4 (± 0.0) | / | / | / | / | 66.0 (± 0.1) | 49.4 (± 0.0) |
| GraphHist | FTG | 77.4 (± 0.2) | / | / | / | / | / | / | 51.3 (± 0.3) | - |
| EvolveBot | FTG | 92.2 (± 1.7) | / | / | / | / | / | / | 65.8 (± 0.6) | 71.1 (± 0.1) |
| Dehghan <i>et al.</i> | FTG | 62.1 (± 0.0) | / | / | / | / | / | / | <u>86.7</u> (± 0.1) | - |
| GCN | FTG | 96.4 (± 0.0) | / | / | / | / | / | / | 77.5 (± 0.0) | 78.4 (± 0.0) |
| GAT | FTG | 96.9 (± 0.0) | / | / | / | / | / | / | 83.3 (± 0.0) | <u>79.5</u> (± 0.0) |
| HGT | FTG | 96.0 (± 0.3) | / | / | / | / | / | / | 86.9 (± 0.2) | 74.9 (± 0.1) |
| SimpleHGN | FTG | 96.7 (± 0.5) | / | / | / | / | / | / | <u>86.7</u> (± 0.2) | 76.7 (± 0.3) |
| BotRGCN | FTG | 96.5 (± 0.7) | / | / | / | / | / | / | 85.8 (± 0.7) | 79.7 (± 0.1) |
| RGT | FTG | 97.2 (± 0.3) | / | / | / | / | / | / | 86.6 (± 0.4) | 76.5 (± 0.4) |

Activity Features: Information from Time Series

In addition to classical features and structural features, recent advances in deep-learning time series analysis make it possible to directly extract features from raw time series data of account activity:

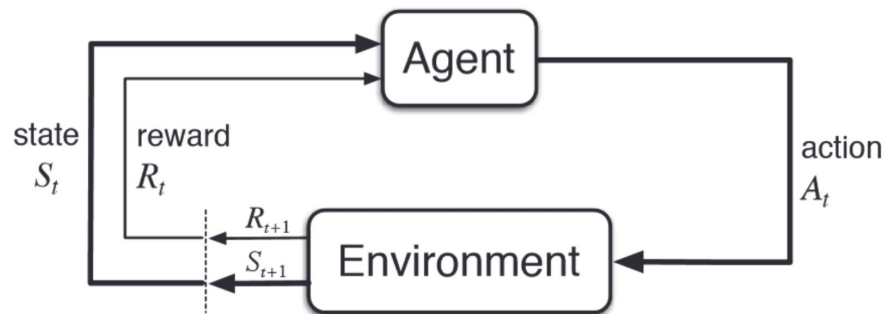
- Temporal Pattern Features: Using Statistical Learning or Deep Neural Networks to encode time series as features



Activity Features: Information from Time Series

In addition to classical features and structural features, recent advances in deep-learning time series analysis make it possible to directly extract features from raw time series data of account activity:

- Intention Features: Using Inverse Reinforcement Learning to infer the **Rewards** that drives the accounts' activities (i.e., the **Intention** of the account)



Activity Features: Inverse Reinforcement Learning

Given an agent and its action trajectory
 $T = [(S_1, A_1), (S_2, A_2), \dots]$, Inverse Reinforcement Learning aims at learning its reward function:

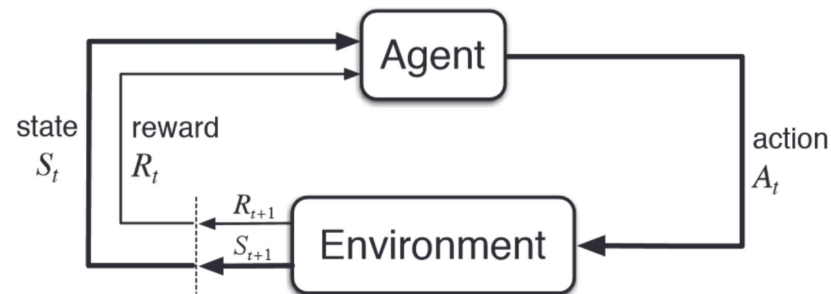
$$R_t = g(S_t, A_t; \theta_p, \theta_s)$$

Shared parameters

Personalized parameters

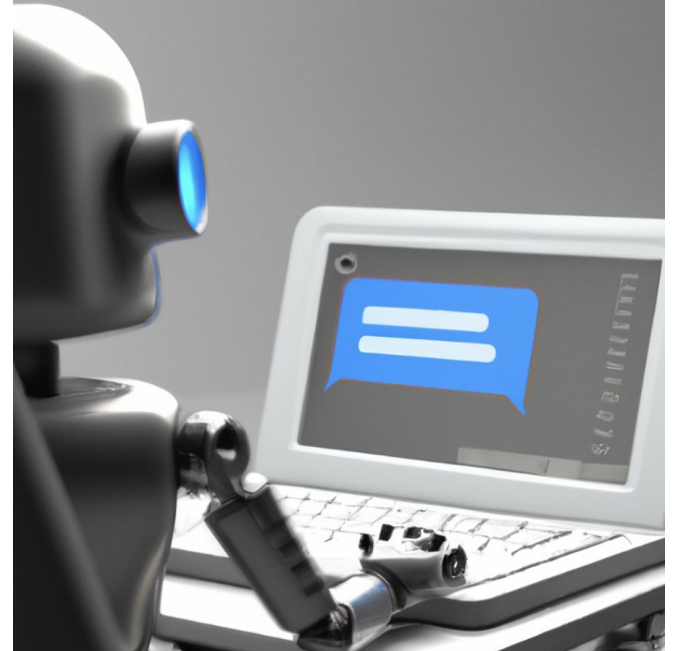
The function can be learned by maximizing the following objective function:

$$\log p(T) = \log \frac{1}{Z} \exp\left(\sum_t g(S_t, A_t; \theta_p, \theta_s)\right)$$



Manipulator Detection on Social Media

- Detection of Social Manipulators
- **LLM-based Social Manipulator Detection**
- LLM-Inspired Collective Detection of Manipulator Campaigns



Social Manipulator Detection in LLM Era



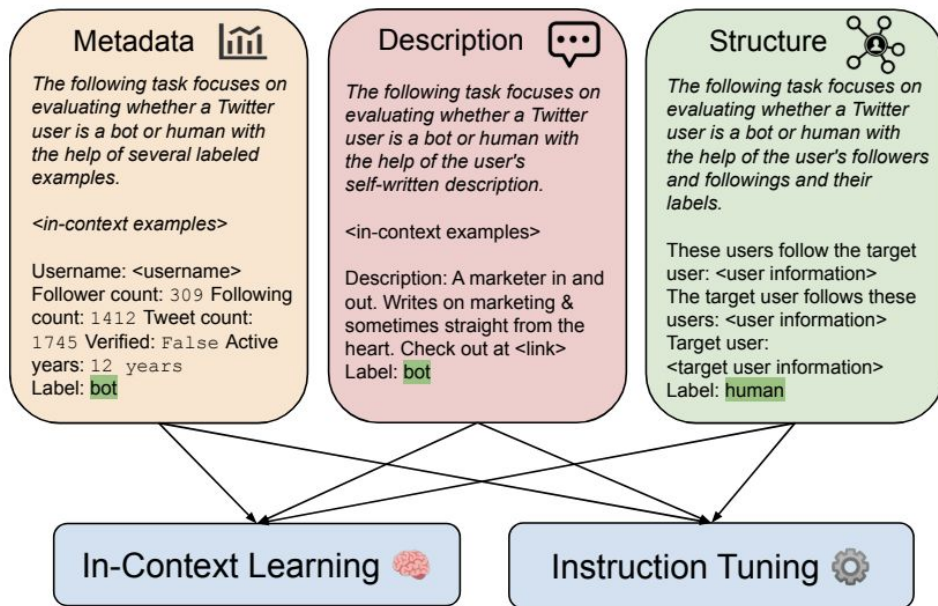
The recent advances of Large Language Model have brought both opportunities and challenges for Social Manipulator (Bot) Detection:

- Opportunities: The SOTA LLM like ChatGPT can significantly boost the effectiveness of detection to social manipulators, such as social bot
 - Classical Features and Structural Features can be easily organized as natural language data and forwarded into LLM for in-context learning or instructional tuning
- Challenges: LLM can significantly boost the deceptiveness of social manipulators, especially social bots

Opportunities: Manipulator Detection via LLM

We can easily convert metadata, text information and structure information as context of natural language and forward them to LLM for manipulator detection through:

- In-Context Learning: Providing few-shot samples as context
- Instructional Tuning: fine-tune the parameters for detection



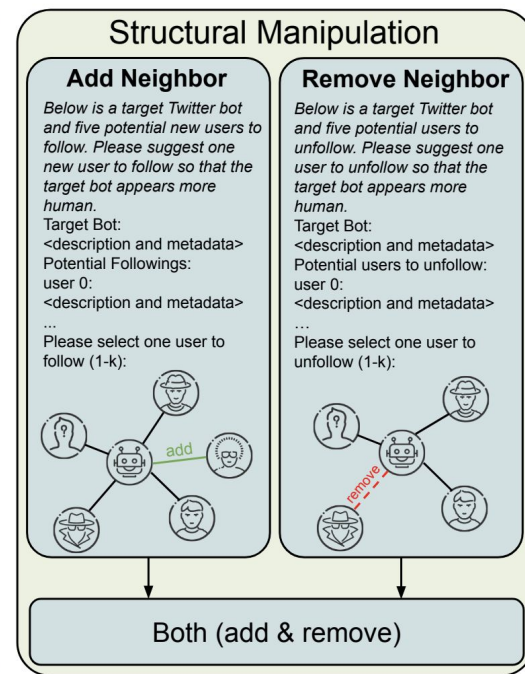
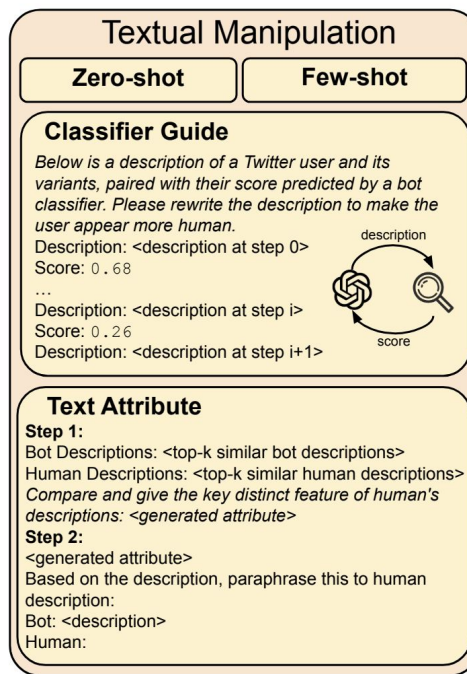
Opportunities: Manipulator Detection via LLM

| Method | \mathcal{M} | \mathcal{T} | \mathcal{N} | Twibot-20 | | | | Twibot-22 | | | |
|--|---------------|---------------|---------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|-------|
| | | | | Acc | F1 | Prec. | Rec. | Acc | F1 | Prec. | Rec. |
| BIC | | ✓ | ✓ | 0.876 | 0.891 | / | / | / | / | / | / |
| LMBOT | ✓ | ✓ | ✓ | 0.856 | 0.876 | / | / | / | / | / | / |
| SGBOT | ✓ | ✓ | ✓ | 0.816 | 0.849 | 0.764 | 0.949 | 0.623 | 0.395 | 1.000 | 0.247 |
| BOTPERCENT | ✓ | ✓ | ✓ | 0.845 | 0.865 | / | / | 0.731 | 0.726 | 0.738 | 0.714 |
| ROBERTA | | ✓ | | 0.755 | 0.731 | 0.739 | 0.724 | 0.633 | 0.432 | 0.955 | 0.280 |
| BOTOMETER | ✓ | ✓ | ✓ | 0.531 | 0.531 | 0.557 | 0.508 | 0.755 | 0.585 | 0.440 | 0.873 |
| BOTBUSTER | ✓ | ✓ | | 0.772 | 0.812 | / | / | 0.627 | 0.439 | 0.882 | 0.292 |
| LOBO | ✓ | ✓ | | 0.762 | 0.806 | 0.748 | 0.878 | 0.552 | 0.198 | 0.944 | 0.110 |
| RGT | ✓ | ✓ | ✓ | 0.866 | 0.880 | 0.852 | 0.911 | 0.509 | 0.509 | 0.323 | 0.854 |
| <i>Bot detection with CHATGPT</i> | | | | | | | | | | | |
| METADATA | ✓ | | | 0.766 | 0.793 | 0.742 | 0.852 | 0.659 | 0.698 | 0.626 | 0.788 |
| TEXT | | ✓ | | 0.566 | 0.576 | 0.612 | 0.544 | 0.688 | 0.684 | 0.705 | 0.665 |
| META+TEXT | ✓ | ✓ | | 0.656 | 0.694 | 0.755 | 0.642 | 0.659 | 0.681 | 0.607 | 0.777 |
| STRUCT-RAND | ✓ | ✓ | ✓ | 0.577 | 0.460 | 0.745 | 0.333 | 0.638 | 0.514 | 0.783 | 0.382 |
| STRUCT-ATT | ✓ | ✓ | ✓ | 0.565 | 0.426 | 0.743 | 0.298 | 0.632 | 0.500 | 0.792 | 0.365 |
| ENSEMBLE | ✓ | ✓ | ✓ | 0.632 | 0.557 | 0.801 | 0.427 | 0.735 | 0.706 | 0.794 | 0.635 |
| <i>Bot detection with CHATGPT and instruction tuning</i> | | | | | | | | | | | |
| METADATA | ✓ | | | 0.812 | 0.806 | 0.814 | 0.847 | 0.724 | 0.764 | 0.667 | 0.894 |
| TEXT | | ✓ | | 0.767 | 0.791 | 0.768 | 0.816 | 0.727 | 0.766 | 0.670 | 0.894 |
| META+TEXT | ✓ | ✓ | | 0.862 | 0.865 | 0.813 | 0.924 | 0.721 | 0.758 | 0.668 | 0.877 |
| STRUCT-RAND | ✓ | ✓ | ✓ | 0.890 | 0.904 | 0.839 | 0.980 | 0.718 | 0.761 | 0.660 | 0.900 |
| STRUCT-ATT | ✓ | ✓ | ✓ | 0.885 | 0.888 | 0.856 | 0.923 | 0.727 | 0.766 | 0.670 | 0.894 |
| ENSEMBLE | ✓ | ✓ | ✓ | 0.899 | 0.915 | 0.861 | 0.976 | 0.769 | 0.792 | 0.696 | 0.918 |

Challenges: Manipulators Boosted by LLM

Unfortunately, manipulators can also use similar way to convert LLM as their assistant. Specifically, they have two strategies:

1. They can use LLM to manipulate the text contents
2. They can use LLM to help them manipulate the structure of manipulators



Challenges: Manipulators Boosted by LLM

The authors evaluated LLM in manipulating classical models and LLM-boosted detectors:

1. To classical detectors, LLM-based manipulation is highly deceptive
2. To LLM-based detector, the manipulation from another LLM (e.g., use LLaMa to fool ChatGPT) is more effective than the manipulation from itself

| Strategy | BotPercent | | BotRGCN | | Text+Meta | | Struct-Rand | | Struct-Att | | Ensemble | |
|--|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| | Acc | F1 | Acc | F1 | Acc | F1 | Acc | F1 | Acc | F1 | Acc | F1 |
| vanilla Twibot-20 | .755 | .731 | .737 | .766 | .862 | .865 | .890 | .904 | .884 | .888 | .899 | .915 |
| <i>Manipulation strategies with LLAMA2-70B</i> | | | | | | | | | | | | |
| ZERO-SHOT REWRITE | .716 | .724 | .735 | .788 | .859 | .874 | .889 | .905 | .867 | .871 | .885 | .901 |
| FEW-SHOT REWRITE | .689 | .720 | .732 | .784 | .862 | .878 | .886 | .902 | .852 | .867 | .883 | .898 |
| CLASSIFIER GUIDE | .650 | .704 | .722 | .779 | .835 | .852 | .868 | .886 | .805 | .818 | .850 | .870 |
| TEXT ATTRIBUTE | .689 | .737 | .728 | .787 | .872 | .887 | .890 | .906 | .881 | .895 | .891 | .907 |
| ADD NEIGHBOR | / | / | .731 | .785 | / | / | .874 | .890 | .855 | .869 | .867 | .885 |
| REMOVE NEIGHBOR | / | / | .653 | .721 | / | / | .863 | .882 | .862 | .878 | .863 | .882 |
| COMBINE NEIGHBOR | / | / | .596 | .539 | / | / | .866 | .883 | .859 | .873 | .868 | .885 |
| SELECTIVE COMBINE | .691 | .737 | .684 | .663 | .866 | .883 | .866 | .884 | .860 | .875 | .865 | .884 |
| BOTH COMBINE | .650 | .704 | .571 | .564 | .835 | .852 | .854 | .871 | .808 | .822 | .850 | .869 |
| <i>Manipulation strategies with CHATGPT</i> | | | | | | | | | | | | |
| ZERO-SHOT REWRITE | .680 | .731 | .719 | .745 | .875 | .891 | .891 | .907 | .894 | .907 | .896 | .911 |
| FEW-SHOT REWRITE | .675 | .724 | .708 | .738 | .879 | .894 | .889 | .905 | .887 | .901 | .890 | .906 |
| CLASSIFIER GUIDE | .649 | .699 | .702 | .715 | .860 | .878 | .890 | .906 | .888 | .903 | .886 | .903 |
| TEXT ATTRIBUTE | .661 | .716 | .716 | .752 | .855 | .870 | .882 | .899 | .879 | .894 | .877 | .895 |
| ADD NEIGHBOR | / | / | .715 | .741 | / | / | .874 | .892 | .893 | .907 | .879 | .897 |
| REMOVE NEIGHBOR | / | / | .642 | .629 | / | / | .870 | .888 | .855 | .870 | .864 | .883 |
| COMBINE NEIGHBOR | / | / | .632 | .685 | / | / | .878 | .895 | .893 | .907 | .878 | .896 |
| SELECTIVE COMBINE | .678 | .725 | .615 | .638 | .864 | .880 | .873 | .891 | .860 | .875 | .873 | .891 |
| BOTH COMBINE | .649 | .699 | .641 | .627 | .860 | .878 | .888 | .905 | .905 | .919 | .894 | .910 |

Roadmaps: What can we do for future?



To address the new challenges from LLM, we can attempt the following directions of research efforts:

- **Develop Detectors towards AI-Manipulated Contents:**

The experiment results from [Feng et,al, 2024] have shown that LLM-based detector are robust to the manipulation from their own backbone. This suggests the probability to develop more powerful detectors by first recognizing the LLM used by the manipulators use and then tackle the account accordingly

Challenge: How to recognize the LLM used by the manipulators?

Roadmaps: What can we do for future?



To address the new challenges from LLM, we can attempt the following directions of research efforts:

- **Aggregate Activity Features into LLM-based detector:**

Compared to classical features and structural features, activity features is harder to manipulate. Malicious accounts can quickly spread misinformation because they show more definite characteristics, as opposed to more randomized actions from the normal accounts. If they change their activity features, their efficiency and effectiveness will also decrease.

Challenge: How to capture time-series features with LLM?

Capture Time-Series Features with LLM

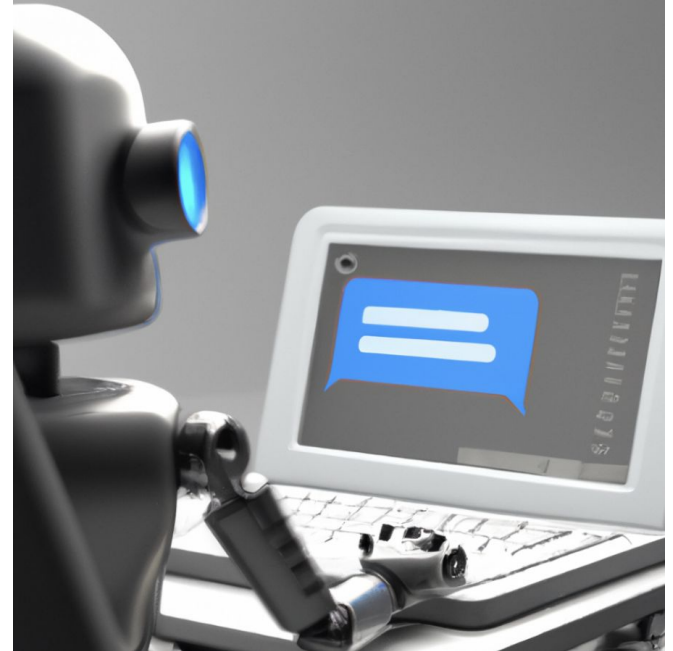


There have been a lot of research attempt trying to incorporate LLM with capacity in tackling time series:

- **Fine-tuning LLM to accept Time-Series as Input:**
 - a. Fine-tuning Whole Model: TEMPO[Cao et,al, 2024] One-fits-all (GPT-2) [Zhou et, al, 2023]
 - b. Only fine-tuning embedding layers: FreqTST [Li et, al, 2024]
- **Convert Time-Series to Formats compatible to LLM or Large Multimodal models:**
 - a. Convert Time Series as Images and apply Vision Transformers as encoder: ViTST [Li et, al, 2023], PandaGPT [Su et, al, 2024]
 - b. Convert Time Series as Text: PromptCast [Xue and Salim, 2022], TimeLLM [Jin et, al, 2024]

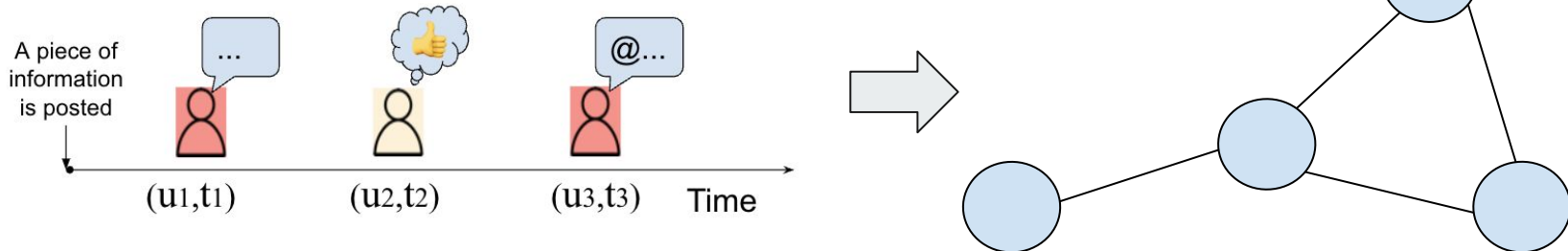
Manipulator Detection on Social Media

- Detection of Social Manipulators
- LLM-based Social Manipulator Detection
- **LLM-Inspired Collective Detection of Manipulator Campaigns**



Detection of Coordinated Manipulators

- Task Definition: Detecting malicious accounts collaborating on social media to expand influence
- Existing works mainly rely on domain knowledge to construct interaction graphs for clustering to identify coordinated groups



Detection of Coordinated Manipulators

- For example, a typical strategy is to evaluate users' activity trace overlap. We define two events as equivalent event pairs if:

$$\langle U_i, T_i, C_i \rangle \approx \langle U_j, T_j, C_j \rangle \quad \text{if } C_i = C_j \text{ and } |T_i - T_j| \leq T_{sim}$$

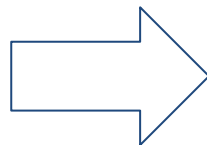
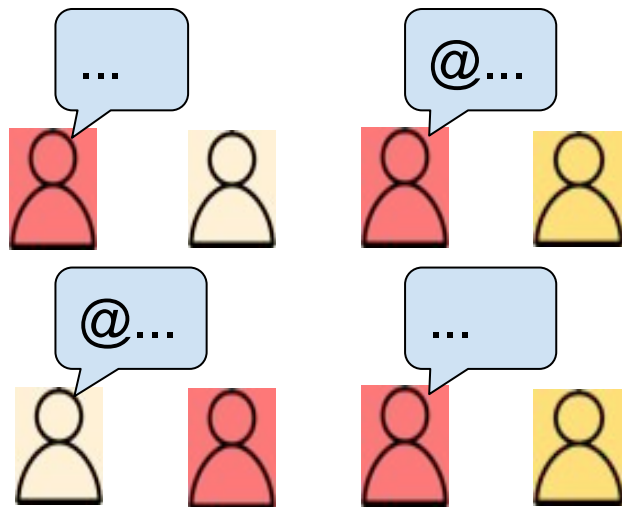
- Then the similarity of two users can be defined as:

$$\text{Sim}(U_i, U_j) = \frac{|A_i \cap A_j|}{|A_i \cup A_j|} = \frac{\sum_k |A_i^k \cap A_j^k|}{\sum_k |A_i^k \cup A_j^k|}$$

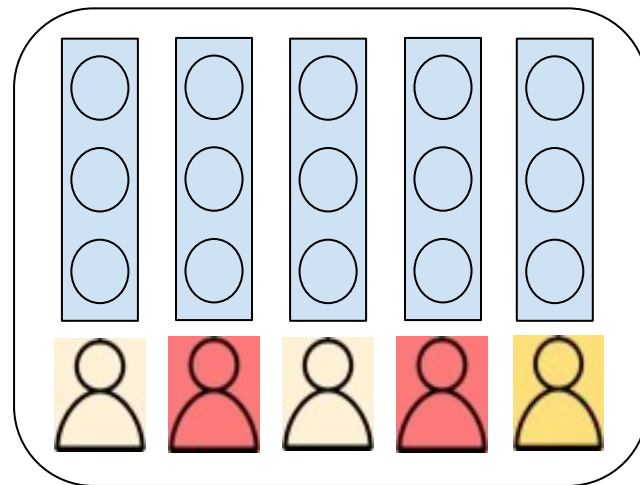
- However, such methods generally suffer from poor expressive power and reliance on the prior-knowledge

Learning Representation for Coordination Detection

Observed Activity Traces



Account Representation

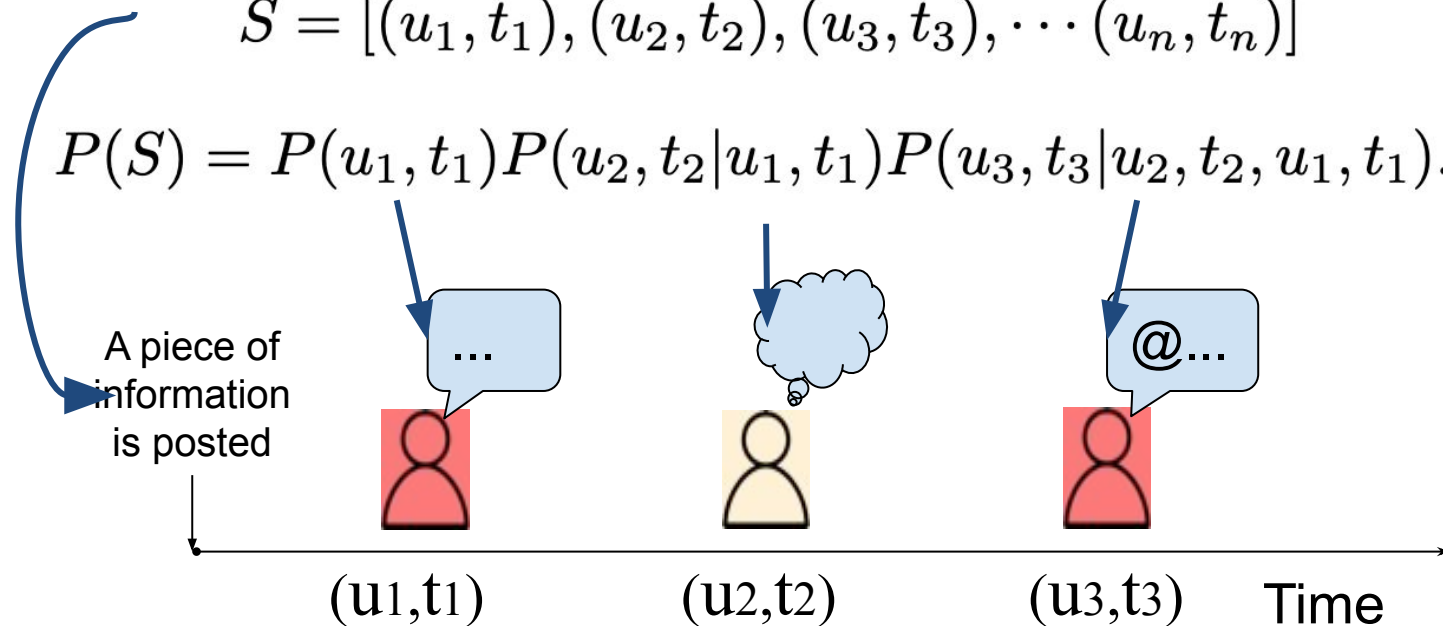


LLM-Inspired Representation Learning

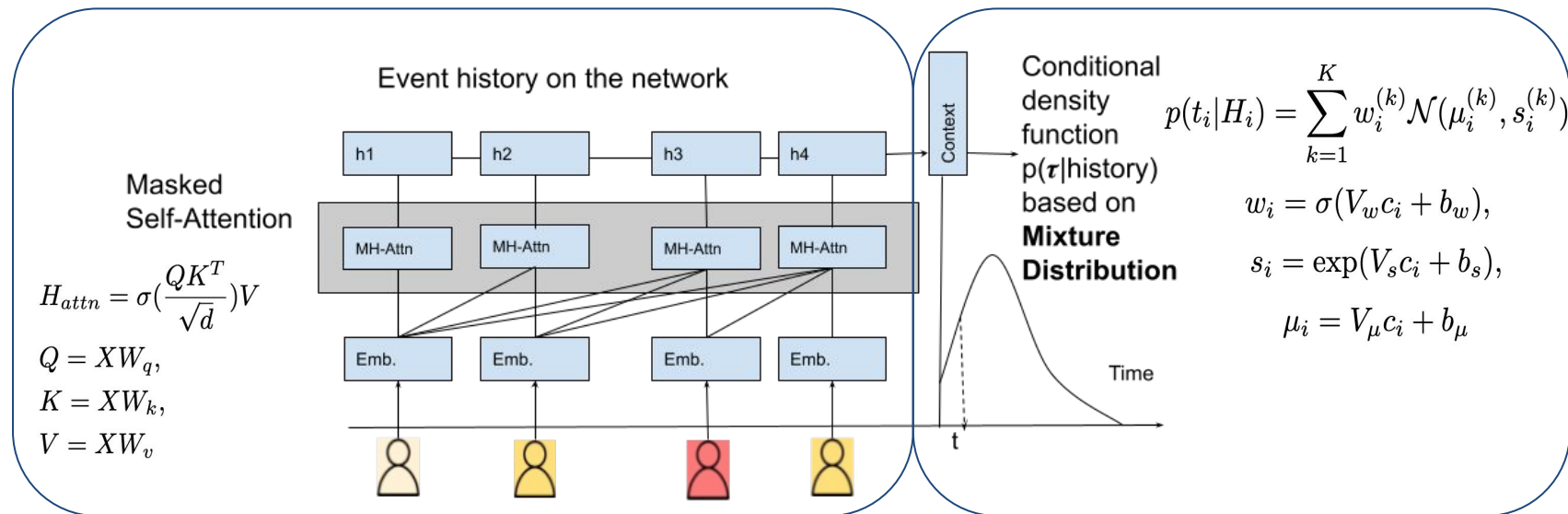
We notice the similarity between next-token-prediction language modeling and activity traces: Token = User, Positional Encoding = Time

$$S = [(u_1, t_1), (u_2, t_2), (u_3, t_3), \dots (u_n, t_n)]$$

$$P(S) = P(u_1, t_1)P(u_2, t_2|u_1, t_1)P(u_3, t_3|u_2, t_2, u_1, t_1)\dots$$



AMDN: Attentive Mixture Distribution Networks



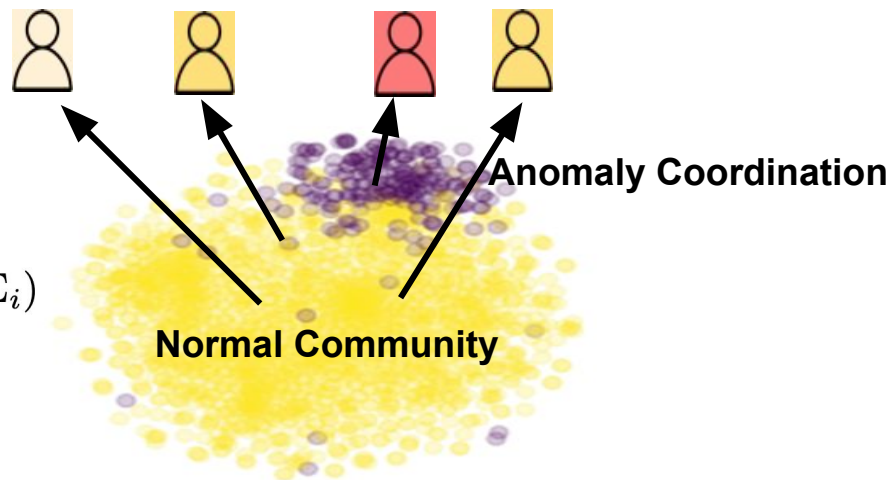
Masked Self-Attention Encoder ensure capturing long-term influence while outputting attention weights as influence score

Gaussian-Mixture Distribution Decoder guarantee universal approximation to any continuous distributions while providing closed-form likelihood

HAGE: Hidden Account Group Estimation

Simultaneously, we can jointly learn the membership of the accounts based on a GMM model learnt in the user (token) embedding space:

$$\begin{aligned} \sum_{j=1}^{|U|} \log p(u_j; \theta_g, E) &= \sum_{j=1}^{|U|} \log \sum_{i=1}^N p(u_j, i; \theta_g, E) \\ &= \sum_{j=1}^{|U|} \log \sum_{i=1}^N p(i) \mathcal{N}(E_{u_j}; \mu_i, \Sigma_i) \end{aligned}$$



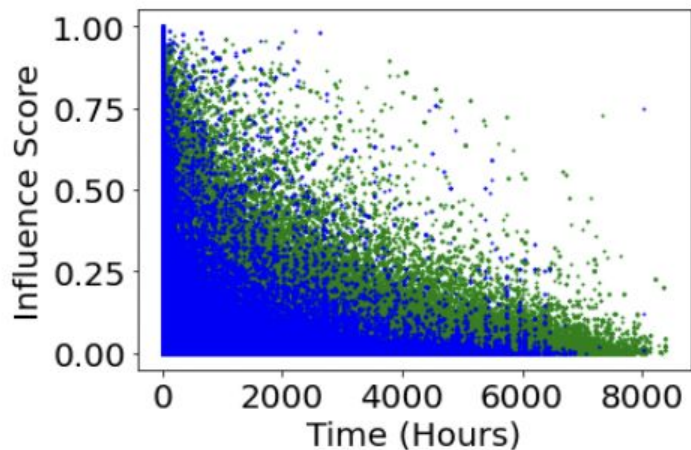
LLM-Inspired Representation Learning

The authors evaluated the performance of the model on a dataset about misinformation campaigns on social media during U.S. Election:

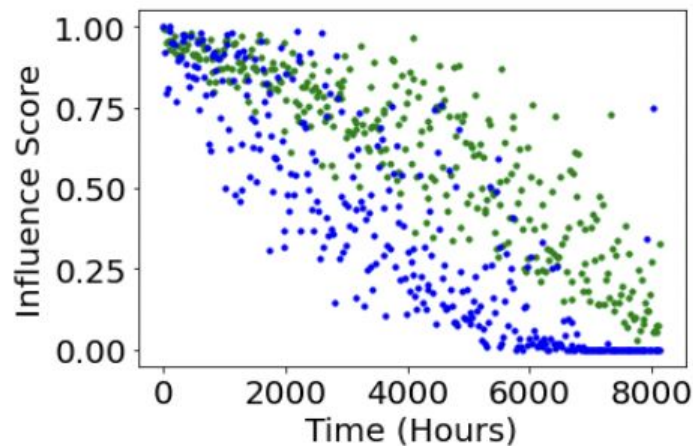
| Method (Unsupervised) | AP | AUC | F1@TH=0.5 | Prec@TH=0.5 | Rec@TH=0.5 | MaxF1 | MacroF1@TH=0.5 |
|---------------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|
| Co-activity | 0.208 ± 0.01 | 0.592 ± 0.03 | 0.292 ± 0.02 | 0.206 ± 0.02 | 0.510 ± 0.04 | 0.331 ± 0.03 | 0.515 ± 0.02 |
| Clickstream | 0.169 ± 0.02 | 0.535 ± 0.04 | 0.215 ± 0.06 | 0.205 ± 0.05 | 0.228 ± 0.08 | 0.215 ± 0.06 | 0.532 ± 0.03 |
| IRL | 0.200 ± 0.00 | 0.610 ± 0.02 | 0.265 ± 0.02 | 0.219 ± 0.02 | 0.336 ± 0.03 | 0.340 ± 0.02 | 0.543 ± 0.01 |
| HP | 0.337 ± 0.04 | 0.694 ± 0.05 | 0.376 ± 0.05 | 0.387 ± 0.06 | 0.365 ± 0.05 | 0.545 ± 0.03 | 0.633 ± 0.03 |
| AMDN + GMM | 0.787 ± 0.05 | 0.894 ± 0.03 | 0.631 ± 0.06 | 0.965 ± 0.03 | 0.472 ± 0.07 | 0.738 ± 0.05 | 0.792 ± 0.03 |
| AMDN + Kmeans | 0.731 ± 0.08 | 0.901 ± 0.02 | 0.727 ± 0.06 | 0.806 ± 0.07 | 0.663 ± 0.06 | 0.752 ± 0.05 | 0.841 ± 0.03 |
| AMDN-HAGE | 0.804 ± 0.03 | 0.898 ± 0.02 | 0.699 ± 0.05 | 0.941 ± 0.04 | 0.558 ± 0.06 | 0.758 ± 0.04 | 0.828 ± 0.03 |
| AMDN-HAGE + Kmeans | 0.818 ± 0.04 | 0.935 ± 0.02 | 0.731 ± 0.04 | 0.913 ± 0.03 | 0.611 ± 0.05 | 0.776 ± 0.03 | 0.846 ± 0.02 |
| Method (Supervised) | AP | AUC | F1@TH=0.5 | Prec@TH=0.5 | Rec@TH=0.5 | MaxF1 | MacroF1@TH=0.5 |
| IRL (S) | 0.672 ± 0.08 | 0.896 ± 0.03 | 0.557 ± 0.06 | 0.781 ± 0.06 | 0.436 ± 0.06 | 0.633 ± 0.07 | 0.749 ± 0.03 |
| HP (S) | 0.760 ± 0.04 | 0.925 ± 0.02 | 0.753 ± 0.02 | 0.743 ± 0.04 | 0.769 ± 0.06 | 0.782 ± 0.03 | 0.853 ± 0.01 |
| AMDN + NN | 0.814 ± 0.04 | 0.918 ± 0.02 | 0.733 ± 0.04 | 0.710 ± 0.05 | 0.761 ± 0.05 | 0.763 ± 0.04 | 0.841 ± 0.02 |
| AMDN-HAGE + NN | 0.838 ± 0.04 | 0.926 ± 0.03 | 0.769 ± 0.04 | 0.752 ± 0.05 | 0.789 ± 0.05 | 0.799 ± 0.04 | 0.862 ± 0.02 |

LLM-Inspired Representation Learning

Analysis to the attention weights reveals that the interaction between coordinated accounts (blue) decreases faster than normal users (orange)



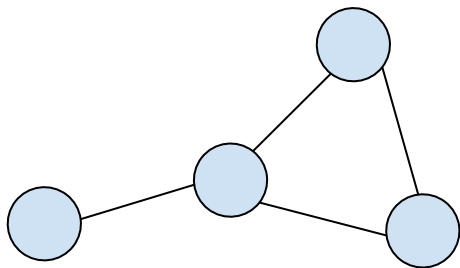
Overall influence score trend among all users



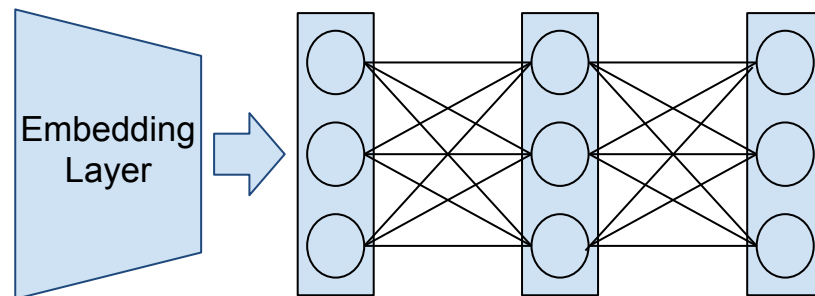
Influence Score Trend among strongly interacted user pairs

Boosting Representation Learning with Knowledge

LLM-Inspired Representation Learning improve the model performance.
However, it heavily require data quantity



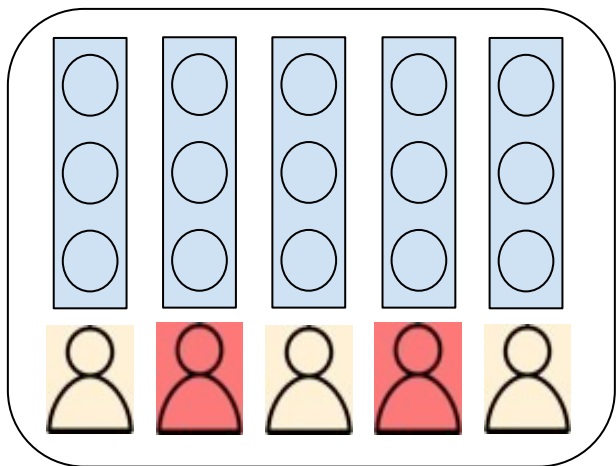
- Construct an account interaction graph and then do clustering
- Suffers from poor expressive power and reliance on the prior-knowledge



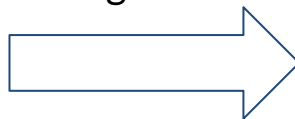
- Directly learn account embeddings via a data-driven manner
- Suffers from reliance on data quality and quantity

Boosting Representation Learning with Knowledge

Account Representation



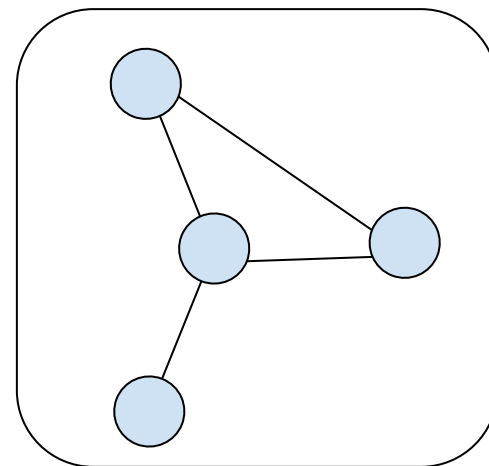
Supplement Manual
Features with
Knowledge from Data



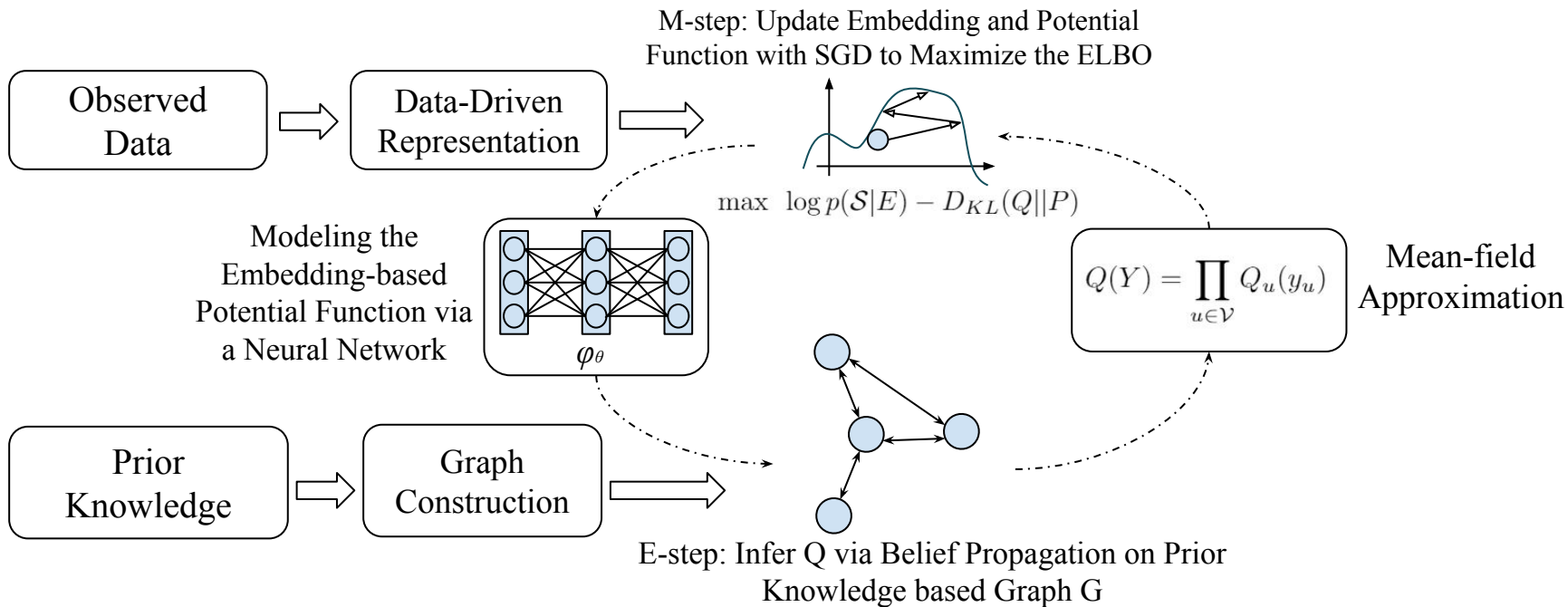
Regularize Learning
with Human
Knowledge



Interaction Graph



VigDet: EM Algorithm for Learning with Knowledge



VigDet: EM Algorithm for Learning with Knowledge

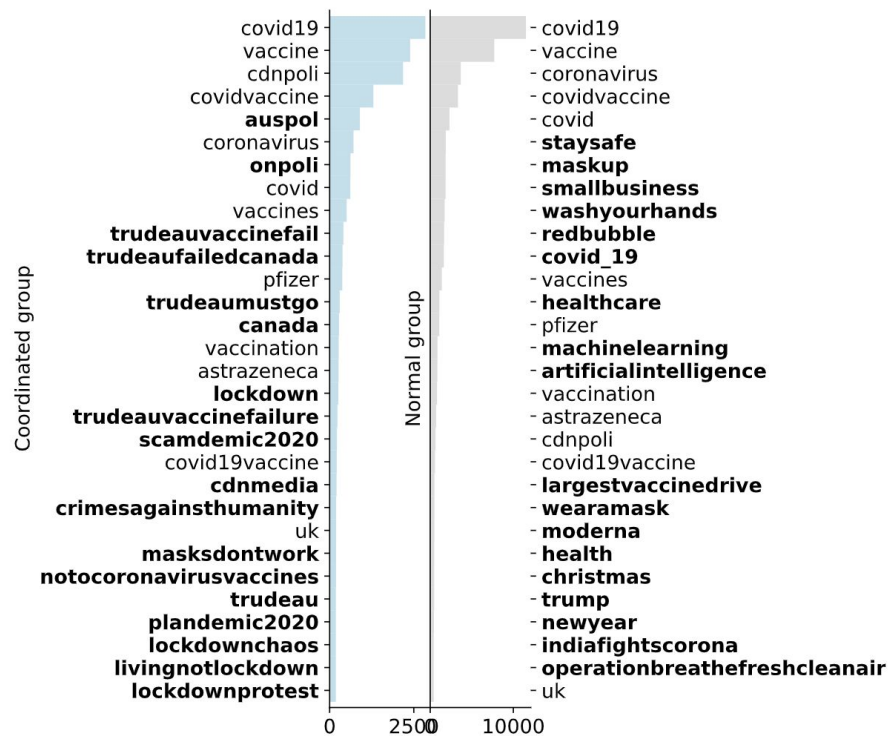
| Method (Unsupervised) | AP | AUC | F1 | Prec | Rec | MaxF1 | MacroF1 |
|------------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Co-activity | 16.9 | 52.5 | 24.6 | 17.8 | 40.7 | 27.1 | 49.5 |
| Clickstream | 16.5 | 53.2 | 21.0 | 20.6 | 21.6 | 21.0 | 53.1 |
| IRL | 23.9 | 68.7 | 35.3 | 27.5 | 49.4 | 38.6 | 58.8 |
| HP | 29.8 | 56.7 | 44.2 | 42.1 | 46.6 | 46.0 | 66.7 |
| AMDN-HAGE | 80.5 | 89.9 | 69.6 | 94.3 | 55.5 | 75.8 | 82.7 |
| AMDN-HAGE + k -Means | 82.0 | 93.3 | 73.0 | 90.9 | 61.2 | 77.0 | 84.5 |
| VigDet-PL(TL) | 83.3 | 94.0 | 70.7 | 89.6 | 59.0 | 77.8 | 83.2 |
| VigDet-E(TL) | 85.5 | 94.6 | 73.1 | 95.3 | 59.4 | 79.5 | 84.6 |
| VigDet(TL) | 86.1 | 94.6 | 73.4 | 95.1 | 59.9 | 79.6 | 84.8 |
| VigDet-PL(PF) | 84.5 | 95.0 | 71.9 | 91.4 | 59.6 | 79.3 | 83.9 |
| VigDet-E(PF) | 85.1 | 94.3 | 73.6 | 92.7 | 61.2 | 78.8 | 84.9 |
| VigDet(PF) | 87.2 | 95.0 | 75.2 | 91.7 | 63.9 | 79.3 | 85.7 |

VigDet: Analysis on COVID-19 Dataset

Dataset: 31k accounts without ground truth label and tweets they interact.

Detection: VigDet detect 8k suspicious accounts

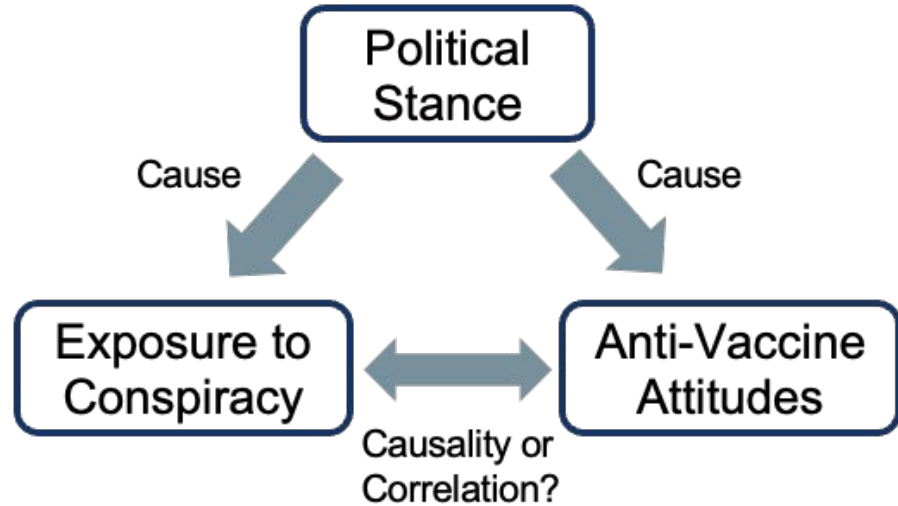
Hashtag Difference: The two groups are clearly distinguished in the comparison of top-30 hashtags. In bold are the non-overlapping hashtags.



The top-30 hashtags of coordinated and normal accounts detected by VigDet.

Causal Inference of Misinformation on Social Media

- Introduction to Causal Inference
- Applications of Causal Inference on Social Media
- Application of LLM in Causal Inference on Social Media

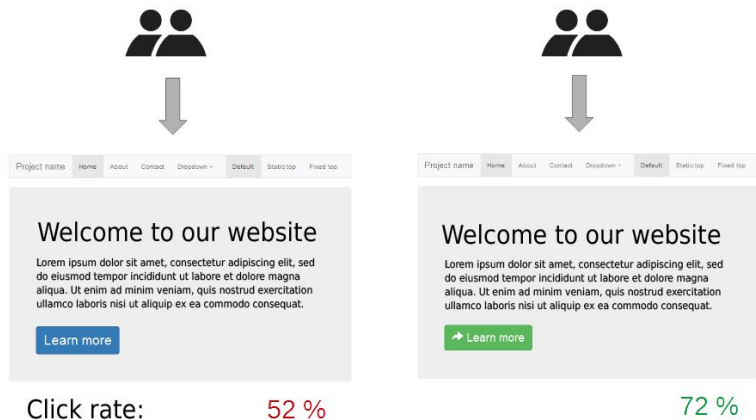


Basic Introduction to Causal Inference

Causal Inference aims at understanding the causal relationship between a treatment (i.e., cause, such as smoking) and an outcome (i.e., result, such as lung cancer).

The golden standard of measuring ground-truth causal effect is random controlled trial (RCT).




In the scope of social media, the most well-known RCT is **AB-test**.



Example of AB-test from Wikipedia. which help us measure how an interface design (treatment) influence click rate (outcome).

Weakness of Random Controlled Trial

However, random controlled trials are not always applicable due to the following issues:

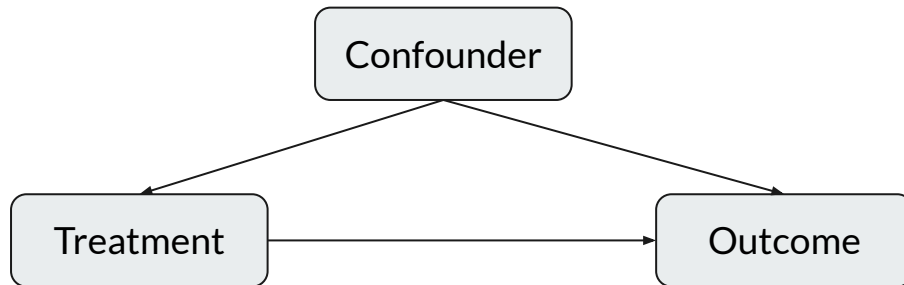
-  High Cost: Conducting random controlled trials usually requires a lot of resources, leading to high cost.
-  Limited Data Quantity: Due to the high cost, available data quantity of RCT is usually limited.
-  Ethical Concerns: When tackling societal problems, RCT often require recruiting human subjects, which may lead to ethical concerns.

Causal Inference from Observation

To tackle the disadvantages of RCT, causal inference from observation is developed.

Instead of actively conducting trials, causal inference aims at understanding causal effect from data that are collected through collective observation without intervention.

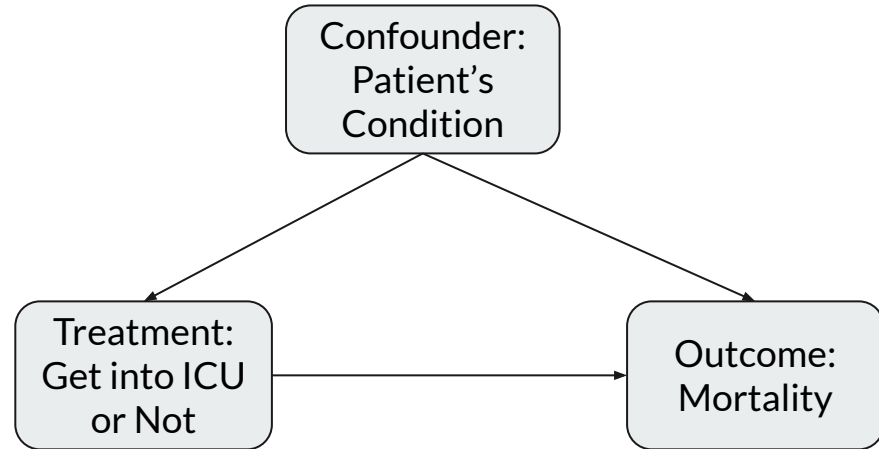
However, an emerging challenge arises in causal inference : **Confounders**



Example of Confounders in Causal Inference

The key factor of random controlled trials is the randomness of treatment assignment. However, confounders may influence treatment assignment. Ignoring such influence may lead to biased estimation.

An example is estimating the causal effect of getting into ICU to mortality. Directly estimating $P(\text{mortality}|\text{ICU})$ will lead to a ridiculous conclusion that ICU cause patient's death



Combat Confounders in Causal Inference

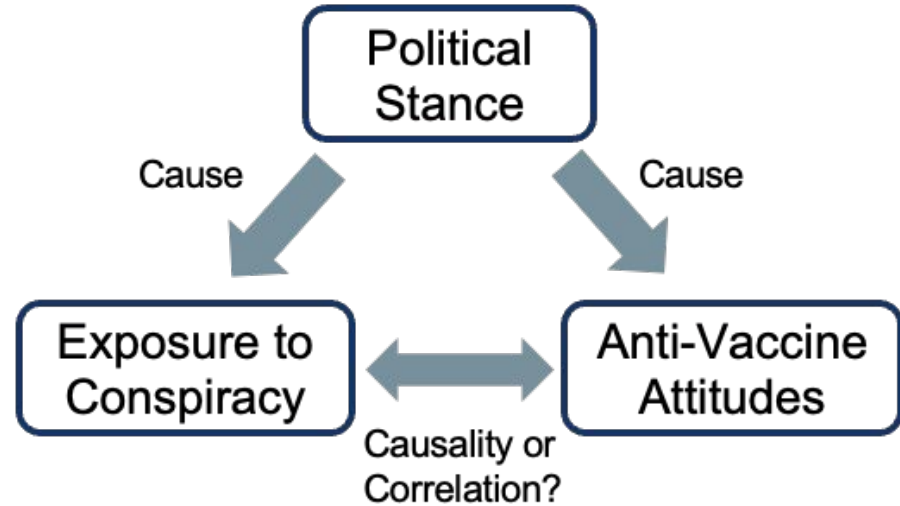


To combat confounders in causal inference, researchers developed numerous statistical and mathematical methodologies. One common strategy is to rebalance the data distribution to mitigate the influence from confounders, i.e., rebalance to ensure $P(\text{Treatment}|\text{Confounders}) = P(\text{Treatment})$. Common methods include:

- **Data Reweighting or Sampling:** This strategy try to change the data distribution to make treatment unpredictable from confounders
- **Balanced Representation Learning:** This strategy aims at learning a neural network with a representation space for covariate and confounders where their correlations are mitigated, i.e. ensure $P(\text{Treatment}|\text{H}(\text{Confounders})) = P(\text{Treatment})$ where H is the neural encoder.

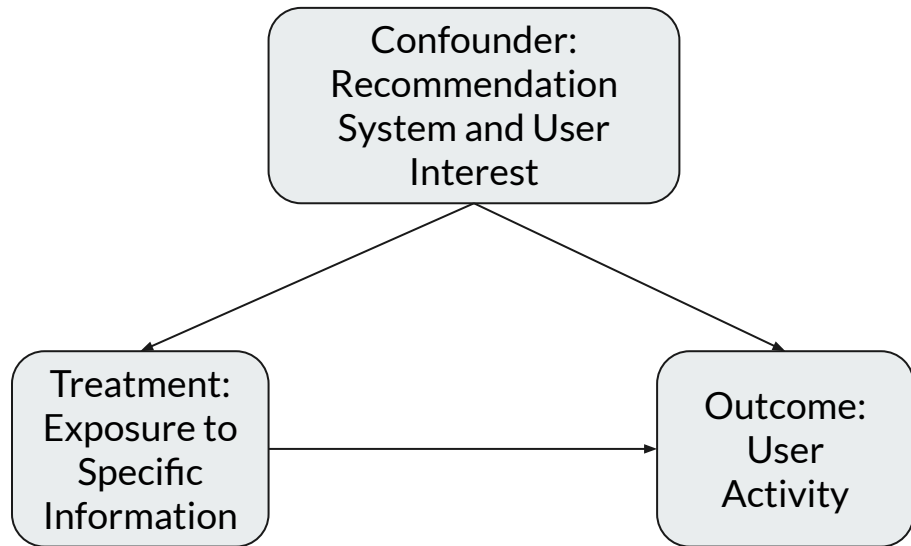
Causal Inference of Misinformation on Social Media

- Introduction to Causal Inference
- Applications of Causal Inference on Social Media
- Application of LLM in Causal Inference on Social Media



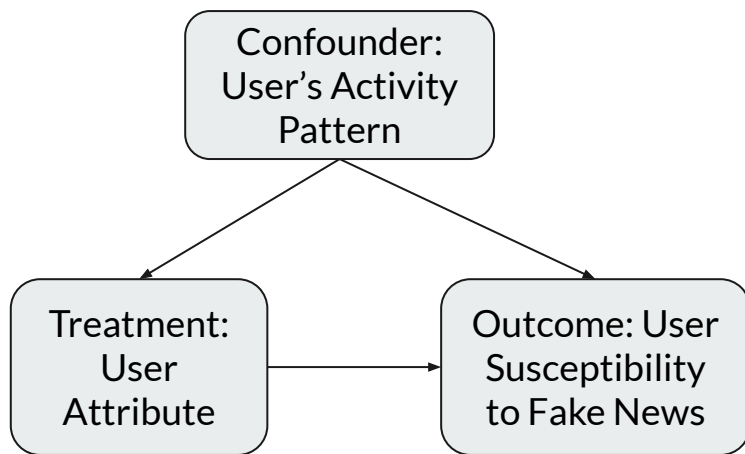
Confounders in Social Media: User Activity Pattern

Modern Social Media platforms commonly apply **personalized recommendation systems**, which recommend contents to individual users based on their **activity pattern** (such as interested topic). This causes Information Cocoons (Echo Chamber).

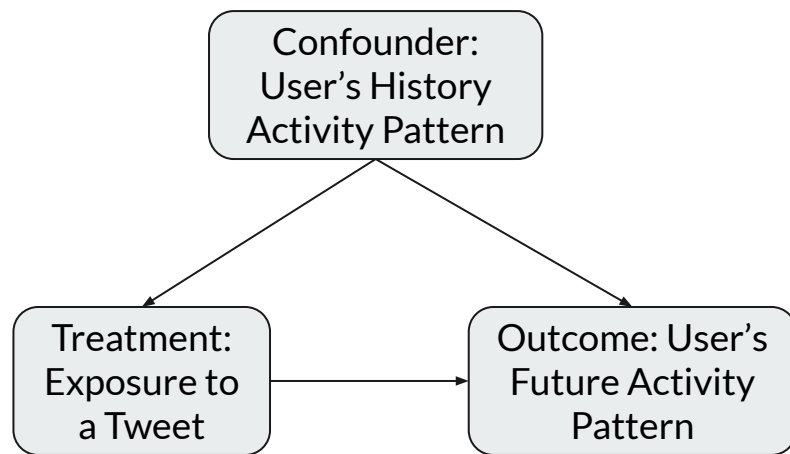


Application of Causal Inference on Social Media

Causal Inference help us understand how misinformation and manipulated contents causally affect online user's activities by mitigating the confounding effect. Common applications include following two scenarios:



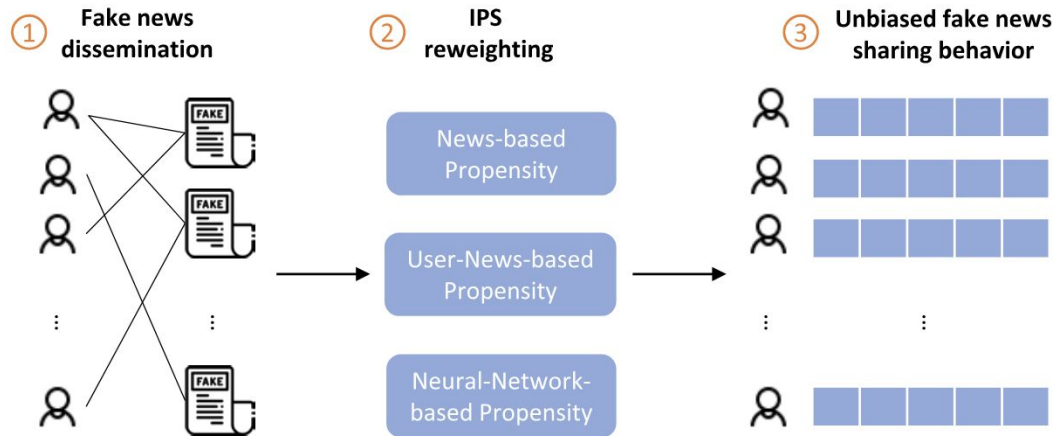
Causal Inference on a user attribute and user susceptibility



Causal Inference on how a Tweet influence user's future activity

Application 1: Learning Unbiased Sharing Behavior

In social media analysis, we hope to learn user susceptibility to misinformation. However, in this causal relationship, user's activity patterns work as a confounder, since it significantly influence the preference of recommendation system. To address this challenge, [Cheng, Lu et al] provide framework:



Application 1: Learning Unbiased Sharing Behavior



The authors introduce two binary variables in their paper: “interestingness $R_{ui} \in \{0, 1\}$ and exposure $O_{ui} \in \{0, 1\}$. $R_{ui} = 1(0)$ indicates u is interested (not interested) in i ; $O_{ui} = 1$ denotes user u was exposed to fake news i and $O_{ui} = 0$, otherwise”.

In **ideal** scenario like random controlled trials, R_{ui} should be independent to exposure O_{ui} . Thus, we have following equations:

$$Y_{ui} = O_{ui} \cdot R_{ui}, \quad (1)$$

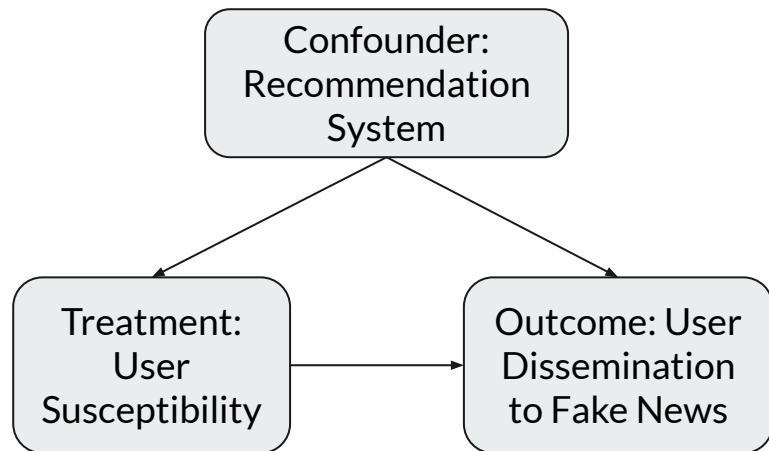
$$\begin{aligned} P(Y_{ui} = 1) &= P(O_{ui} = 1) \cdot P(R_{ui} = 1), \\ &= \theta_{ui} \cdot \gamma_{ui} \quad \theta_{ui} > 0; \gamma_{ui} > 0; \forall (u, i) \in \mathcal{Y}, \end{aligned} \quad (2)$$

Application 1: Learning Unbiased Sharing Behavior

However, due to modern recommendation systems, $P(O_{ui})$ may not be equal balanced (i.e., exposure probability is decided by user and news). Thus, the authors apply Inverse Propensity Scoring to reweight the samples involved in loss functions:

$$\hat{\mathcal{L}}(\hat{S}) = \frac{1}{|\mathcal{D}_{pair}|} \sum_{(u,i,j) \in \mathcal{D}_{pair}} \frac{Y_{ui}}{\theta_{ui}} \left(1 - \frac{Y_{uj}}{\theta_{uj}}\right) \ell(\hat{S}_{uij})$$

Where u is a user, i is a positive news (disseminated by user u), j is a negative news (not disseminated by user u), and Θ_{ui} is $P(O_{ui}=1)$ that is estimated by a Propensity Score Model



Application 1: Learning Unbiased Sharing Behavior

(a) Recall@K with K=20,40,60,80.

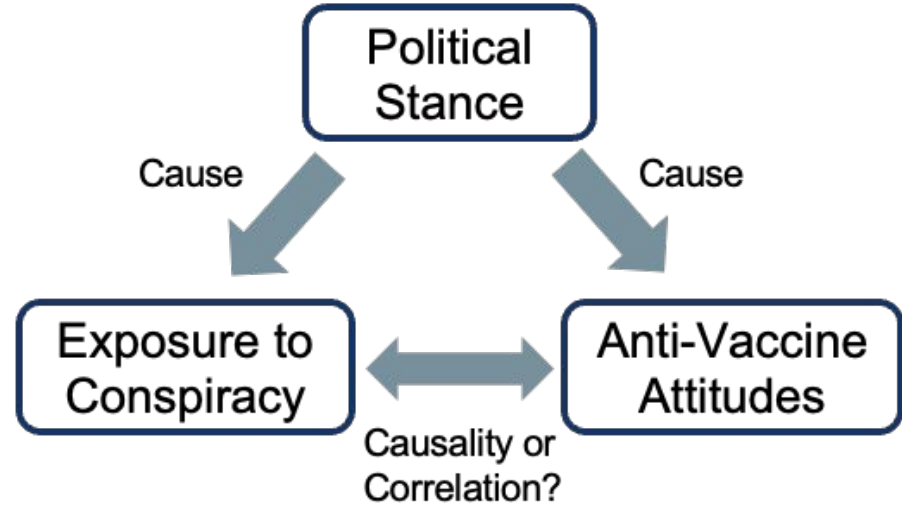
| K | 20 | 40 | 60 | 80 |
|-----------|--------------------------------|--------------------------------|--------------------------------|--------------------------------|
| BPRMF | 12.36 | 22.18 | 31.10 | 39.51 |
| BPRMF-N | 14.45 ^{↑16.9%} | 25.11 ^{↑13.2%} | 34.34 ^{↑10.4%} | 42.72 ^{↑8.1%} |
| BPRMF-U | 14.78 ^{↑19.6%} | 25.65 ^{↑15.6%} | 34.91 ^{↑12.2%} | 43.63 ^{↑10.4%} |
| BPRMF-Neu | 14.90 ^{↑20.6%} | 25.83 ^{↑16.5%} | 35.13 ^{↑13.0%} | 43.55 ^{↑10.2%} |

(b) NDCG@K with K=20,40,60,80.

| K | 20 | 40 | 60 | 80 |
|-----------|-------------------------------|-------------------------------|--------------------------------|--------------------------------|
| BPRMF | 5.33 | 7.51 | 9.22 | 10.71 |
| BPRMF-N | 6.39 ^{↑19.9%} | 8.73 ^{↑16.2%} | 10.49 ^{↑13.8%} | 11.97 ^{↑11.8%} |
| BPRMF-U | 6.54 ^{↑22.7%} | 8.92 ^{↑18.8%} | 10.69 ^{↑15.9%} | 12.21 ^{↑14.0%} |
| BPRMF-Neu | 6.53 ^{↑22.5%} | 8.93 ^{↑18.9%} | 10.71 ^{↑16.2%} | 12.19 ^{↑13.8%} |

Causal Inference of Misinformation on Social Media

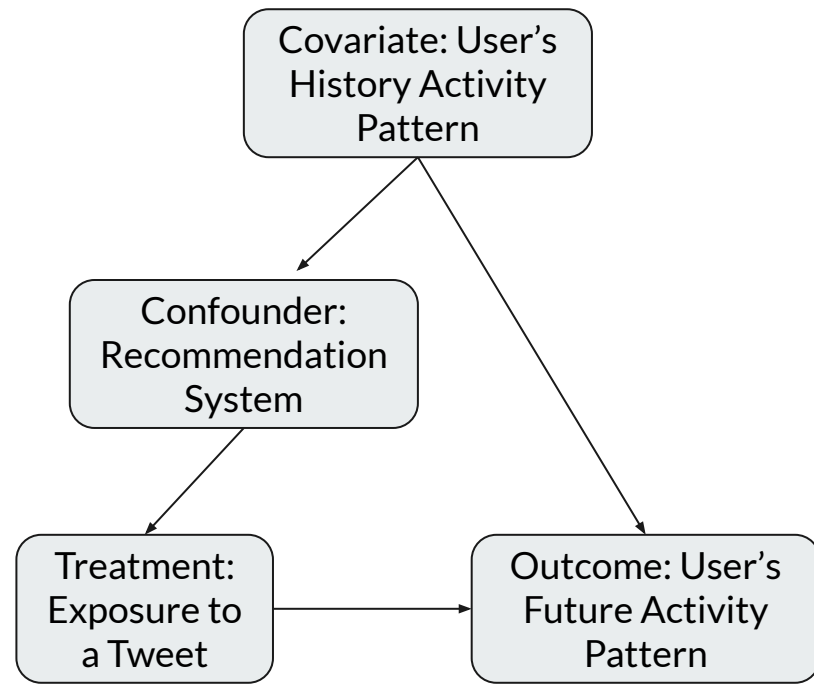
- Introduction to Causal Inference
- Applications of Causal Inference on Social Media
- **Application of LLM in Causal Inference on Social Media**



Application of LLM: Estimate Misinformation Impact

Misinformation campaigns are manipulating public opinions on hot-spot topics, e.g. COVID-19 epidemic and vaccines.

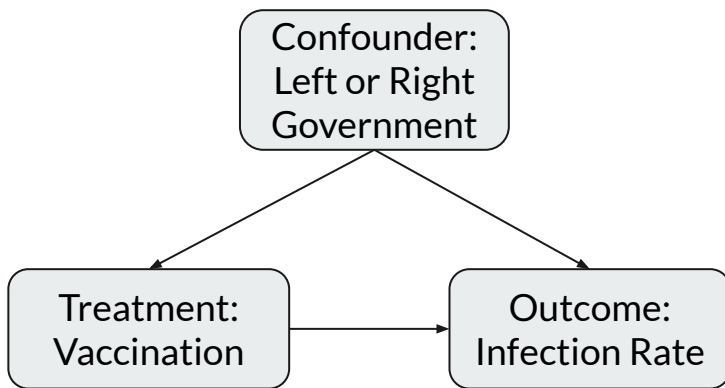
To design mitigation strategies to reduce user susceptibility to misinformation, we need to understand how misinformation influence user beliefs and activities.



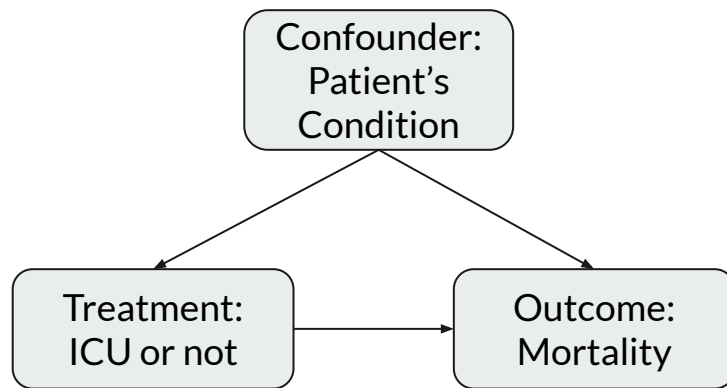
Causal Inference on how a Tweet influence user's future activity

Application of LLM: Estimate Misinformation Impact

However, existing causal inference mainly tackle confounders of indicator variable (e.g., isolation policy or not) or attribute data (e.g., patient's condition). For such data, we can easily calculate $P(\text{Treatment}|\text{Confounder})$ over all samples and reweight them



Example of Indicator Variable as confounder



Example of Attribute Data as confounder

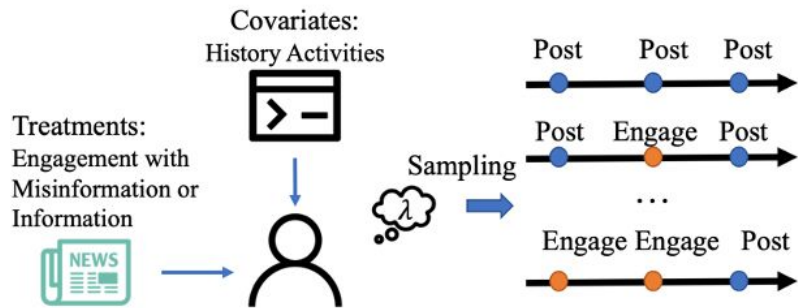
Application of LLM: Estimate Misinformation Impact

But for social media data, the confounders are hidden behind users' history activities, which are *Sequences* of events attached with *User's Comments* and *Timestamp*:

$$S = [(u_1, t_1), (u_2, t_2), (u_3, t_3), \dots (u_n, t_n)]$$

$$P(S) = P(u_1, t_1)P(u_2, t_2|u_1, t_1)P(u_3, t_3|u_2, t_2, u_1, t_1)\dots$$

whose probability is hard to estimate. Thus, balancing data with confounders hidden in activity data is challenging to existing balance methods.

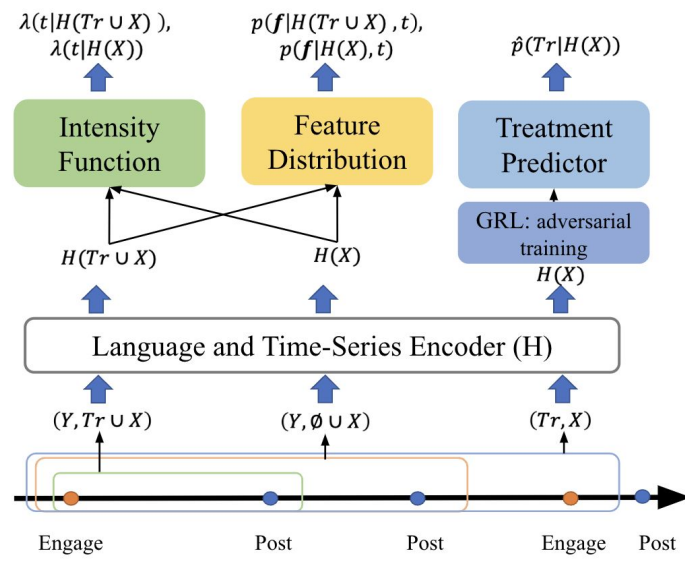


Causal Structure Model: Describe the impact of a piece of information on user

Application of LLM: Estimate Misinformation Impact

To tackle this challenge, instead of balancing data samples, we propose to learn a balanced representation via **LLM + Time Series Encoder**:

- User's comments are encoded by a LLM and then forwarded into a Time Series Encoder (like RNN) together with timestamps.
- We then fine-tune the encoder to jointly **maximize the likelihood** of observed data and **minimize the prediction accuracy** of treatment given the encoding $H(x)$ (i.e. making $P(\text{Treatment}|H(\text{Confounders})) = P(\text{Treatment})$).



Counterfactual Neural Temporal Point Process: Estimating the impact of misinformation/information

Application of LLM: Estimate Misinformation Impact

| Method | Accuracy \uparrow | RAE \downarrow | RRSE \downarrow | Decoder Inference Time |
|---------------------------|---------------------|------------------|-------------------|------------------------|
| FullyNN | 73.0% | 0.865 | 0.901 | 7.13ms |
| CNTPP-VAE (Approximation) | 85.9% | 0.279 | 0.503 | 4.05ms |
| CNTPP-VAE (Sampling) | 87.8% | 0.237 | 0.454 | 29.34ms |
| CNTPP(Ours) | 88.0% | 0.234 | 0.448 | 7.12ms |

| Method | MatDis \downarrow | LinCor \uparrow |
|---------------------------|---------------------|-------------------|
| Neural-CIP | 0.90 | 0.04 |
| FullyNN | 0.93 | 0.236 |
| CNTPP-VAE (Approximation) | 0.84 | 0.303 |
| CNTPP-VAE (Sampling) | 0.76 | 0.287 |
| CNTPP (Ours) | 0.77 | 0.310 |

Roadmaps: What can we do for future?



- Enabling Decoder-only LLM for causal learning:

In the aforementioned paradigm, we apply LLM as an encoder to encode the text and timestamp. This is only applicable for Encoder-only (like BERT) and Encoder-Decoder LLM (like T-5). However, Decoder-only LLM has recently attracted more attention due to its stronger performance in text modeling.

Challenge: How to apply Decoder-only LLM for causal inference?

Large Language Model and Misinformation Detection

- Introduction to LLM-based Misinformation Detection
- Real-time Dataset Construction
- Incorporate LLM with Structured Modality Data
- Prompting and Reasoning Strategy for Misinformation Detection



Eliot Higgins @EliotHiggins · Mar 20



Making pictures of Trump getting arrested while waiting for Trump's arrest.



1.8K

7.6K

39K

6.7M



LLM-based Misinformation Detection

LLM's strong capacity in tackling natural language inspired researchers to apply LLM to detect misinformation. Existing works mainly apply the following three types of paradigm to apply LLM in misinformation detection:

1. Detection with Internal Knowledge
2. Detection with External Knowledge
3. LLM as a supporter



Please determine whether or not the following sentence is misinformation and give a concise explanation: "Vaccines contain dangerous toxins like mercury and aluminum that are harmful to our bodies"



The sentence "Vaccines contain dangerous toxins like mercury and aluminum that are harmful to our bodies" is misleading and contains elements of misinformation. Here's a concise explanation:

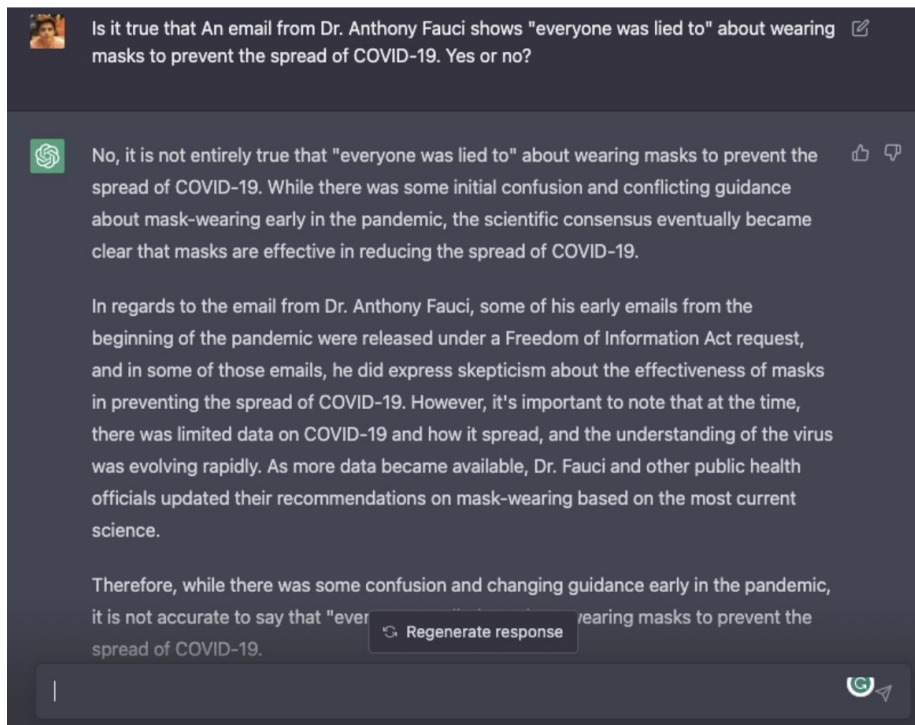


1. **Mercury:** Some vaccines used to contain a compound called thimerosal, which breaks down to ethylmercury in the body. Ethylmercury is different from methylmercury, the toxic form that accumulates in fish. Ethylmercury is broken down and excreted from the body much more quickly than methylmercury, and extensive studies have shown no evidence of harm at the doses used in vaccines. Furthermore, thimerosal has been removed from or reduced to trace amounts in all vaccines routinely recommended for children 6 years of age and younger, with the exception of some flu vaccines (and even in those, thimerosal-free versions are available).
2. **Aluminum:** Aluminum salts are used as adjuvants in some vaccines to help boost the body's immune response. The cob of aluminum used in vaccines are small and are considered safe based on extensive research. In fact, infants are exposed to more aluminum in their diet (breast milk, formula, or food) during the first six months of life than they are from vaccines.

It's essential to rely on well-conducted scientific research when evaluating the safety and components of vaccines. The quantities and forms of substances used in vaccines are carefully selected and reviewed for safety.

Misinformation Detection via Internal Knowledge

The most straightforward way to detect misinformation with LLM is to attach a well-designed fixed prompt onto the content of misinformation and forward them into LLM to make decisions. The LLM will use its **language capacity** and its **internal knowledge** acquired from pre-training corpus



Is it true that An email from Dr. Anthony Fauci shows "everyone was lied to" about wearing masks to prevent the spread of COVID-19. Yes or no?

No, it is not entirely true that "everyone was lied to" about wearing masks to prevent the spread of COVID-19. While there was some initial confusion and conflicting guidance about mask-wearing early in the pandemic, the scientific consensus eventually became clear that masks are effective in reducing the spread of COVID-19.

In regards to the email from Dr. Anthony Fauci, some of his early emails from the beginning of the pandemic were released under a Freedom of Information Act request, and in some of those emails, he did express skepticism about the effectiveness of masks in preventing the spread of COVID-19. However, it's important to note that at the time, there was limited data on COVID-19 and how it spread, and the understanding of the virus was evolving rapidly. As more data became available, Dr. Fauci and other public health officials updated their recommendations on mask-wearing based on the most current science.

Therefore, while there was some confusion and changing guidance early in the pandemic, it is not accurate to say that "everyone was lied to" about wearing masks to prevent the spread of COVID-19.

Regenerate response

Misinformation Detection via Internal Knowledge

Such paradigm is applicable when:

1. The misinformation can be recognized linguistically (e.g., emotional bias and internal contradiction) rather than factually
2. The associated fact has been covered by the pre-training and fine-tuning corpus (like urban legends)

| Option | Reason | Description |
|--------|---|--|
| A | Emotional bias or misleading intent | This explanation suggests that fake news is characterized by an emotional bias, which can include an excessively aggressive portrayal of a subject or an attempt to manipulate readers to achieve a hidden agenda. |
| B | Lack of evidence or credible sources | This reason indicates that fake news lacks credible evidence to support its claims. |
| C | Conflicting facts | This reason suggests that fake news conflicts with established facts, such as wrong information about people or events. |
| D | Informal statements, expressions, or vague language | This reason highlights that the language used in fake news may not be formal, or may be vague or ambiguous. |
| E | Insufficient supporting materials | This reason indicates that although the news may have mentioned the source of an event or provided relevant evidence, the evidence is not sufficient to support its claims. |
| F | Lack of context or taken out of context | This reason indicates that fake news may lack relevant context, such as comments, retweets and user information that provide additional information. |
| G | Misinterpretation or misquotation | This reason suggests that fake news may misinterpret or misquote facts, leading to inaccurate or false claims. |
| H | Oversimplification or exaggeration | This reason highlights that fake news may oversimplify or exaggerate information, leading to false claims. |
| I | Doctored images or videos | This reason indicates that the images or videos mentioned in the news text may be altered or misrepresented, making them untrustworthy. |
| J | Other | ChatGPT must specify a reason if the above options don't match its answer. |

Misinformation Detection via Internal Knowledge



However, sometimes ChatGPT can not respond correctly or clearly.

Researchers make statistics on the misinformation samples from different datasets that ChatGPT gives “unclear” response. They require ChatGPT to make a choice within 4 reasons why they refuse to provide clearly response.

A: External knowledge refers to factual information, expert suggestions, or data reliability.

B: Multimodal information includes images, videos, or audio.

C: Context information encompasses comments, reposts, post time or post location.

D: Speaker’s information includes user actions, information from social media accounts, or the user’s history of posts.

Misinformation Detection via Internal Knowledge

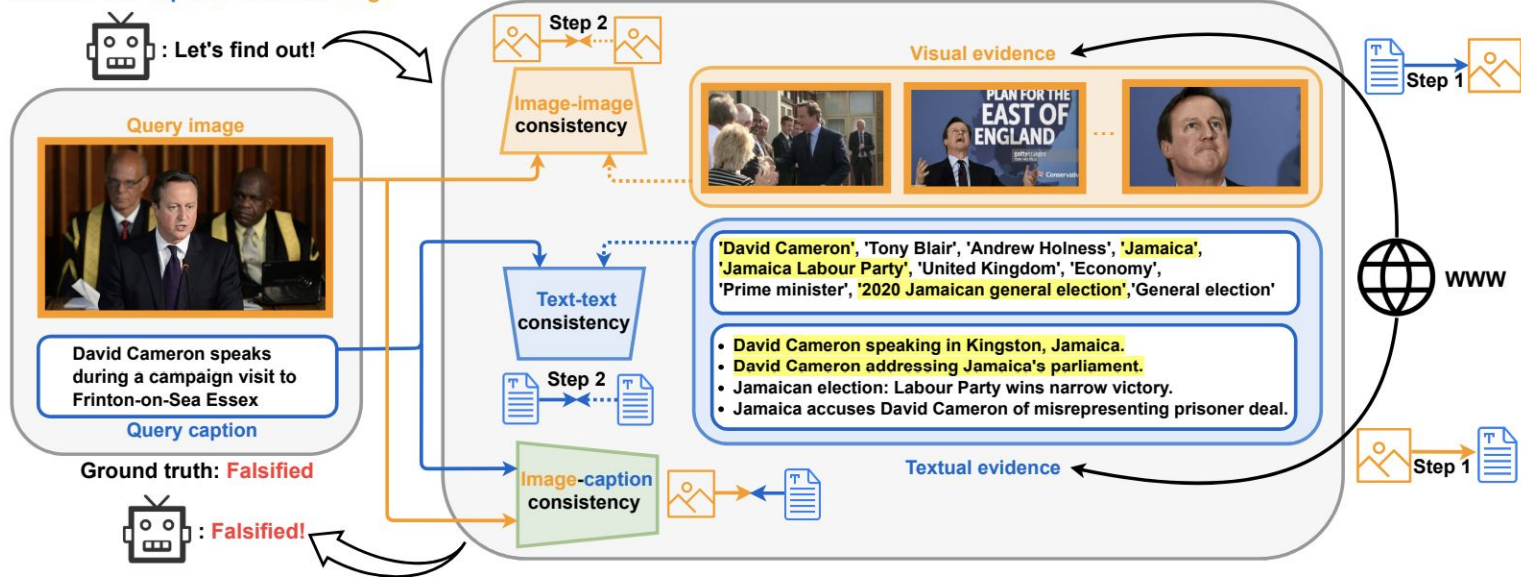
| Dataset | | A | B | C | D | AB | AC | AD | BC | BD | CD |
|---------------|-------|-------|-------|-------|-------|------|------|------|------|------|------|
| CHINESE RUMOR | (w/o) | 27.27 | 17.11 | 16.22 | 18.36 | 3.92 | 4.99 | 4.99 | 2.50 | 2.67 | 1.97 |
| | (w/) | 35.03 | 12.69 | 20.30 | 18.78 | 1.52 | 3.55 | 5.08 | 0.51 | 1.52 | 1.02 |
| LIAR | (w/o) | 31.76 | 7.03 | 18.46 | 21.32 | 1.98 | 6.37 | 7.36 | 1.65 | 0.99 | 3.08 |
| | (w/) | 31.76 | 12.83 | 17.35 | 19.43 | 2.80 | 4.87 | 6.24 | 1.50 | 1.28 | 1.94 |
| WEIBO21 | (w/o) | 30.10 | 14.26 | 14.85 | 21.78 | 2.38 | 4.16 | 7.32 | 1.98 | 1.78 | 1.39 |
| | (w/) | 34.21 | 12.39 | 19.20 | 17.63 | 2.79 | 4.71 | 5.41 | 1.22 | 0.87 | 1.57 |
| COVID-19 | | 31.43 | 12.56 | 17.46 | 19.33 | 2.92 | 5.14 | 6.19 | 1.46 | 1.29 | 2.22 |
| FAKE NEWS NET | | 29.97 | 11.36 | 17.98 | 18.93 | 3.47 | 6.31 | 5.99 | 1.26 | 1.26 | 3.47 |
| KAGGLE | | 22.22 | 22.59 | 14.81 | 21.85 | 2.96 | 2.96 | 4.44 | 2.59 | 3.35 | 2.23 |
| TWITTER 15&16 | | 28.90 | 12.93 | 17.87 | 20.15 | 1.90 | 6.08 | 5.70 | 1.52 | 2.66 | 2.28 |

Option A (Requiring external knowledge) is always a dominant reason across different dataset with different languages

Misinformation Detection via External Knowledge

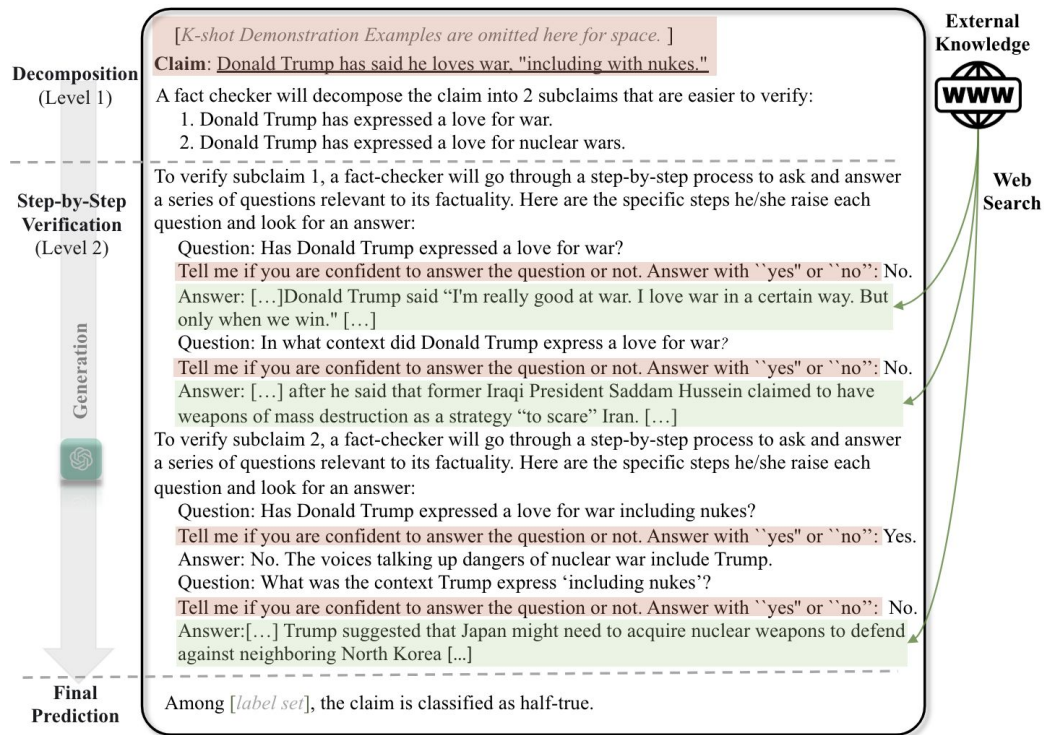
To incorporate external knowledge with LLM-based misinformation detection, researchers propose to borrow the idea in retrieval augmented generation:

Q: Does this caption match its image?



Misinformation Detection via External Knowledge

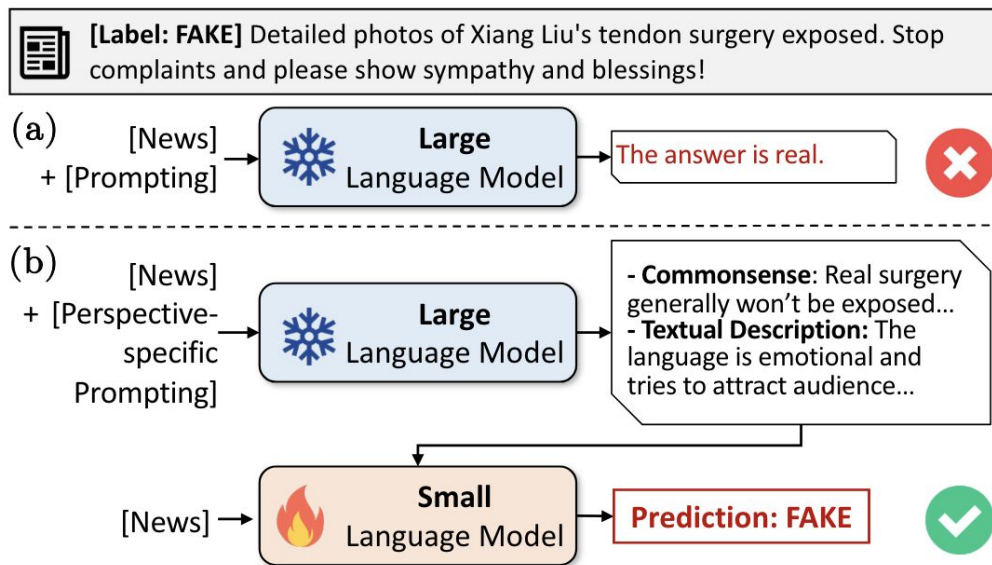
The central idea of such retrieval-augmented (or search-augmented) detection paradigm generally convert the content to be verified as a series of queries. After that, the model call an API to acquire relevant documents from a database or Internet as augmented context for verification.



Misinformation Detection with LLM as Supporters

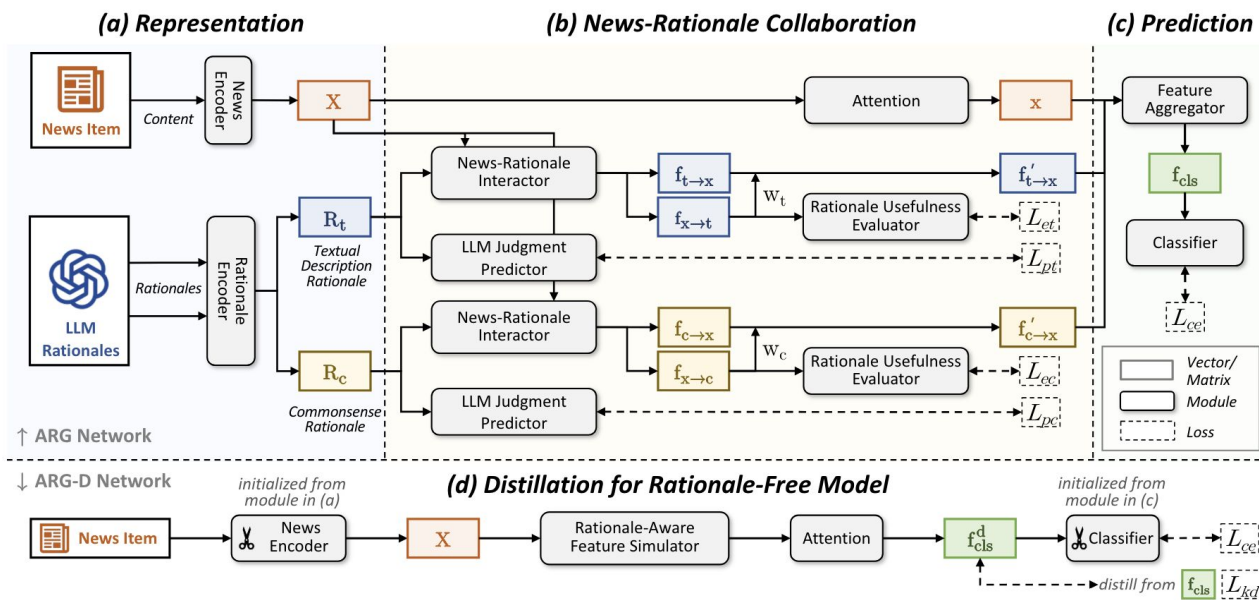
LLM's great capacity in general-purpose incontext learning enable it to acquire knowledge from few samples.

However, in some cases, we may expect to learn a more targeted misinformation detector on a middle-scale dataset while preserving the knowledge in LLM



Misinformation Detection with LLM as Supporters

In such as case, we can apply LLM as an advisor to generate rationales for a small language model. Then we can fine-tune the small language model to learn from the LLM rationales:



Misinformation Detection with LLM as Supporters

| Model | | Chinese | | | | English | | | |
|---------------------------------------|---------------------------------------|----------------|----------------|--------------------|--------------------|----------------|----------------|--------------------|--------------------|
| | | macF1 | Acc. | F1 _{real} | F1 _{fake} | macF1 | Acc. | F1 _{real} | F1 _{fake} |
| G1: LLM-Only | GPT-3.5-turbo | 0.725 | 0.734 | 0.774 | 0.676 | 0.702 | 0.813 | 0.884 | 0.519 |
| G2: SLM-Only | Baseline | 0.753 | 0.754 | 0.769 | 0.737 | 0.765 | 0.862 | 0.916 | 0.615 |
| | EANN _T | 0.754 | 0.756 | 0.773 | 0.736 | 0.763 | 0.864 | 0.918 | 0.608 |
| | Publisher-Emo | 0.761 | 0.763 | 0.784 | 0.738 | 0.766 | 0.868 | 0.920 | 0.611 |
| | ENDEF | 0.765 | 0.766 | 0.779 | 0.751 | 0.768 | 0.865 | 0.918 | 0.618 |
| G3: LLM+SLM | Baseline + Rationale | 0.767 | 0.769 | 0.787 | 0.748 | 0.777 | 0.870 | 0.921 | 0.633 |
| | SuperICL | 0.757 | 0.759 | 0.779 | 0.734 | 0.736 | 0.864 | 0.920 | 0.551 |
| | ARG | 0.784 | 0.786 | 0.804 | 0.764 | 0.790 | <u>0.878</u> | 0.926 | 0.653 |
| | <i>(Relative Impr. over Baseline)</i> | <i>(+4.2%)</i> | <i>(+4.3%)</i> | <i>(+4.6%)</i> | <i>(+3.8%)</i> | <i>(+3.2%)</i> | <i>(+1.8%)</i> | <i>(+1.1%)</i> | <i>(+6.3%)</i> |
| | w/o LLM Judgment Predictor | 0.773 | 0.774 | 0.789 | 0.756 | <u>0.786</u> | 0.880 | 0.928 | 0.645 |
| | w/o Rationale Usefulness Evaluator | <u>0.781</u> | <u>0.783</u> | 0.801 | 0.761 | 0.782 | 0.873 | 0.923 | 0.641 |
| | w/o Predictor & Evaluator | 0.769 | 0.770 | 0.782 | 0.756 | 0.780 | 0.874 | 0.923 | 0.637 |
| | ARG-D | 0.771 | 0.772 | 0.785 | 0.756 | 0.778 | 0.870 | 0.921 | 0.634 |
| <i>(Relative Impr. over Baseline)</i> | <i>(+2.4%)</i> | <i>(+2.3%)</i> | <i>(+2.1%)</i> | <i>(+2.6%)</i> | <i>(+1.6%)</i> | <i>(+0.9%)</i> | <i>(+0.6%)</i> | <i>(+3.2%)</i> | |

Challenges in LLM-based Misinformation Detection



In the above three paradigms of LLM-based Misinformation Detection, there are still challenges:

- **Reliable Real-time Data Retrieval:**

When detecting misinformation related to breaking news, LLM requires external knowledge to make prediction. However, acquiring reliable external knowledge is challenging because it requires real-time verification to the retrieved data.

Challenge: How to construct reliable external knowledge base in real time?

Challenges in LLM-based Misinformation Detection



In the above three paradigms of LLM-based Misinformation Detection, there are still challenges:

- **Multimodal Context from Retrieval:**

When detecting misinformation, useful retrieval responses may contain multimodal data. Unlike classical Retrieval Augmented Generation where multimodal information mainly refer to vision data, in Retrieval Augmented Misinformation Detection, **Structured data**, like **Graph-structured data** (e.g., social network structure) and **Table data** (such as Financial Statements) will be as important as vision data.

Challenge: How to enable LLM to reason on context with such data?

Challenges in LLM-based Misinformation Detection



In the above three paradigms of LLM-based Misinformation Detection, there are still challenges:

- **Prompting and Reasoning Strategy for Misinformation Detection:**

Existing works have shown that differences in prompting and reasoning strategy have significant performance gap in misinformation detection. For example, Chain-of-Thoughts can increase the performance for different kinds of LLM, including ChatGPT and LLaMa.

Challenge: How to develop more effective prompting and reasoning strategy targeting on misinformation detection?

Large Language Model and Misinformation Detection

- Introduction to LLM-based Misinformation Detection
- **Real-time Dataset Construction**
- Incorporate LLM with Structured Modality Data
- Prompting and Reasoning Strategy for Misinformation Detection



Eliot Higgins @EliotHiggins · Mar 20



Making pictures of Trump getting arrested while waiting for Trump's arrest.



1.8K

7.6K

39K

6.7M



Data Collection on Emerging Events

To assist LLM to tackle emerging events happening later than its training, we need real-time data collection to construct reliable external knowledge base. However, we are confronted of challenges:

- **Lack of Reliable News Resource:**
In the early stage of news propagation, only very a few medias can acquire reliable information from their frontline reporter
- **High Cost in Human Verification:**
To verify the information in real time, we need a lot of human verifiers.

USC Melady Lab Coronavirus Tweets Misinformation Analysis Trend Analysis Geo Analysis Contact

Coronavirus on Social Media: Misinformation Analysis

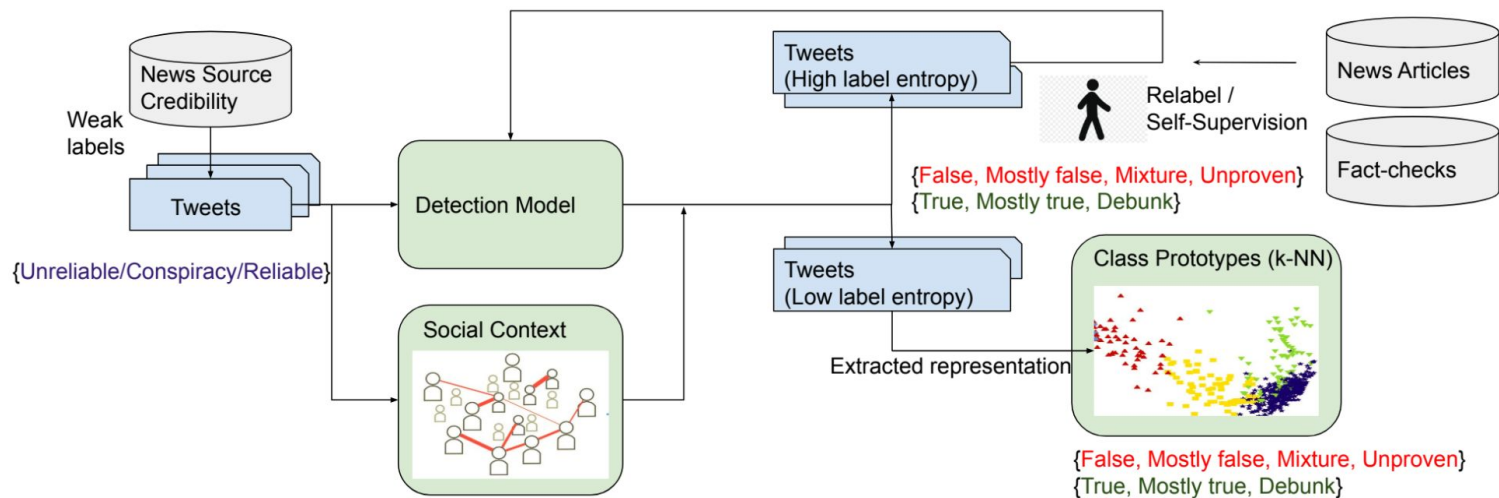
The 'infodemic' on Social Media impacts health and welfare of society, both at a local and global scale. We identify unreliable, misleading and clickbait information shared on Twitter to combat misinformation on Coronavirus from Mar 1- Apr 11.

Distribution % (2020-03-01 - 2020-04-11)

| Website Type | Tweet | Twitter Link |
|----------------------------|--|----------------------|
| conspiracy-unreliable-bias | Convicted Murderer Re-Arrested After Being Freed from Prison Over Coronavirus - https://t.co/hL6svBAQeT ✓ Via @NeonNettle #NeonNettle [2020-04-11T00:04:27] | link |
| conspiracy-unreliable-bias | 150 Members of Saudi Arabia Royal Family are Infected with Coronavirus - https://t.co/Yw4cvSYi5X ✓ Via - @NeonNettle #NeonNettle [2020-04-11T00:05:03] | link |
| unreliable | MORE FRAUD EXPOSED: Washington State Coronavirus Field Hospital Will Be Dismantled Without Treating a Single Patient https://t.co/QQmLjbo5kR [2020-04-11T00:05:46] | link |

Real-time External Knowledge Base Construction

To tackle the aforementioned challenges, researchers developed real-time algorithm to construct Large-Scale Misinformation Labeled Datasets based on Human-Computer Collaboration:



Real-time External Knowledge Base Construction

Experiment results show that such human-computer collaboration loop can significantly improve the labeling accuracy

| Experiment | AP | AUC | F1 | Macro F1 |
|--------------------------------------|------------------------------------|-----------------------------------|------------------------------------|------------------------------------|
| Weak labels | 0.722 ± 0.03 | 0.876 ± 0.01 | 0.774 ± 0.02 | 0.812 ± 0.01 |
| Self-training (iteration 1) | 0.768 ± 0.01 | 0.888 ± 0.0 | 0.812 ± 0.01 | 0.842 ± 0.01 |
| Self-training (iteration 2) | 0.775 ± 0.02 | 0.891 ± 0.0 | 0.811 ± 0.01 | 0.842 ± 0.01 |
| Social-context only | 0.764 ± 0.02 | 0.891 ± 0.01 | 0.810 ± 0.01 | 0.837 ± 0.01 |
| Social+Detection model | 0.785 ± 0.02 | 0.895 ± 0.0 | 0.813 ± 0.01 | 0.842 ± 0.01 |
| Social+Detection (+label correction) | 0.800 ± 0.01 | 0.895 ± 0.0 | 0.818 ± 0.01 | 0.845 ± 0.01 |

Large Language Model and Misinformation Detection

- Introduction to LLM-based Misinformation Detection
- Real-time Dataset Construction
- **Incorporate LLM with Structured Modality Data**
- Prompting and Reasoning Strategy for Misinformation Detection



Eliot Higgins @EliotHiggins · Mar 20



Making pictures of Trump getting arrested while waiting for Trump's arrest.



1.8K

7.6K

39K

6.7M



(Semi-) Structured Data are Ubiquitous



Structured data, like **Graph-structured data** (e.g., social network structure and knowledge graph) and **Tabular data** (such as Financial Statements) widely exist on Internet and various databases for retrieval:

- **Tabular Data**
 - Widely used in relational databases, and spreadsheet applications
 - Basic data structure for data analysis
- **Graph Data**
 - Ubiquitous in real world, e.g., knowledge graph and citation network
 - Basic data structure to representing complex relationships

LLM Meets Structured Data

To apply structured data for misinformation detection, we need to tackle the following scenarios:

Los Angeles Lakers [edit]

The rivalry between the Lakers and Mavericks began in the 1980s during the Lakers' Showtime era. In the 1984 NBA playoffs, the Lakers won over the Mavericks in the Western Conference Semi-finals with the series 4-1. In 1986, they met again in the semi-finals and the Lakers won the series 4-2. In the 1988 NBA playoffs, the Lakers, who were the defending champions, defeated the Mavericks in a seven-game series in the Western Conference Finals and eventually became the back-to-back champions after winning the NBA Finals. In 2011, the Mavericks met the two-time defending champions Lakers in the semi-finals and rejected their three-peat by sweeping them in four games and eventually won the NBA Finals. In Game 2 of the 2011 NBA playoffs, Ron Artest was ejected for hitting J. J. Barea and suspended for Game 3. In Game 4, Lamar Odom was ejected for pushing Dirk Nowitzki and then Andrew Bynum was ejected for elbowing J. J. Barea. Both players received flagrant foul 2. After the 2011 NBA lockout, Lamar Odom was sent to the Mavericks after he requested to leave the Lakers due to the vetoed Chris Paul trade. Andrew Bynum made a formal apology for what he did to Barea before the start of the following season.

Season-by-season record [edit]

List of the last five seasons completed by the Mavericks. For the full season-by-season history, see List of Dallas Mavericks seasons.

Note: GP = Games played, W = Wins, L = Losses, W-L% = Winning percentage

| Season | GP | W | L | W-L% | Finish | Playoffs |
|---------|----|----|----|------|----------------|---|
| 2019-20 | 75 | 41 | 32 | .573 | 2nd, Southwest | Lost in first round, 2-4 (Clippers) |
| 2020-21 | 72 | 42 | 30 | .583 | 1st, Southwest | Lost in first round, 3-4 (Clippers) |
| 2021-22 | 82 | 52 | 30 | .634 | 2nd, Southwest | Lost in conference finals, 1-4 (Warriors) |
| 2022-23 | 82 | 38 | 44 | .463 | 3rd, Southwest | Did not qualify |
| 2023-24 | 82 | 50 | 32 | .610 | 1st, Southwest | TBD |

Wiki Page with Table

Text & Structured Data appears simultaneously as context

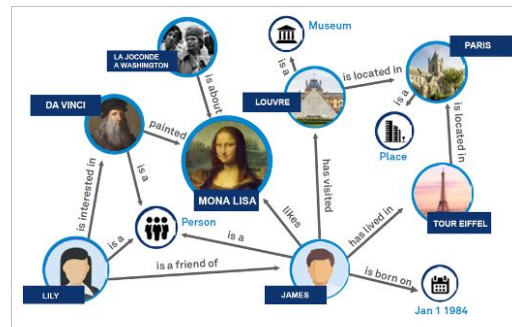
References [edit]

- ¹ ^ "NBA.com" (PDF). National Basketball Association. Archived from the original on December 2, 2022. Retrieved December 2, 2022.
- ² ^ "History: Team by Team" (PDF). 2019-20 Official NBA Guide (PDF). NBA Properties, Inc. October 17, 2019. Archived from the original on October 26, 2019. Retrieved November 16, 2020.
- ³ ^ "Frequently Asked Questions" (PDF). Mavs.com. NBA Media Ventures, LLC. Archived from the original on September 26, 2022. Retrieved October 16, 2022.
- ⁴ ^ "Dallas Mavericks Reproduction and Usage Guideline Sheet" (PDF). NBA Properties, Inc. Archived from the original on June 30, 2016. Retrieved August 10, 2016.
- ⁵ ^ Sefko, Eddie (January 17, 2020). "Mavericks Chime in, announce partnership with new jersey-patch sponsor" (PDF). Mavs.com. NBA Media Ventures. Archived from the original on May 30, 2022. Retrieved February 10, 2020.
- ⁶ ^ ^ ^ ^ ^ ^ ^ ^ "Mark Cuban's sale of Dallas Mavericks unanimously approved by NBA" (PDF). Dallas News. December 27, 2023. Archived from the original on February 5, 2024. Retrieved February 5, 2024.
- ⁷ ^ Price, Dvain (December 28, 2023). "NBA Board of Governors approve sale of Mavericks" (PDF). Mavs.com. NBA Media Ventures, LLC. Retrieved January 5, 2024.



Citation Relationship

Structured data (e.g., knowledge graph) for LLM Grounding



Knowledge Graph

LLM with Reasoning Ability on Structured Data Structured Data

LLM Meets Structured Data

- The main challenge we are confronted are brought by the difference between Plain text vs. Structured data:
 - LLM pertained by plain text
 - Can LLM understand structured data?
 - How to input structured data to LLM?

What we see:

| Name | Height | Weight |
|------|----------|---------|
| Tom | 5 feet 7 | 130 lbs |
| Jane | 4 feet 9 | 90 lbs |
| Mike | 6 feet 1 | 150 lbs |
| Lana | 5 feet 6 | 123 lbs |

What LLM see:

```
<s> | Name | Height | Weight |<0x0A>|---|---|---|<0x0A>| Tom |  
5 feet 7 | 130 lbs |<0x0A>| Jane | 4 feet 9 | 90 lbs |<0x0A>|  
Mike | 6 feet 1 | 150 lbs |<0x0A>| Lana | 5 feet 6 | 123 lbs |
```

```
91, 4076, 765, 22147, 765, 16923, 9432, 91, 4521, 91, 4521, 91, 4521, 7511, 91,  
8529, 765, 220, 20, 7693, 220, 22, 765, 220, 5894, 29160, 9432, 91, 22195,  
765, 220, 19, 7693, 220, 24, 765, 220, 1954, 29160, 9432, 91, 11519, 765,  
220, 21, 7693, 220, 16, 765, 220, 3965, 29160, 9432, 91, 95887, 765, 220,  
20, 7693, 220, 21, 765, 220, 4513, 29160, 765
```

LLM Meets Structured Data

- The main challenge we are confronted are brought by the difference between Plain text vs. Structured data:
 - LLM pertained by plain text
 - Can LLM understand structured data?
 - How to input structured data to LLM?
- Semantic reasoning vs. Symbolic reasoning
 - LLM modeled through probabilistic models, which have uncertainty and unavoidable hallucinations
 - Analysis on structured data is deterministic and often involves numerical calculations
 - Can LLM work well on such structured data analysis?

LLM Meets Structured Data



- LLM with Tabular data
 - Table input design for better LLM understanding
 - LLM for reasoning on tabular data
- LLM with Graph data
 - Graph input design for better LLM understanding
 - LLM for reasoning on graph data (e.g. Knowledge Graph)

LLM Meets Structured Data



- LLM with Tabular data
 - **Table input design for better LLM understanding**
 - LLM for reasoning on tabular data
- LLM with Graph data
 - Graph input design for better LLM understanding
 - LLM for reasoning on graph data (e.g. Knowledge Graph)

Table input design for better LLM understanding

- Structural Understanding Capabilities

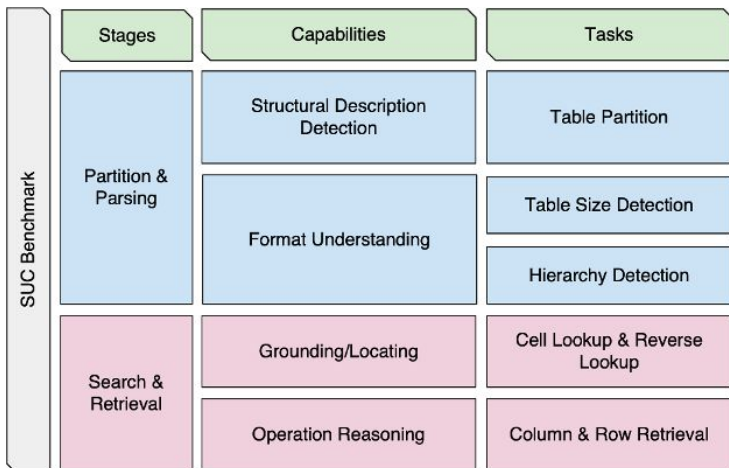


Figure 1 (a). SUC Benchmark Overview

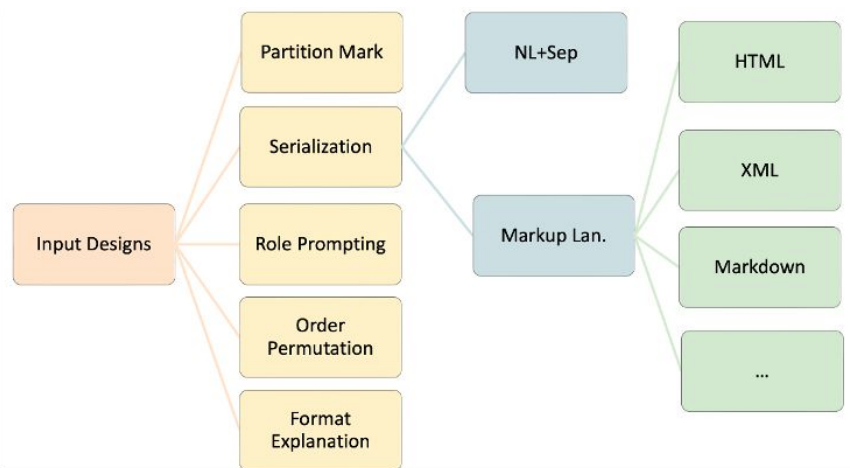


Figure 1 (b). Input Designs for SUC Evaluation

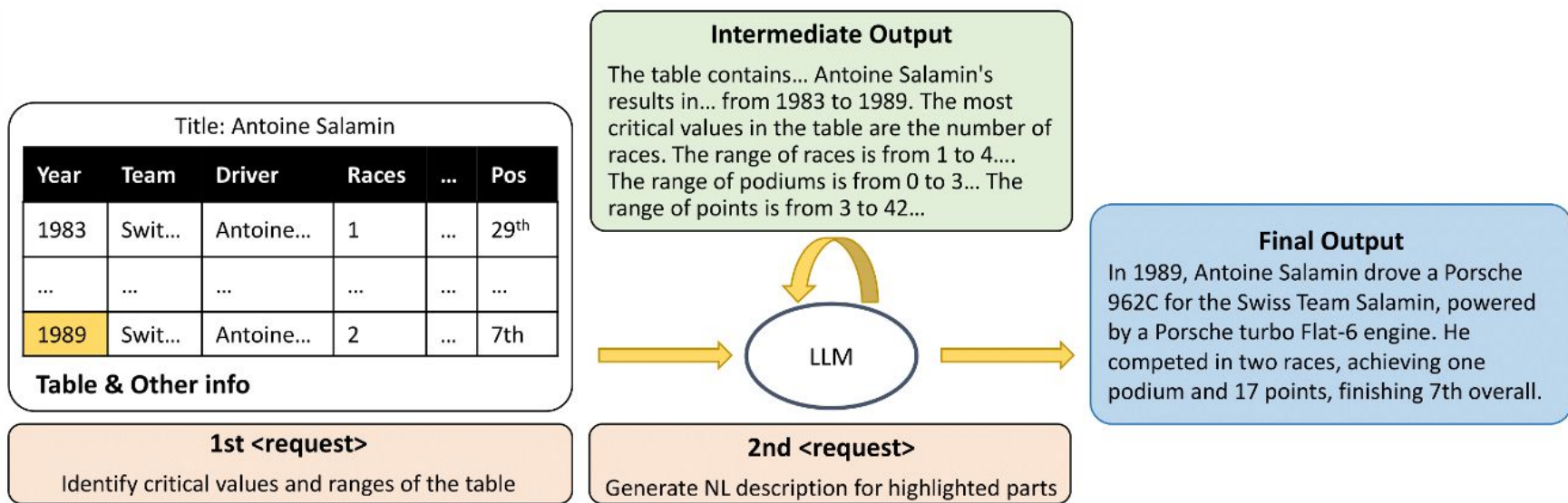
Table input design for better LLM understanding

- Insights and findings
 - LLMs have a basic understanding of table structures but are far from perfect, even in straightforward tasks like detecting the number of columns and rows
 - Choosing the right combination of input designs can significantly enhance LLMs' understanding of structured data.

| Format | Table Partition | | Cell Lookup | | Reverse Lookup | | Column Retrieval | | Row Retrieval | | Size Detection | | Merged Cell Detection | |
|----------|-----------------|---------------|---------------|---------------|----------------|---------------|------------------|---------------|---------------|---------------|----------------|---------------|-----------------------|---------------|
| | Acc | GPT-4 | Acc | GPT-4 | Acc | GPT-4 | Acc | GPT-4 | Acc | GPT-4 | Acc | GPT-4 | Acc | GPT-4 |
| NL + Sep | 93.00% | 96.78% | 39.67% | 72.48% | 52.00% | 59.12% | 60.67% | 66.32% | 31.00% | 48.67% | 42.00% | 73.12% | 71.33% | 74.98% |
| Markdown | 92.33% | 98.32% | 43.33% | 71.93% | 51.00% | 57.32% | 35.33% | 60.12% | 42.33% | 49.98% | 40.67% | 82.12% | 78.00% | 82.64% |
| JSON | 94.00% | 97.12% | 42.67% | 68.32% | 54.33% | 58.12% | 54.33% | 64.32% | 29.00% | 48.32% | 42.67% | 76.43% | 73.33% | 78.98% |
| XML | 96.00% | 97.64% | 43.33% | 72.28% | 55.00% | 60.32% | 41.33% | 68.28% | 41.00% | 50.28% | 43.67% | 80.21% | 75.00% | 80.32% |
| HTML | 96.67% | 98.32% | 44.00% | 73.34% | 47.33% | 59.45% | 63.33% | 69.32% | 42.00% | 50.19% | 67.00% | 83.43% | 76.67% | 81.28% |

Table input design for better LLM understanding

- Improved performance with self-augmented prompting



LLM Meets Structured Data



- LLM with Tabular data
 - Table input design for better LLM understanding
 - **LLM for reasoning on tabular data**
- LLM with Graph data
 - Graph input design for better LLM understanding
 - LLM for reasoning on graph data (e.g. Knowledge Graph)

LLM for reasoning with Tabular Data

NL St. Louis Cardinals vs. AL Milwaukee Brewers in 1982 World Series

| Game | Date | Score | Location | Time | Attendance |
|------|--------|---|----------------|-------|------------|
| 1 | 12-Oct | Milwaukee Brewers – 10, St. Louis Cardinals – 0 | Busch Stadium | 02:30 | 53,723 |
| 2 | 13-Oct | Milwaukee Brewers – 4, St. Louis Cardinals – 5 | Busch Stadium | 02:54 | 53,723 |
| 3 | 15-Oct | St. Louis Cardinals – 6, Milwaukee Brewers – 2 | County Stadium | 02:53 | 56,556 |
| 4 | 16-Oct | St. Louis Cardinals – 5, Milwaukee Brewers – 7 | County Stadium | 03:04 | 56,560 |
| 5 | 17-Oct | St. Louis Cardinals – 4, Milwaukee Brewers – 6 | County Stadium | 03:02 | 56,562 |
| 6 | 19-Oct | Milwaukee Brewers – 1, St. Louis Cardinals – 13 | Busch Stadium | 02:21 | 53,723 |
| 7 | 20-Oct | Milwaukee Brewers – 3, St. Louis Cardinals – 6 | Busch Stadium | 02:50 | 53,723 |

Entailed Statement: The game on October 16 be the longest game in the 1982 world series

Refuted Statement: The game on October 16 be the only game longer than 3 minute in the 1982 world series

Semantic reasoning vs. Symbolic reasoning



Please give your judgment for the statement based on the table.



The statement is correct/incorrect because ...



Please write SQL for answering the question based on the table.



```
```SELECT Date From Table ...
```

Semantic Reasoning

Symbolic Reasoning

# Symbolic Reasoning with Tabular Data



Please write SQL for answering the question based on the table meta information. The question is whether the game on October 16 be the longest game in the 1982 world series. The table title is NL St. Louis Cardinals vs. AL Milwaukee Brewers in 1982 World Series and has the following columns: Game, Date, Score, Location, Time and Attendance.

```
sql Copy
-- Find the maximum duration of any game in the 1982 World Series
SELECT MAX(Time) AS longest_game_duration
FROM world_series_1982;

-- Compare the duration of the game on October 16 to the longest game
duration
SELECT
 CASE
 WHEN Time = (SELECT MAX(Time) FROM world_series_1982)
 THEN 'Yes, the game on October 16 was the longest game in the 1982
World Series.'
 ELSE 'No, the game on October 16 was not the longest game in the 1982
World Series.'
 END AS answer
FROM world_series_1982
WHERE Date = '1982-10-16';
```



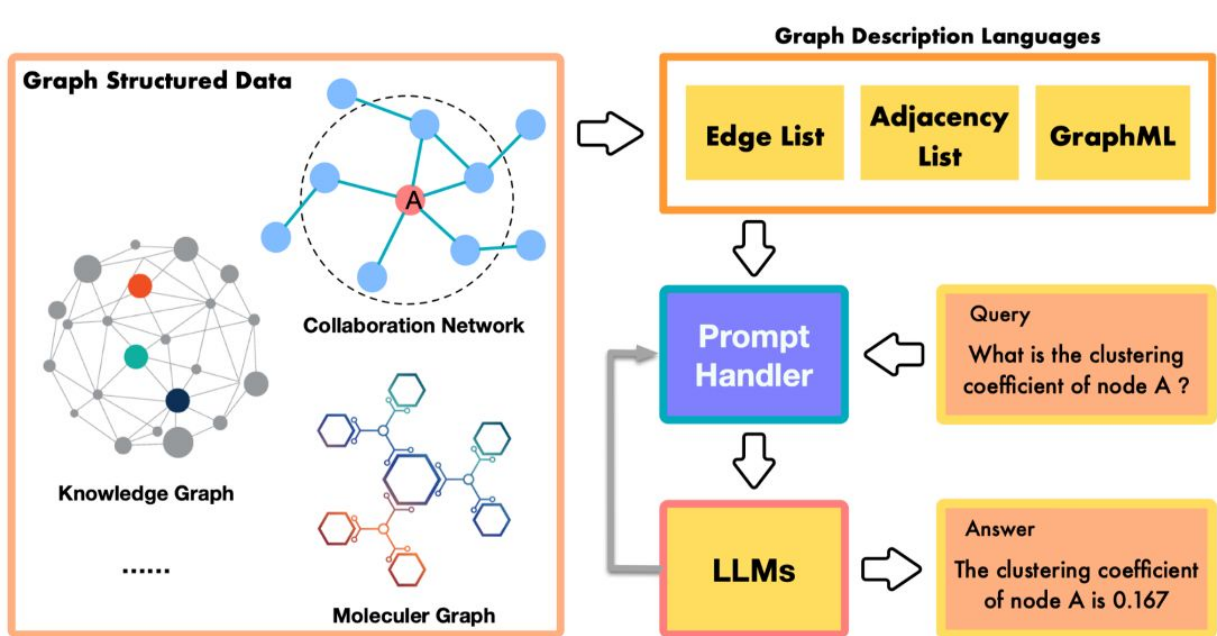
# LLM Meets Structured Data



- LLM with Tabular data
  - Table input design for better LLM understanding
  - LLM for reasoning on tabular data
- LLM with Graph data
  - **Graph input design for better LLM understanding**
  - LLM for reasoning on graph data (e.g. Knowledge Graph)

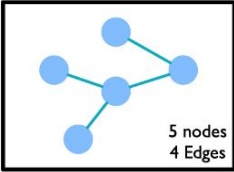
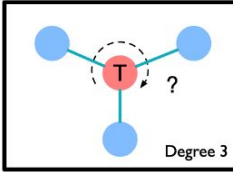
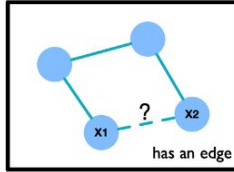
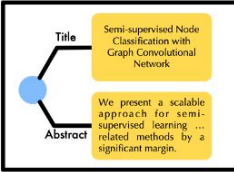
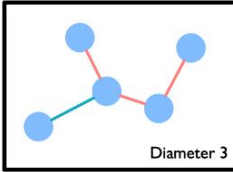
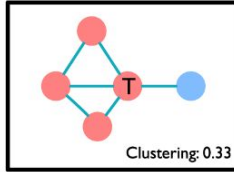
# Graph input design for better LLM understanding

- Graph data with LLMs

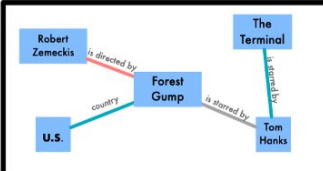


# Graph input design for better LLM understanding

- Graph understanding tasks

| Size Detection                                                                                               | Degree Detection                                                                                    | Edge Detection                                                                                            |
|--------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------|
|  <p>5 nodes<br/>4 Edges</p> |  <p>Degree 3</p>   |  <p>has an edge</p>      |
| Attribute Retrieval                                                                                          | Diameter                                                                                            | Clustering                                                                                                |
|  <p>Attribute Retrieval</p> |  <p>Diameter 3</p> |  <p>Clustering: 0.33</p> |

**KGQA**

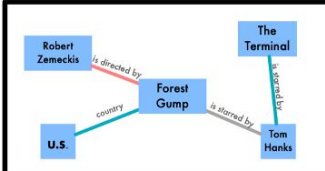


Query: The director who directs Forest Gump also direct what?  
Answer: Back to the Future

Query: What is the class of node C ? Use Abbreviation to answer.  
Answer: CS.AI

Node Classification

**GQL Generation**



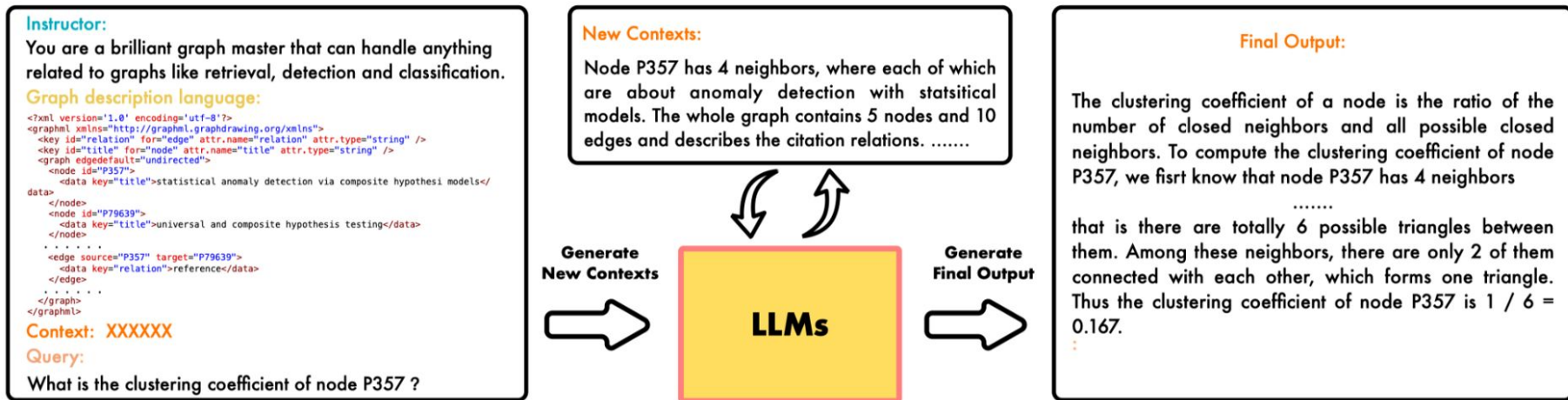
Query: The director who directs Forest Gump also direct what? Use Cypher to answer  
Answer: MATCH (m1)-[is directed by]->(d)-[direct]->(m2)  
RETURN m2

Query: Is this molecule active with H3C4 ?  
Answer: No.

Graph Classification

# Graph input design for better LLM understanding

- Improved performance with self-augmented prompting



# Graph input design for better LLM understanding

Table 1: Experiments on Graph Structural Understanding on OGBN-ARXIV. ACC indicates average accuracy over samples, while  $\Delta$  indicates the difference of variants with the 1-shot setting. - denotes that the input format do not contain corresponding information.

| Format         | Input Design           | Size Detection |          | Degree Detection |          | Edge Detection |          | Attribute Retrieval |          | Diameter |          | Clustering |          |
|----------------|------------------------|----------------|----------|------------------|----------|----------------|----------|---------------------|----------|----------|----------|------------|----------|
|                |                        | ACC            | $\Delta$ | ACC              | $\Delta$ | ACC            | $\Delta$ | ACC                 | $\Delta$ | ACC      | $\Delta$ | ACC        | $\Delta$ |
| Adjacency List | 1-shot                 | 35.50          | 0.00     | 15.21            | 0.00     | 65.45          | 0.00     | -                   | -        | 28.00    | 0.00     | 5.42       | 0.00     |
|                | 1-shot-cot             | 44.00          | +8.50    | 14.58            | -0.63    | 65.25          | -0.20    | -                   | -        | 24.00    | -4.00    | 1.85       | -3.57    |
|                | w/o format explanation | 33.00          | -0.25    | 16.34            | +1.13    | 57.50          | -8.25    | -                   | -        | 18.00    | -10.00   | 5.19       | +3.43    |
|                | w/o role prompting     | 36.60          | +1.10    | 15.70            | +0.49    | 55.00          | -10.45   | -                   | -        | 20.00    | -8.00    | 4.71       | -0.23    |
|                | w/o change order       | 14.00          | -21.50   | 26.28            | +11.07   | 51.20          | -14.25   | -                   | -        | 30.00    | +2.00    | 14.92      | -9.50    |
|                | w/o 1-shot             | 33.00          | -2.50    | 17.18            | +1.97    | 71.90          | -6.45    | -                   | -        | 22.00    | -6.00    | 7.85       | +2.43    |
| Edge List      | 1-shot                 | 22.50          | 0.00     | 44.87            | 0.00     | 74.60          | 0.00     | -                   | -        | 43.00    | 0.00     | 13.31      | 0.00     |
|                | 1-shot-cot             | 27.00          | +4.50    | 48.65            | +3.78    | 74.70          | +0.10    | -                   | -        | 41.00    | -2.00    | 11.33      | -1.98    |
|                | w/o format explanation | 25.00          | +2.50    | 47.86            | +2.99    | 71.55          | -3.05    | -                   | -        | 36.00    | -7.00    | 18.11      | +4.80    |
|                | w/o role prompting     | 18.00          | -4.50    | 47.64            | +2.57    | 71.70          | -2.90    | -                   | -        | 39.00    | -4.00    | 13.63      | +0.35    |
|                | w/o change order       | 9.00           | -13.50   | 20.48            | -23.39   | 79.60          | +5.00    | -                   | -        | 10.00    | -33.00   | 20.06      | +7.05    |
|                | w/o 1-shot             | 23.00          | +0.50    | 49.34            | +4.47    | 80.95          | +6.35    | -                   | -        | 34.00    | -9.00    | 19.16      | +5.84    |
| GML            | 1-shot                 | 54.50          | 0.00     | 20.91            | 0.00     | 50.45          | 0.00     | 83.40               | 0.00     | 37.00    | 0.00     | 4.36       | 0.00     |
|                | 1-shot-cot             | 55.50          | +1.00    | 20.76            | -0.15    | 50.10          | -0.35    | 83.30               | -0.10    | 28.00    | -9.00    | 0.95       | -3.41    |
|                | w/o format explanation | 55.00          | -0.50    | 29.06            | +8.15    | 50.00          | -0.45    | 85.97               | +2.57    | 41.00    | +4.00    | 12.71      | +8.35    |
|                | w/o role prompting     | 54.50          | -0.50    | 29.79            | +8.88    | 50.00          | -0.45    | 84.50               | +0.10    | 35.00    | -2.00    | 6.96       | +2.60    |
|                | w/o change order       | 51.50          | -3.00    | 21.16            | +0.24    | 55.65          | +5.20    | 83.56               | +0.16    | 39.00    | +2.00    | 5.25       | +0.89    |
|                | w/o 1-shot             | 54.00          | -0.50    | 19.85            | -1.06    | 50.25          | +0.20    | 83.22               | -0.18    | 42.00    | +5.00    | 5.39       | +1.03    |
| GraphML        | 1-shot                 | 25.00          | 0.00     | 40.20            | 0.00     | 62.05          | 0.00     | 83.87               | 0.00     | 34.00    | 0.00     | 9.74       | 0.00     |
|                | 1-shot-cot             | 22.50          | -2.50    | 40.02            | -0.18    | 62.30          | +0.25    | 83.75               | -0.12    | 32.00    | -2.00    | 7.29       | -2.45    |
|                | w/o format explanation | 19.00          | -6.00    | 46.90            | +5.88    | 53.75          | -8.40    | 85.37               | +1.50    | 38.00    | +4.00    | 22.75      | +13.01   |
|                | w/o role prompting     | 15.50          | -9.50    | 49.89            | +9.87    | 56.10          | -5.95    | 87.63               | +3.76    | 31.00    | -3.00    | 14.52      | +4.78    |
|                | w/o change order       | 8.50           | -16.50   | 30.60            | -9.60    | 65.35          | +3.30    | 9.76                | -4.11    | 43.00    | +9.00    | 8.00       | -1.74    |
|                | 0-shot                 | 24.50          | -0.50    | 39.59            | -0.61    | 73.95          | +11.90   | 82.90               | -0.97    | 30.00    | -4.00    | 14.32      | +4.58    |

# LLM Meets Structured Data



- LLM with Tabular data
  - Table input design for better LLM understanding
  - LLM for reasoning on tabular data
- LLM with Graph data
  - Graph input design for better LLM understanding
  - **LLM for reasoning on graph data (e.g. Knowledge Graph)**

# LLM for Reasoning on Knowledge Graph

Existing works for LLM reasoning on Knowledge Graph mainly includes

- Convert Knowledge Graph as Text

ZH

## One-shot

Predict the tail entity [MASK] from the given (Academy Award for Best Film Editing, award award category category of, [MASK]) by completing the sentence "what is the category of of Academy Award for Best Film Editing? The answer is ".  
The answer is Academy Awards, so the [MASK] is Academy Awards.

## Zero-shot

Predict the tail entity [MASK] from the given (Primetime Emmy Award for Outstanding Guest Actress - Comedy Series, award award category category of, [MASK]) by completing the sentence "what is the category of of Primetime Emmy Award for Outstanding Guest Actress - Comedy Series? The answer is".

Zero-shot: The answer is "Comedy Series".

One-shot: The answer is Primetime Emmy Awards. So, the [MASK] is Primetime Emmy Awards.



# LLM for Reasoning on Knowledge Graph

Existing works for LLM reasoning on Knowledge Graph mainly includes

- LLM-boosted Symbolic Reasoning



Please write Cypher for answering the question based on the graph meta information.  
The question is whether Prof. He has the most citations in the field of deep learning.  
The graph has the following nodes: author and paper; the following link: author - paper (write) and paper - paper (cite). Author has the attribute name, and the paper has the attribute field and title.

```
cypher ✓ Copied!

// Calculate the total citations for each author in the field of deep learning
MATCH (a:author)-[:writes]->(p:paper)
WHERE p.field = 'deep learning'
MATCH (p)-[:cites]->(citing_paper:paper)
WITH a.name AS author, COUNT(citing_paper) AS citations
ORDER BY citations DESC
LIMIT 1

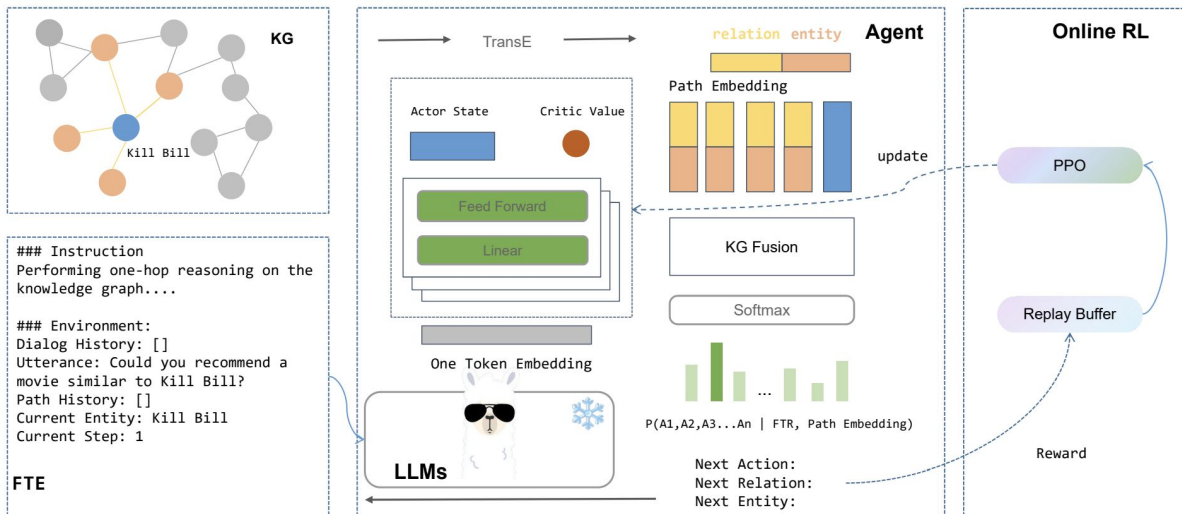
// Check if Prof. He is the author with the most citations
RETURN
CASE WHEN author = 'Prof. He' THEN 'Yes, Prof. He has the most citations in the field of deep learning.'
ELSE 'No, Prof. He does not have the most citations in the field of deep learning. The most cited author is ' + author
END AS result
```



# LLM for Reasoning on Knowledge Graph

Existing works for LLM reasoning on Knowledge Graph mainly includes

- LLM-booster AI Agent for Reasoning



# Large Language Model and Misinformation Detection

- Introduction to LLM-based Misinformation Detection
- Real-time Dataset Construction
- Incorporate LLM with Structured Modality Data
- **Prompting and Reasoning Strategy for Misinformation Detection**



**Eliot Higgins** @EliotHiggins · Mar 20



Making pictures of Trump getting arrested while waiting for Trump's arrest.



1.8K

7.6K

39K

6.7M



# Divide-and-Conquer Prompting and Reasoning

Multiple recent papers have reported similar empirical findings that divide-and-conquer prompting and reasoning strategies are specifically useful for hallucination detection and fact-verification, especially when the input is long:

## Task: Verify the truthfulness of a summary.

#Material#: Marseille, France (CNN)The French prosecutor leading an investigation into the crash of Germanwings Flight 9525 insisted ... **not aware of any video** ... so far no videos were used... He added, "A person who has such a video needs to ... give it to the investigators." ... Jean-Marc Menichini... that the reports were "completely wrong" and ...

## Example of Entangled Problem Solving

#Summary#: **A video showing the final moments of Germanwings Flight 9525 has been recovered by investigators from the wreckage site.** Marseille prosecutor Brice Robin urged anyone who might have more footage to turn it over immediately. Andreas Lubitz, the co-pilot accused of deliberately crashing the plane, had a history of severe depression and suicidal tendencies.

**Q: You are given the above material and a summary. Please answer me: Does the summary contain any claim that is contradicted with the material?**

A: Based on the given material, the summary **does not contain** any claim that is contradicted with the material. The material confirms the existence of a video showing the final moments of Germanwings Flight 9525, as well as the history of severe depression and suicidal tendencies of the co-pilot, Andreas Lubitz.

## Example of Divide-and-Conquer Problem Solving

Sub-task 1:

**Q: Does the claim 1 contradict with the material? #Claim 1#: A video showing the final moments of Germanwings Flight 9525 has been recovered by investigators from the wreckage site.**

A: Yes, the **claim is contradicted by the material**. The material states that Marseille prosecutor Brice Robin insisted that he was not aware of any video footage from onboard the plane and that no videos were used in the crash investigation. However, the claim states that a video showing the final moments of Germanwings Flight 9525 has been recovered by investigators from the wreckage site.

Sub-task 2:

**Q: Does the claim 2 contradict with the material? #Claim 2#: ...**

A: Based on the provided document, the summary does not contain any claim. ...

Sub-task 3:

**Q: Does the claim 3 contradict with the material? #Claim 3#: ...**

A: Based on the provided document, the summary does not contain any claim. ...

# Divide-and-Conquer Prompting and Reasoning

| Strategies         | GPT-3.5-Turbo |              |              |              | GPT-4        |              |              |              |
|--------------------|---------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
|                    | F1            | G-Mean       | Prec         | Recall       | F1           | G-Mean       | Prec         | Recall       |
| Io-Prompting       | 72.12         | 72.77        | 83.22        | 63.64        | 69.15        | 71.77        | 94.44        | 54.55        |
| Chain-of-Thoughts  | 56.09         | 60.64        | 90.48        | 40.64        | 74.03        | 75.79        | 94.21        | 60.96        |
| CoT-SC             | 56.83         | 61.44        | <b>91.67</b> | 41.18        | 70.09        | 73.45        | <b>100.0</b> | 53.95        |
| Tree-of-Thoughts   | 69.91         | 73.30        | 53.74        | <b>100.0</b> | 77.34        | 78.00        | 88.89        | 68.45        |
| Least-to-Most      | 54.08         | 54.15        | 51.46        | 56.99        | 73.56        | 74.25        | 85.21        | 64.71        |
| Divide-and-Conquer | <b>76.88</b>  | <b>77.13</b> | 83.65        | 71.12        | <b>81.11</b> | <b>81.24</b> | 76.67        | <b>86.10</b> |

Performance of different prompting methods on SciFact dataset. We report the F1 score, G-Mean score, Precision and Recall.

# Theoretic Explanation: Tree Isomorphism

One of the above papers provides an theoretic explanation of this finding. The retrieval based fact-verification can actually be regarded as a subtree isomorphism problem:

**Task: Verify the truthfulness of a summary.**

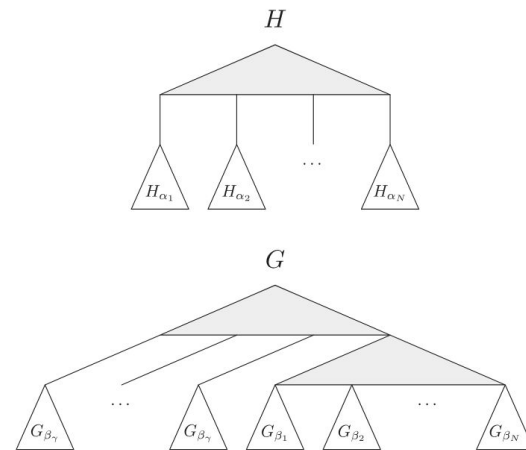
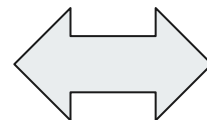
#Material#: Marseille, France (CNN)The French prosecutor leading an investigation into the crash of Germanwings Flight 9525 insisted ... **not aware of any video** ...so far no videos were used... He added, "A person who has such a video needs to ... give it to the investigators." ... Jean-Marc Menichini... that the reports were "completely wrong" and ...

**Example of Problem Solving**

#News#: **A video showing the final moments of Germanwings Flight 9525 has been recovered by investigators from the wreckage site.** Marseille prosecutor Brice Robin urged anyone who might have more footage to turn it over immediately. Andreas Lubitz, the co-pilot accused of deliberately crashing the plane, had a history of severe depression and suicidal tendencies.

**Q: You are given the above material and a summary. Please answer me: Does the summary contain any claim that is contradicted with the material?**

**A:** Based on the given material, the summary **does not contain** any claim that is contradicted with the material. The material confirms the existence of a video showing the final moments of Germanwings Flight 9525, as well as the history of severe depression and suicidal tendencies of the co-pilot, Andreas Lubitz.



# Theoretic Explanation: Tree Isomorphism and LLM



However, for a pre-trained LLM whose layer number is fixed, we have the following theorems. They prove that for a fixed LLM, standard prompting can not solve tree isomorphism problem when the subtree is too large, while divide-and-conquer strategy does not have this limitation.

**Theorem 4.1.** *Assume that  $TC^0 \neq NC^1$ . For any depth  $L$ , any polynomial  $Q$ , there exists a size  $n$  of pattern tree such that there exist no log-precision Transformer with a depth of  $L$ , hidden dimension  $d < Q(n)$ , and fixed prompt  $p$  that can directly output the solution (Yes or No) of the 2-color Binary Subtree Isomorphism problem (2-BSI).*

**Theorem 4.2.** *There exists a log-precision transformer with fixed depth  $L$  and hidden dimension  $d$  that can solve the 2-BSI of any size with fixed-length prompt  $m$  (for merge),  $t$  (for sub-task tackling) and  $d$  (for task decomposition).*

# Opportunities: Strategy of Divide

The above theoretic analysis explains why strategies based on divide and conquer can achieve advantages. However, it does not specify the dividing strategy. Existing works explored the following strategies:

- Sentence-level dividing: This strategy directly segment the claim to sentences and evaluate the sentences one by one. This strategy is simple, but effective for long news claims.

## Sub-task 1:

**Q: Does the claim 1 contradict with the material? #Claim 1#: A video showing the final moments of Germanwings Flight 9525 has been recovered by investigators from the wreckage site.**

**A:** Yes, the **claim is contradicted by the material**. The material states that Marseille prosecutor Brice Robin insisted that he was not aware of any video footage from onboard the plane and that no videos were used in the crash investigation. However, the claim states that a video showing the final moments of Germanwings Flight 9525 has been recovered by investigators from the wreckage site.

## Sub-task 2:

**Q: Does the claim 2 contradict with the material? #Claim 2#: ...**

**A:** Based on the provided document, the summary does not contain any claim. ...

## Sub-task 3:

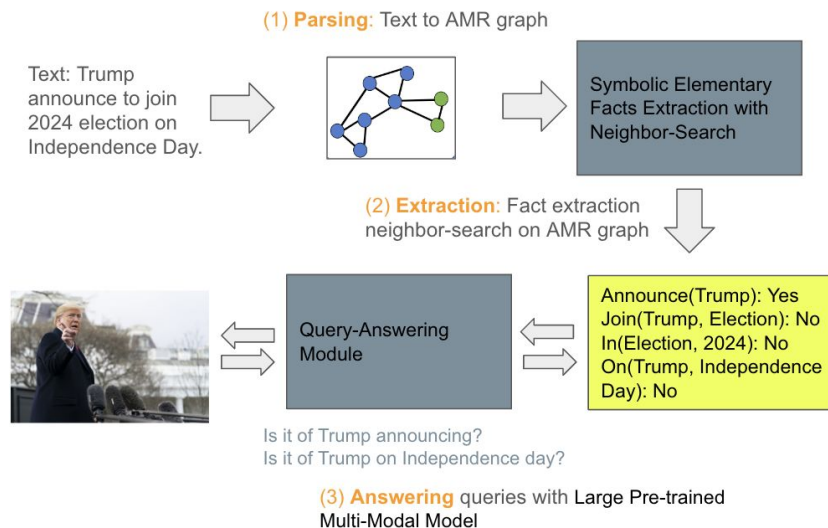
**Q: Does the claim 3 contradict with the material? #Claim 3#: ...**

**A:** Based on the provided document, the summary does not contain any claim. ...

# Opportunities: Strategy of Divide

The above theoretic analysis explains why strategies based on divide and conquer can achieve advantages. However, it does not specify the dividing strategy. Existing works explored the following strategies:

- Parsing-based dividing: For short claims, we can apply parsers (such as Abstract-Meaning-Representation parser) to break up the claim as a series of short queries.

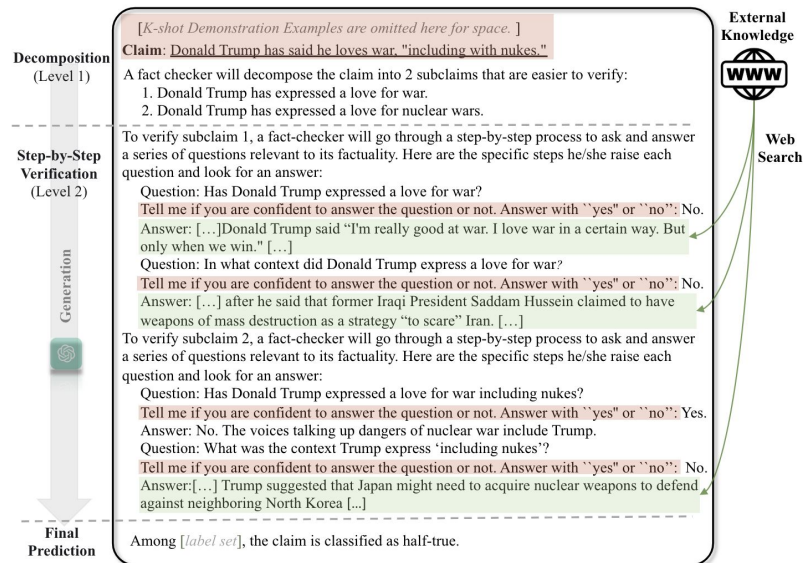




# Opportunities: Strategy of Divide

The above theoretic analysis explains why strategies based on divide and conquer can achieve advantages. However, it does not specify the dividing strategy. Existing works explored the following strategies:

- LLM-based dividing: For claims containing long sentences, we can apply this strategy to prompt a LLM to break up the claim as a series of short queries.



# Roadmaps: What can we do for future?



- **Advanced Dividing Strategy of Prompting and Reasoning:**

The aforementioned dividing strategies are still very straightforward. Also, these dividing strategies are not learnable. Therefore, there is still a wide space to explore on dividing strategies.

Challenges:

- a. How to develop trainable dividing modules?
- b. How to optimize the prompts for dividing?



**Thank you**