

T²Agent: A Tool-augmented Multimodal Misinformation Detection Agent with Monte Carlo Tree Search

Xing Cui¹, Yueying Zou¹, Zekun Li², Peipei Li^{1*}, Xinyuan Xu¹, Xuannan Liu¹, Huaibo Huang³

¹Beijing University of Posts and Telecommunications

²University of California, Santa Barbara

³MAIS & NLPR, Institute of Automation, Chinese Academy of Sciences
{cuixing, zouyueying2001, lipeipei, xuxinyuan, liuxuannan}@bupt.edu.cn,
zekunli@cs.ucsb.edu, huaibo.huang@cripac.ia.ac.cn

Abstract

Real-world multimodal misinformation often arises from mixed forgery sources, requiring dynamic reasoning and adaptive verification. However, existing methods mainly rely on static pipelines and limited tool usage, limiting their ability to handle such complexity and diversity. To address this challenge, we propose T²Agent, a novel misinformation detection agent that incorporates an extensible toolkit with Monte Carlo Tree Search (MCTS). The toolkit consists of modular tools such as web search, forgery detection, and consistency analysis. Each tool is described using standardized templates, enabling seamless integration and future expansion. To avoid inefficiency from using all tools simultaneously, a greedy search-based selector is proposed to identify a task-relevant subset. This subset then serves as the action space for MCTS to dynamically collect evidence and perform multi-source verification. To better align MCTS with the multi-source nature of misinformation detection, T²Agent extends traditional MCTS with multi-source verification, which decomposes the task into coordinated subtasks targeting different forgery sources. A dual reward mechanism containing a reasoning trajectory score and a confidence score is further proposed to encourage a balance between exploration across mixed forgery sources and exploitation for more reliable evidence. We conduct ablation studies to confirm the effectiveness of the tree search mechanism and tool usage. Extensive experiments further show that T²Agent consistently outperforms existing baselines on challenging mixed-source multimodal misinformation benchmarks, demonstrating its strong potential as a training-free detector.

Code — <https://github.com/cuixing100876/T2Agent>

Introduction

The remarkable progress in Artificial Intelligence Generated Content (AIGC) technologies (Ouyang et al. 2022; Touvron et al. 2023; Dhariwal and Nichol 2021; Rombach et al. 2022; Cui et al. 2024a,b) has lowered the barrier to generating sophisticated multimodal misinformation, posing severe threats to information integrity, public governance, economic stability, and societal well-being (Zannettou et al.

2019; Apuke and Omar 2021; Allcott and Gentzkow 2017). Addressing this issue is not only a technical imperative but also a societal necessity. Therefore, developing approaches for multimodal misinformation detection is critical to safeguard the integrity of digital information ecosystems.

Automated multimodal misinformation detection is inherently complex, requiring capabilities in reasoning, information retrieval, and cross-modal verification. While recent automated approaches have made progress (Liu et al. 2024c; Braun et al. 2025; Beigi et al. 2025), they still fall short of emulating the dynamic strategies employed by human experts. This gap can be attributed to two key factors. First, the diversity of forgery sources (Guo et al. 2025) necessitates tailored tools for different scenarios. For instance, the AMG benchmark (Guo et al. 2025) considers temporal inconsistencies, whereas MMfakebench (Liu et al. 2024c) incorporates counterfactual misinformation. However, most existing LLM-based methods rely on fixed and limited toolsets, which lack the flexibility and scalability required to handle such a wide range of forgery types. Second, real-world multimodal misinformation often arises from mixed forgery sources, such as textual inaccuracies, image manipulations, or cross-modal inconsistencies (Liu et al. 2024c). Reliable detection requires not only the ability to perform in-depth exploitation to retrieve evidence for each forgery source but also to adaptively explore multiple potential forgery sources. However, existing LLM-based methods (Liu et al. 2024c; Braun et al. 2025; Beigi et al. 2025; Li, Zhang, and Malt-house 2024; Lakara et al. 2024b) typically rely on rigid designs that fail to strike this critical balance, limiting their effectiveness in complex scenarios.

To address these challenges, we introduce T²Agent, a novel misinformation detection agent that incorporates an extensible toolkit with Monte Carlo Tree Search (MCTS). As illustrated in Fig. 1, the extensible toolkit includes a range of functional tools such as web searching, time detection, forgery detection, counterfactual detection, image understanding, and entity recognition. Each tool is described using standardized templates, enabling seamless integration and future expansion for new tasks. Given that using all available tools simultaneously may overwhelm the agent and reduce efficiency (Lu et al. 2025), we introduce a tool selection mechanism based on greedy search. This mechanism

*Corresponding Author

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

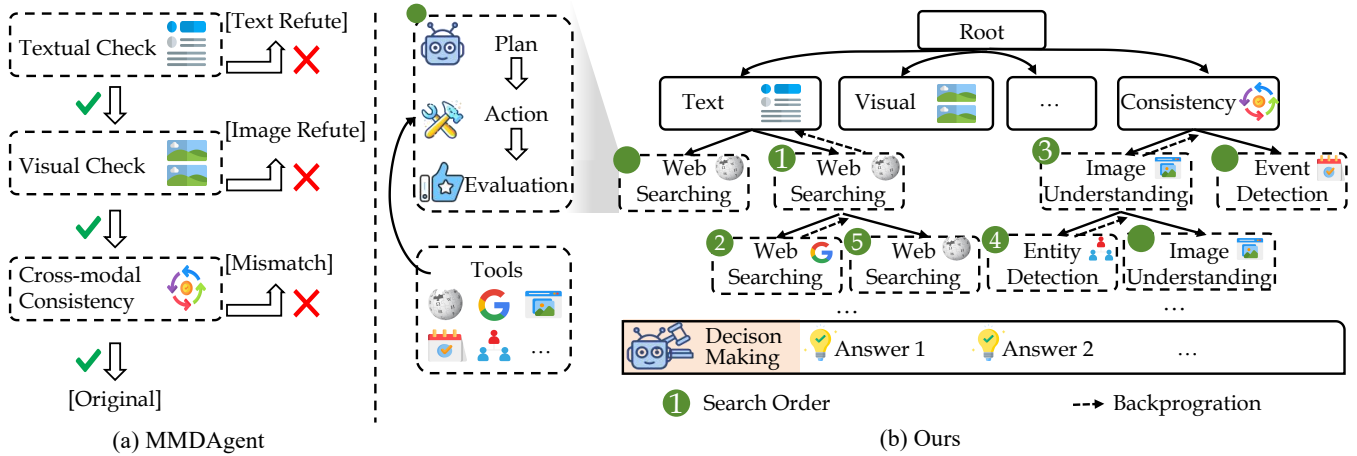


Figure 1: (1) MMDAgent adopts a fixed verification process. (2) Our T²Agent builds a multi-source verification framework inspired by MCTS, enabling dynamic verification through adaptive tool selection and evidence integration.

identifies the most relevant subset of tools for each task type, forming a task-specific subset. Then, the optimized subset serves as an action space for MCTS to dynamically collect evidence and perform multi-source verification. Different from the previous MCTS, which is primarily designed for tasks with a single target, T²Agent extends MCTS with multi-source verification. Specifically, T²Agent first decomposes the misinformation detection task into multiple subtasks, each corresponding to a potential forgery source. Subsequently, it dynamically performs verification. To guide this process effectively, we propose a dual evaluation function that combines two components: the reasoning trajectory score, which evaluates the quality of the reasoning path, and the result confidence score, which measures the certainty of the final decision. By balancing exploration across mixed forgery sources with the exploitation of high-confidence evidence, T²Agent enhances robustness and generalization in complex detection scenarios.

Our experiments demonstrate that T²Agent presents a promising performance in mixed-source multimodal misinformation detection. On the MMfakebench (Liu et al. 2024c), T²Agent improves accuracy of the baseline MMDAgent (Liu et al. 2024c) by 28.7% with GPT4-o (OpenAI 2022), achieving a new state-of-the-art (SOTA). On AMG (Guo et al. 2025), T²Agent is competitive with existing training-based approaches, offering a promising direction for enhancing misinformation detection without the need of additional training. Ablation studies further confirm that the performance gains stem from the use of MCTS and tool integration. The primary contributions of this paper can be summarized as follows:

- We propose T²Agent, a novel misinformation detection agent that integrates an extensible toolkit with Monte Carlo Tree Search (MCTS). By dynamically planning verification paths, leveraging tools for evidence collection, and performing structured multi-source validation, T²Agent enables adaptive reasoning across diverse forgery sources.

- We design an extensible toolkit based on modularized, standardized tool templates. This design facilitates rapid adaptation to various forgery patterns and allows seamless integration of new verification capabilities, improving system scalability and generalization.
- We extend traditional MCTS with a multi-source verification framework by decomposing the detection task into subtasks, where each subtask targets a specific forgery source. A dual evaluation function further guides cross-subtask reasoning, balancing exploration across sources with exploitation of reliable evidence.

Related work

Misinformation Detection

Early misinformation detection datasets primarily focus on a single source, such as textual veracity distortion (Wang 2017; Thorne et al. 2018; Shu et al. 2020; Hanselowski et al. 2019; Yao et al. 2023a; Chen and Shu 2024), or cross-modal inconsistency (Luo, Darrell, and Rohrbach 2021; Aneja, Bregler, and Nießner 2023; Shao, Wu, and Liu 2023; Suryavardan et al. 2023; Nielsen and McConville 2022). Some methods focus on mining forgery features from given content with specifically designed networks (Hu et al. 2023; Papadopoulos et al. 2025; Shao, Wu, and Liu 2023). Some recent works (Cheung and Lam 2023; Zeng et al. 2024; Liu et al. 2024b; Qi et al. 2024; Yang and Rocha 2024) fine-tune LLMs to enhance their detection capabilities for specific tasks. Benefiting from the powerful reasoning ability of LLMs, some works (Kakizaki, Matsunaga, and Furukawa 2025; Li, Zhang, and Malthouse 2024; Beigi et al. 2025; Lin et al. 2025; Liu et al. 2025) directly leverage LLM to build reasoning-driven misinformation detection pipelines. Some methods (Khaliq et al. 2024; Tahmasebi, Müller-Budack, and Ewerth 2024; Braun et al. 2025; Chen et al. 2024b) further retrieve evidence to better handle the evolving nature of news content.

Despite being effective, these methods are unable to handle the situation where there are multiple sources of forged

information in the real world. Several benchmarks (Liu et al. 2024c; Guo et al. 2025; Wang et al. 2024a) are designed to focus on the scenarios where misinformation stems from combined sources. Although some works (Lakara et al. 2024a; Dey et al. 2025) experiment on the multi-source forgery detection benchmark MMfakebench (Liu et al. 2024c), they simply regarded the task as a binary classification of real or fake, ignoring the fine-grained identification for different sources of forged information. To solve the challenge of mixed-source multimodal misinformation detection, MMDAgent (Liu et al. 2024c) integrates the reasoning ability of LLMs and relies on a predefined static detection workflow. MGCA (Guo et al. 2025) extracts multi-view features and trains an end-to-end model to detect misinformation. A recent work LRQ-FACT (Beigi et al. 2024) retrieves evidence by generating questions, and can detect multi-source fabricated information. However, these methods are limited by the fixed workflows and a limited set of tools. Our T²Agent, on the other hand, introduces a modified MCTS that adaptively searches across mixed forgery sources. This enables a principled trade-off between exploiting verified evidence and exploring diverse forgery sources, leading to improved performance in complex and heterogeneous misinformation settings.

Monte Carlo Tree Search

Monte Carlo Tree Search (MCTS) (Coulom 2006) is a powerful heuristic search algorithm that has demonstrated exceptional performance in solving complex decision-making problems. Its core mechanism involves iteratively building a search tree and selecting the optimal action based on the outcomes of numerous simulations. The operation of MCTS typically consists of four key steps. **Selection:** Starting from the root node, a child node is recursively selected according to a policy such as the Upper Confidence bound applied to Trees (UCT) (Kocsis and Szepesvári 2006) until a leaf node or a partially expanded node is reached. **Expansion:** If the selected node does not represent a terminal state, one or more new child nodes are created. **Simulation:** From the newly expanded node, a fast policy is executed to simulate a trajectory until a terminal state is reached. This simulation provides an estimate of the potential value of the node. **Backpropagation:** The outcome of the simulation is propagated back up the selection path to the root node, updating statistics such as visit counts and cumulative rewards for each node along the way.

MCTS and its variants have demonstrated broad applicability across a wide range of domains, which have significantly improved decision-making performance in various areas (Anthony, Tian, and Barber 2017; Ontanón 2013). Monte Carlo tree search simulates the human decision-making process. Some MCTS-based systems like AlphaGo (Silver et al. 2016) and AlphaZero (Silver et al. 2017) have even achieved superhuman performance. MCTS has been widely applied to domains such as robotic path planning (Eiffert et al. 2020) and combinatorial optimization (Cazenave 2009). In recent years, researchers integrate MCTS with large language models to enhance the exploration of solution spaces in complex tasks such as question answering (Yao et al. 2023b; Hao

et al. 2023; Xie et al. 2024; Yu, Chen, and Yu 2023; Zhou et al. 2024), mathematics (Zhang et al. 2024; Chen et al. 2024a), websearch (Yu et al. 2024; Koh et al. 2024), prompt optimization (Wang et al. 2024b), code generation (Antonides et al. 2024), video understanding (Yang et al. 2024), etc. These MCTS applications focus on tasks with a single target and are not well-suited for multi-source misinformation detection. In contrast, we adapt MCTS to the mixed-source characteristics of multimodal misinformation detection by introducing a multi-source verification mechanism augmented with extensible tools.

Method

In this section, we first define the problem of multimodal misinformation detection (Sec.). Then, we introduce our proposed framework, T²Agent, which consists of two core components: a Multi-Source Verification Monte Carlo Tree Search (MCTS) and an extensible toolset for adaptive evidence reasoning. Specifically, we extend traditional MCTS with multi-source verification (Sec.) to deal with potential mixed forgery sources in multimodal misinformation. Our approach introduces a dual evaluation function and a custom backpropagation strategy that together maintain a balance between exploration and exploitation during search. Additionally, we propose a collaborative decision-making strategy to aggregate evidence and reach a final authenticity judgment. Complementing the search process, our framework includes an extensible toolset (Section) that supports plug-and-play integration of specialized tools. This enables dynamic and task-aware tool selection across diverse types of multimodal misinformation.

Problem formulation

We address the task of mixed-source multimodal misinformation detection (Liu et al. 2024c), which aims to assess the authenticity of online news content by analyzing multiple mixed sources of information and classifying it into various types of misinformation. Unlike traditional single-source multimodal misinformation detection tasks that focus on generating a single correct answer, this task involves verifying content across different modalities and potential forgery sources. It may require handling several sub-tasks, such as Textual Veracity Distortion, Visual Veracity Distortion, and Cross-modal Consistency Distortion (Liu et al. 2024c). Only through a comprehensive analysis of these multiple, mixed sources of evidence can the authenticity of a piece of news be reliably determined.

The verification process is a transition process among different states, where each state s_t contains the thought, action a_t , and observation o_t . In our T²Agent, the thought and action are determined by LLM. The observation is obtained by executing the action with the corresponding tool. A misinformation detection agent begins by receiving an input instance c that may contain text and visual content. Then, at each time step t , the agent generates an action $a_t \in \mathcal{A}$ based on the previous state s_{t-1} . By executing a_t , the agent obtains a new observation o_t and transitions to a new state s_t . This agent-environment interaction continues until either the

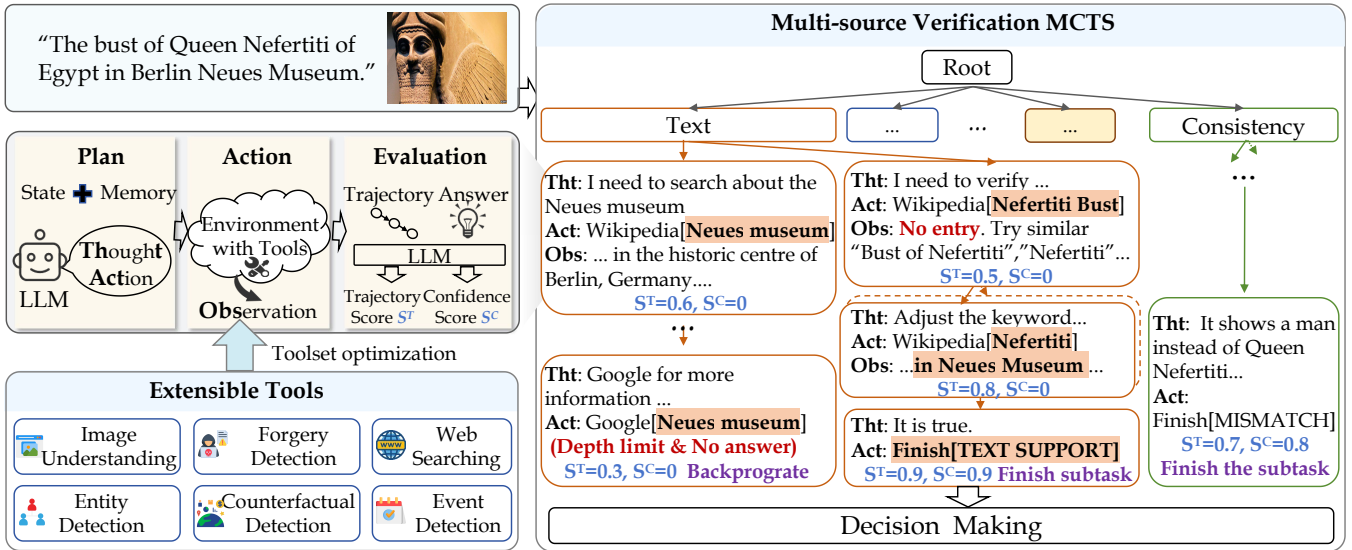


Figure 2: Overview of T^2 Agent. The toolkit acts as the action space, with a greedy search selecting relevant tools. T^2 Agent extends MCTS via multi-source verification, breaking tasks into subtasks targeting different forgery sources. At each node of the tree search process, the agent plans verification paths, selects tools based on task requirements, and evaluates outcomes using a dual reward function that balances exploration across forgery sources with evidence exploitation.

agent makes a final judgment on the authenticity of the input information, or the maximum number of steps T is reached.

Multi-source Verification MCTS

We propose Multi-source Verification MCTS, an extension of classical MCTS tailored for multi-source misinformation detection. During inference, similar to classic MCTS, it iteratively builds a search tree through the selection-expansion-simulation-backpropagation paradigm, predicting rewards at each step to guide the search. Unlike traditional MCTS, we introduce subtask nodes in the search tree, each representing a potential forgery source (e.g., textual inconsistency, visual manipulation, or cross-modal discrepancy). The LVLMM serves as the reasoning controller. At each node, it first makes a plan to generate a thought and corresponding action based on the current state and previous trajectory. We also utilized the trajectories of failures as memory, enabling drawing on past experiences. Once an action is selected, it invokes the corresponding tool from its extensible toolkit to retrieve external evidence or perform specific analyses. To balance exploration across mixed forgery sources with the exploitation of reliable evidence, we introduce a dual reward mechanism, which evaluates both the quality of the reasoning trajectory and the confidence of the final decision. This reward signal is used to update the value of its parent nodes during backpropagation, enabling dynamic adaptation to complex multimodal misinformation scenarios.

Initialization The root of the tree represents the overall task of determining whether a given piece of news is true or false. The first layer of the tree consists of multiple child nodes, each representing a specific sub-task corresponding to mixed forgery sources, which can be adaptively pre-defined by the user based on the requirements of the

task. To facilitate the exploration of the first step, we assign a weight to each child node. We use LVLMM to analyze the input content and estimate the probability distribution over these sub-task categories. This means that the model assigns a likelihood score to each sub-task, reflecting how relevant or necessary it is for verifying the claim. These probabilities guide the expansion of the tree by prioritizing which sub-tasks should be explored further.

Evaluation To evaluate the nodes, we propose a dual scoring mechanism that combines a reasoning trajectory score with a result confidence score. Both scores are calculated by an LLM. We provide corresponding prompts in the appendix. The reasoning trajectory score quantifies the overall quality and coherence of the reasoning process along the path from the root node to the current node. It is utilized in two scenarios. One is to evaluate the quality of the reasoning path during the simulation process and serve as the value of each step node, and the other is to calculate the quality of the reasoning path of the answer node and serve as part of the reward of the node. The input to the reasoning trajectory score function includes all state-action pairs $\{s_i, a_i\}$ up to the current state.

$$S_t^T = LLM(\{s_i, a_i\}_{i=0\dots t}). \quad (1)$$

Upon reaching a leaf node, we evaluate both the quality of the collected evidence and its internal consistency using the proposed confidence score. This score reflects the quality of the collected evidence and the degree to which the evidence supports the conclusion. The confidence score serves as the reward signal for the terminal node in the MCTS framework. The input to the confidence scoring function includes all evidence, i.e., observations $\{o_i\}_{i=0\dots t}$ gathered along the current reasoning path, as well as the news c to be verified.

By analyzing the alignment and coherence among different pieces of evidence, it estimates a confidence value that captures the overall consistency of the verification process:

$$S_t^C = LLM(\{o_i\}_{i=0\dots t}, c). \quad (2)$$

Search Algorithm During the search process, our method automatically selects the sub-tasks that need verification. Specifically, we employ the Upper Confidence Bound for Trees (UCT) criterion (Kocsis and Szepesvári 2006). At each iteration, the node with the highest UCT value is selected for expansion. In this work, we revise the UCT formulation to better balance the trade-off between exploring under-visited sub-task nodes and exploiting those with high confidence scores. Our modification focuses on sub-task nodes that have not been explored, i.e., $N(s_i) = 0$. Traditional methods usually assign an arbitrarily large bonus to unvisited nodes, which could lead to inefficient exploration. Different from them, we introduce a bias term and redefine the UCT function as follows:

$$UCT(s_t) = \frac{V(s_t)}{N(s_t) + 1} + C \sqrt{\frac{\ln(N(s) + 1)}{N(s_t) + 1}} \quad (3)$$

where $V(s_t) = \alpha S_t^T + (1 - \alpha) S_t^C$ is the value of the terminal node. C is a hyperparameter. $N(s_t)$ and $N(s)$ are the visit count for the child node and the parent node, respectively. By adding the bias term 1 to $N(s_t)$, UCT can still be calculated when $N(s_t) = 0$. The term $\sqrt{\frac{\ln(N(s) + 1)}{N(s_t) + 1}}$ allows the UCT values of unexplored nodes to be updated along with the overall situation. In the early stages of the search, in contrast to assigning an arbitrarily large bonus to unvisited nodes, calculating UCT with our function avoids overly prioritized newly expanded sub-task nodes, thereby avoiding unnecessary resource allocation. As the search progresses, the increasing visit count of the parent node gradually raises the UCT value of its unvisited children, thereby encouraging the exploration of unvisited sub-tasks.

Another key component of the search process is back-propagation. Considering the property that a piece of news can only be classified as true if all relevant modalities and sub-tasks independently confirm its authenticity, we introduce a novel pruning strategy. If a sub-task node returns a high-confidence verification result indicating that the forgery source is likely true, we consider this sub-task completed and prune further expansion of its child nodes. By pruning already-confirmed branches, the search process avoids redundant exploration and allocates computational resources toward verifying remaining modalities. This strategy mimics how human experts operate: once a source or modality has been confirmed to be reliable, they tend to focus on validating other uncertain components.

Decision Making After the MCTS search is completed, we perform a final decision-making step based on the verification results collected from all sub-task nodes. If MCTS completes with a sub-task node with a high-confidence score indicating that the content is fake, the system classifies the entire news item as false, without further aggregation. This early-stop mechanism ensures efficient and decisive misinformation detection when strong evidence against authenticity is found in any modality. In cases where no such high-confidence false signal is detected, we proceed to aggregate

the results from all verified sub-task nodes using a heuristic probabilistic fusion strategy. For each sub-task (representing a potential forgery source), we calculate the probability that the model believes the input information belongs to each forgery source i , ($i = visual, text\dots$) by utilizing the confidence score $S^{C,i}$.

$$p(fake^i) = \begin{cases} S^{C,i}, & \text{if } answer(i) = \text{fake}, \\ 1 - S^{C,i}, & \text{if } answer(i) = \text{real}. \end{cases} \quad (4)$$

We assume that different modalities (sub-tasks) provide independent evidence. Therefore, we define the overall likelihood that the news is true as:

$$p(real) = \prod_{i=1}^n \left(1 - p(fake^i)\right)^{(1/n)}, \quad (5)$$

where n is the number of verified sub-tasks. We obtain the final result by comparing the probabilities:

$$answer = arg \max \left(p(real), \left\{ p(fake^i) \right\}_{i=1}^n \right). \quad (6)$$

The proposed multi-source verification MCTS enables the system to explore multiple sub-tasks in a structured manner, assess the quality of evidence collected along different paths, prune verified branches to focus on remaining uncertainties, and integrate each sub-task to obtain the final decision. It emulates the iterative and dynamic reasoning process of a human expert, which dynamically shifts between exploring new sources, validating evidence across modalities, and consolidating findings into a final decision.

Extensible Toolset

Beyond the enhanced MCTS framework, T²Agent incorporates a modular and extensible toolset designed to support diverse misinformation verification tasks across multiple forgery sources. This toolset serves as the action space within the tree search process, enabling the agent to invoke appropriate tools for evidence gathering. Each tool is encapsulated in a tool card, which abstracts its functionality, input-output format, and invocation method into a unified structure. This modular representation allows for seamless integration and easy extension of new tools. The toolset includes web searching tools, time detection tools, forgery detection tools, counterfactual detection tools, image understanding tools, and entity recognition tools. Details are presented in the appendix.

While the full toolset provides comprehensive capabilities, not all tools are equally useful for every task (Lu et al. 2025). To ensure both efficiency and effectiveness, we utilize an adaptive tool selection mechanism tailored to each benchmark. We employ greedy search to efficiently search for the optimal subset of tools. First, we predefine a minimal default toolset, denoted as D_{base} . Then, each candidate tool d_i is evaluated by adding it to the base set. We compute the accuracy improvement as $\Delta_{d_i} = Acc(D_{base} \cup \{d_i\}) - Acc(D_{base})$. If $\Delta_{d_i} > 0$, the tool d_i is considered beneficial for the target task. The updated D_{base} is:

$$D_{base} = D_{base} \cup \{d_i\}.$$

This adaptive configuration ensures that the system maintains high performance while minimizing computational overhead. It also enables T²Agent to generalize across diverse domains.

Experiments

| Approach | Backbone | F1 \uparrow | Accuracy \uparrow |
|-----------|--------------|---------------|---------------------|
| SP | BLIP-2 | 0.167 | 0.328 |
| | LLaVA-1.6 | 0.257 | 0.404 |
| | GPT-4o | 0.492 | 0.609 |
| MMD-agent | GPT-4.1-nano | 0.398 | 0.424 |
| | GPT-4o-mini | 0.478 | 0.485 |
| | GPT-4o | 0.614 | 0.616 |
| LRQ-FACT | GPT-4o | 0.716 | 0.708 |
| Ours | GPT-4.1-nano | 0.568 | 0.569 |
| | GPT-4o-mini | 0.631 | 0.629 |
| | GPT-4o | 0.759 | 0.753 |

Table 1: Comparison results on MMfakebench.

Experiment Setting

Datasets. We evaluate T²Agent on two challenging misinformation detection benchmark: MMFakeBench (Liu et al. 2024c) and AMG (Guo et al. 2025). Both of them contain mixed-source misinformation and categorize news into multiple forgery classes. MMFakeBench (Liu et al. 2024c) contains 11,000 image-text pairs, covering “Real”, “Textual Veracity Distortion (TVD)”, “Visual Veracity Distortion (VVD)”, and “Cross-modal Consistency Distortion (CCD)”, with both human- and machine-generated images. We sample 1,000 validation instances, balanced as 300 Real, 300 TVD, 100 VVD, and 300 CMM. AMG (Guo et al. 2025) classifies news into five forgery categories: “Image Fabrication”, “Non-evidential Image”, “Entity Inconsistency”, “Event Inconsistency”, and “Time Inconsistency”. The dataset contains 4,922 samples in total, divided into 3,532 for training, 517 for validation, and 973 for testing. The test set includes 575 Real samples, 74 Image Fabrication, 69 Non-evidential Image, 29 Entity Inconsistency, 136 Event Inconsistency, and 90 Time Inconsistency.

Evaluation Metrics. To ensure a comprehensive comparison, we adopt the accuracy and the macro-F1 score as evaluation metrics. The macro-F1 score equally weighs performance across all classes by computing the harmonic mean of per-class precision and recall.

Main Results

MMfakebench. On MMfakebench, we compare with Standard Prompt (SP), MMD-Agent (Liu et al. 2024c), and LRQ-FACT (Beigi et al. 2024). For Standard Prompt, we utilize BLIP-2 (Li et al. 2023), LLaVA-1.6 (Liu et al. 2024a), and GPT-4o (OpenAI 2022) as LLM backbones. Since the code of LRQ-FACT is not open-source, the results in the table are those from their paper. As shown in Table 1, our T²Agent consistently outperforms all comparison methods. The results of the standard prompt indicate that the LLM itself has a certain ability to detect false information. This might be due to the fact that the LLM contains knowledge. However, this detection capability is insufficient. Our

T²Agent outperforms the MMD-Agent under all three LLM models (GPT-4.1 Nano, GPT-4o Mini, and GPT-4o), achieving an average 28.7% relative improvement in accuracy. Under a lightweight setting (GPT-4.1 Nano), T²Agent boosts F1 by over 42.7%, demonstrating strong generalization and robustness across model scales. Notably, with GPT-4o Mini, our method achieves an accuracy of 0.629 and F1 of 0.631, which even outperforms MMD-agent with GPT-4o (accuracy 0.616, F1 0.614). T²Agent also outperforms LRQ-FACT, demonstrating the necessity of dynamic verification.

A mismatch forgery case sampled from MM-FakeBench (Fig. 2) showcases the effectiveness of T²Agent in performing multi-source verification through iterative exploration and adaptive pruning. In the first iteration, the agent selects the text subtask node for investigation based on the initial-ization score. It fails to obtain conclusive evidence. We evaluate the node and update the value of itself and its parent nodes via backpropagation to reflect this uncertainty. During the second iteration, the text node is reselected based on the evaluation score, and after tool invocation and simulation, the system confirms that the text is authentic, assigning it a high confidence score. Based on this reliable outcome, the corresponding branch is pruned to avoid redundant verification, demonstrating the method’s efficiency in identifying trustworthy sources. In the third iteration, the agent shifts focus to a consistency forgery source node. Through structured reasoning and evidence aggregation, T²Agent reaches a final judgment with high confidence and terminates the verification process successfully. This example highlights several key advantages of the framework: (1) the ability to iteratively refine node values based on partial evidence, preventing premature decisions; (2) the dynamic expansion of the search tree toward informative forgery sources; and (3) the integration of pruning strategies to improve computational efficiency without sacrificing accuracy.

| Backbone | Approach | F1 \uparrow | Accuracy \uparrow |
|--------------|-----------|---------------|---------------------|
| GPT-4.1-nano | MMD-agent | 0.290 | 0.192 |
| | Ours | 0.402 | 0.503 |
| GPT-4o-mini | MMD-agent | 0.360 | 0.227 |
| | Ours | 0.499 | 0.538 |
| GPT-4o | MMD-agent | 0.365 | 0.306 |
| | Ours | 0.510 | 0.579 |

Table 2: Comparison with MMD-agent on AMG (grouped by backbone for easy comparison).

AMG. We also conduct an evaluation of our method against the MMD-agent on the AMG dataset. As shown in Table 2, our approach consistently outperforms the MMD-agent across all model sizes. In terms of F1 score, our method achieves improvements of 38.6%, 38.6%, and 39.7% over the MMD-agent when using GPT-4.1-nano, GPT-4o-mini, and GPT-4o, respectively. The superior performance of our method can be attributed to its ability to handle the complexity of the AMG dataset, which contains five forgery sources. The MMD-Agent adopts a sequential

decision-making strategy, which tends to suffer from error propagation. This means that incorrect judgments made at earlier stages affect subsequent decisions, leading to a degradation in overall performance. In contrast, our approach explores a more balanced and holistic strategy by incorporating MCTS into the decision process. This allows our model to better assess the likelihood of each forgery type and avoid premature or biased decisions.

The results on MMfakebench and AMG collectively highlight the feasibility of constructing general-purpose, training-free misinformation detection agents using large language models and structured reasoning frameworks, paving the way for more adaptable verification systems.

Ablation Study

We conduct an ablation study on MM-FakeBench using the GPT-4.1-nano backbone over 1,000 samples to evaluate the effectiveness of the multi-source verification MCTS and toolset in our framework. Results are summarized in Table 3.

MCTS search enables flexible evidence verification. MMD-Agent follows a rigid, rule-based verification path with fixed state transitions, limiting its adaptability when handling incomplete or low-confidence evidence across diverse forgery sources. In contrast, introducing MCTS-based tree search allows our agent to dynamically explore and prioritize subtasks based on real-time confidence estimates. This flexibility leads to notable improvements across all metrics: the F1 score increases from 0.398 to 0.535 (+34.4%) and accuracy from 0.424 to 0.534 (+25.9%).

| Method | F1 ↑ | Accuracy ↑ |
|----------------------|--------------|--------------|
| MMD-agent (baseline) | 0.398 | 0.424 |
| TOOLS | 0.413 | 0.459 |
| MV_MCTS | 0.535 | 0.534 |
| MV_MCTS+TOOLS (Ours) | 0.568 | 0.569 |

Table 3: Ablation Study on MMfakebench.

Extensible tools enhance performance through adaptive reasoning. We conduct an ablation study to evaluate the effectiveness of our extensible toolset. The base configuration includes fundamental tools such as image understanding and Wikipedia lookup. Building upon this, we employ our tool selection mechanism to automatically identify and integrate the most beneficial additional tools—specifically, Google Search for up-to-date external knowledge and entity detection via the Baidu API¹ for object categories in complex images. As shown in Table 3, using the selected tools alone yields gains over the baseline (F1: 0.413 vs. 0.398), confirming their utility. More importantly, when integrated into the MCTS framework, they lead to significant performance improvements: our full method achieves an F1 score of 0.568 and accuracy of 0.569, corresponding to relative gains of +6.2% and +6.5% over the MCTS-only variant. This indicates that our tool selection mechanism picks useful tools that integrate well with MCTS.

¹<https://ai.baidu.com/tech/imagerecognition/general>

Cost Analysis

Table 4 compares the inference costs (in USD) between the MMD-Agent and our method on the MMFakeBench dataset, using three GPT-4 variants: GPT-4o, GPT-4o-mini and GPT-4.1-nano. While our method has higher computational costs with the same LLMs, it provides two key advantages: (1) Strong adaptivity: Our T²Agent features task-adaptive reasoning without the need of additional training. This capability is especially important in real-world mixed-source forgery detection. Besides, it allows the system to adapt to novel forgeries without requiring retraining when a new source of forgery emerges. (2) Better cost-effectiveness: As shown in Table 1, when using GPT-4o-mini, our T²Agent achieves an F1 score of 0.631, surpassing MMDAgent (F1 = 0.614) that relies on the larger GPT-4o model. Meanwhile, our method incurs significantly lower resource costs—only 129.4\$ compared to MMDAgent’s 344.4\$. These results clearly indicate that our method offers both higher performance and greater cost-efficiency, showcasing a strong advantage in terms of cost-effectiveness.

| Model | MMD-agent | Ours |
|--------------|-----------|--------|
| GPT-4o | 344.4 | 1637.1 |
| GPT-4o-mini | 14.3 | 129.4 |
| GPT-4.1-nano | 9.5 | 76.2 |

Table 4: Cost comparison (USD) between MMD-agent and our T²Agent on MMfakebench.

Conclusion

In this paper, we introduce T²Agent, a novel misinformation detection agent designed to tackle the complexities and diversities of mixed-source multimodal misinformation. To achieve this goal, we design an extensible toolkit incorporating various modular tools such as web search, forgery detection, and consistency analysis, and propose the integration of Monte Carlo Tree Search (MCTS) with multi-source verification capabilities. The incorporation of these elements achieves a balance between exploration across mixed forgery sources and exploitation for more reliable evidence, enhancing the adaptability and efficiency of the detection process. Extensive results show the efficiency of our approach, suggesting a strong potential for our method as a training-free approach for enhancing detection accuracy in real-world applications.

Limitations and future directions. One notable limitation of our current implementation is the increased computational overhead introduced by the tree search mechanism. Further works may address this limitation by introducing efficient pruning strategies or designing hybrid approaches by combining the proposed method with lightweight expert models to guide the search more effectively. Besides, open-source toolchains introduce new security risks. In practice, these risks may be mitigated by implementing principles such as least privilege and whitelisting tool calls.

Acknowledgement This research is sponsored by National Natural Science Foundation of China (Grant No. 62306041).

References

- Allcott, H.; and Gentzkow, M. 2017. Social media and fake news in the 2016 election. *JEP*.
- Aneja, S.; Bregler, C.; and Nießner, M. 2023. Cosmos: Catching out-of-context misinformation with self-supervised learning. In *AAAI*.
- Anthony, T.; Tian, Z.; and Barber, D. 2017. Thinking fast and slow with deep learning and tree search. In *NeurIPS*.
- Antoniades, A.; Örwall, A.; Zhang, K.; Xie, Y.; Goyal, A.; and Wang, W. 2024. SWE-Search: Enhancing Software Agents with Monte Carlo Tree Search and Iterative Refinement. *arXiv preprint arXiv:2410.20285*.
- Apuke, O. D.; and Omar, B. 2021. Fake news and COVID-19: modelling the predictors of fake news sharing among social media users. *TELEMAT INFORM*.
- Beigi, A.; Jiang, B.; Li, D.; Kumarage, T.; Tan, Z.; Shaeri, P.; and Liu, H. 2025. Lrq-fact: Llm-generated relevant questions for multimodal fact-checking. In *COLING*.
- Beigi, A.; Jiang, B.; Li, D.; Tan, Z.; Shaeri, P.; Kumarage, T.; Bhattacharjee, A.; and Liu, H. 2024. Can LLMs Improve Multimodal Fact-Checking by Asking Relevant Questions? *arXiv preprint arXiv:2410.04616*.
- Braun, T.; Rothermel, M.; Rohrbach, M.; and Rohrbach, A. 2025. DEFAME: Dynamic Evidence-based FAct-checking with Multimodal Experts. In *ICML*.
- Cazenave, T. 2009. Nested Monte-Carlo search. In *IJCAI*.
- Chen, C.; and Shu, K. 2024. Can llm-generated misinformation be detected? In *ICLR*.
- Chen, G.; Liao, M.; Li, C.; and Fan, K. 2024a. AlphaMath Almost Zero: Process Supervision without Process. In *NeurIPS*.
- Chen, J.; Kim, G.; Sriram, A.; Durrett, G.; and Choi, E. 2024b. Complex Claim Verification with Evidence Retrieved in the Wild. In *ACL*.
- Cheung, T.-H.; and Lam, K.-M. 2023. Factllama: Optimizing instruction-following language models with external knowledge for automated fact-checking. In *APSIPA ASC*.
- Coulom, R. 2006. Efficient selectivity and backup operators in Monte-Carlo tree search. In *ICCG*.
- Cui, X.; Li, P.; Li, Z.; Liu, X.; Zou, Y.; and He, Z. 2024a. Localize, understand, collaborate: Semantic-aware dragging via intention reasoner. In *NeurIPS*.
- Cui, X.; Li, Z.; Li, P.; Huang, H.; Liu, X.; and He, Z. 2024b. Instastyle: Inversion noise of a stylized image is secretly a style adviser. In *ECCV*.
- Dey, A. U.; Awan, M. J.; Channing, G.; de Witt, C. S.; and Collomosse, J. 2025. Fact-checking with contextual narratives: Leveraging retrieval-augmented llms for social media analysis. *arXiv preprint arXiv:2504.10166*.
- Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. In *NeurIPS*.
- Eiffert, S.; Kong, H.; Pirmarzdashti, N.; and Sukkarieh, S. 2020. Path planning in dynamic environments using generative rnns and monte carlo tree search. In *ICRA*.
- Guo, H.; Ma, Z.; Zeng, Z.; Luo, M.; Zeng, W.; Tang, J.; and Zhao, X. 2025. Each Fake News is Fake in its Own Way: An Attribution Multi-Granularity Benchmark for Multimodal Fake News Detection. In *AAAI*.
- Hanselowski, A.; Stab, C.; Schulz, C.; Li, Z.; and Gurevych, I. 2019. A richly annotated corpus for different tasks in automated fact-checking. In *CoNLL*.
- Hao, S.; Gu, Y.; Ma, H.; Hong, J.; Wang, Z.; Wang, D.; and Hu, Z. 2023. Reasoning with Language Model is Planning with World Model. In *EMNLP*.
- Hu, B.; Sheng, Q.; Cao, J.; Zhu, Y.; Wang, D.; Wang, Z.; and Jin, Z. 2023. Learn over Past, Evolve for Future: Forecasting Temporal Trends for Fake News Detection. In *ACL*.
- Kakizaki, K.; Matsunaga, Y.; and Furukawa, R. 2025. MAFT: Multimodal Automated Fact-Checking via Textualization. In *AAAI*.
- Khalid, M. A.; Chang, P. Y.-C.; Ma, M.; Pflugfelder, B.; and Miletić, F. 2024. RAGAR, Your Falsehood Radar: RAG-Augmented Reasoning for Political Fact-Checking using Multimodal Large Language Models. In *FEVER*.
- Kocsis, L.; and Szepesvári, C. 2006. Bandit based monte-carlo planning. In *ECML*.
- Koh, J. Y.; McAleer, S.; Fried, D.; and Salakhutdinov, R. 2024. Tree search for language model agents. *arXiv preprint arXiv:2407.01476*.
- Lakara, K.; Channing, G.; Sock, J.; Rupprecht, C.; Torr, P.; Collomosse, J.; and de Witt, C. S. 2024a. LLM-Consensus: Multi-Agent Debate for Visual Misinformation Detection. *arXiv preprint arXiv:2410.20140*.
- Lakara, K.; Sock, J.; Rupprecht, C.; Torr, P.; Collomosse, J.; and de Witt, C. S. 2024b. MAD-Sherlock: Multi-Agent Debates for Out-of-Context Misinformation Detection. *arXiv preprint arXiv:2410.20140*.
- Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*.
- Li, X.; Zhang, Y.; and Malthouse, E. C. 2024. Large language model agent for fake news detection. *arXiv preprint arXiv:2405.01593*.
- Lin, H.; Deng, Y.; Gu, Y.; Zhang, W.; Ma, J.; Ng, S.-K.; and Chua, T.-S. 2025. FACT-AUDIT: An Adaptive Multi-Agent Framework for Dynamic Fact-Checking Evaluation of Large Language Models. *arXiv preprint arXiv:2502.17924*.
- Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024a. Llava-next: Improved reasoning, ocr, and world knowledge. <https://llava-vl.github.io/blog/2024-01-30-llava-next>.
- Liu, X.; Li, P.; Huang, H.; Li, Z.; Cui, X.; Liang, J.; Qin, L.; Deng, W.; and He, Z. 2024b. Fka-owl: Advancing multimodal fake news detection through knowledge-augmented lvlms. In *ACM MM*.
- Liu, X.; Li, Z.; Li, P.; Huang, H.; Xia, S.; Cui, X.; Huang, L.; Deng, W.; and He, Z. 2024c. Mmfakebench: A mixed-source multimodal misinformation detection benchmark for lvlms. In *ICLR*.

- Liu, Y.; Sun, H.; Guo, W.; Xiao, X.; Mao, C.; Yu, Z.; and Yan, R. 2025. Bidev: Bilateral defusing verification for complex claim fact-checking. In *AAAI*.
- Lu, P.; Chen, B.; Liu, S.; Thapa, R.; Boen, J.; and Zou, J. 2025. OctoTools: An Agentic Framework with Extensible Tools for Complex Reasoning. In *ICLRW*.
- Luo, G.; Darrell, T.; and Rohrbach, A. 2021. Newsclippings: Automatic generation of out-of-context multimodal media. In *EMNLP*.
- Nielsen, D. S.; and McConville, R. 2022. Mumin: A large-scale multilingual multimodal fact-checked misinformation social network dataset. In *SIGIR*.
- Ontanon, S. 2013. The combinatorial multi-armed bandit problem and its application to real-time strategy games. In *AAAI*.
- OpenAI. 2022. Chatgpt. <https://openai.com/blog/chatgpt>.
- Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.
- Papadopoulos, S.-I.; Koutlis, C.; Papadopoulos, S.; and Peprantonakis, P. C. 2025. Red-dot: Multimodal fact-checking via relevant evidence detection. *IEEE TCSS*.
- Qi, P.; Yan, Z.; Hsu, W.; and Lee, M. L. 2024. Sniffer: Multimodal large language model for explainable out-of-context misinformation detection. In *CVPR*.
- Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*.
- Shao, R.; Wu, T.; and Liu, Z. 2023. Detecting and grounding multi-modal media manipulation. In *CVPR*.
- Shu, K.; Mahudeswaran, D.; Wang, S.; Lee, D.; and Liu, H. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big Data*.
- Silver, D.; Huang, A.; Maddison, C. J.; Guez, A.; Sifre, L.; Van Den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; et al. 2016. Mastering the game of Go with deep neural networks and tree search. *Nature*.
- Silver, D.; Schrittwieser, J.; Simonyan, K.; Antonoglou, I.; Huang, A.; Guez, A.; Hubert, T.; Baker, L.; Lai, M.; Bolton, A.; et al. 2017. Mastering the game of go without human knowledge. *nature*.
- Suryavardan, S.; Mishra, S.; Patwa, P.; Chakraborty, M.; Rani, A.; Reganti, A. N.; Chadha, A.; Das, A.; Sheth, A. P.; Chinnakotla, M.; et al. 2023. Factly 2: A multimodal fake news and satire news dataset. In *AAAIW*.
- Tahmasebi, S.; Müller-Budack, E.; and Ewerth, R. 2024. Multimodal misinformation detection using large vision-language models. In *CIKM*.
- Thorne, J.; Vlachos, A.; Christodoulopoulos, C.; and Mittal, A. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *NAACL*.
- Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Wang, S.; Lin, H.; Luo, Z.; Ye, Z.; Chen, G.; and Ma, J. 2024a. Mfc-bench: Benchmarking multimodal fact-checking with large vision-language models. *arXiv preprint arXiv:2406.11288*.
- Wang, W. Y. 2017. “Liar, Liar Pants on Fire”: A New Benchmark Dataset for Fake News Detection. In *ACL*.
- Wang, X.; Li, C.; Wang, Z.; Bai, F.; Luo, H.; Zhang, J.; Jovic, N.; Xing, E.; and Hu, Z. 2024b. PromptAgent: Strategic Planning with Language Models Enables Expert-level Prompt Optimization. In *ICLR*.
- Xie, Y.; Goyal, A.; Zheng, W.; Kan, M.-Y.; Lillicrap, T. P.; Kawaguchi, K.; and Shieh, M. 2024. Monte Carlo Tree Search Boosts Reasoning via Iterative Preference Learning. In *NeurIPS*.
- Yang, J.; and Rocha, A. 2024. Take it easy: Label-adaptive self-rationalization for fact verification and explanation generation. In *WIFS*.
- Yang, Z.; Chen, G.; Li, X.; Wang, W.; and Yang, Y. 2024. DoraemonGPT: Toward Understanding Dynamic Scenes with Large Language Models (Exemplified as A Video Agent). In *ICML*.
- Yao, B. M.; Shah, A.; Sun, L.; Cho, J.-H.; and Huang, L. 2023a. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. In *SIGIR*.
- Yao, S.; Yu, D.; Zhao, J.; Shafran, I.; Griffiths, T.; Cao, Y.; and Narasimhan, K. 2023b. Tree of thoughts: Deliberate problem solving with large language models. In *NeurIPS*.
- Yu, X.; Chen, M.; and Yu, Z. 2023. Prompt-Based Monte-Carlo Tree Search for Goal-oriented Dialogue Policy Planning. In *EMNLP*.
- Yu, X.; Peng, B.; Vajipey, V.; Cheng, H.; Galley, M.; Gao, J.; and Yu, Z. 2024. Improving Autonomous AI Agents with Reflective Tree Search and Self-Learning. In *ICLR*.
- Zannettou, S.; Sirivianos, M.; Blackburn, J.; and Kourtellis, N. 2019. The web of false information: Rumors, fake news, hoaxes, clickbait, and various other shenanigans. *JDIQ*.
- Zeng, F.; Li, W.; Gao, W.; and Pang, Y. 2024. Multimodal Misinformation Detection by Learning from Synthetic Data with Multimodal LLMs. In *EMNLP*.
- Zhang, D.; Huang, X.; Zhou, D.; Li, Y.; and Ouyang, W. 2024. Accessing gpt-4 level mathematical olympiad solutions via monte carlo tree self-refine with llama-3 8b. *arXiv preprint arXiv:2406.07394*.
- Zhou, A.; Yan, K.; Shlapentokh-Rothman, M.; Wang, H.; and Wang, Y.-X. 2024. Language agent tree search unifies reasoning, acting, and planning in language models. In *ICML*.