Can LLMs Help Decentralized Dispute Arbitration? A Case Study of UMA-Resolved Markets on Polymarket

Junhao Wen
Southwest University
Chongqing, China
wen264132454@email.swu.edu.cn

Juncen Zhou
University of Sydney
Sydney, Australia
zhou_2002510@163.com

Junjie Huang*
Southwest University
Chongqing, China
junjiehuang@swu.edu.cn

ABSTRACT

Web3 prediction markets, exemplified by Polymarket, have gained prominence for leveraging collective intelligence to forecast a wide range of social, political, and sports events. However, among the thousands of prediction market events, consensus `disputes` still arise due to imperfections in market mechanisms. On Polymarket alone, the trading volume involving disputed events has reached $972,370,804.71, underscoring the critical need for objective and efficient dispute resolution. In this study, we introduce large language models (LLMs) to: (1) evaluate whether web-enabled LLMs can reproduce the decision quality of UMA's on-chain voting process once a dispute has been raised, and (2) predict, based on event rules, which market events are likely to face future `disputes` before they occur. Our findings show that LLMs are unable to reliably predict which events will become disputed in advance; however, once a dispute is initiated, web-enabled LLMs achieve 89.58% agreement with UMA's final resolutions and demonstrate strong stability.

CCS CONCEPTS

• **Information systems → Collaborative and social computing systems and tools**; **Web mining**.

KEYWORDS

Web3, Prediction Market, Large Language Models

ACM Reference Format:
Junhao Wen, Juncen Zhou, and Junjie Huang. 2026. Can LLMs Help Decentralized Dispute Arbitration? A Case Study of UMA-Resolved Markets on Polymarket. In *The Web Conference '26: ACM Web Conference, April 13–17, 2026, Dubai, UAE.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Decentralized prediction markets have emerged as one of the most prominent applications of Web3, offering a mechanism for aggregating collective intelligence to forecast real-world events. Platforms such as Polymarket allow users to trade on outcomes across political, economic, and sports domains, with market prices reflecting collective expectations grounded in publicly available information [16]. Despite their scalability and transparency, these systems remain vulnerable to `disputes` (i.e., disagreements over event outcomes raised by users or resolvers) caused by ambiguous event wording, inconsistent external reporting, or evolving real-world developments. To address such disagreements, Polymarket relies on **UMA's Optimistic Oracle (OO)**[1], which escalates contested outcomes to a decentralized token-holder vote. While effective in many cases, this governance process has faced criticism for potential subjectivity, coordination challenges, and susceptibility to interpretive ambiguity. Our data indicates that `disputed` markets on Polymarket—whether current or historical—represent $972,370,804.71 in trading volume, demonstrating the vital need for an effective resolution mechanism.

In parallel, recent advances [4, 15, 17] in large language models (LLMs) have demonstrated strong capabilities in retrieving, interpreting, and synthesizing real-world information. Their increasing reliability raises an important question: Can LLMs serve as impartial evaluators in decentralized governance systems, either by predicting which markets are likely to become `disputed` or by resolving `disputes` once they occur? If effective, LLM-based adjudication could offer a scalable and transparent complement to traditional, human-driven oracle mechanisms.

This study offers the first systematic examination of LLMs within the lifecycle of Web3 prediction-market governance. We investigate two core research questions: **RQ1** evaluates **whether LLMs can reproduce UMA's final decisions** in `disputed` Polymarket events using only information available before the UMA vote is finalized, thereby assessing whether LLMs can effectively substitute for the token-holder voting stage. **RQ2** examines **whether LLMs can predict `dispute` formation in advance** based solely on the semantic content of event rules. Our results show that although LLMs struggle to anticipate which events will eventually be `disputed`, they achieve high pre-vote agreement with UMA's final outcomes once a `dispute` has been raised—achieving 89.58% for DeepSeek V3.1 and 89.19% for Qwen Max, with Qwen Max further demonstrating exceptionally strong internal stability (96.14%). These findings indicate that while LLMs exhibit limited predictive capability for identifying `disputes`, they serve as reliable and consistent evaluators during the `post-dispute` resolution phase.

Overall, this work highlights both the promise and the limitations of LLMs in decentralized governance systems. Rather than replacing existing oracle mechanisms, LLMs may function as consistent, auditable, and efficient assistants in `dispute` resolution, helping to enhance transparency and reduce the burdens associated with human-driven voting processes.

*Junjie Huang is the Corresponding author.

[1] https://uma.xyz

The Web Conference '26, April 13–17, 2026, Dubai, United Arab Emirates

ACM ISBN 978-1-4503-XXXX-X/26/04…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

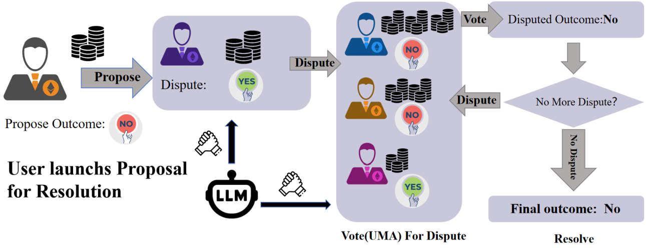

Figure 1: The Polymarket dispute lifecycle: a proposed outcome can be challenged via staking, triggering stake-weighted voting rounds until resolution. We employ LLMs to predict and resolve these `disputes`.

2 BACKGROUND

2.1 Polymarket

Polymarket[2] is a leading decentralized prediction market platform that enables users to trade on the outcomes of real-world events across politics, economics, and culture. It leverages market mechanisms to aggregate public information and quantify collective beliefs through prices that reflect event probabilities. Built on polygon blockchain technology, Polymarket ensures transparency, liquidity, and censorship resistance, attracting significant trading volumes and mainstream attention. To securely resolve market outcomes, it relies on **UMA's Optimistic Oracle (OO)** for decentralized and verifiable truth determination.

2.2 UMA

UMA (short for *Universal Market Access*) is a decentralized protocol deployed on Ethereum and other blockchains. One of its core components is the **Optimistic Oracle (OO)**, a mechanism designed to securely and efficiently bring verifiable real-world data or event outcomes (e.g., *whether a person won a competition*).

The **OO** operates under an ASSUME HONESTY FIRST paradigm: when a participant submits an *assertion* about a real-world fact, they must post a bond as collateral, and the claim then enters a predetermined *challenge period*. If no one `disputes` it during this window, the assertion is considered correct by default and finalized automatically. However, if challenged, the claim escalates to a `dispute phase`, where UMA's Data Verification Mechanism (DVM) based dispute resolution life circle is invoked(see Figure 1). The **DVM** relies on UMA token holders to vote on the correct outcome, aligning economic incentives to ensure honest and accurate resolution. This design allows most queries to be resolved efficiently and inexpensively, while only the small fraction of `disputed` cases invoke the more resource-intensive voting process, striking a balance between decentralization and scalability.

Nevertheless, the **OO** system also carries certain risks and limitations. Its optimistic design relies on rational and well-incentivized participants; if voting power becomes overly concentrated or if community engagement is low, outcomes could be influenced by a small subset of token holders. Moreover, because many event statements are expressed in natural language, ambiguity in question phrasing can lead to inconsistent interpretations or contentious resolutions. These risks have manifested in real-world applications such as Polymarket.

2.3 `Disputed` Example

A notable example of these risks occurred in the Polymarket event *Will Zelenskyy wear a suit before July?*, which drew over $200 million in trading volume. The market asked whether Ukrainian President Volodymyr Zelenskyy would appear publicly in a "suit" between May 22 and June 30, 2025. Controversy emerged after he attended a NATO summit wearing a dark jacket, shirt, and trousers, which some participants described as a suit while others rejected because it lacked traditional elements such as a matching set or a tie. Both "Yes" and "No" outcomes were proposed and `disputed`, prompting **OO** to resolve the question through its voting mechanism. The final decision of "No" triggered backlash [6] from traders who claimed governance centralization and subjective interpretation had compromised fairness. UMA's co-founder, Hart Lambur, rejected manipulation claims, but the controversy revealed how even simple, natural-language questions can challenge decentralized truth systems reliant on collective judgment.

These controversies highlight that while **OO** represents a groundbreaking innovation in decentralized data verification, challenges remain in governance, incentive alignment, and the interpretation of natural-language conditions.

2.4 Related Work

Prediction markets have long been studied as mechanisms for aggregating distributed beliefs, with foundational work on automated market makers, illustrating how liquidity and incentives shape information aggregation [2, 9]. Decentralized variants, such as Polymarket, extend these ideas to trust-minimized environments, relying on oracle-based governance to finalize outcomes [3, 7]. UMA's Optimistic Oracle represents a prominent design in this space, using challenge-based verification and token-holder voting to resolve ambiguous or `disputed` events [14]. Despite their scalability, existing oracle systems also face critical limitations. For example, implementation surveys have shown that trust-minimized oracle designs cannot fully eliminate integrity and freshness risks, especially when aggregating data from heterogeneous sources [10].

Concurrently, rapid advances in large language models (LLMs) have spurred research into autonomous agents and LLMs' web-search for decentralized governance [8, 11, 13]. Recent work shows that web-enabled LLM agents can evaluate DAO proposals and align with collective voting behavior [1], while self-sovereign LLM agents demonstrate the feasibility of trust-minimized autonomous decision-makers in Web3 systems [5]. Complementary studies in AI-assisted governance and `dispute` resolution [12, 18] further suggest that LLMs can perform structured fact interpretation and rule-based reasoning. These findings collectively indicate that LLM-driven agents may augment or partially automate governance workflows in decentralized environments.

In this paper, we present the first attempt to apply large language models to decentralized dispute arbitration and investigate whether LLMs can assist in resolving on-chain disputes.

[2] https://polymarket.com

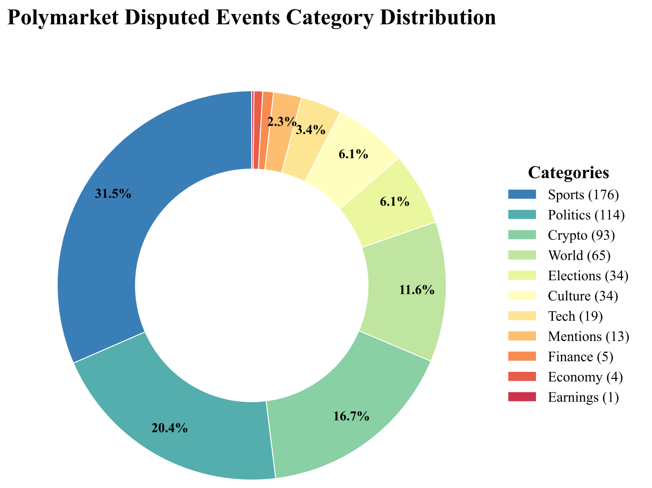

Figure 2: Distribution of Polymarket disputed Events by Category. Sports (31.5%), Politics (20.4%), and Crypto (16.7%) are the three largest categories.

3 DATASET AND DATA PREPROCESSING

We retrieved all available event data from Polymarket using the official **API**[3] endpoint, iteratively accessing paginated results to ensure complete coverage of all active and historical markets. The API was queried three times at regular intervals, and we observed that each subsequent query returned all previously obtained events along with newly created ones, indicating that the official endpoint is stable and consistent[4]. Each record represents a prediction market event with detailed metadata including its identifier, question content, `resolution statuses`, market outcomes, prices, and sub-event grouping information.Using the `umaResolutionStatuses` , we filtered for records containing the keyword `disputed`, resulting in a total of 558 user-disputed events.Based on these events, we will conduct further filtering and analysis in Section 4 and Section 5.

A multi-LLM classification employing three distinct models was performed on the disputed events. The outcome demonstrated high classification stability, with 457 events (81.90%) achieving full consistency (*Very Stable*) and 98 events (17.56%) showing large consistency (*Stable*), resulting in nearly 99.5% of events exhibiting robust agreement. Only three events (0.54%) were inconsistent (*Unstable*), which were subsequently resolved through manual review and assigned to the **Mentions** category (IDs: 542935, 534923, 570937). As depicted in Figure 2, the distribution of event categories post-classification reveals a significant concentration of disputed events within the **Sports**, **Politics**, and **Crypto** domains, collectively accounting for approximately 68.6% of all disputes. This pronounced asymmetry, reliably identified via highly stable LLM classification, suggests a clear direction for subsequent research, primarily focusing on characterizing the unique factors driving disputes specifically within these three predominant categories.

Table 1: Consistency reflects how closely LLM predictions align with UMA's final resolutions across 259 events.

Model	Consistency (259)	Very-Stable	Stable	Unstable
DeepSeekV3.1	232 (89.58%)	227 (87.64%)	32 (12.36%)	0 (0.00%)
Qwen Max	231 (89.19%)	249 (96.14%)	10 (3.86%)	0 (0.00%)
Claude-4.5-Sonnet	228 (88.03%)	248 (95.75%)	11 (4.25%)	0 (0.00%)
gpt4o	203 (78.38%)	166 (64.09%)	89 (34.36%)	4 (1.54%)
gpt-4o-search-preview	185 (71.43%)	223 (86.10%)	35 (13.51%)	1 (0.39%)

4 CAN LLMS REPRODUCE UMA'S FINAL DECISIONS?

A Polymarket event's complete `dispute` trajectory (see Figure 1) is encoded in the `umaResolutionStatuses` field, which records each state transition the market undergoes. A typical sequence may look like:

```
["proposed", "disputed", "proposed", "resolved"].
```

Using this field in our collected API data, we first identify all individual occurrences of the `disputed` state, yielding a total of 558 user-initiated dispute events. Next, we restrict attention to markets that both (1) contain at least one `disputed` transition and (2) ultimately reach a terminal `resolved` state. Applying these criteria results in a final set of 259 `disputed-and-resolved` markets, which form the basis of our subsequent analysis.Our goal is to investigate whether LLMs equipped with web-search capabilities can independently reproduce UMA's final resolutions. If LLMs can match UMA' outcomes, this would suggest the feasibility of a more transparent and objective mechanism for resolving decentralized prediction markets—potentially mitigating subjective biases or manipulation risks in the UMA voting process.To evaluate this, we prompt five frontier LLMs to search only for information available before the UMA decision timestamp and to infer the correct market resolution according to the official Polymarket rulebook. Each model is queried three independent times to measure stability. For each event, a model's final answer is determined by majority vote across its three runs. Consistency is then computed as the agreement rate between the model's majority-vote prediction and the ground-truth UMA resolution across the 259 `disputed` markets.

The results, including both model–UMA consistency and model internal stability (very-stable, stable, unstable across the three runs), are shown in Table 1. Across the five evaluated frontier models, we observe a clear separation in both alignment with UMA outcomes and internal stability. DeepSeekV3.1, Qwen Max, and Claude-4.5-Sonnet achieve the highest consistency (88–90%) and very-stable rates (exceeding 95% for the latter two), demonstrating that frontier LLMs can reproduce UMA's final resolutions with high fidelity when limited to information available at decision time. While the two OpenAI variants show comparatively lower consistency—especially `gpt-4o-search-preview` (71.43%)—their performance still exceeds chance by a wide margin, indicating nontrivial signal extraction under temporal constraints. Overall, these results affirm that LLMs are capable of approximating UMA arbitration outcomes, though persistent model-specific reliability gaps highlight the need for careful validation before deploying LLM-assisted adjudication in decentralized governance settings.

[3] https://gamma-api.polymarket.com/markets

[4] The most recent retrieval was performed on UTC `2025-10-27 11:58:47`, yielding a total of 140,582 Polymarket events.

Table 2: Model performance when predicting dispute outcomes using only pre-dispute semantic information.

Model	Accuracy	Precision	Recall	F1 Score
Claude-Sonnet-4.5	0.5701	0.6305	0.3388	0.4408
DeepSeek-V3.1	0.5410	0.5680	0.3424	0.4273
Qwen-Max	0.4731	0.4601	0.3157	0.3745
GPT-4o	0.3998	0.3618	0.2623	0.3041

5 CAN LLMS PREDICT UMA DISPUTES IN ADVANCE?

To evaluate this hypothesis, we construct a structured few-shot prompt using twelve events: nine disputed cases covering diverse dispute types (semantic ambiguity, definitional uncertainty, evidentiary limitations, reporting inconsistencies, and temporal instability), plus three non-disputed events. The choice of twelve exemplars balances coverage of major dispute categories while keeping the prompt size tractable for all models. Models receive only the event rule text and timestamps, ensuring that predictions rely solely on semantic features of the rule descriptions.

The dataset for this experiment is derived from all Polymarket events that contain at least one user-initiated disputed transition, yielding 558 total dispute events. From these, nine disputed events are selected as in-context exemplars for the prompt, leaving 549 remaining disputed cases. For evaluation, we pair these 549 cases with 549 randomly sampled non-disputed events, forming a balanced test set of 1,098 samples. Because the dataset is balanced, a random or majority-class baseline corresponds to an accuracy of 0.50. We evaluate four LLMs on this test set, and their performance is shown in Table 2.

From Table 2, we observe that predictive performance remains limited across all models. Although some models perform slightly above the random baseline (for example, Claude with an accuracy of 0.5701), none achieve strong discriminative ability. This indicates that semantic information derived solely from event rules and pre-dispute descriptions provides only weak signals for forecasting future disputes.

This observation leads to two insights. First, within our dataset, neither the clarity nor the structural quality of Polymarket rule wording appears to play a decisive role in determining whether an event will later become disputed. Textual cues alone are not sufficient for reliably anticipating dispute formation. Second, the emergence of disputes seems to be shaped mainly by external real-world developments, including evolving facts, unstable or incomplete evidence, conflicting reports, media dynamics, and delays in authoritative confirmation.

6 CONCLUSION

Our analysis provides clear answers to the two research questions posed in this work. For **RQ1**, we find that advanced LLMs with controlled access to public information can closely align with UMA's final dispute resolution outcomes, showing strong agreement and internal stability within our experimental setting. This result indicates that the evidence interpretation process, which is traditionally carried out through decentralized voting by token holders, can in part be approximated by an autonomous agent equipped with retrieval capabilities. Rather than replacing human governance, LLMs may function as complementary adjudication tools that deliver consistent and auditable reasoning and help promote transparency and scalability in optimistic oracle frameworks.

For **RQ2**, we show that LLMs cannot reliably predict which Polymarket events will become disputed when limited to the semantic content of event rules. Textual clarity or ambiguity in the rule descriptions does not provide a sufficient signal for forecasting dispute formation. Instead, disputes tend to emerge from external real-world uncertainty, evolving evidence, reporting inconsistencies, and delays in authoritative confirmation, all of which fall outside the rule text itself.

Taken together, these findings point to a natural division of labor. The emergence of disputes is shaped by the evolving state of real-world information, whereas dispute resolution depends on interpreting that information once it becomes available. Within this lifecycle, LLMs appear well-suited for assisting in the resolution stage but are not effective for anticipating disputes based solely on event semantics. This reveals a realistic and bounded role for LLM-based agents in decentralized governance: providing stable and consistent evaluations in the post-dispute stage while leaving pre-dispute forecasting to market behavior.

REFERENCES

[1] Agostino Capponi et al. 2025. DAO-AI: Evaluating Collective Decision-Making with Web-Enabled LLM Agents. *arXiv preprint arXiv:2510.21117* (2025).
[2] Yiling Chen and David Pennock. 2010. A survey of prediction market design. In *Algorithmic Game Theory*. 1–33.
[3] Andrea Chiarelli et al. 2023. A Systematic Literature Review of Blockchain Oracles. *Aalto University Publication Series* (2023).
[4] Zheng Chu, Jingchang Chen, Qianglong Chen, Haotian Wang, Kun Zhu, Xiyuan Du, Weijiang Yu, Ming Liu, and Bing Qin. 2024. BeamAggR: Beam Aggregation Reasoning over Multi-source Knowledge for Multi-hop Question Answering. In *ACL*. 1229–1248.
[5] Botao Amber Hu, Yuhan Liu, and Helena Rong. 2025. Trustless Autonomy: Self-Sovereign Large Language Model Agents in Decentralized Systems. *arXiv preprint arXiv:2505.09757* (2025).
[6] Joel Khalili and Kate Knibbs. 2025. *Volodymyr Zelensky's Clothing Has Sparked a Polymarket Rebellion.* https://www.wired.com/story/volodymyr-zelensky-suit-polymarket-rebellion
[7] Kartikay Kumar and Muhammad Khan. 2021. Decentralized oracles: A comprehensive survey. *IEEE Access* 9 (2021), 92272–92294.
[8] Ollie Liu, Deqing Fu, Dani Yogatama, and Willie Neiswanger. 2025. DeLLMa: Decision Making Under Uncertainty with Large Language Models. In *ICLR*.
[9] Othman et al. 2013. A practical liquidity-sensitive automated market maker. *TEAC* 1, 3 (2013), 1–25.
[10] Amir Pasdar and Young Choon Lee. 2023. A Survey on Blockchain Oracle Implementation. *Comput. Surveys* 55, 12 (2023), 1–36.
[11] Oriane Peter and Kate Devlin. 2025. Decentralising LLM Alignment: A Case for Context, Pluralism, and Participation. In *AAAI*.
[12] Jaromir Savelka et al. 2023. Large Language Models in Legal Reasoning: A Study on Real-World Case Interpretation. *Journal of Artificial Intelligence and Law* (2023).
[13] Sofia Eleni Spatharioti et al. 2025. Effects of LLM-based Search on Decision Making: Speed, Accuracy, and Overreliance. In *CHI*. ACM.
[14] UMA Protocol. 2024. Understanding UMA's Optimistic Oracle and Its Governance Mechanisms.
[15] Zhepei Wei, Wei-Lin Chen, and Yu Meng. 2025. InstructRAG: Instructing Retrieval-Augmented Generation via Self-Synthesized Rationales. In *ICLR*.
[16] Justin Wolfers and Eric Zitzewitz. 2004. Prediction Markets. *Journal of Economic Perspectives* 18, 2 (2004), 107–126.
[17] Qianqian Xie et al. 2024. FinBen: A Holistic Financial Benchmark for Large Language Models. In *NeurIPS*.
[18] Shunyu Yao et al. 2024. Lawyer GPT: A Legal Large Language Model with Enhanced Domain Knowledge and Reasoning Capabilities. In *Proceedings of the 2024 3rd International Symposium on Robotics, Artificial Intelligence and Information Engineering*. 108–112.