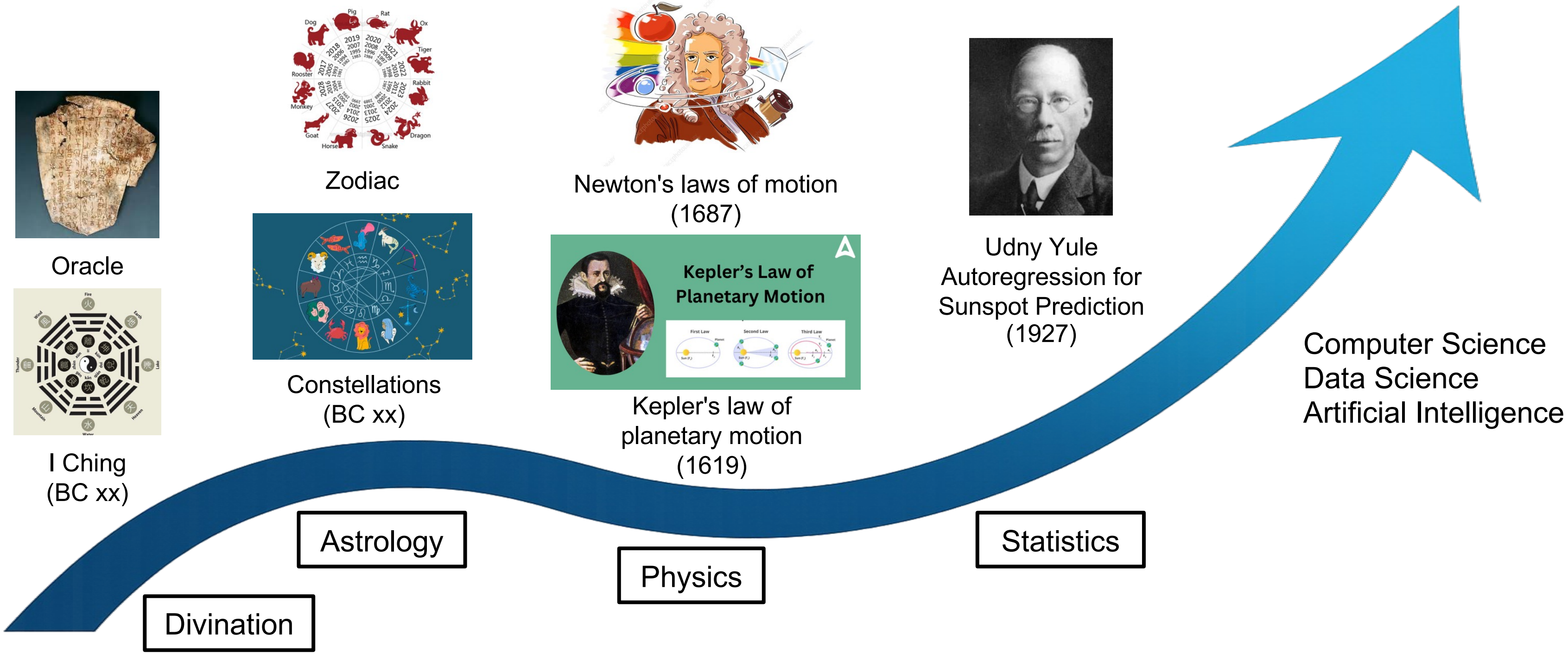
# AI That Sees the Future:

## Multimodal LLMs for Open-World Forecasting

Assistant Prof. MA Yunshan
HMCS Cluster, SCIS SMU

12 Sep, 2025

# History of temporal forecasting



Oracle

I Ching
(BC xx)

Zodiac

Constellations
(BC xx)

Newton's laws of motion
(1687)

Kepler's law of
planetary motion
(1619)

Udny Yule
Autoregression for
Sunspot Prediction
(1927)

Computer Science
Data Science
Artificial Intelligence

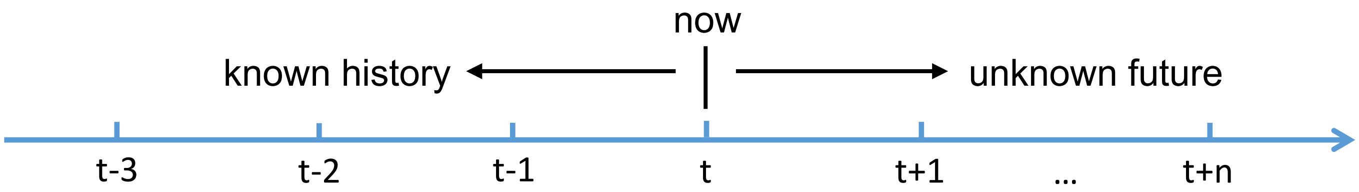
Divination

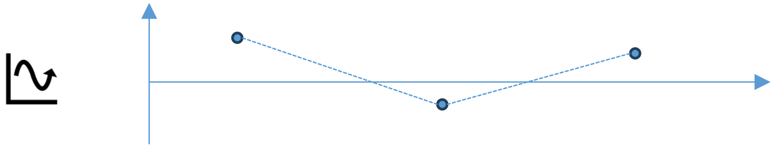
Astrology

Physics

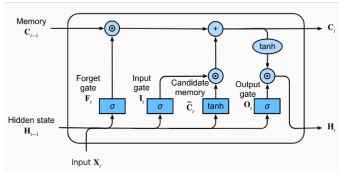
Statistics

# Data-driven approaches

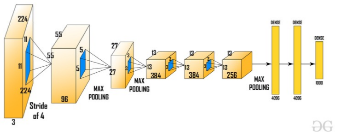




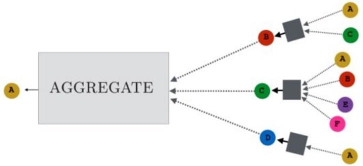
Numerical value: predict a number



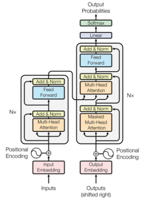
RNN



CNN

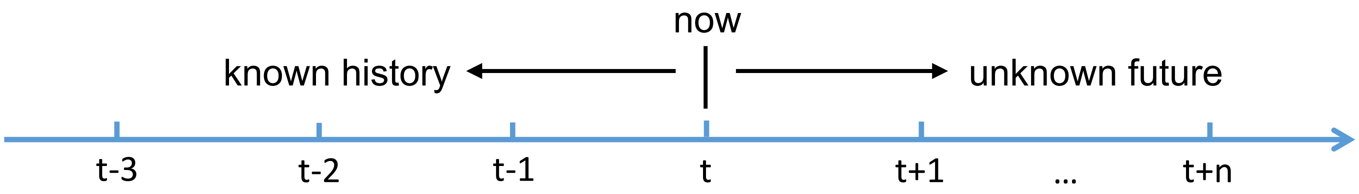


GNN



Transformer

# Data-driven approaches

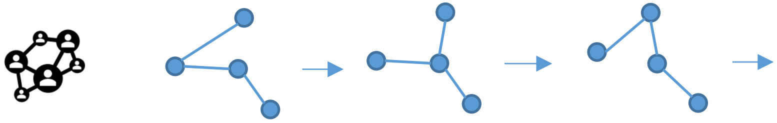


now

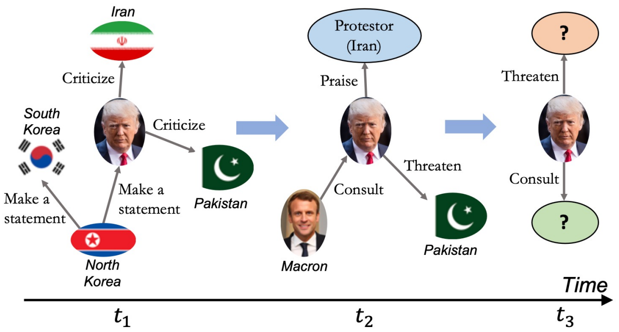
known history

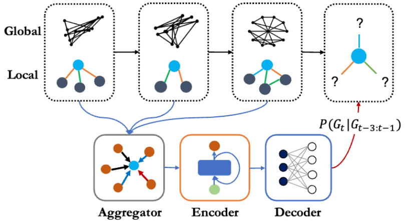
unknown future

t-3 t-2 t-1 t t+1 … t+n

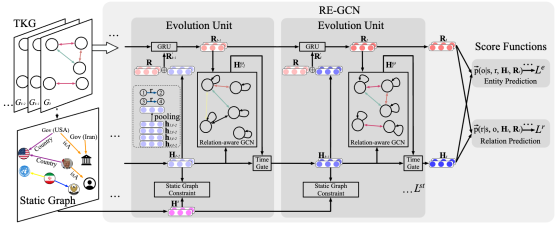


Triplet or graph: retrieve an entity or relation





RE-NET[1]



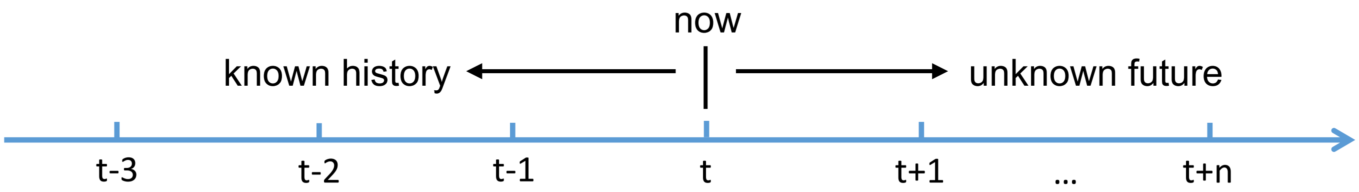
REGCN[2]

A spatial module (e.g., GCN) + A temporal module (e.g., RNN)

[1] Jin et al. Recurrent Event Network: Autoregressive Structure Inference over Temporal Knowledge Graphs. ACL 2020.

[2] Li et al. Temporal Knowledge Graph Reasoning Based on Evolutional Representation Learning. SIGIR 2021.

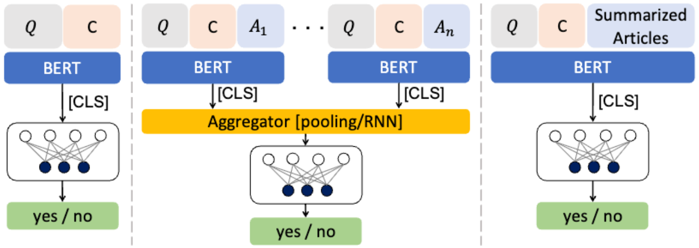
# Data-driven approaches



Hamas-led militants storm across Israel's southern border with Gaza. Israel says the attack killed 1,200 people

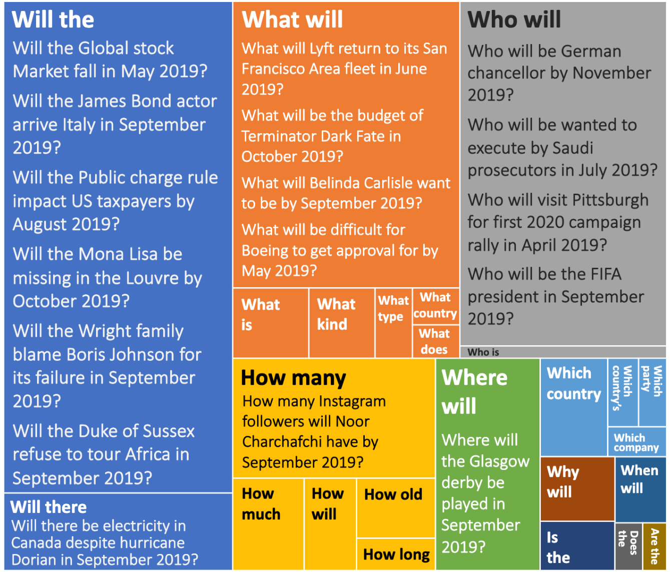
Israeli military begins ordering residents north of Wadi Gaza, including Gaza City, to evacuate south

The first aid trucks enter Gaza since the start of the war through the Rafah crossing after President Biden visits Israel.



BERT + Classification Head

Natural language QA: MCQ



**Will the**

Will the Global stock Market fall in May 2019?

Will the James Bond actor arrive Italy in September 2019?

Will the Public charge rule impact US taxpayers by August 2019?

Will the Mona Lisa be missing in the Louvre by October 2019?

Will the Wright family blame Boris Johnson for its failure in September 2019?

Will the Duke of Sussex refuse to tour Africa in September 2019?

**Will there**

Will there be electricity in Canada despite hurricane Dorian in September 2019?

**What will**

What will Lyft return to its San Francisco Area fleet in June 2019?

What will be the budget of Terminator Dark Fate in October 2019?

What will Belinda Carlisle want to be by September 2019?

What will be difficult for Boeing to get approval for by May 2019?

**What is**

**What kind**

**What type**

**What country**

**What does**

**Who will**

Who will be German chancellor by November 2019?

Who will be wanted to execute by Saudi prosecutors in July 2019?

Who will visit Pittsburgh for first 2020 campaign rally in April 2019?

Who will be the FIFA president in September 2019?

Who is

**How many**

How many Instagram followers will Noor Charchafchi have by September 2019?

**How much**

**How will**

**How old**

**How long**

**Where will**

Where will the Glasgow derby be played in September 2019?

**Which country**

Which country's

Which party

Which company

**Why will**

**When will**

**Is the**

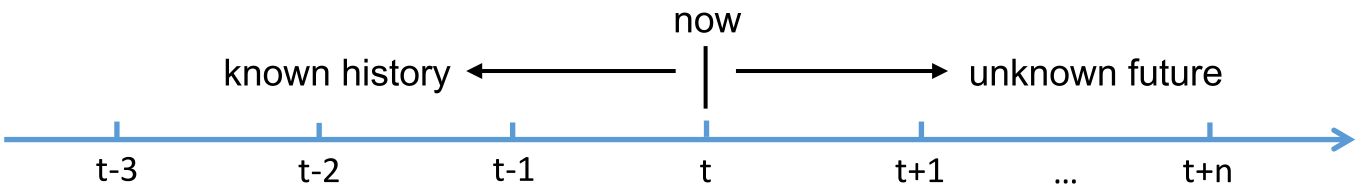
Does the

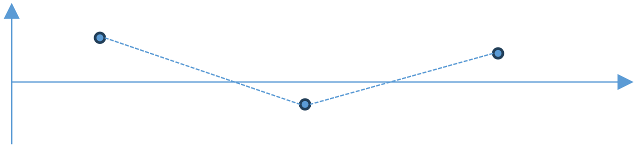
Are the

Statistics of the first two words of the questions [ForecastQA]

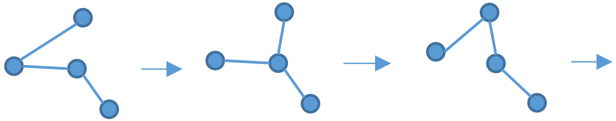
Jin et al. FORECASTQA: A Question Answering Challenge for Event Forecasting with Temporal Text Data. ACL 2021.

# Data-driven approaches





Numerical value: predict a number



Triplet or graph: retrieve an entity or relation

Hamas-led militants storm across Israel's southern border with Gaza. Israel says the attack killed 1,200 people

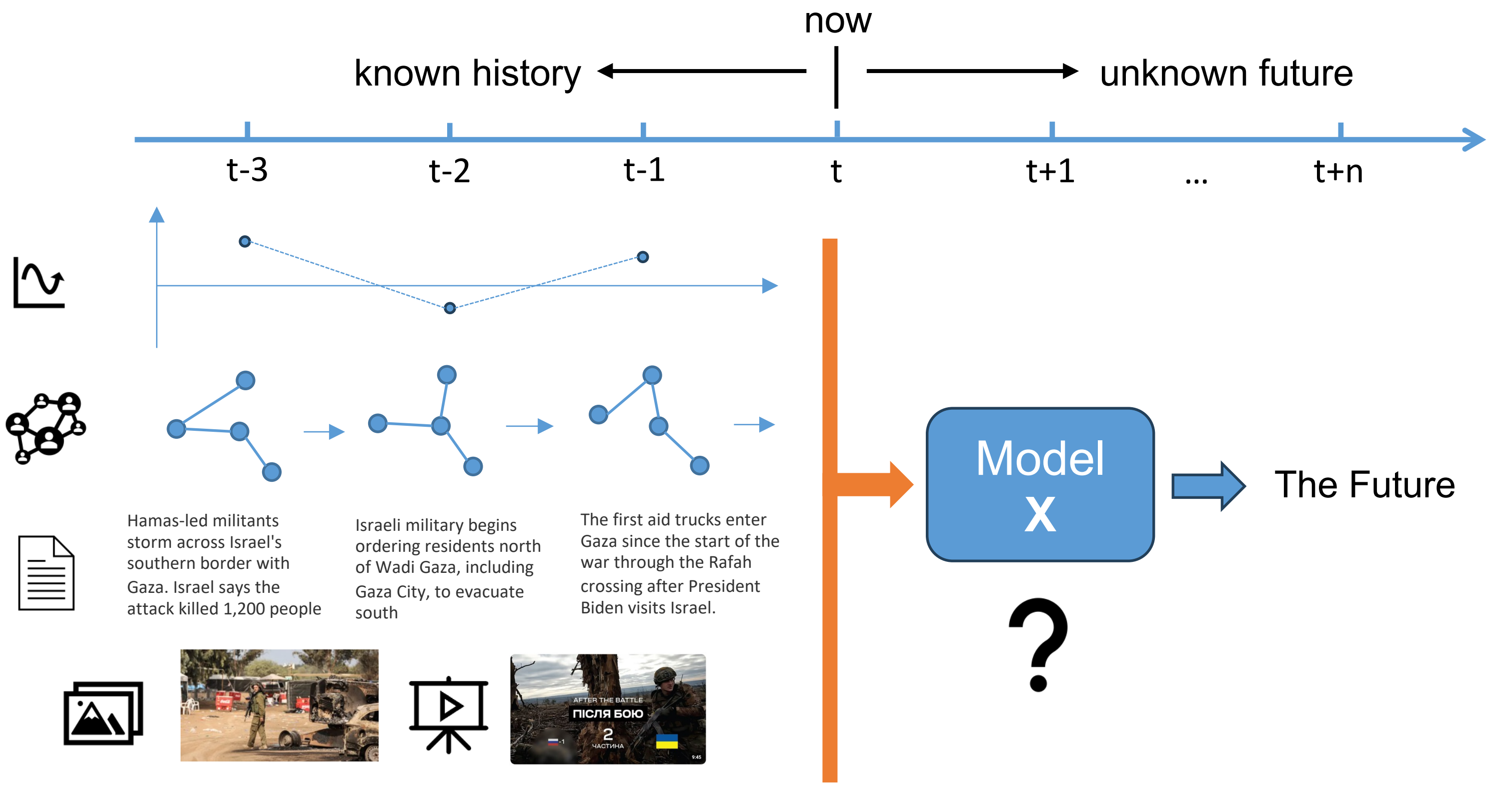
Israeli military begins ordering residents north of Wadi Gaza, including Gaza City, to evacuate south

The first aid trucks enter Gaza since the start of the war through the Rafah crossing after President Biden visits Israel.

Natural language QA: MCQ

More modalities input: image, video, etc.
Open-world question: can be any format

# Can we have one model for all?

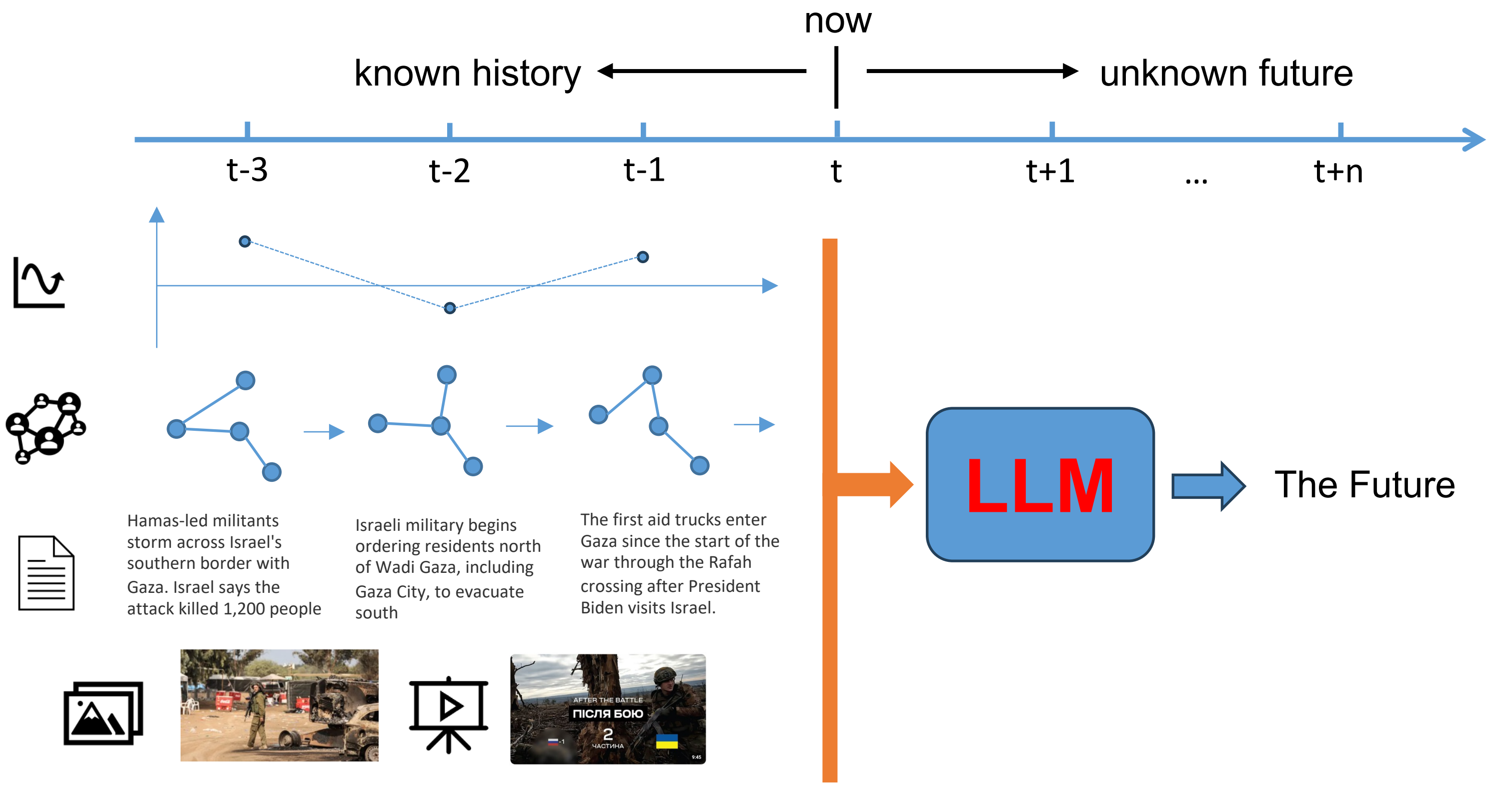


Hamas-led militants storm across Israel's southern border with Gaza. Israel says the attack killed 1,200 people

Israeli military begins ordering residents north of Wadi Gaza, including Gaza City, to evacuate south

The first aid trucks enter Gaza since the start of the war through the Rafah crossing after President Biden visits Israel.

# Can we have one model for all? - Yes



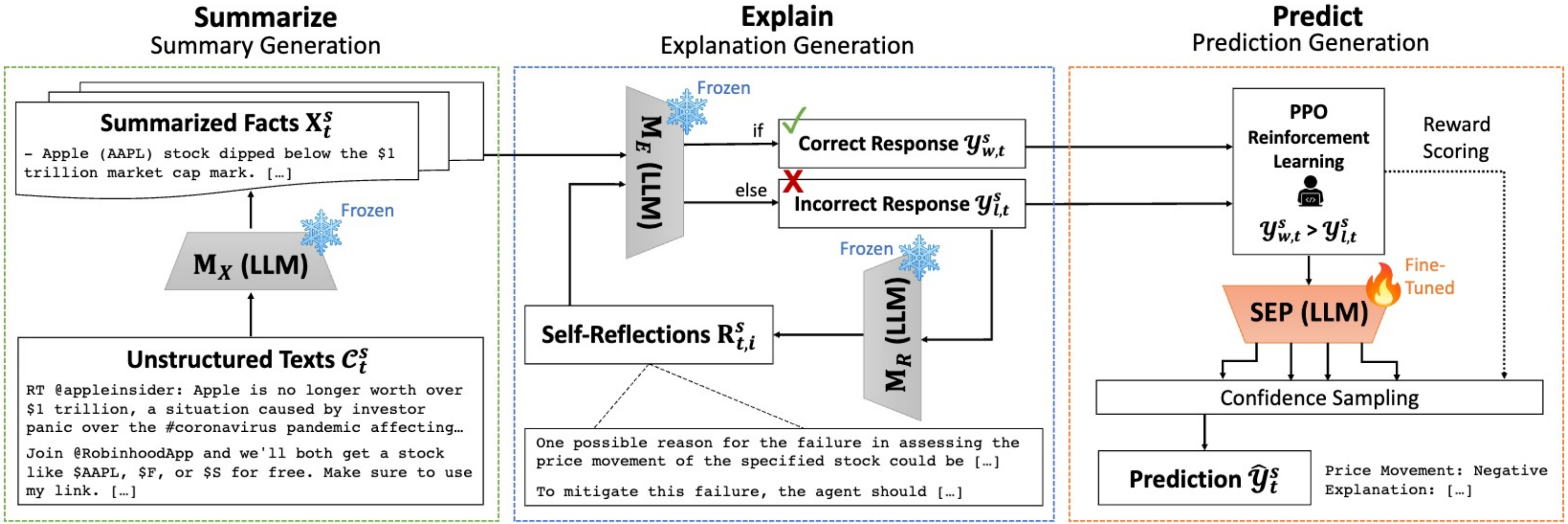
Hamas-led militants storm across Israel's southern border with Gaza. Israel says the attack killed 1,200 people

Israeli military begins ordering residents north of Wadi Gaza, including Gaza City, to evacuate south

The first aid trucks enter Gaza since the start of the war through the Rafah crossing after President Biden visits Israel.

# LLM for multimodal open-world forecasting

- **LLM for stock forecasting (1/3)**
  - Predict the movement (up or down) of a stock, given its historical prices and relevant news.
  - Provide verbal explanations along with the movement prediction.



- Apple (AAPL) stock dipped below the $1 trillion market cap mark. [...]

RT @appleinsider: Apple is no longer worth over $1 trillion, a situation caused by investor panic over the #coronavirus pandemic affecting…

Join @RobinhoodApp and we'll both get a stock like $AAPL, $F, or $S for free. Make sure to use my link. [...]

One possible reason for the failure in assessing the price movement of the specified stock could be [...]

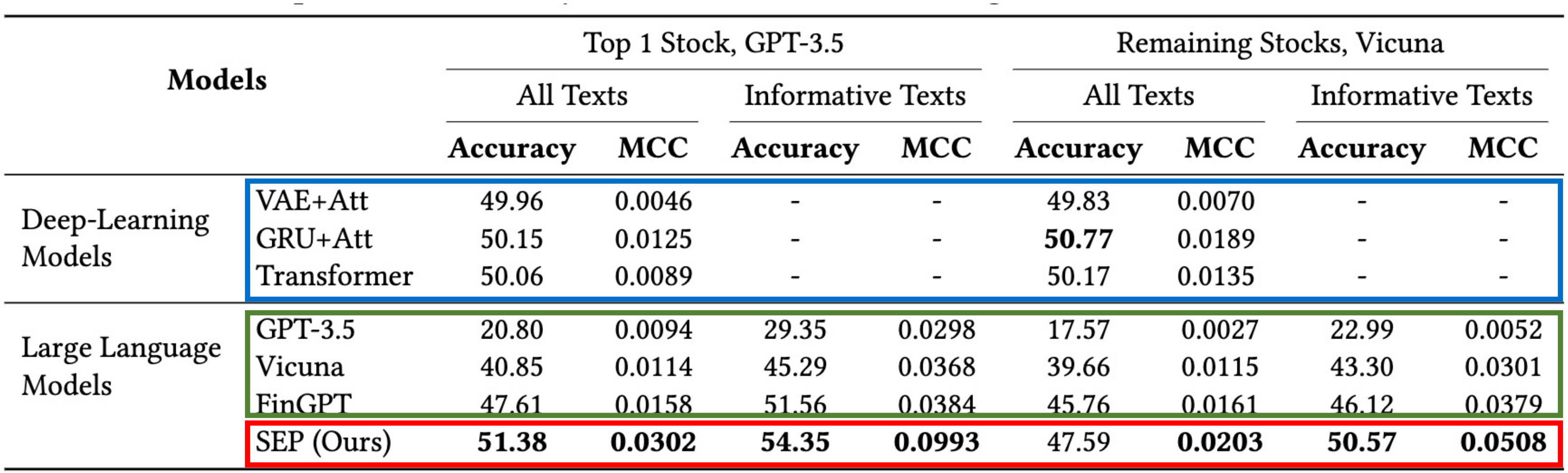
To mitigate this failure, the agent should [...]

Price Movement: Negative
Explanation: [...]

Koa et al. Learning to generate explainable stock predictions using self-reflective large language models. WWW 2024.

# LLM for multimodal open-world forecasting

- **LLM for stock forecasting (1/3)**
  - Dataset: ACL18 StockNet dataset, updated for the year 2020–2022
  - Evaluation metrics: prediction accuracy and Matthews Correlation Coefficient (MCC)
  - Overall performance



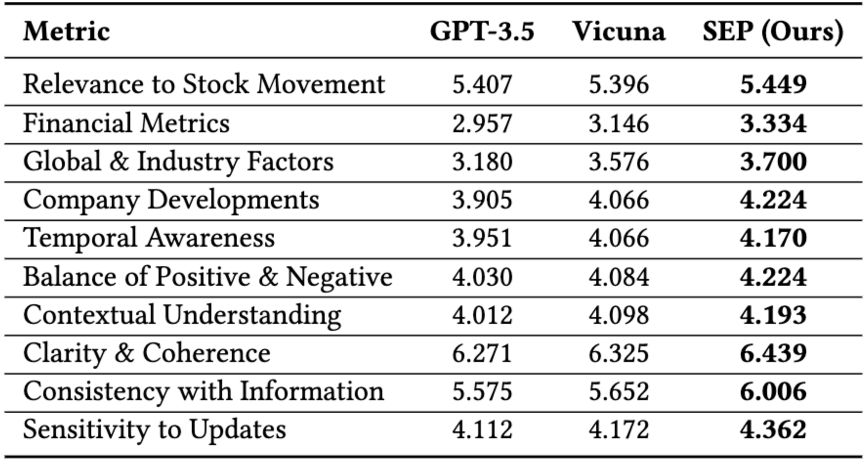
| **Models** | | Top 1 Stock, GPT-3.5 | | | | Remaining Stocks, Vicuna | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | All Texts | | Informative Texts | | All Texts | | Informative Texts | |
| | | **Accuracy** | **MCC** | **Accuracy** | **MCC** | **Accuracy** | **MCC** | **Accuracy** | **MCC** |
| Deep-Learning Models | VAE+Att | 49.96 | 0.0046 | - | - | 49.83 | 0.0070 | - | - |
| | GRU+Att | 50.15 | 0.0125 | - | - | **50.77** | 0.0189 | - | - |
| | Transformer | 50.06 | 0.0089 | - | - | 50.17 | 0.0135 | - | - |
| Large Language Models | GPT-3.5 | 20.80 | 0.0094 | 29.35 | 0.0298 | 17.57 | 0.0027 | 22.99 | 0.0052 |
| | Vicuna | 40.85 | 0.0114 | 45.29 | 0.0368 | 39.66 | 0.0115 | 43.30 | 0.0301 |
| | FinGPT | 47.61 | 0.0158 | 51.56 | 0.0384 | 45.76 | 0.0161 | 46.12 | 0.0379 |
| | SEP (Ours) | **51.38** | **0.0302** | **54.35** | **0.0993** | 47.59 | **0.0203** | **50.57** | **0.0508** |

- Our method outperforms both deep-learning methods and LLM methods
- LLMs without finetuning perform worse in forecasting accuracy due to mixed or neutral prediction

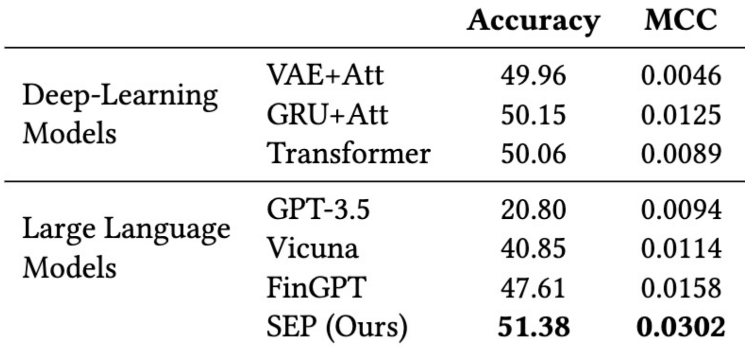
Koa et al. Learning to generate explainable stock predictions using self-reflective large language models. WWW 2024.

# LLM for multimodal open-world forecasting

- **LLM for stock forecasting (1/3)**
  - Explanation quality



| Metric | GPT-3.5 | Vicuna | SEP (Ours) |
|---|---|---|---|
| Relevance to Stock Movement | 5.407 | 5.396 | **5.449** |
| Financial Metrics | 2.957 | 3.146 | **3.334** |
| Global & Industry Factors | 3.180 | 3.576 | **3.700** |
| Company Developments | 3.905 | 4.066 | **4.224** |
| Temporal Awareness | 3.951 | 4.066 | **4.170** |
| Balance of Positive & Negative | 4.030 | 4.084 | **4.224** |
| Contextual Understanding | 4.012 | 4.098 | **4.193** |
| Clarity & Coherence | 6.271 | 6.325 | **6.439** |
| Consistency with Information | 5.575 | 5.652 | **6.006** |
| Sensitivity to Updates | 4.112 | 4.172 | **4.362** |



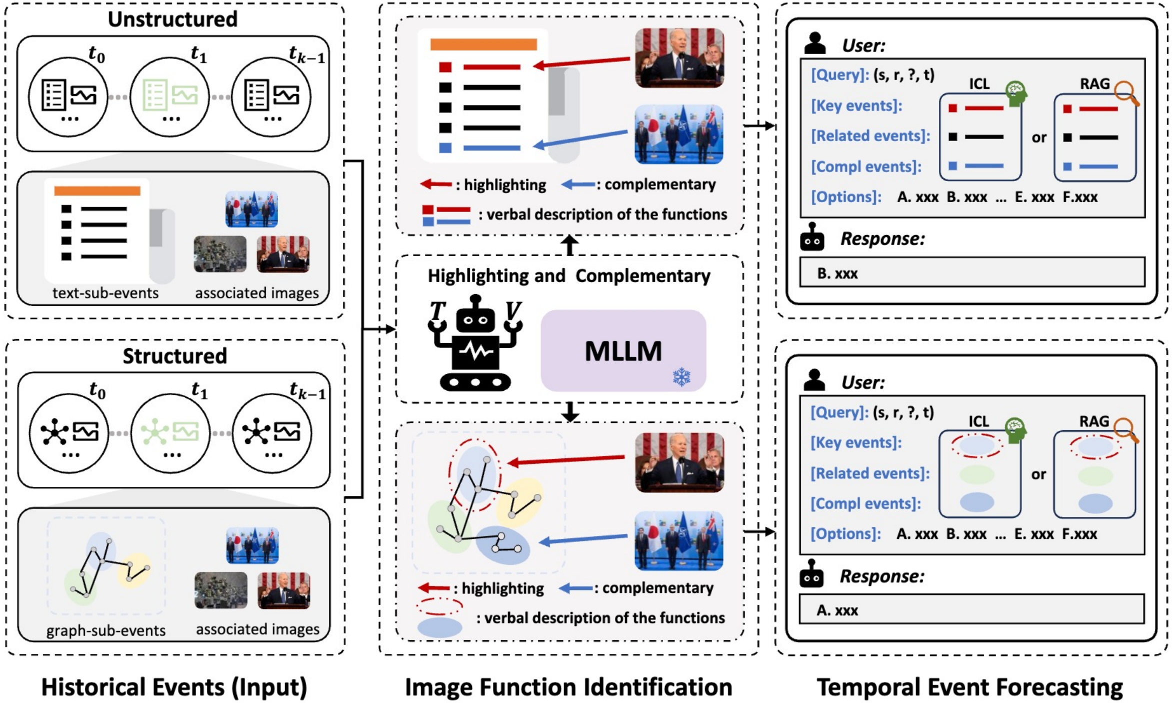
| | | Accuracy | MCC |
|---|---|---|---|
| Deep-Learning Models | VAE+Att | 49.96 | 0.0046 |
| | GRU+Att | 50.15 | 0.0125 |
| | Transformer | 50.06 | 0.0089 |
| Large Language Models | GPT-3.5 | 20.80 | 0.0094 |
| | Vicuna | 40.85 | 0.0114 |
| | FinGPT | 47.61 | 0.0158 |
| | SEP (Ours) | **51.38** | **0.0302** |

- We derive a set of evaluation metrics for explanation quality and use GPT-4 to score.
- All LLMs give good-quality explanations even if the prediction is wrong.
- Our method got the highest scores among all the metrics, even though it is not trained directly following these metrics.

Koa et al. Learning to generate explainable stock predictions using self-reflective large language models. WWW 2024.

# LLM for multimodal open-world forecasting

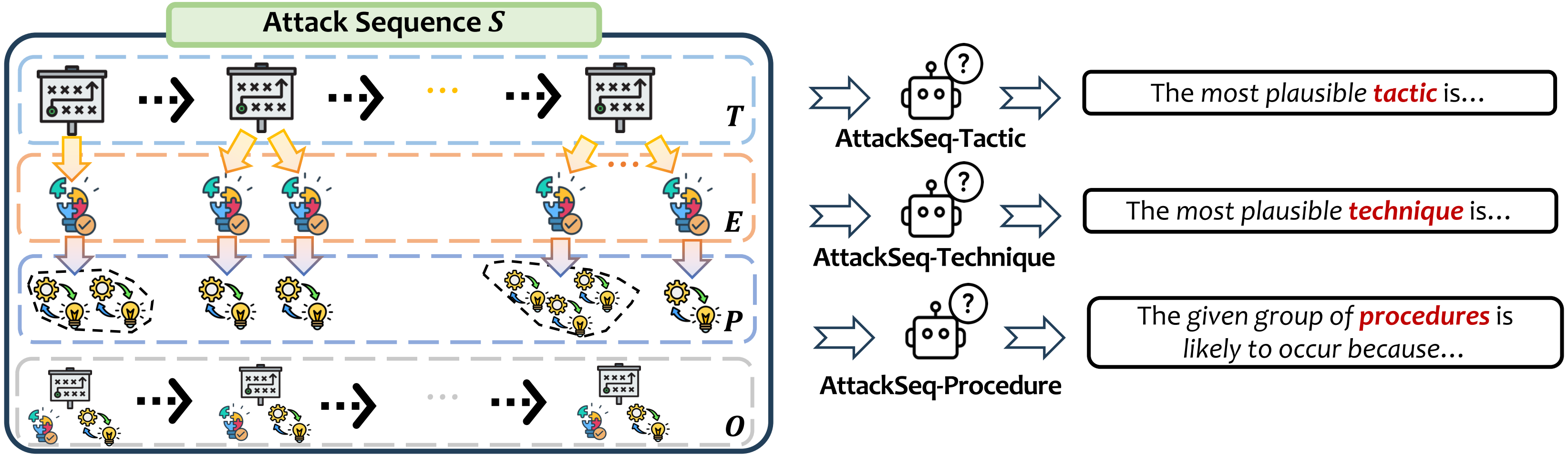
- **LLM for geo-political event forecasting (2/3)**
  - Leverage both image and text



Li et al. MM-forecast: A multimodal approach to temporal event forecasting with large language models. ACM MM 2024.

# LLM for multimodal open-world forecasting

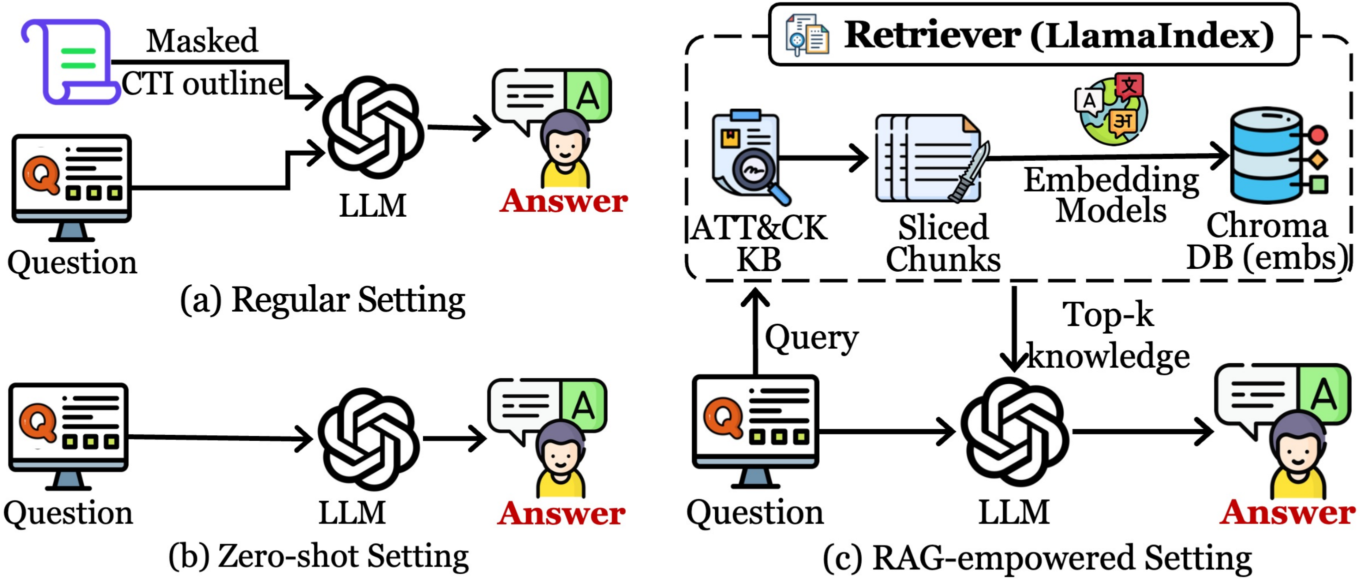
- **LLM for cyber security event forecasting (3/3)**
  - Cyber attacks often involves multiple consecutive steps, forming an attack sequence (attack flow).
  - Understanding the sequential patterns and making accurate prediction are essential for cyber attack analysis.


Attack Sequence *S*
*T*
*E*
*P*
*O*
AttackSeq-Tactic
AttackSeq-Technique
AttackSeq-Procedure
The most *plausible* **tactic** is…
The most *plausible* **technique** is…
The *given group of* **procedures** is *likely to occur because…*

Yong et al. AttackSeqBench: Benchmarking Large Language Models' Understanding of Sequential Patterns in Cyber Attacks. arXiv 2025.

# LLM for multimodal open-world forecasting

- **LLM for cyber security event forecasting (3/3)**
  - We leverage LLM and RAG for attack sequence prediction .



Yong et al. AttackSeqBench: Benchmarking Large Language Models' Understanding of Sequential Patterns in Cyber Attacks. arXiv 2025.

# Take-aways

- History of temporal forecasting
  - From ancient times to modern data-driven approaches
- Conventional data-driven approaches
  - Time series, temporal knowledge graph, temporal QA
  - Deep-learning methods with supervised training
- LLM for multimodal open-world forecasting
  - Multiple domains: stock, geo-political event, cyber attack sequence
  - Multimodal inputs: time series, graph, text, image

# Future works

There’re a bunch of works released recently, most are benchmark works:

- Tsinghua: OpenEP: Open-Ended Future Event Prediction 2025
- ByteDance: FutureX: An Advanced Live Benchmark for LLM Agents in Future Prediction 2025
- CAS: OpenForecast: A Large-Scale Open-Ended Event Forecasting Dataset 2025

What shall we do beyond benchmarking, simple fine-tuning and RAG?

- Study the in-depth mechanism of reasoning LLMs in temporal forecasting
  - Time series reasoning in finance forecasting
  - Multi-threaded hypothetical thinking in geo-political event forecasting
  - Mitigate the over-thinking problem in attack sequence prediction

Disclaim: **All models are wrong, but some are useful. -- George E. P. Box**

Thanks & QA

Contact: MA Yunshan
Email: ysma@smu.edu.sg