SMU
SINGAPORE MANAGEMENT UNIVERSITY

# Open-world Event Forecasting using Large Language Models

**Presenter: Ma Yunshan (马云山)**
**Assistant. Prof. Singapore Management University**
**July 30, 2026**

# History of temporal event forecasting

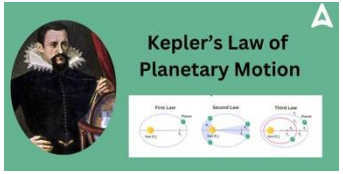
Oracle

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

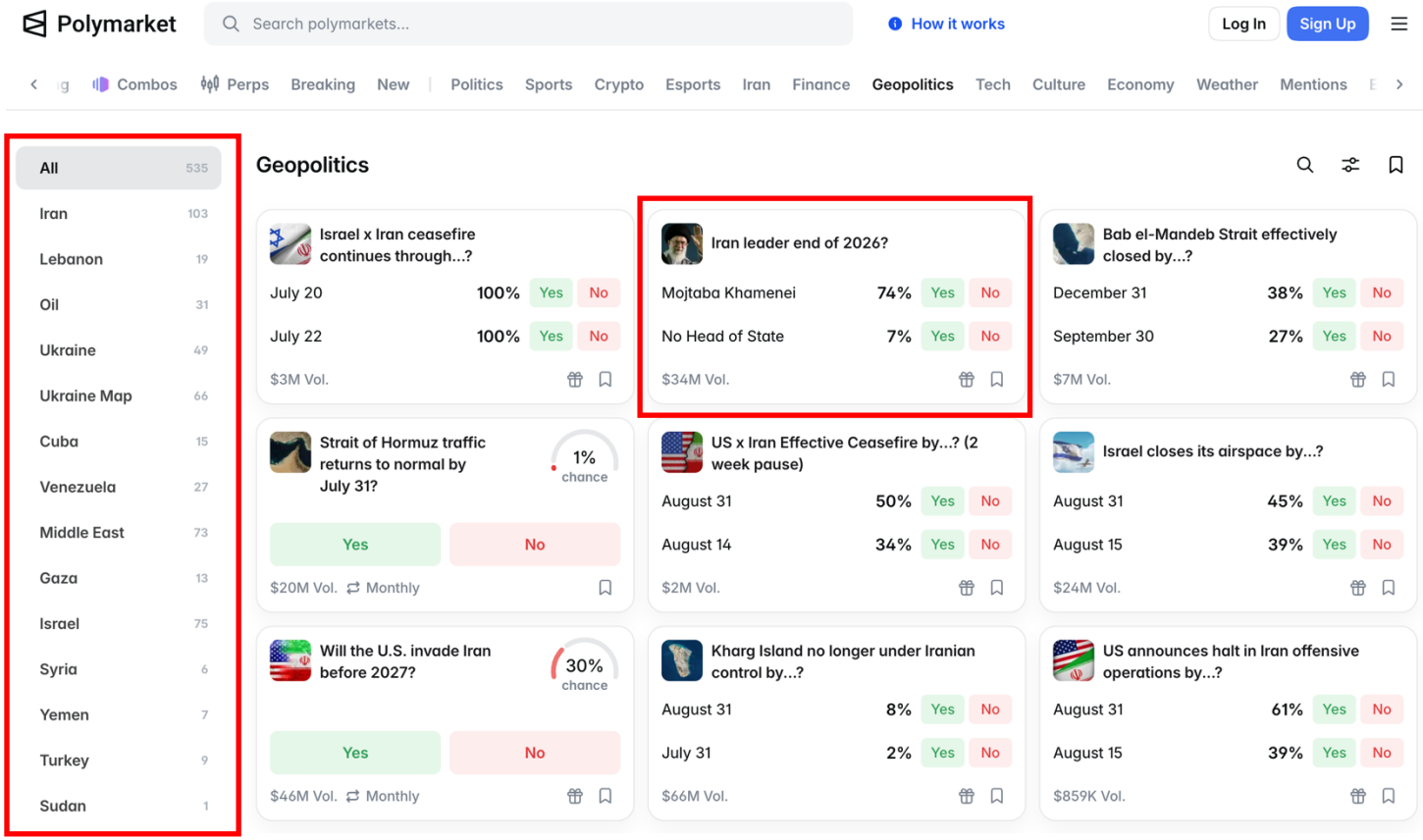
Divination

Astrology

Physics

Statistics

# An example of event forecasting



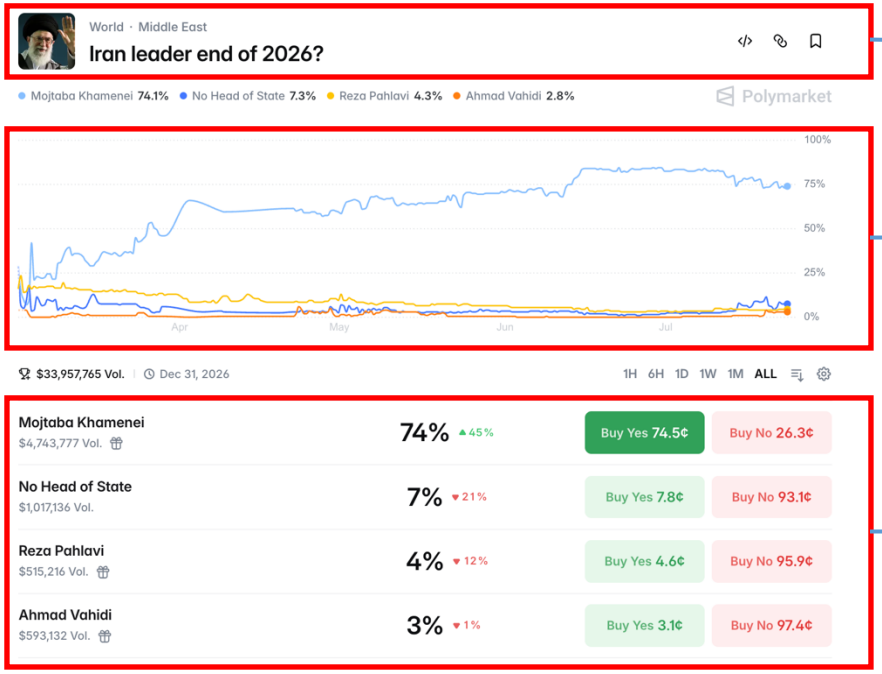
# Event forecasting formulations (1/2)

Natural Language MCQ

Polymarket

Metaculus

Kalshi



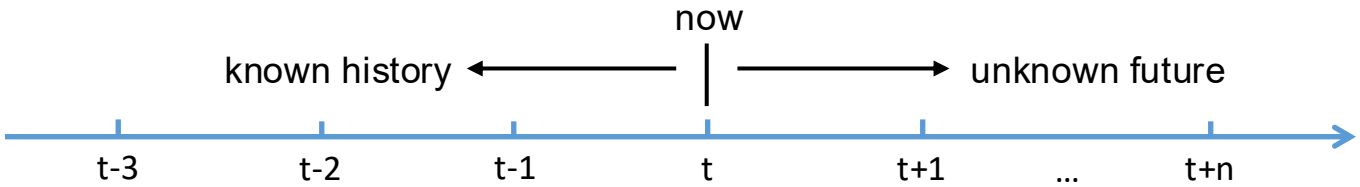

→ Forecasting Question

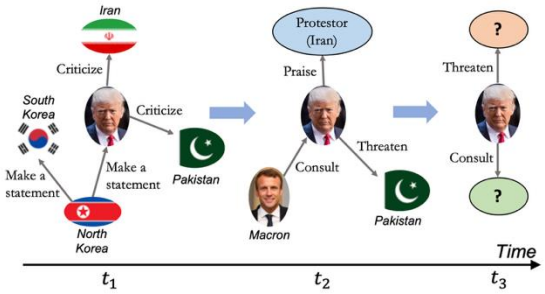
→ The Market Prediction and Trend

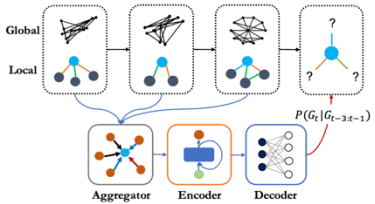
→ The Options (possible outcomes)

# Event forecasting formulations (2/2)

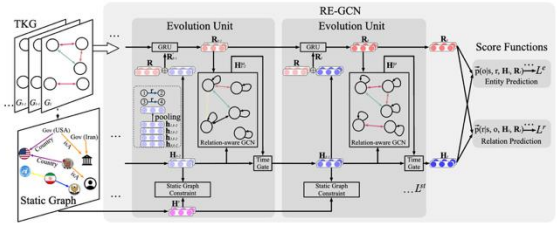
## Temporal Knowledge Graph




RE-NET[1]



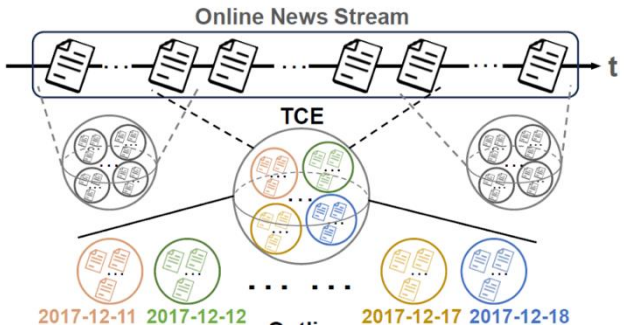

REGCN[2]

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

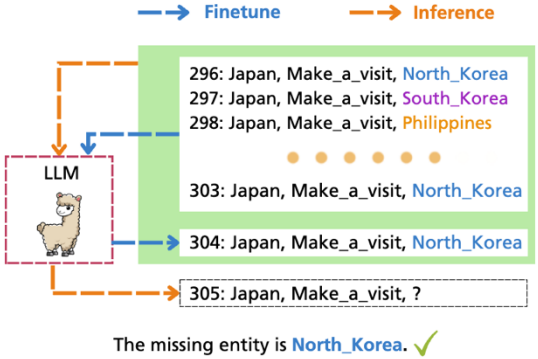
[1] Jin et al. Recurrent Event Network: Autoregressive Structure Inference over Temporal Knowledge Graphs. ACL 2020.

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

# Large language models for event forecasting




TCELongBench (RAG & Long Context LLM)[2]



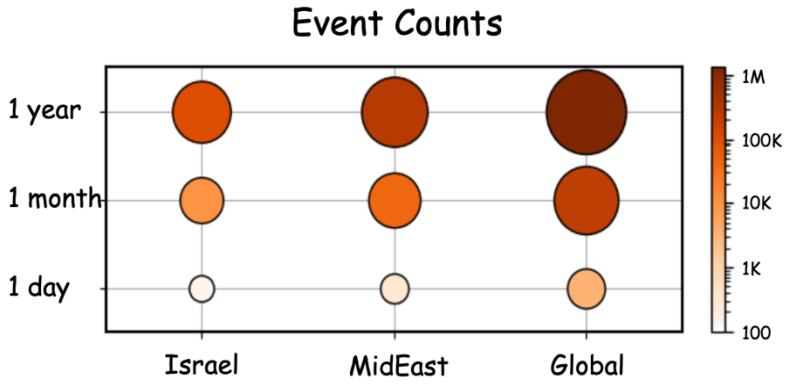

Chain of History (CoT & Finetune)[2]

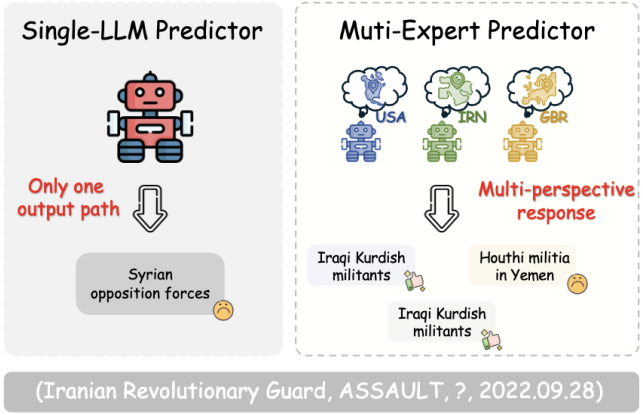
[1] Zhang et al. Analyzing temporal complex events with large language models? a benchmark towards temporal, long context understanding. ACL 2024.

[2] Luo et al. Chain of history: Learning and forecasting with llms for temporal knowledge graph completion. ACL 2024 Findings.

# Challenges and motivation

- C1: The number of events grows significantly with the increasing of regional and temporal span.
- C2: Events in different regions demonstrate distinct evolutionary patterns.



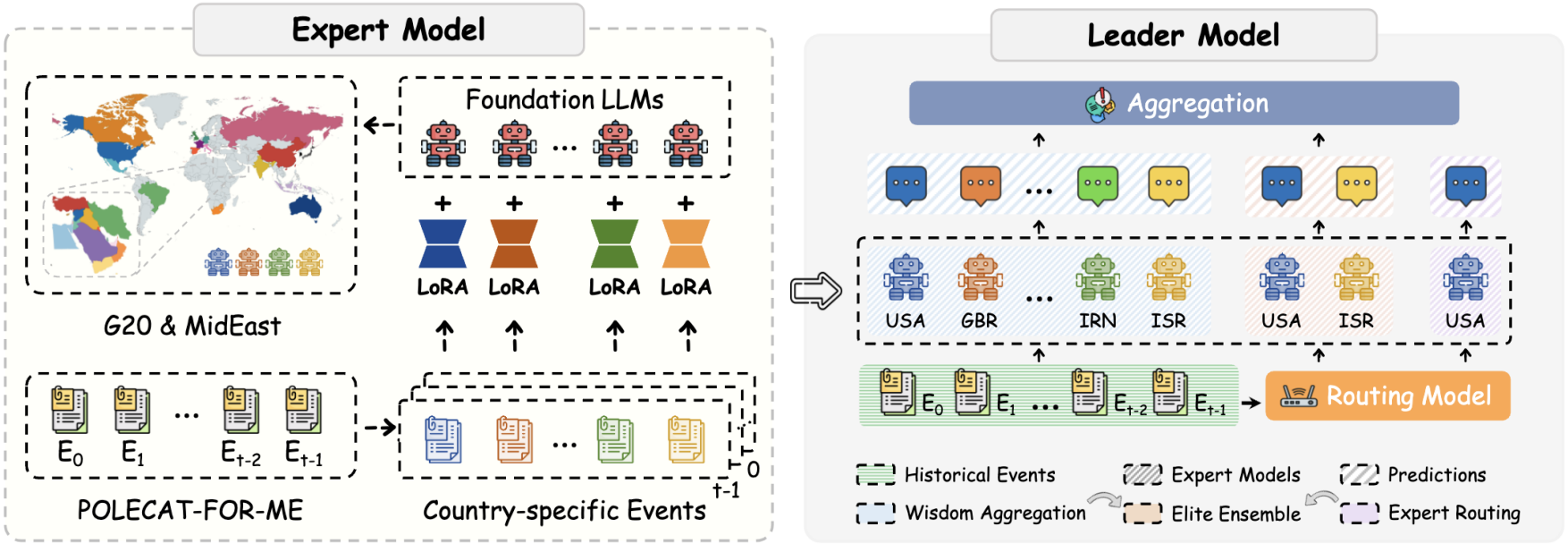




**Relying on a single model for prediction is limited**

**We need multiple experts, each focusing on a specific domain/area**

# ThinkTank-ME: A Multi-Expert Framework for Middle East Event Forecasting

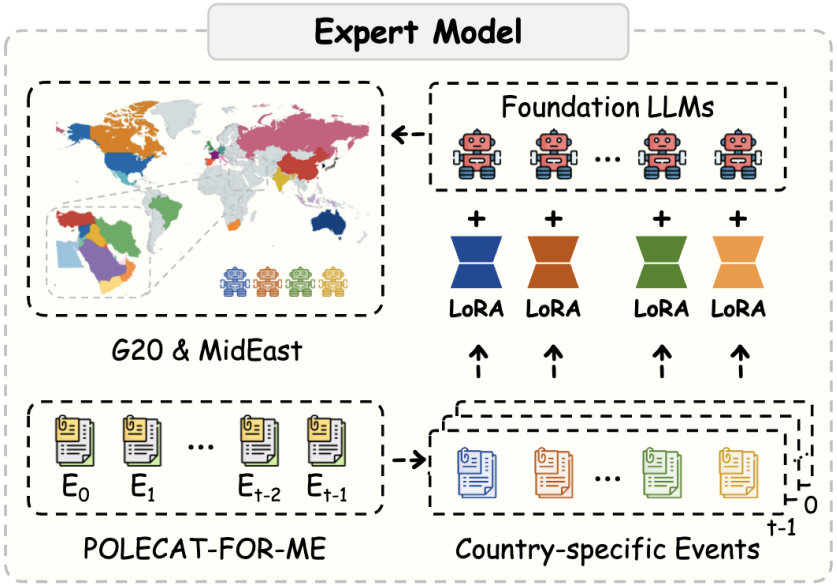
- Inspired by the real-world think tank models, where a group of experts collectively think on problems, we propose ThinkTank-ME, which consists of multiple expert models and one leader model




ThinkTank-ME: A Multi-Expert Framework for Middle East Event Forecasting. Haoxuan Li, He Chang, and Yunshan Ma et al. WWW 2026.

# Methodology (1/2)

## Expert Model

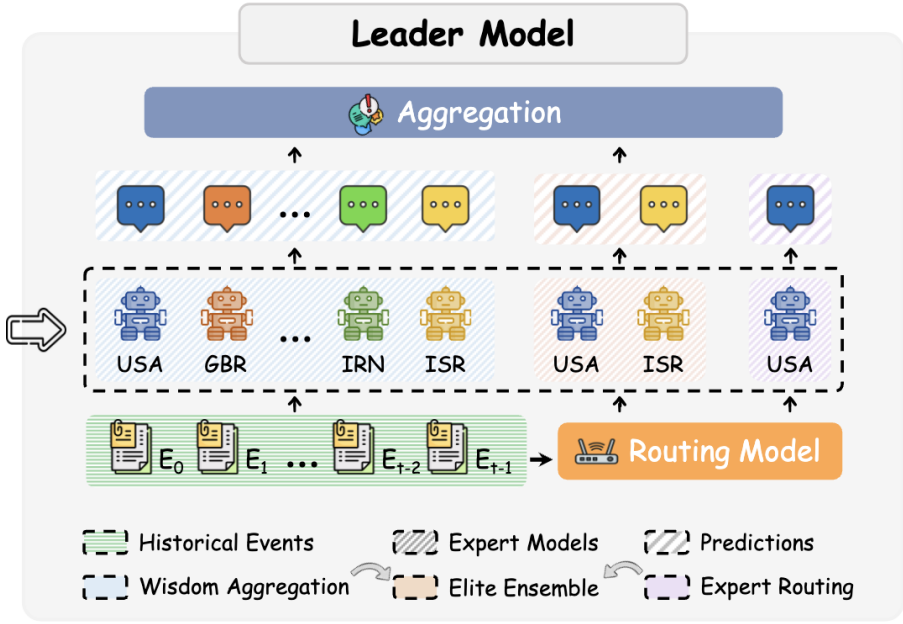



- ✓ We finetune one expert model for each region (i.e., country), simulating multidisciplinary think-tank expertise for context-aware geopolitical forecasting
- ✓ Knowledge Acquisition: LLMs are fine-tuned on country-specific subsets of the POLECAT-FOR-ME to capture unique regional event patterns and dynamics.
- ✓ Prompt Engineering: we design a structured prompt to instruct the LLMs to assume the role of a geopolitical expert and predict the missing object based on historical context.

# Methodology (2/2)

**Leader Model:** How to select the proper expert model given a forecasting question?

- **Expert Routing:** we train a multi-class classifier to predict which expert model to use
- **Wisdom Aggregation Strategies:**
  - Majority voting
  - Best-of-N
    - Vanilla Best-of-N: selects the prediction with the highest individual confidence
    - Weighted Best-of-N, aggregate all the predictions using the confidence as weights
- **Elite Ensemble:** we use the expert routing model to select the top-k experts, which will then be aggregated by Wisdom Aggregation Strategies



# Datasets


ICEWS[1]


GDELT[2]

**POLECAT**

- Refined ontology
- Better information extraction methods

**POLECAT-FOR-ME:** A subset specifically for Middle East countries

- **Data cleaning**: event deduplication and entity validation to ensure consistency and reliability
- **Historical sequence formation**: group semantically related events into historical event sequences
- **Country dataset construction**: events are partitioned by the attribute ***Country*** to create country-specific dataset. Considering geopolitical relevance and data sufficiency, we select 35 countries and regions[4]. Test set partitioning accounts for the knowledge cutoff of Llama 3.1 (December 2023) to prevent leakage.

[1] https://dataverse.harvard.edu/dataverse/icews
[2] https://www.gdeltproject.org/about.html
[3] Andrew et al. PLOVER and POLECAT: A New Political Event Ontology and Dataset. https://dataverse.harvard.edu/dataverse/POLECAT
[4] Middle East: IRN, ISR, EGY, SAU, TUR, IRQ, YEM, SYR, JOR, ARE, LBN, OMN, KWT, QAT, BHR, CYP, PSE. Global (G20): CHN, USA, RUS, GBR, FRA, DEU, KOR, JPN, IND, CAN, ITA, AUS, ESP, ARG, BRA, IDN, MEX, ZAF. Country codes based on ISO 3166.

# Experimental Results

- Performance (accuracy) comparison across baseline methods, proprietary LLMs, and ThinkTank-ME. For country-specific accuracy, we report results on representative countries from the Middle East. The best results among non-proprietary models are marked in bold, while the overall best results are underlined. Mi and Ma denote micro- and macro-average accuracy.

| Methods | Variants | Country-specific Accuracy | | | | | | | | Total Accuracy | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | ISR | EGY | IRQ | YEM | SYR | LBN | QAT | PSE | Mi | Ma |
| Baselines | No Training | 11.1 | 7.8 | 9.6 | 34.0 | 17.8 | 12.4 | 10.1 | 8.3 | 15.5 | 15.7 |
| | All-data Training | 17.4 | 10.7 | 16.2 | 49.9 | 19.3 | 25.3 | 8.5 | 15.8 | 22.7 | 22.6 |
| Proprietary LLMs | GPT-4o [9] | 22.2 | 17.3 | 15.3 | 37.9 | 18.6 | 15.6 | 20.8 | 15.8 | 23.5 | 24.0 |
| | GPT-4o-mini [9] | 20.3 | 21.3 | 15.3 | 37.2 | 17.7 | 16.3 | 19.2 | 14.1 | 22.7 | 22.5 |
| Expert Routing | - | 13.9 | 8.4 | 11.4 | 39.0 | 15.6 | 23.7 | 10.7 | 13.5 | 18.6 | 19.4 |
| Wisdom Aggregation | Majority Voting | 17.7 | 11.9 | 13.9 | 50.4 | 21.6 | 26.9 | 11.2 | 18.1 | 23.6 | 23.9 |
| | Vanilla Best-of-N | 16.3 | 13.1 | 9.6 | 48.4 | 18.2 | 19.6 | 9.0 | 15.5 | 21.2 | 22.2 |
| | Weighted Best-of-N | 18.0 | 13.9 | 14.4 | 50.7 | 22.2 | 26.6 | 10.9 | 17.5 | 23.9 | 24.4 |
| Elite Ensemble | Majority Voting | 18.0 | 12.8 | 13.5 | 50.7 | 22.6 | 26.9 | 11.2 | 17.0 | 23.9 | 24.2 |
| | Vanilla Best-of-N | 17.7 | 14.1 | 11.8 | 50.9 | 19.2 | 24.8 | 8.0 | 16.1 | 22.8 | 23.9 |
| | Weighted Best-of-N | **18.3** | **15.7** | **14.4** | **51.6** | **23.4** | **26.9** | **11.5** | **17.5** | **24.6** | **25.0** |

ThinkTank-ME: A Multi-Expert Framework for Middle East Event Forecasting. Haoxuan Li, He Chang, and Yunshan Ma et al. WWW 2026.

# Experimental Results

- Elite Ensemble (Weighted Best-of-N) substantially outperforms All-data Training and No Training.
- Our method consistently outperforms the proprietary LLMs (GPT-4o and GPT-4o-mini).
- For country-specific results, proprietary LLMs perform better on high-profile (popular) countries, while our expert models perform better on long-tail (low-resource) countries.

| Methods | Variants | Country-specific Accuracy | | | | | | | | Total Accuracy | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | ISR | EGY | IRQ | YEM | SYR | LBN | QAT | PSE | Mi | Ma |
| Baselines | No Training | 11.1 | 7.8 | 9.6 | 34.0 | 17.8 | 12.4 | 10.1 | 8.3 | 15.5 | 15.7 |
| | All-data Training | 17.4 | 10.7 | 16.2 | 49.9 | 19.3 | 25.3 | 8.5 | 15.8 | 22.7 | 22.6 |
| Proprietary LLMs | GPT-4o [9] | 22.2 | 17.3 | 15.3 | 37.9 | 18.6 | 15.6 | 20.8 | 15.8 | 23.5 | 24.0 |
| | GPT-4o-mini [9] | 20.3 | 21.3 | 15.3 | 37.2 | 17.7 | 16.3 | 19.2 | 14.1 | 22.7 | 22.5 |
| Expert Routing | - | 13.9 | 8.4 | 11.4 | 39.0 | 15.6 | 23.7 | 10.7 | 13.5 | 18.6 | 19.4 |
| Wisdom Aggregation | Majority Voting | 17.7 | 11.9 | 13.9 | 50.4 | 21.6 | 26.9 | 11.2 | 18.1 | 23.6 | 23.9 |
| | Vanilla Best-of-N | 16.3 | 13.1 | 9.6 | 48.4 | 18.2 | 19.6 | 9.0 | 15.5 | 21.2 | 22.2 |
| | Weighted Best-of-N | 18.0 | 13.9 | 14.4 | 50.7 | 22.2 | 26.6 | 10.9 | 17.5 | 23.9 | 24.4 |
| Elite Ensemble | Majority Voting | 18.0 | 12.8 | 13.5 | 50.7 | 22.6 | 26.9 | 11.2 | 17.0 | 23.9 | 24.2 |
| | Vanilla Best-of-N | 17.7 | 14.1 | 11.8 | 50.9 | 19.2 | 24.8 | 8.0 | 16.1 | 22.8 | 23.9 |
| | Weighted Best-of-N | **18.3** | **15.7** | **14.4** | **51.6** | **23.4** | **26.9** | **11.5** | **17.5** | **24.6** | **25.0** |

ThinkTank-ME: A Multi-Expert Framework for Middle East Event Forecasting. Haoxuan Li, He Chang, and Yunshan Ma et al. WWW 2026.

# From MCQ to open-ended forecasting

## Open-ended forecasting

What will happen in the next timestamp?

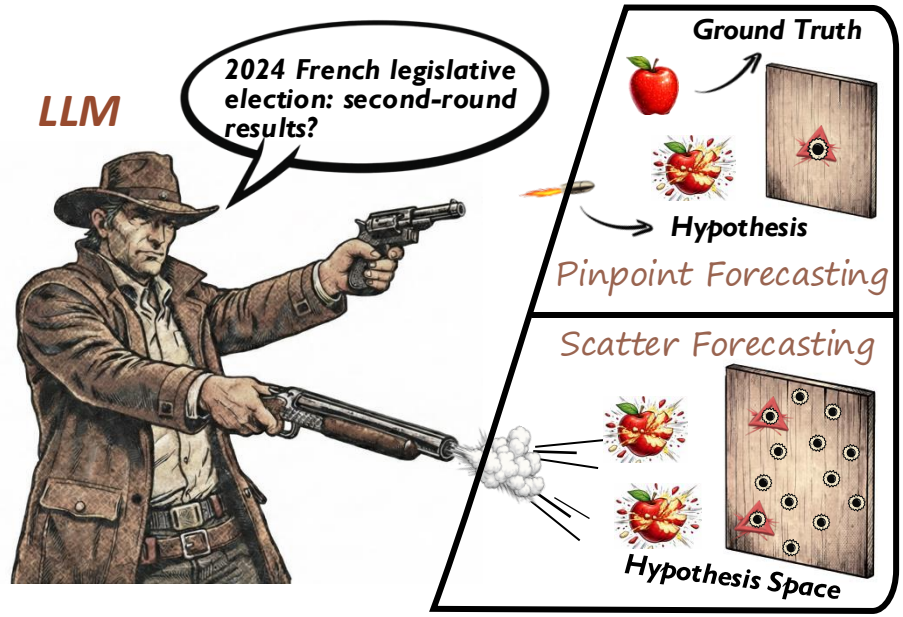
World · Middle East

**Iran leader end of 2026?** (No options provided)

| MCQ / TKG | Open-ended |
|---|---|
| Discriminative:<br>classification or ranking | Generative:<br>directly generate the results |

# Research gaps and motivations

An analogy between open-ended event forecasting and shooting




**Single Prediction (existing)**

- *The New Popular Front became the largest bloc by seats after the second round.*

**Multiple Predictions**

- *The New Popular Front became the largest bloc by seats after the second round.*
- ...
- *NFP and Macron-aligned centrists coordinated withdrawals to block RN, and it worked.*

# Challenges in hypothesis generation

Objective: to achieve both **inclusiveness** and **diversity** of the generated hypotheses

Solution 1: directly prompt LLMs to generate multiple non-redundant hypotheses

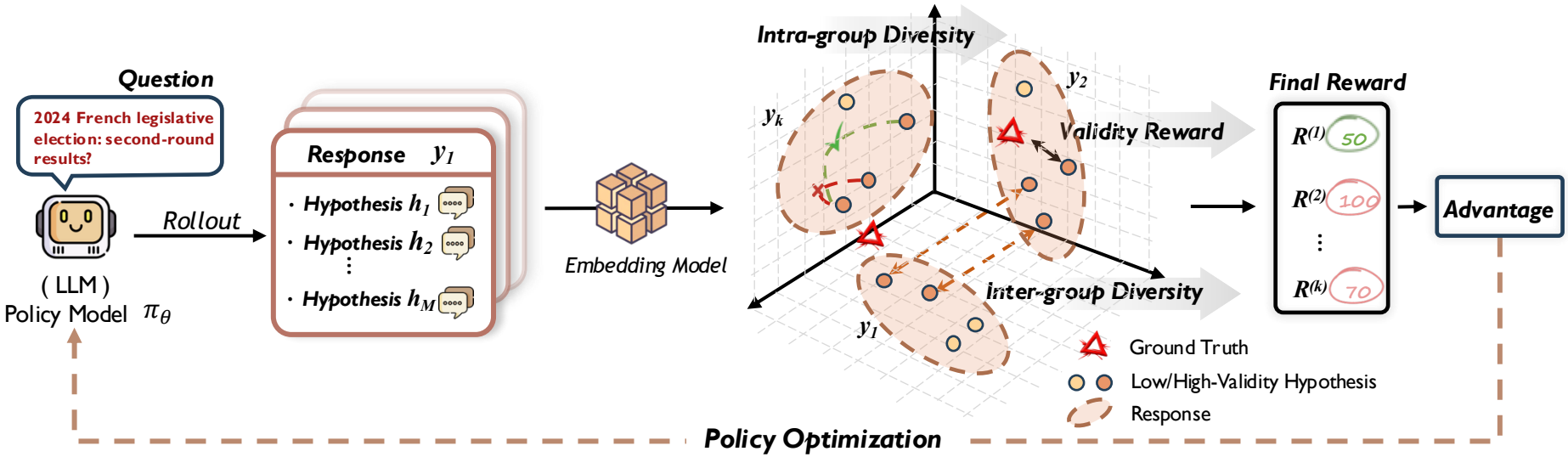
⚠ LLMs struggle to consistently produce distinct yet valid long-tail events

Solution 2: post-training of LLMs using RL, such as GRPO

Optimizing for both inclusiveness and diversity is inherently conflicting

⚠
- Emphasizing inclusiveness tends to favor conservative hypotheses and is prone to mode collapse, ultimately resulting in a single outcome.
- Emphasizing diversity often yields noisy gradients and unstable optimization, thereby hampering convergence.

# Our approach: SCATTER




$$R^{(k)} = R^{(k)}_{\text{validity}} + R^{(k)}_{\text{intra}} + R^{(k)}_{\text{inter}}$$

Scattered Hypothesis Generation for Open-Ended Event Forecasting. He Chang et al. ACL 2026 Findings.

# Experiments

## Overall performance comparison

| Model | OpenForecast | | | | OpenEP | | | |
|---|---|---|---|---|---|---|---|---|
| | Pass@1/Pass@16 | SP@1/SP@16 | SR@1/SR@16 | VR@16 | Pass@1/Pass@16 | SP@1/SP@16 | SR@1/SR@16 | VR@16 |
| **GPT-4o-mini** | $7.55^{\pm0.73}$/25.44 | $9.39^{\pm0.80}$/22.81 | $4.96^{\pm0.37}$/12.76 | 4.23 | $18.82^{\pm1.44}$/32.22 | $6.87^{\pm1.06}$/17.78 | $3.06^{\pm0.81}$/9.28 | 4.08 |
| ***Llama3.2-3B-Instruct*** | | | | | | | | |
| **Base** | $4.40^{\pm0.72}$/21.71 | $9.40^{\pm0.94}$/27.19 | $4.85^{\pm0.53}$/15.42 | 5.73 | **$18.26^{\pm1.76}$/33.33** | $7.29^{\pm1.57}$/18.89 | $2.97^{\pm0.76}$/9.72 | 5.67 |
| **+ SFT** | $4.50^{\pm1.05}$/20.39 | $7.10^{\pm0.66}$/21.27 | $3.64^{\pm0.45}$/11.72 | **7.74** | $6.81^{\pm2.00}$/25.56 | $9.10^{\pm1.85}$/30.00 | $2.37^{\pm0.54}$/11.43 | **6.92** |
| **+ GRPO** | **$11.90^{\pm1.05}$**/28.29 | $13.72^{\pm0.74}$/21.05 | $7.08^{\pm0.42}$/11.61 | 0.71 | $2.85^{\pm1.30}$/5.56 | $11.11^{\pm1.52}$/20.00 | $3.14^{\pm0.70}$/6.35 | 0.62 |
| **+ SCATTER** | $7.73^{\pm0.80}$/**31.36** | **$17.45^{\pm0.81}$/42.32** | **$9.47^{\pm0.55}$/25.64** | 6.54 | $7.36^{\pm2.29}$/25.56 | **$15.69^{\pm2.63}$/36.67** | **$4.67^{\pm1.03}$/15.30** | 4.90 |
| ***Qwen2.5-3B-Instruct*** | | | | | | | | |
| **Base** | $5.96^{\pm1.01}$/25.88 | $6.95^{\pm0.58}$/20.61 | $3.67^{\pm0.26}$/10.72 | 5.07 | **$15.00^{\pm2.22}$**/32.22 | $4.93^{\pm1.71}$/17.78 | $2.51^{\pm0.95}$/10.20 | 5.56 |
| **+ SFT** | $5.51^{\pm0.46}$/23.68 | $8.55^{\pm0.91}$/24.56 | $4.48^{\pm0.45}$/13.85 | 7.10 | $5.00^{\pm2.19}$/20.00 | $6.46^{\pm1.72}$/21.11 | $2.02^{\pm0.71}$/8.38 | 5.52 |
| **+ GRPO** | **$11.84^{\pm0.98}$**/25.66 | $17.60^{\pm0.66}$/27.85 | $9.70^{\pm0.46}$/15.94 | 0.87 | $5.49^{\pm1.73}$/13.33 | $19.65^{\pm1.51}$/26.67 | **$6.92^{\pm0.53}$**/10.58 | 0.76 |
| **+ SCATTER** | $8.55^{\pm0.87}$/**30.26** | **$17.61^{\pm1.12}$/41.89** | **$9.77^{\pm0.59}$/24.29** | **9.14** | $9.65^{\pm2.71}$/**35.56** | **$16.25^{\pm2.75}$/38.89** | $4.88^{\pm0.93}$/**17.59** | **8.21** |

- Datasets: open-ended event forecasting benchmarks, OpenForecast and OpenEP.
- Evaluation metrics: Pass@K, SoftPass@K, SoftRecall@K, VR@K, K in {1, 16}
  - Pass@K: LLM as a judge
  - Soft-xx: using embedding model and cosine similarity to compare the prediction and ground-truth
  - VR@K: ValidRatio, the proportion of valid, non-redundant hypotheses

# Experiments

## Overall performance comparison

| Model | OpenForecast | | | | OpenEP | | | |
|---|---|---|---|---|---|---|---|---|
| | Pass@1/Pass@16 | SP@1/SP@16 | SR@1/SR@16 | VR@16 | Pass@1/Pass@16 | SP@1/SP@16 | SR@1/SR@16 | VR@16 |
| **GPT-4o-mini** | $7.55^{\pm0.73}$/25.44 | $9.39^{\pm0.80}$/22.81 | $4.96^{\pm0.37}$/12.76 | 4.23 | $18.82^{\pm1.44}$/32.22 | $6.87^{\pm1.06}$/17.78 | $3.06^{\pm0.81}$/9.28 | 4.08 |
| ***Llama3.2-3B-Instruct*** | | | | | | | | |
| **Base** | $4.40^{\pm0.72}$/21.71 | $9.40^{\pm0.94}$/27.19 | $4.85^{\pm0.53}$/15.42 | 5.73 | **18.26**$^{\pm1.76}$/**33.33** | $7.29^{\pm1.57}$/18.89 | $2.97^{\pm0.76}$/9.72 | 5.67 |
| **+ SFT** | $4.50^{\pm1.05}$/20.39 | $7.10^{\pm0.66}$/21.27 | $3.64^{\pm0.45}$/11.72 | **7.74** | $6.81^{\pm2.00}$/25.56 | $9.10^{\pm1.85}$/30.00 | $2.37^{\pm0.54}$/11.43 | **6.92** |
| **+ GRPO** | **11.90**$^{\pm1.05}$/28.29 | $13.72^{\pm0.74}$/21.05 | $7.08^{\pm0.42}$/11.61 | 0.71 | $2.85^{\pm1.30}$/5.56 | $11.11^{\pm1.52}$/20.00 | $3.14^{\pm0.70}$/6.35 | 0.62 |
| **+ SCATTER** | $7.73^{\pm0.80}$/**31.36** | **17.45**$^{\pm0.81}$/**42.32** | **9.47**$^{\pm0.55}$/**25.64** | 6.54 | $7.36^{\pm2.29}$/25.56 | **15.69**$^{\pm2.63}$/**36.67** | **4.67**$^{\pm1.03}$/**15.30** | 4.90 |
| ***Qwen2.5-3B-Instruct*** | | | | | | | | |
| **Base** | $5.96^{\pm1.01}$/25.88 | $6.95^{\pm0.58}$/20.61 | $3.67^{\pm0.26}$/10.72 | 5.07 | **15.00**$^{\pm2.22}$/32.22 | $4.93^{\pm1.71}$/17.78 | $2.51^{\pm0.95}$/10.20 | 5.56 |
| **+ SFT** | $5.51^{\pm0.46}$/23.68 | $8.55^{\pm0.91}$/24.56 | $4.48^{\pm0.45}$/13.85 | 7.10 | $5.00^{\pm2.19}$/20.00 | $6.46^{\pm1.72}$/21.11 | $2.02^{\pm0.71}$/8.38 | 5.52 |
| **+ GRPO** | **11.84**$^{\pm0.98}$/25.66 | $17.60^{\pm0.66}$/27.85 | $9.70^{\pm0.46}$/15.94 | 0.87 | $5.49^{\pm1.73}$/13.33 | $19.65^{\pm1.51}$/26.67 | **6.92**$^{\pm0.53}$/10.58 | 0.76 |
| **+ SCATTER** | $8.55^{\pm0.87}$/**30.26** | **17.61**$^{\pm1.12}$/**41.89** | **9.77**$^{\pm0.59}$/**24.29** | **9.14** | $9.65^{\pm2.71}$/**35.56** | **16.25**$^{\pm2.75}$/**38.89** | $4.88^{\pm0.93}$/**17.59** | **8.21** |

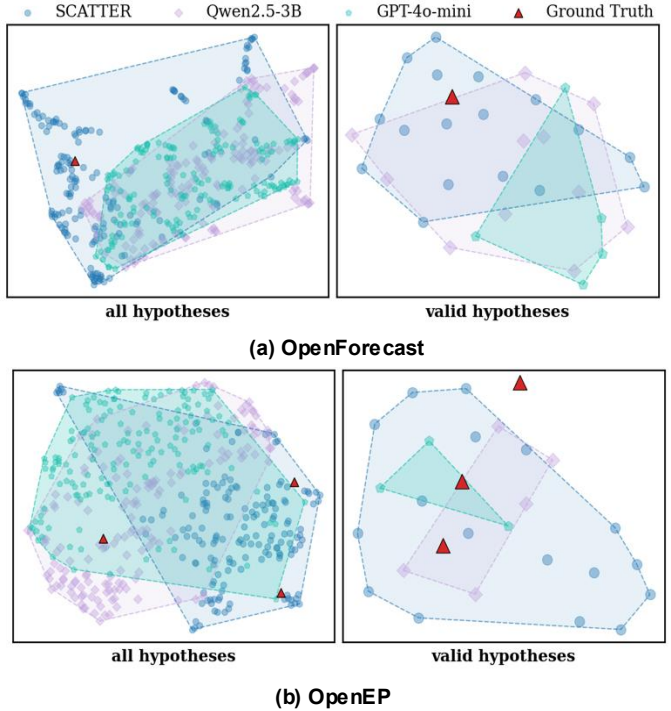
- SCATTER generally performs the best against most of the evaluation metrics
- GRPO performs best on Pass@1, while it has near-zero valid ratio, indicating a severe modal collapse
  - It only optimizes the inclusiveness while sacrificing diversity

# In-depth analysis

- Ablation study of the rewards

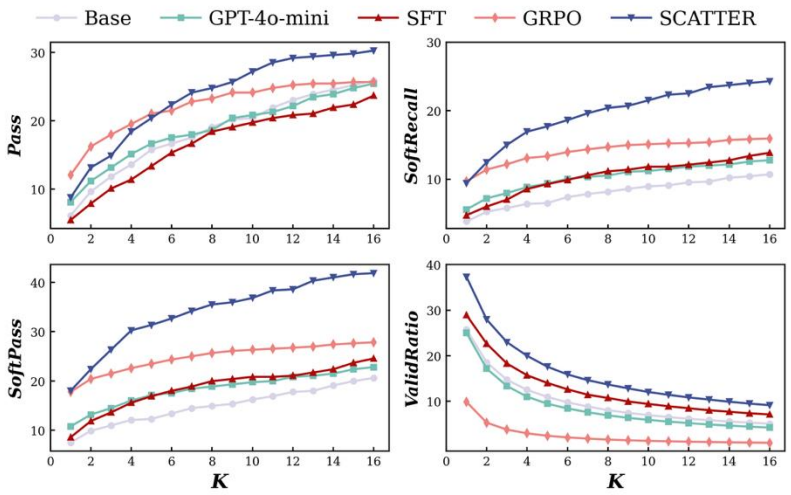
| Model | OpenForecast | | | | OpenEP | | | |
|---|---|---|---|---|---|---|---|---|
| | Pass | SP | SR | VR | Pass | SP | SR | VR |
| *Llama3.2-3B-Instruct* | | | | | | | | |
| **SCATTER** | 31.36 | 42.32 | 25.64 | **6.54** | **25.56** | **36.67** | **15.30** | **4.90** |
| *-inter* | **34.43** | **44.74** | **25.91** | 4.36 | 22.22 | 33.33 | 14.42 | 3.97 |
| *-intra* | 31.14 | 42.11 | 25.36 | 2.64 | 23.33 | 30.00 | 12.45 | 1.93 |
| **GRPO** | 28.29 | 21.05 | 11.61 | 0.71 | 5.56 | 20.00 | 6.35 | 0.62 |
| *Qwen2.5-3B-Instruct* | | | | | | | | |
| **SCATTER** | 30.26 | **41.89** | 24.29 | **9.14** | **35.56** | 38.89 | 17.59 | **8.21** |
| *-inter* | **33.11** | 41.01 | **24.44** | 4.74 | 33.33 | **40.00** | **18.84** | 4.78 |
| *-intra* | 33.11 | 41.45 | 24.14 | 4.19 | 32.22 | 40.0 | 18.7 | 4.14 |
| **GRPO** | 25.66 | 27.85 | 15.94 | 0.87 | 13.33 | 26.67 | 10.58 | 0.76 |

- Visualization of the results

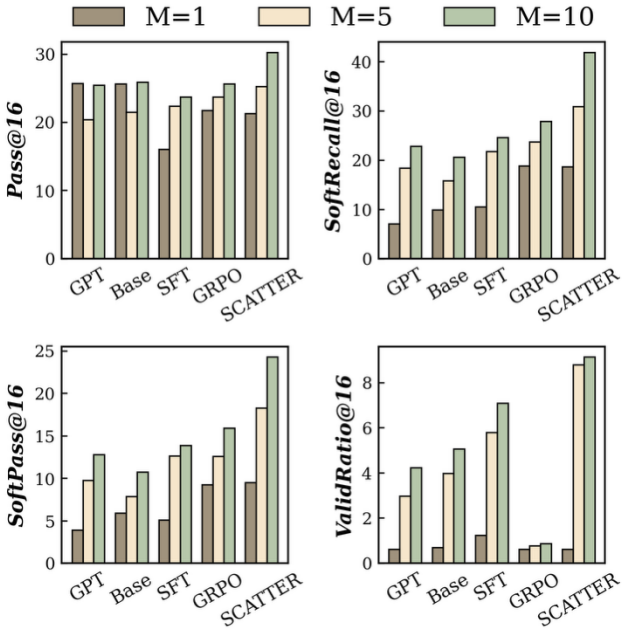


# Hyperparameter analysis

- Performance w.r.t. sampling rounds K




- Performance w.r.t. #hypothesis per round



# Key take-aways

ThinkTank-ME – Multi-Agent Framework

- A multi-expert framework for temporal event forecasting
- Including multiple expert models and one leader model, inspired by real-world ThinkTank
- Applied in the middle-east geopolitical event forecasting scenario

SCATTER – LLM Post Training

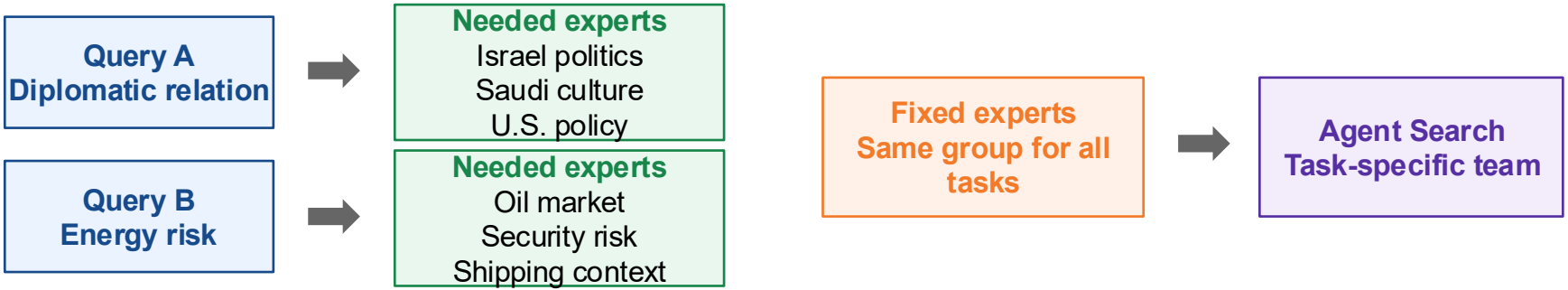
- A paradigm shift from single prediction to scatter forecasting
- Introduce hypothesis generation as the proxy task to forecasting
- Design post-training framework SCATTER to achieve both inclusiveness and diversity
- Evaluated on open-ended event forecasting benchmarks: OpenForecast and OpenEP

? What's next beyond multi-agent and post-training

# Future work: from multi-expert selection to agent search

## Motivation: why do we need Agent Search?

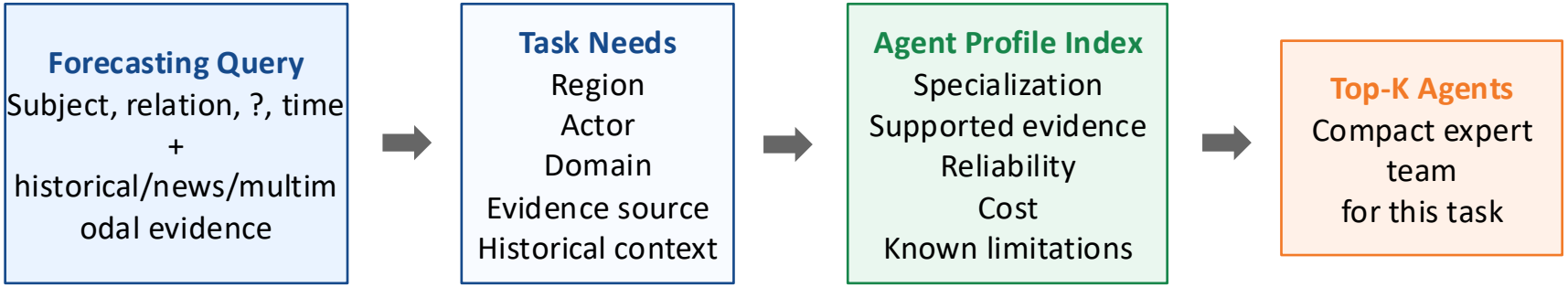
- One expert per country is still too limited, we need to define more fine-grained experts w.r.t. region, domain, time, celebrity, and more importantly, composition of experts. Large number of agents make it a must for agent search.
- Finetuning expert model becomes more and more challenging to small groups, the cutting-edge LLMs are more and more powerful, we need to probe the internal knowledge using agents rather than inject more knowledge constrained by limited resources and scale.
- Convert a central LLM to agent search offer more controllability for human to audit and monitor the AI decision making process, increasing trust and mitigating bias.



# Future work: from multi-expert selection to agent search

**A conceptual design:**

- Forecast Agent Search treats expert agents as searchable, rankable, and composable resources.
- Each agent is described by a lightweight profile, so it can be indexed and retrieved for a specific query.




# THANK YOU & QA

Website: https://mysbupt.github.io/