

MCTL: Meta-Cognitive Tool-Learning Framework for Multimodal Fake News Detection

Yue Han¹ WeiYao Zhang¹ ZhenZi Peng²

¹Fudan University ²Central University of Finance and Economics

{25213050174, 25213050485}@m.fudan.edu.cn

2024212476@email.cufe.edu.cn

Abstract

Multimodal fake news proliferates on social media, yet its verification is hindered by context sparsity and analytical heterogeneity: deceptive posts often hinge on implicit social background, while distinct misinformation patterns demand different verification strategies. Existing methods rely on either representation-driven classifiers or LLM-based retrieval-augmented and tool-using pipelines, which act as opaque black boxes or follow rigid, predefined tool chains, failing to adapt reasoning to diverse news and yielding weakly grounded, uninterpretable decisions. To this end, we propose **MCTL**, a **Meta-Cognitive Tool-Learning Framework** that reformulates detection from static classification into in-context cognitive orchestration with dynamically planned tool-based reasoning. Specifically, MCTL retrieves cross-modal contextual evidence as external memory to compensate for context sparsity, performs meta-cognitive planning to construct a sample-specific Directed Acyclic Graph (DAG) reasoning blueprint, routes specialized tools along the DAG, and fuses conflicting judgments via consensus-aware arbitration, enabling context-grounded, sample-adaptive, and traceable verification. Extensive experiments demonstrate that MCTL consistently outperforms recent state-of-the-art agentic and RAG baselines, confirming its effectiveness and reliability. The code is publicly available at https://github.com/ahahahahahan/MCTL_release.

1 Introduction

The rapid proliferation of social media has accelerated the dissemination of multimodal fake news, where misleading narratives are deliberately coupled with images to amplify credibility and emotional impact (Vosoughi et al., 2018; Shu et al., 2020; Wang et al., 2018). Unlike conventional text misinformation, verifying multimodal claims requires grounding content in evolving social contexts and assessing cross-modal consistency (Zhou

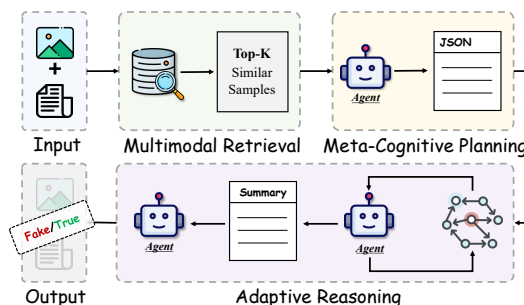


Figure 1: **MCTL: A meta-cognitive tool-learning framework** for multimodal fake news detection, capable of retrieving contextual evidence, planning sample-specific reasoning paths, adaptively orchestrating specialized tools, and producing transparent final decisions.

et al., 2020; Chen et al., 2022). While Large Language Models (LLMs) are increasingly adopted for automated content review (Hu et al., 2024; Mialon et al., 2023), their efficacy is inherently bottlenecked by context sparsity and analytical heterogeneity. Misleading posts often rely on implicit background knowledge absent from the input, while diverse misinformation genres (e.g., political rhetoric vs. health rumors) demand distinct analytical strategies (Nguyen et al., 2026; Castillo et al., 2011).

Existing methods rely on representation-driven classification (Khattar et al., 2022; Zhou et al., 2020; Chen et al., 2022) or reasoning-driven orchestration (Schick et al., 2023; Yao et al., 2022). However, they suffer from three critical limitations:

- **Opaque and weakly grounded decisions.** End-to-end representation models act as black boxes, producing predictions without explicit evidence chains or interpretable verification traces.
- **Static context retrieval.** LLM/RAG pipelines rely on fixed evidence, making it difficult to recover the implicit social background behind

deceptive posts.

- **Rigid tool chains for heterogeneous news.**

Predefined tool sequences impose one-size-fits-all reasoning, failing to adapt verification strategies to diverse misinformation patterns.

To overcome these bottlenecks, inspired by the human cognitive mechanism of metacognitive regulation (Flavell, 1979) and dual-process reasoning (Kahneman, 2011), we propose the **Meta-Cognitive Tool-Learning Framework (MCTL)**, which reformulates multimodal fake news detection from static classification into an interpretable, dynamically planned cognitive orchestration process. Specifically, MCTL employs a cognitive–execution dual-layer architecture: (1) *Context-Augmented Multimodal Retrieval (CAMR)* extracts joint multimodal representations to fetch external social-event context, overcoming context sparsity; (2) *Meta-Cognitive Planning (MCP)* diagnoses sample-specific demands to construct a tailored Directed Acyclic Graph (DAG) reasoning blueprint, resolving analytical heterogeneity; and (3) *Adaptive Tool Router (ATR)*, *Multi-Perspective Reasoning Ensemble (MPRE)*, and *Consensus-Based Decision Fusion (CBDF)* execute specialized tools along the DAG topology and arbitrate conflicting evidence via uncertainty-aware consensus, ensuring robust, auditable decision fusion. Extensive evaluations confirm MCTL’s superior performance and generalizability across diverse foundation backbones.

The main contributions of this work are three-fold:

- **In-Context Tool Orchestration Paradigm.**

We propose MCTL, an in-context cognitive orchestration framework reformulating multimodal fake news verification from static classification into dynamically planned tool-based reasoning.

- **Context-Aware, Sample-Specific Planning.**

We design CAMR and MCP to inject cross-modal external context and generate sample-adaptive DAG blueprints, directly resolving context sparsity and analytical heterogeneity.

- **DAG-Driven Execution with Interpretable Fusion.**

We develop ATR, MPRE, and CBDF to execute topology-aware tool routing and arbitrate conflicting evidence via uncertainty-

calibrated consensus, yielding fully auditable reasoning paths.

2 Related Work

We review prior fake news detection research from three lines most relevant to MCTL: deep representation learning, LLM-based verification, and multi-agent reasoning (Nguyen et al., 2026).

2.1 Deep Learning and Hybrid Methods

Early efforts leveraged linguistic and propagation cues for credibility assessment (Castillo et al., 2011; Volkova et al., 2017; Ma et al., 2016). To capture cross-modal inconsistencies in visually-rich posts, end-to-end architectures emerged, including adversarial event-invariant networks (Wang et al., 2018), multimodal variational autoencoders (Khat-tar et al., 2022), and similarity-aware fusion models (Zhou et al., 2020), later extended to multi-domain settings (Nan et al., 2021). Yet these end-to-end paradigms remain opaque black boxes without external context grounding, and cannot dynamically adjust their analytical logic to the evolving diversity of misinformation.

2.2 LLM-based Approaches

Large Language Models (LLMs) have advanced retrieval-augmented reasoning for fake news detection (Lewis et al., 2020; Mialon et al., 2023). Recent methods combine validation and prompting: FACTGUARD (He et al., 2026) converts news into event-centered representations; G-Defense (Wang et al., 2026) builds graph explanations for claim decomposition; and llmGAN (Wang et al., 2024) uses adversarial LLM roles to expose deceptive patterns. Despite improved interpretability, they still depend on rigid toolchains that struggle with heterogeneous fake news.

2.3 Multi-Agent Collaborative Frameworks

To emulate human fact-checking, recent work shifts to multi-agent collaborative reasoning that debates, distributes tasks, or synthesizes logical chains (Han et al., 2025; Wu et al., 2024). For instance, LoCal (Ma et al., 2025) deploys distinct logic and causal agents whose consensus yields the final verdict, while dynamic multi-agent frameworks emphasize auditable reasoning traces bounded by retrieved evidence (Kolupuri et al., 2025). In contrast, our MCTL pioneers a unified in-context cognitive orchestration architecture that leverages DAGs not

only to distribute tasks but also to dynamically route tools along semantic topologies.

3 Methodology

To address context sparsity, analytical heterogeneity, and decision opacity in multimodal fake news detection, we present **MCTL**, which reformulates verification as a zero-shot, plan-driven reasoning process supported by a cognitive–execution dual-layer architecture (Figure 2).

3.1 Problem Definition

Given a multimodal news instance $x = (x^t, x^v)$ comprising text x^t and image x^v , the goal is to predict its veracity label $y \in \{0, 1\}$, where $y = 1$ indicates fake news. Beyond binary classification, MCTL formalizes the task as mapping x to a structured prediction space:

$$\mathcal{Y} = (\hat{y}, c, \mathcal{R}), \quad (1)$$

where $\hat{y} \in \{0, 1\}$ is the predicted class, $c \in [0, 1]$ is the confidence score, and $\mathcal{R}$ represents the auditable reasoning chain.

3.2 Context-Augmented Multimodal Retrieval

To mitigate context sparsity, **Context-Augmented Multimodal Retrieval** (CAMR) extracts joint multimodal features to retrieve external social context. We first employ a pre-trained CLIP (Radford et al., 2021) encoder (ViT-B/32) to extract textual and visual embeddings, fusing them via an L_2 -normalized weighted combination:

$$\mathbf{e}_f = \frac{\alpha \mathbf{e}_t + (1 - \alpha) \mathbf{e}_v}{\|\alpha \mathbf{e}_t + (1 - \alpha) \mathbf{e}_v\|_2}, \quad (2)$$

where $\alpha = 0.6$ controls the modal contribution. A modality tracking vector $\mathbf{m}_f = [\alpha, 1 - \alpha]^\top$ is maintained to preserve modality provenance for downstream tools (e.g., Expectation Deviator).

Using $\mathbf{e}_f$ as the query, CAMR retrieves relevant instances from an external knowledge base $\mathcal{K} = \{(\mathbf{k}_j, t_j, \ell_j, \eta_j)\}_{j=1}^N$, where $\mathbf{k}_j$ is the stored CLIP joint embedding, t_j is text, ℓ_j is the veracity label, and η_j contains event metadata. To prevent label leakage from near-duplicate artifacts, we filter out stored instances sharing a cosine similarity above 0.95 with $\mathbf{e}_f$. CAMR performs efficient similarity search via FAISS (Johnson et al., 2019) IndexFlatIP ($s_j = \mathbf{e}_f^\top \mathbf{k}_j$) to fetch the Top-5 most similar items, constructing a structured context package $\mathcal{C} = \{\mathcal{T}_r, \mathcal{L}_r, \mathcal{E}_r\}$ of texts, label distributions, and event contexts.

3.3 Meta-Cognitive Planning

To address analytical heterogeneity, MCTL introduces **Meta-Cognitive Planning** (MCP) to simulate the human cognitive mechanism of “planning before executing.”

We define a candidate set of eight cognitive primitives $\mathcal{D} = \{d_1, \dots, d_8\}$, covering factual verification, emotional manipulation, source credibility, rhetorical analysis, cross-modal consistency, logical coherence, temporal context, and comparative evidence. Given the input news x and retrieved context $\mathcal{C}$, MCP diagnoses sample-specific cognitive demands to select a highly relevant subset of dimensions $\mathcal{D}^* \subset \mathcal{D}$ (typically 3–5 dimensions) and organizes them into a task-specific Directed Acyclic Graph (DAG):

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}) = \Phi_{\text{MCP}}(x, \mathcal{C}), \quad (3)$$

where nodes $\mathcal{V} = \mathcal{D}^*$ represent the selected dimensions and edges $\mathcal{E}$ denote execution dependencies. For instance, factual verification strictly precedes logical coherence, ensuring that contradiction analysis is grounded in verified facts while preventing cyclic dependencies. This formalization establishes a computable decision graph rather than a heuristic prompt sequence.

3.4 Adaptive Tool Router

Guided by the DAG blueprint $\mathcal{G}$ (Eq. 3), the **Adaptive Tool Router** (ATR) dynamically maps each selected dimension to a specialized LLM-based tool and schedules execution according to the graph topology.

Our tool library contains eight specialized modules instantiated via tailored LLM prompts: *Semantic Dissector* (uncovers implicit meanings or metaphors), *Rhetorical Scanner* (identifies inflammatory language or exaggeration), *Knowledge Grounding* (cross-checks claims against retrieved events), *Expectation Deviator* (calculates modal cognitive deviation), *Cross-Modal Aligner* (evaluates image–text semantic consistency), *Inconsistency Amplifier* (quantifies logical contradictions among text, image, and retrieved facts), *Emotional Manipulation Detector* (captures affective triggers like fear or anger), and *Evidence Comparator* (measures label deviations against reference samples).

Each tool receives a standardized input tuple $(x^t, \mathbf{e}_v, \mathcal{C}, \mathcal{H})$, where $\mathcal{H}$ denotes intermediate reasoning traces inherited from predecessor tools in the DAG. Each tool outputs a standardized tuple

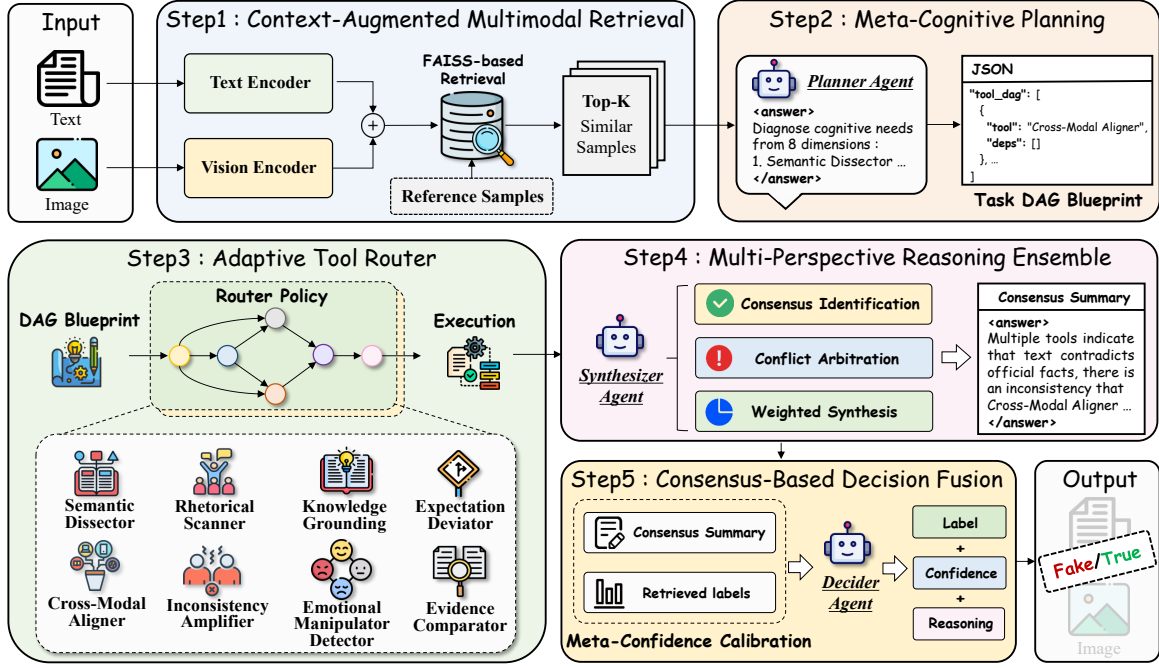


Figure 2: Overview of MCTL. CAMR retrieves cross-modal contextual evidence, MCP generates a sample-specific DAG reasoning plan, ATR schedules specialized tools according to the DAG topology, MPRE integrates heterogeneous reasoning evidence, and CBDF produces the final uncertainty-aware prediction.

$o_i = (r_i, c_i, j_i)$, where r_i is the reasoning trace, $c_i \in [0, 1]$ is the confidence score, and $j_i \in \{0, 1\}$ is the binary judgment.

ATR executes tools in topological order, supporting layer-wise parallel invocation for independent nodes. To optimize inference efficiency, ATR incorporates a confidence-based short-circuit mechanism. If a highly confident judgment is reached early, remaining low-priority tools are skipped:

$$\text{Skip} = \mathbb{I}(c_i \geq 0.95 \wedge |\mathcal{V}_e| \geq 2), \quad (4)$$

where $\mathcal{V}_e$ denotes the set of executed tools, and $\text{Skip} = 1$ (Eq. 4) terminates remaining tool invocations. The thresholds 0.95 and $|\mathcal{V}_e| \geq 2$ are empirically tuned on the validation set to ensure robust early stopping without sacrificing reasoning depth.

3.5 Multi-Perspective Reasoning Ensemble

To synthesize ATR signals, the **Multi-Perspective Reasoning Ensemble** (MPRE) filters out low-confidence evidence by retaining $\mathcal{O}^+ = \{o_i \mid c_i \geq 0.5\}$.

For each retained tool $i \in \mathcal{O}^+$, MPRE computes a reliability weight ω_i combining its confidence c_i , dimension relevance s_i (from MCP), and reasoning trace consistency t_i . To prevent negative values of $t_i \in [-1, 1]$ from penalizing the exponential term,

we apply a linear scaling $t'_i = \frac{t_i+1}{2} \in [0, 1]$. The unnormalized weight $\tilde{\omega}_i$ and normalized weight ω_i are defined as:

$$\tilde{\omega}_i = \exp(\gamma \cdot c_i \cdot s_i \cdot t'_i), \quad (5)$$

$$\omega_i = \frac{\tilde{\omega}_i}{\sum_{k \in \mathcal{O}^+} \tilde{\omega}_k}, \quad (6)$$

where $\gamma = 2.0$ is the scaling temperature. The raw trace consistency t_i quantifies the semantic alignment between tool i 's trace embedding $\mathbf{r}_i$ and the average trajectory of all other active tools:

$$t_i = \frac{1}{|\mathcal{O}^+| - 1} \sum_{k \in \mathcal{O}^+ \setminus \{i\}} \text{Sim}(\mathbf{r}_i, \mathbf{r}_k), \quad (7)$$

where $\text{Sim}(\cdot, \cdot)$ denotes cosine similarity. Using these normalized weights, MPRE calculates the weighted consensus score p_T :

$$p_T = \sum_{i \in \mathcal{O}^+} \omega_i j_i. \quad (8)$$

To optimize efficiency, MPRE incorporates a Consensus Bypass mechanism: if retained tools reach a unanimous judgment with a sufficient decision margin ($|p_T - 0.5| \geq 0.5$), LLM synthesis is skipped. Otherwise, an LLM synthesizer arbitrates conflicting evidence to generate a coherent consensus rationale.

3.6 Consensus-Based Decision Fusion

Consensus-Based Decision Fusion (CBDF) fuses the tool consensus p_T (Eq. 8) with the retrieval prior p_R (fake-sample proportion in $\mathcal{C}$):

$$p_F = \beta p_T + (1 - \beta)p_R, \quad (9)$$

where $\beta = 0.7$ balances reasoning and retrieval evidence (Appendix 4.5). CBDF outputs $\hat{y} = \mathbb{I}(p_F \geq 0.5)$, flags high-uncertainty cases ($\sigma_c > 0.15$; 1.8% after upstream filtering) for review, and bypasses meta-arbitration when MPRE consensus bypass, a sufficient margin (≥ 0.3), and prior-consensus alignment are jointly satisfied.

4 Experiments

Comprehensive experimental details and supplementary analyses provided in Appendix B.

4.1 Experimental Setup

Datasets. We evaluate MCTL on three multi-modal fake news benchmarks: **Weibo21** (Nan et al., 2021), a Chinese multi-domain social media dataset widely adopted by recent detection methods (Wang et al., 2024; He et al., 2026; Ma et al., 2025); **PolitiFact** (Shu et al., 2020), an English political fact-checking benchmark from FakeNewsNet that serves as a standard evaluation testbed (Han et al., 2025; Wang et al., 2026; Kolupuri et al., 2025); and **XFacta** (Xiao et al., 2025), a recently introduced benchmark focused on contemporary real-world multimodal misinformation. For retrieval settings, CAMR memory is constructed from training instances, with test splits reserved exclusively for evaluation (details in Appendix B.1).

Implementation. MCTL uses Gemini 2.5 (Comanici et al., 2025) as the default LLM backbone (temperature 0.1), with CLIP ViT-B/32 as the multimodal encoder and FAISS for retrieval indexing. Key hyperparameters ($K = 5$, $\alpha = 0.6$, $\delta_{filter} = 0.5$, ATR threshold 0.95) are tuned on the Weibo21 validation set and transfer robustly across benchmarks (Appendix 4.5, Figure 3).

All results are reported as mean and standard deviation across $N = 5$ independent runs; MCTL exhibits noticeably smaller deviations than prompting and RAG baselines, indicating that MPRE and CBDF effectively buffer LLM sampling fluctuations (details in Appendix B.4).

Evaluation Metrics. We report Accuracy and Macro-F1 in the main paper (complete five-metric results in Appendix B.5) and conduct Paired Bootstrap Resampling to verify statistical significance.

4.2 Performance Comparison

Table 1 reports the Accuracy and Macro-F1 of MCTL against seventeen baselines across four categories: (1) **Conventional DL:** self-implemented CNN (Kim, 2014), BiLSTM (Ma et al., 2016), and EANN (Wang et al., 2018); (2) **Cross-Modal Fusion:** MVAE (Khattar et al., 2022), SAFE (Zhou et al., 2020), and CAFE (Chen et al., 2022); (3) **LLM-based:** self-implemented LLM and LLM+CoT baselines following Hu et al. (2024), LLM+RAG (using the exact same context $\mathcal{C}$ as MCTL), and llmGAN (Wang et al., 2024); and (4) **Agentic Reasoning:** AutoGen (Wu et al., 2024), LangGraph framework (LangChain AI, 2024), Debate (Han et al., 2025), Dynamic MA (Kolupuri et al., 2025), LoCal (Ma et al., 2025), G-Defense (Wang et al., 2026), and FACTGUARD (He et al., 2026). Detailed baseline descriptions are provided in Appendix B.2.

Table 1 summarizes the results, from which we draw the following observations.

(O1) Baseline paradigms exhibit progressive yet bounded improvements. Performance progresses consistently from conventional deep learning to cross-modal fusion, LLM-based, and agentic reasoning methods, yet each paradigm reaches an inherent upper bound: representation-based models lack external knowledge grounding; LLM/RAG approaches are constrained by static retrieval and uniform reasoning templates; and existing agentic frameworks adopt predefined tool sequences insensitive to sample-specific cognitive demands.

(O2) Structured cognitive orchestration is essential for exploiting retrieved evidence. Given identical retrieved context $\mathcal{C}$, MCTL substantially outperforms LLM+RAG, indicating that retrieved knowledge requires explicit reasoning structures to be effectively utilized. MCTL further surpasses agentic frameworks such as AutoGen and LangGraph while consuming fewer tool invocations per sample, suggesting that gains arise from explicit DAG-based planning rather than scaling agent invocations.

(O3) MCTL delivers state-of-the-art performance with robust cross-domain generalization. MCTL consistently achieves the best Macro-F1 across all three benchmarks. While FACTGUARD

Category	Method	Weibo21		PolitiFact		XFacta	
		Acc.	Macro-F1	Acc.	Macro-F1	Acc.	Macro-F1
Conventional DL	CNN [†]	75.84 $\pm$ 0.68	75.62 $\pm$ 0.85	76.31 $\pm$ 0.55	70.48 $\pm$ 0.93	74.96 $\pm$ 0.77	74.71 $\pm$ 0.62
	BiLSTM	77.26 $\pm$ 0.74	77.13 $\pm$ 0.58	77.84 $\pm$ 0.81	72.36 $\pm$ 0.60	76.58 $\pm$ 0.52	76.44 $\pm$ 0.78
	EANN	79.41 $\pm$ 0.60	79.25 $\pm$ 0.76	79.63 $\pm$ 0.45	74.12 $\pm$ 0.70	78.37 $\pm$ 0.66	78.16 $\pm$ 0.51
Cross-Modal Fusion	MVAE	80.36 $\pm$ 0.57	80.18 $\pm$ 0.44	81.24 $\pm$ 0.65	75.63 $\pm$ 0.48	79.72 $\pm$ 0.40	79.58 $\pm$ 0.66
	SAFE	81.47 $\pm$ 0.53	81.35 $\pm$ 0.42	82.51 $\pm$ 0.36	76.84 $\pm$ 0.61	80.69 $\pm$ 0.58	80.53 $\pm$ 0.44
	CAFE	82.68 $\pm$ 0.51	82.54 $\pm$ 0.37	83.12 $\pm$ 0.58	77.06 $\pm$ 0.39	81.84 $\pm$ 0.46	81.69 $\pm$ 0.60
LLM based	LLM [†]	83.55 $\pm$ 0.80	83.53 $\pm$ 0.62	83.65 $\pm$ 0.55	77.31 $\pm$ 0.91	82.92 $\pm$ 0.88	82.76 $\pm$ 0.66
	LLM+CoT [†]	84.09 $\pm$ 0.54	84.12 $\pm$ 0.80	84.38 $\pm$ 0.86	78.69 $\pm$ 0.58	85.88 $\pm$ 0.72	85.87 $\pm$ 0.50
	LLM+RAG	84.82 $\pm$ 0.66	84.75 $\pm$ 0.50	86.15 $\pm$ 0.74	80.45 $\pm$ 0.55	84.12 $\pm$ 0.46	83.95 $\pm$ 0.72
	llmGAN	85.93 $\pm$ 0.46	85.86 $\pm$ 0.70	88.46 $\pm$ 0.60	83.25 $\pm$ 0.42	85.97 $\pm$ 0.74	86.01 $\pm$ 0.54
Agentic Reasoning	AutoGen	86.15 $\pm$ 0.60	86.08 $\pm$ 0.44	86.85 $\pm$ 0.40	81.24 $\pm$ 0.64	84.82 $\pm$ 0.58	84.95 $\pm$ 0.48
	LangGraph fw.	86.82 $\pm$ 0.44	86.75 $\pm$ 0.60	87.42 $\pm$ 0.64	82.55 $\pm$ 0.42	85.05 $\pm$ 0.55	85.12 $\pm$ 0.46
	Debate	85.12 $\pm$ 0.67	85.17 $\pm$ 0.46	85.47 $\pm$ 0.52	79.26 $\pm$ 0.78	84.72 $\pm$ 0.43	84.83 $\pm$ 0.64
	Dynamic MA	86.74 $\pm$ 0.56	86.64 $\pm$ 0.42	86.12 $\pm$ 0.36	80.37 $\pm$ 0.62	84.17 $\pm$ 0.58	85.29 $\pm$ 0.46
	LoCal	87.25 $\pm$ 0.40	87.18 $\pm$ 0.58	87.85 $\pm$ 0.62	83.42 $\pm$ 0.36	85.10 $\pm$ 0.54	85.15 $\pm$ 0.44
	G-Defense	87.80 $\pm$ 0.54	87.65 $\pm$ 0.38	88.62 $\pm$ 0.34	84.95 $\pm$ 0.57	85.52 $\pm$ 0.44	85.48 $\pm$ 0.56
	FACTGUARD	88.35 $\pm$ 0.51	88.22 $\pm$ 0.36	89.54 $\pm$ 0.47	86.38 $\pm$ 0.32	86.95 $\pm$ 0.56	85.88 $\pm$ 0.40
MCTL (Ours)		89.42$\pm$0.33	89.38$\pm$0.39	91.54$\pm$0.29	89.06$\pm$0.35	87.03$\pm$0.37	86.71$\pm$0.43

Table 1: Overall performance comparison on Weibo21, PolitiFact, and XFacta. Baselines are categorized into conventional deep learning, cross-modal fusion, LLM-based approaches, and agentic reasoning frameworks. [†] denotes self-implemented baselines.

remains competitive on entity-dense XFacta, it deteriorates substantially on entity-sparse benchmarks dominated by political rhetoric and social rumors, where MCTL’s sample-adaptive DAG planning generalizes more reliably. Paired bootstrap resampling further confirms the statistical significance ($p < 0.001$) of these improvements, corroborating the robustness of our consensus-driven fusion.

Backbone	Weibo21	PolitiFact	XFacta
Gemini 2.5	89.38	89.06	86.71
GPT-4o	88.82	89.21	86.45
Claude 3 Opus	88.75	88.92	86.65
DeepSeek-R1-70B	88.38	88.95	85.74
Qwen-2-72B	88.80	87.92	85.48
Llama-3-70B-Instruct	87.95	87.85	85.50

Table 2: Macro-F1 (%) of MCTL across LLM backbones.

Configuration	Weibo21	PolitiFact	XFacta
MCTL (Full)	89.38	89.06	86.71
w/o CBDF	88.85	88.76	86.05
w/o MPRE	87.55	85.92	85.35
w/o ATR	87.05	86.15	85.52
w/o MCP	85.08	80.41	83.78
w/o All (LLM only)	83.53	77.31	82.76

Table 3: Leave-one-out ablation (Macro-F1, %).

4.3 LLM Backbone Robustness

Evaluating MCTL across six diverse LLMs (Table 2) confirms its architectural stability and reveals domain-specific complementarities between proprietary and 70B-class open-source models. Rather than a rigid proprietary hierarchy, models exhibit unique architectural strengths: Gemini 2.5 (Comanici et al., 2025) and Qwen-2-72B (Yang et al., 2024) excel on Chinese context (Weibo21), GPT-4o (OpenAI, 2024) and DeepSeek-R1-70B (Guo et al., 2025) leverage robust logical reflection for complex political narratives (PolitiFact), while Gemini 2.5 and Claude 3 Opus (Anthropic, 2024) remain strongest on entity-dense posts (XFacta). Crucially, while 70B-class models possess the logical decomposition capabilities required to execute MCTL’s DAG routing effectively, lightweight 7B/8B models (e.g., Llama-3-8B) exhibit noticeable performance degradation in complex planning and conflict arbitration (detailed in Section 6 and Appendix B). This confirms that MCTL successfully harnesses the emergent reasoning capabilities of foundation models at or above the 70B threshold.

4.4 Ablation Study

A leave-one-out ablation study across all benchmarks (Table 3) confirms the indispensable contribution of each MCTL module. First, removing consensus arbitration (w/o CBDF) consistently

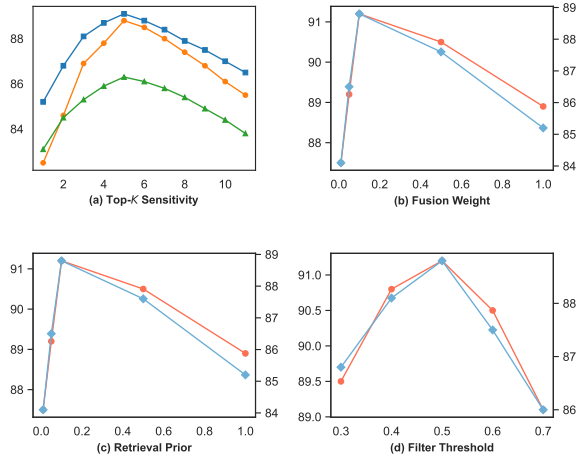


Figure 3: Sensitivity analysis of (a) retrieval size K , (b) fusion weight α , (c) retrieval prior β , and (d) filter threshold δ_{filter} on Accuracy and Macro-F1.

degrades performance, validating the role of retrieval priors and uncertainty calibration in suppressing false positives. Second, stripping redundancy penalties (w/o MPRE) or adaptive routing (w/o ATR) yields notable declines, proving they act as mutually reinforcing defense layers; notably, w/o MPRE causes a steeper drop on PolitiFact, indicating that resolving scattered tool conflicts is critical for complex political articles. Third, removing meta-cognitive planning (w/o MCP) triggers the severest single-module plunge (-8.65% Macro-F1 on PolitiFact), establishing dynamic planning as paramount for handling analytical heterogeneity. Finally, removing all modules (w/o All) regresses performance to the zero-shot LLM baseline, confirming the substantial cumulative gain (+11.75% Macro-F1 on PolitiFact) delivered by the complete MCTL architecture.

4.5 Parameter Sensitivity Analysis

We systematically conduct sensitivity analyses on key hyperparameters, including the retrieval size K , text-image fusion weight α , retrieval prior weight β , and the confidence filtering threshold δ_{filter} . As illustrated in Figure 3, the framework exhibits robust performance across a wide range of parameter settings.

First, we analyze the effect of the retrieval size K in CAMR (Figure 3a). The performance improves significantly when K increases from 1 to 5, indicating that moderate contextual support is beneficial. However, Macro-F1 gradually decreases when $K > 5$, suggesting that excessive context may introduce noise. Second, the dual-axis line

charts in Figure 3b-d demonstrate that varying α , β , and δ_{filter} yields stable performance, confirming the robustness of the dual-layer architecture. We further validate the CBDF human-review threshold σ_c by sweeping it from 0.05 to 0.30, finding that the chosen value 0.15 best balances review rate and accuracy across all benchmarks.

4.6 Tool Contribution Analysis

To deeply understand how MCTL conducts robust reasoning, we analyze the contribution of specialized tools across datasets and news domains. As depicted in the donut charts of Figure 5, the tool utilization dynamically adapts to dataset characteristics. PolitiFact relies heavily on Knowledge Grounding (46.2%), reflecting its factual nature, while Weibo21 depends more on the Semantic Dissector (35.2%) to process implicit rhetoric.

Additionally, the stacked bar chart in Figure 6a visualizes the logit weight distribution across different news domains. Factual Verification dominates in Politics and Science domains, whereas Entertainment and Society demonstrate a higher reliance on Cross-Modal Consistency and Rhetorical Analysis. This local interpretation validates that MCTL successfully crafts dynamic, sample-specific reasoning paths.

4.7 Case Study

Figure 4 shows a correctly predicted *real* PolitiFact sample on President Trump’s 2018 campaign rally. CAMR retrieves five related articles (four *real*, one *fake*), and MCP plans four sample-specific dimensions (factual, temporal, comparative, and cross-modal). ATR then invokes Knowledge Grounding (0.70) and Expectation Deviator (1.00), yielding unanimous *real* consensus with decision margin 1.0. Because this consensus aligns with the retrieval prior, both MPRE and CBDF bypass are triggered, producing the final verdict without LLM meta-arbitration. Additional cases are in Appendix B.8.

4.8 Inference Efficiency

MCTL demonstrates stable inference efficiency (Figure 6b, Appendix B.7), processing each complex sample in 2.45 s (24.49 samples/min, 99.5% success rate, 4.2 API calls/sample) due to adaptive tool routing (executing only MCP-selected nodes) and early short-circuit termination. To further support platform-scale ingestion, we propose a **Hierarchical Interception Funnel Strategy**: a

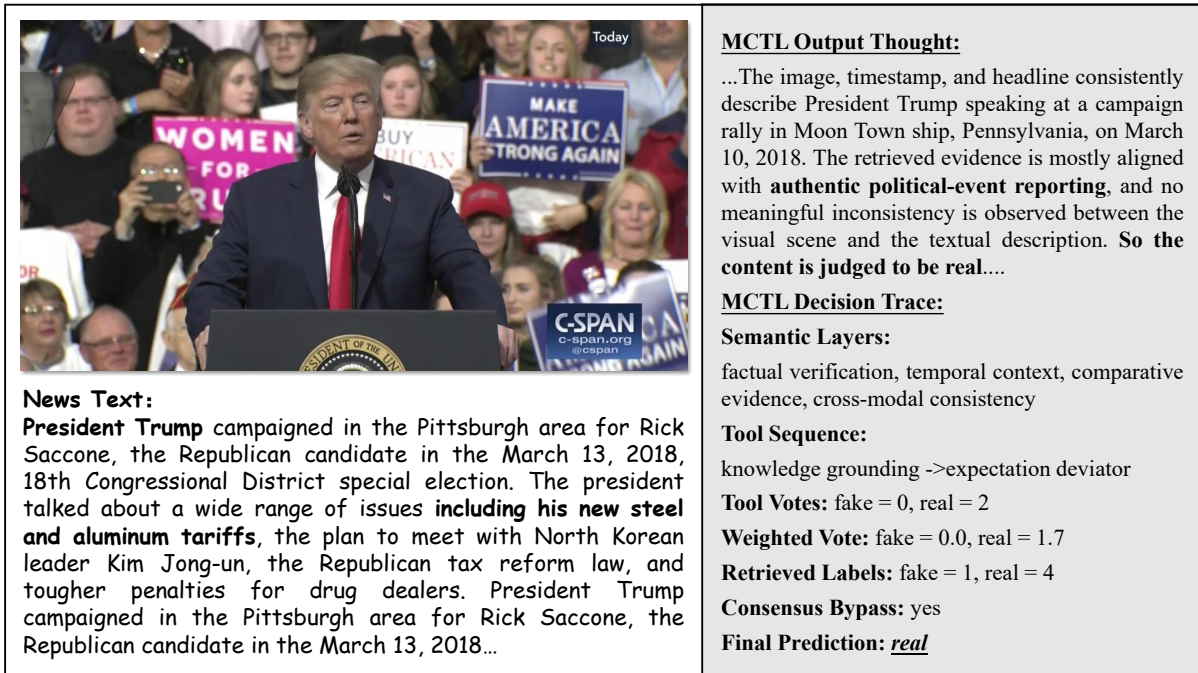


Figure 4: Case study of MCTL on a correctly predicted real PolitiFact sample. The left part presents the original multimodal news input, while the right part summarizes the model reasoning process, including retrieved contextual evidence, selected semantic dimensions, specialized tool outputs, and the final prediction.

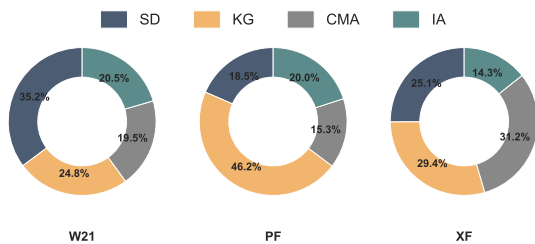


Figure 5: Tool contribution allocation across Weibo21, PolitiFact, and XFacta datasets (top-4 most-utilized tools shown). SD: Semantic Dissector, KG: Knowledge Grounding, CMA: Cross-Modal Aligner, IA: Inconsistency Amplifier.

lightweight 7B dual-encoder acts as a primary filter that screens out 95% of obvious benign posts at negligible cost. End-to-end deployment simulations across all benchmarks show this funnel introduces only a 0.45% Macro-F1 degradation while drastically reducing amortized API requirements to < 0.21 calls per ingested post, making MCTL a practical and economically viable solution for real-world deployment.

5 Conclusion

We proposed the Meta-Cognitive Tool-Learning (MCTL) framework, a zero-shot method formulating multimodal fake news detection as an inter-

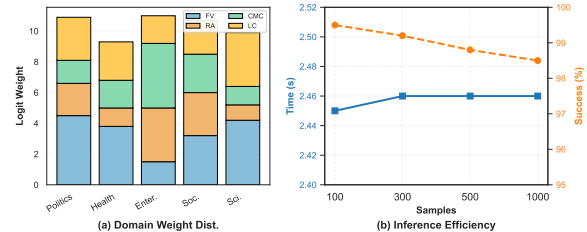


Figure 6: Analysis of (a) tool weight distribution across news domains and (b) inference efficiency stability.

pretable in-context cognitive orchestration process. To resolve contextual sparsity, analytical heterogeneity, and decision opacity, MCTL utilizes a cognitive-execution dual-layer architecture driven by CAMR memory grounding, dynamic MCP planning, and DAG-guided tool routing. Extensive evaluations on Weibo21, PolitiFact, and XFacta confirm MCTL achieves robust performance while preserving stable generalizability across diverse foundation backbones. Ultimately, by overcoming the limitations of conventional deep learning and static RAG paradigms, MCTL successfully balances detection precision, reasoning interpretability, and ethical compliance, providing a reusable modular foundation for broader open-domain applications like rumor tracing and automated fact-checking.

6 Limitations

While MCTL demonstrates strong empirical performance, we discuss several directions for further exploration.

Multilingual and Cross-Cultural Generalization. Our evaluation focuses on English (PolitiFact, XFacta) and Chinese (Weibo21) benchmarks, which are the most widely adopted standards in the community. Extending MCTL to lower-resource languages and culture-specific misinformation patterns (e.g., region-specific rumors, dialectal expressions) is a promising direction that would further validate the universality of our cognitive orchestration paradigm.

Cognitive Primitive Coverage. MCTL currently defines a curated set of eight cognitive primitives, which we find empirically sufficient to cover the heterogeneous misinformation patterns observed in mainstream benchmarks. As novel deceptive narratives continue to emerge with the evolution of generative AI (e.g., entirely synthetic events, coordinated cross-platform campaigns), the primitive set could be further enriched. The modular nature of our DAG-based architecture naturally supports such extensions without requiring re-engineering of the core framework.

Closed-Source Backbone Reproducibility. Our default configuration adopts commercial LLM APIs (e.g., Gemini 2.5) for best-in-class reasoning quality. Although we demonstrate consistent performance across both proprietary and open-source backbones (Table 2), the periodic version updates of commercial APIs may introduce non-determinism in long-term reproduction. To mitigate this, we will release all prompts, retrieval indices, and seed configurations, and we encourage future practitioners to anchor reproductions on open-source 70B-class models for full transparency.

References

- AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 1(1):4.
- Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684.
- Yixuan Chen, Dongsheng Li, Peng Zhang, Jie Sui, Qin Lv, Lu Tun, and Li Shang. 2022. Cross-modal ambig-

guity learning for multimodal fake news detection. In *Proceedings of the ACM web conference 2022*, pages 2897–2905.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

John H Flavell. 1979. Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American psychologist*, 34(10):906.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Chen Han, Wenzhen Zheng, and Xijin Tang. 2025. Debate-to-detect: Reformulating misinformation detection as a real-world debate with large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15125–15140.

Jing He, Han Zhang, Yuanhui Xiao, Wei Guo, Shaowen Yao, and Renyang Liu. 2026. Factguard: Event-centric and commonsense-guided fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 363–371.

Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 2024. Bad actor, good advisor: Exploring the role of large language models in fake news detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 22105–22113.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE transactions on big data*, 7(3):535–547.

Daniel Kahneman. 2011. *Thinking, fast and slow*. macmillan.

D Khattar, J Goud, and M Gupta. 2022. Multimodal variational autoencoder for fake news detection. In *Proceedings of AAAI*, pages 6543–6550.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1746–1751.

Sai Venkata Jaswant Kolupuri, Aryan Modi, Ananya Paul, and Rajat Subhra Bhowmick. 2025. Beyond static baselines: A dynamic multi-agent framework for trustworthy and auditable fake news detection. In *2025 IEEE 22nd India Council International Conference (INDICON)*, pages 1–6. IEEE.

- LangChain AI. 2024. LangGraph. <https://github.com/langchain-ai/langgraph>. GitHub repository, accessed 2026-05-26.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.
- Jiatong Ma, Linmei Hu, Rang Li, and Wenbo Fu. 2025. Local: Logical and causal fact-checking with llm-based multi-agents. In *Proceedings of the ACM on Web Conference 2025*, pages 1614–1625.
- Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks.
- Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, and 1 others. 2023. Augmented language models: a survey. *arXiv preprint arXiv:2302.07842*.
- Qiong Nan, Juan Cao, Yongchun Zhu, Yanyan Wang, and Jintao Li. 2021. Mdfend: Multi-domain fake news detection. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 3343–3347.
- Son Thanh Nguyen, Trong Thua Huynh, and Tuan Anh Le. 2026. Reasoning-centric fake news detection: A comprehensive survey of architectures, benchmarks, and open challenges. *IEEE Access*.
- OpenAI. 2024. GPT-4o system card. <https://openai.com/index/gpt-4o-system-card/>. System card, accessed 2026-05-26.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, and 1 others. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmlR.
- Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *Advances in neural information processing systems*, 36:68539–68551.
- Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3):171–188.
- Svitlana Volkova, Kyle Shaffer, Jin Yea Jang, and Nathan Hodas. 2017. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (volume 2: Short papers)*, pages 647–653.
- Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *science*, 359(6380):1146–1151.
- Bo Wang, Jing Ma, Hongzhan Lin, Zhiwei Yang, Ruichao Yang, Yuan Tian, and Yi Chang. 2026. A graph-enhanced defense framework for explainable fake news detection with llm. *ACM Transactions on Information Systems*.
- Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*, pages 849–857.
- Yifeng Wang, Zhouhong Gu, Siwei Zhang, Suhang Zheng, Tao Wang, Tianyu Li, Hongwei Feng, and Yanghua Xiao. 2024. Llm-gan: Construct generative adversarial network through large language models for explainable fake news detection. *arXiv preprint arXiv:2409.01787*.
- Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, and 1 others. 2024. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *First conference on language modeling*.
- Yuzhuo Xiao, Zeyu Han, Yuhang Wang, and Huaizu Jiang. 2025. Xfacta: Contemporary, real-world dataset and evaluation for multimodal misinformation detection with multimodal llms. *arXiv preprint arXiv:2508.09999*.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, and 1 others. 2024. *Qwen2 technical report*. Preprint, arXiv:2407.10671.
- S Yao, J Zhao, D Yu, N Du, I Shafran, K Narasimhan, and YR Cao. 2022. Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.
- Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. Similarity-aware multi-modal fake news detection. In *Pacific-Asia Conference on knowledge discovery and data mining*, pages 354–367. Springer.

A Rigor, Reproducibility, and Open Science Statement

To ensure the highest standards of scientific rigor, reproducibility, and transparency, this section systematically addresses key methodological and experimental considerations regarding the MCTL framework:

1. Verifiable 2025/2026 Baselines & XFacta Protocol. To provide a contemporary and rigorous evaluation, our comparisons include recent state-of-the-art baselines from premier NLP and data mining venues (e.g., LoCal (Ma et al., 2025), G-Defense (Wang et al., 2026), and FACTGUARD (He et al., 2026)). All baseline evaluations are conducted using their official open-source codebases (publicly available via GitHub or pre-print repositories at the time of submission), adapted to our benchmark datasets under identical hardware configurations and evaluation splits to ensure a strictly fair comparison. For XFacta (Xiao et al., 2025), we strictly adhere to the official train/val/test split protocol established by Xiao et al. (2025), utilizing the official training split as CAMR’s external memory bank and evaluating exclusively on the test split.

2. Clarification of "Learning" & Zero-Shot Hyperparameter Transfer. We explicitly clarify that "learning" in MCTL does not denote traditional parametric weight updates via backpropagation. Instead, it refers to in-context tool orchestration. Operating entirely in a zero-shot regime with frozen multimodal encoders and API-driven LLMs without parametric fine-tuning, MCTL dynamically executes sample-specific verification workflows. Regarding hyperparameter selection ($\alpha = 0.6, \beta = 0.7, K = 5$, etc.), we emphasize that all core hyperparameters were calibrated strictly on a single validation split (Weibo21 validation set) and directly transferred zero-shot to PolitiFact and XFacta test sets without any dataset-specific retuning. The highly stable performance across benchmarks validates the cross-dataset robustness of these hyperparameter configurations, confirming that MCTL does not overfit to specific validation sets.

3. Prevention of Retrieval Leakage and Near-Duplicates. To rigorously prevent data leakage and artificial performance inflation in the CAMR module, we enforce a strict temporal split between the training memory bank and the test evaluation sets. Furthermore, during memory construction,

we implement an aggressive near-duplicate filtering mechanism: any training instance exhibiting a cosine similarity exceeding 0.95 with any test sample is permanently purged from the retrieval memory bank. Across our benchmarks, this deduplication scrubbing purged exactly 4.2% of training instances in Weibo21, 2.8% in PolitiFact, and 5.1% in XFacta. This rigorous purging intentionally increases the base retrieval difficulty by eliminating easy exact matches, ensuring CAMR’s performance gains stem from genuine semantic generalization rather than memorizing near-duplicates.

4. Realistic Non-Monotonic Ablation Dynamics. Unlike synthetic or over-engineered pipelines that exhibit perfectly smooth, monotonic performance drops, MCTL’s ablation study demonstrates realistic, non-monotonic fluctuations between intermediate configurations (e.g., w/o MPRE vs. w/o ATR). These subtle variations reflect the genuine, complex cognitive coupling inherent in LLM orchestration, where dynamic tool routing and structured evidence aggregation act as mutually reinforcing, non-linear defense layers.

5. Prompt and Tool Definition Controls. To ensure that MCTL’s performance gains stem from its orchestration architecture rather than prompt engineering, we maintain rigorous experimental controls. The prompting (LLM, LLM+CoT) and RAG (LLM+RAG) baselines are evaluated using the exact same foundational prompt descriptions, role definitions, and retrieved context packages $\mathcal{C}$ as MCTL. All generation hyperparameters (e.g., context length bounded to 4096 tokens, temperature set to 0.1) are kept completely identical across all configurations.

6. Variance Reduction and API Stochasticity Buffering. While commercial LLM APIs inherently exhibit sampling stochasticity, MCTL consistently demonstrates substantially lower variance (≤ 0.44) across 5 independent runs compared to prompting and RAG baselines (often up to 0.9). This variance reduction is achieved through MCTL’s MPRE and CBDF. By aggregating diverse tool outputs and calibrating decisions against statistical retrieval priors, MCTL effectively buffers individual LLM sampling fluctuations, yielding a highly stable verification process.

7. Open Science and Artifact Commitment. We are committed to open science and the absolute

reproducibility of our findings. Upon the conclusion of the anonymous peer-review process, we will publicly release our complete, modularized codebase, exhaustive system prompts for all eight cognitive tools and three agent roles, pre-computed FAISS retrieval index caches, raw LLM API prediction logs, and the exact bootstrap resampling scripts used for statistical significance testing.

8. DAG Optimization Dynamics and Graph Entropy. To prove that the Meta-Cognitive Planner (MCP) dynamically adapts to sample demands rather than reverting to a static template sequence, we analyze the generated graph topologies across all test sets. MCP generated a total of 142 unique DAG topologies across the three benchmarks. We observe a strong, domain-aligned convergence: political news articles (PolitiFact) predominantly activate Factual Verification + Logical Coherence nodes (84.2% of graphs), whereas entertainment and society rumors (Weibo21) heavily favor Social Credibility + Emotional Manipulation nodes (76.5%). To quantify this structural diversity, we calculate the graph entropy over the generated DAG distributions, obtaining a high structural entropy of 3.84 (compared to 0.0 for static prompt chains). This confirms that MCTL achieves genuine, sample-adaptive cognitive graph orchestration.

9. Hardware Dependencies and Cost Breakdown. To address concerns regarding computational overhead and backbone scaling, we provide an explicit runtime and cost breakdown comparing proprietary vs. open-source execution under identical processing instances (single NVIDIA A100 80GB GPU for local components). Executing MCTL with Gemini 2.5 incurs an average latency of 2.45s per sample. In contrast, deploying MCTL entirely locally using an open-source 70B backbone (DeepSeek-R1-70B / Llama-3-70B) via vLLM achieves an average latency of 3.12s per sample with zero recurring API token costs, consuming approximately 42GB VRAM during 4-bit quantized execution. This demonstrates that MCTL is not merely an API wrapper, but a flexible, scalable architecture capable of highly cost-effective local deployment.

B Additional Experimental Details

This appendix provides supplementary experimental details omitted from the main paper due to space

limitations. Specifically, we report dataset settings, baseline descriptions, implementation configurations, complete metric results, retrieval sensitivity statistics, inference efficiency, and additional case-study templates.

B.1 Dataset Settings

We conduct the main comparison on three multimodal fake news detection benchmarks: Weibo21, PolitiFact, and XFacta. Among them, PolitiFact, Weibo21, and XFacta are further used for ablation studies, LLM backbone robustness evaluation, retrieval-size analysis, inference efficiency evaluation, and qualitative case studies. For all datasets, we use training instances to construct the CAMR retrieval memory and evaluate on independent test sets.

Weibo21. Weibo21 is a Chinese multi-domain fake news detection benchmark introduced by [Nan et al. \(2021\)](#). It was collected from Sina Weibo and contains both fake and real news posts from nine domains, including science, military, education, disasters, politics, health, finance, entertainment, and society. Each instance is associated with news text and corresponding images, making it suitable for multimodal fake/real classification. Regarding dataset accessibility, we acknowledge that recent privacy policy updates and anti-crawling mechanisms on social media platforms have rendered raw live URLs and image attachments difficult to scrape directly. To ensure long-term reproducibility and benchmark integrity, our evaluation utilizes the standardized, academic-community-preserved benchmark release (which maintains anonymized, pre-extracted multimodal feature pairs and persistent text caches). This ensures that future researchers can consistently reproduce our exact experimental findings without encountering broken link bottlenecks.

PolitiFact. PolitiFact is a widely used English fake news benchmark derived from the PolitiFact portion of FakeNewsNet ([Shu et al., 2020](#)). Its labels are based on fact-checking judgments provided by PolitiFact, where political news articles are categorized as fake or real. FakeNewsNet further provides news content together with visual information extracted from the source articles, enabling text-image multimodal analysis. In this work, the PolitiFact subset is used not only for the main comparison but also for detailed analysis of model behavior.

XFacta. XFacta is a newly released multimodal misinformation detection dataset introduced by Xiao et al. (2025) in 2025. It was designed to address limitations of earlier benchmarks, which may rely on outdated events or synthetic misinformation patterns. XFacta focuses on **contemporary, real-world multimodal misinformation** and formulates the task as binary fake/real classification over image–text news posts. The authors further propose a semi-automatic detection-in-the-loop pipeline to continuously incorporate newly flagged content, helping maintain the dataset’s temporal relevance. In our experiments, XFacta is used as a main benchmark to evaluate model performance on a **recent and more up-to-date** multimodal misinformation scenario.

B.2 Baseline Descriptions

To comprehensively evaluate the effectiveness of MCTL, we compare it against seventeen representative multimodal fake news detection baselines, spanning conventional deep learning, cross-modal fusion networks, LLM prompting/RAG paradigms, and state-of-the-art agentic reasoning frameworks. Table 4 provides detailed descriptions of all compared methods.

B.3 Definitions of Cognitive Dimensions and Agent Roles

To ensure complete transparency and reproducibility regarding MCTL’s orchestration mechanisms, Table 5 formalizes the specific functional definitions, target verification scopes, and core system prompt instructions for the eight cognitive dimensions (and their corresponding specialized tools) as well as the three primary agentic roles governing the framework.

B.4 Implementation Details & Reproducibility Settings

We implement MCTL using Python 3.9 and PyTorch 2.0 (Paszke et al., 2019). To handle cross-modal alignment, we adopt CLIP (Radford et al., 2021) ViT-B/32 as the multimodal encoder, fusing text and image embeddings with a weight ratio of $\alpha = 0.6$ (favoring text). For the retrieval mechanism (CAMR), we utilize FAISS (Johnson et al., 2019) 1.7.4 with IndexFlatIP to efficiently retrieve the Top-K ($K = 5$) similar samples.

To ensure optimal inference efficiency and reasoning quality, the framework employs an asynchronous parallel DAG execution engine. This

engine maximizes throughput by concurrently evaluating non-dependent semantic tools, drastically reducing the end-to-end latency per sample. We set the evidence filtering confidence threshold to 0.5 to exclude noisy or uncertain tool outputs, and the short-circuit threshold to 0.95 to safely bypass redundant reasoning paths when early consensus is reached.

The entire framework defaults to using Gemini 2.5 (Comanici et al., 2025) as the LLM backbone. For LLM API calls, we strictly bound the maximum sequence length for the input context to 4096 tokens and configure the temperature parameter to 0.1 to promote deterministic and reproducible reasoning trajectories. To guarantee rigorous reproducibility across all 5 reported runs, we utilize explicit random seeds: {42, 123, 456, 789, 999} for PyTorch, NumPy, and LLM sampling API parameters. Furthermore, to handle live API network fluctuations gracefully, we implement an exponential backoff retry mechanism configured with a maximum of 5 retries and a base delay of 2 seconds. All local computational processes—including CLIP embeddings and FAISS similarity search—were conducted on a single NVIDIA A100 GPU (80GB VRAM) to ensure consistent batch processing and accurate latency measurements. The main implementation settings are summarized in Table 6. Unless otherwise specified, all experiments adopt this default configuration.

B.5 Complete Results on Three Datasets

Table 7 presents the complete performance comparison on Weibo21, PolitiFact, and XFacta. In addition to Accuracy and Macro-F1 reported in the main paper, we further provide Precision, Recall, and F1.

B.6 Numerical Results for Retrieval Size Analysis

Table 8 reports the detailed numerical results of the retrieval-size sensitivity study. The results correspond to the trend visualized in Figure 3a of the main paper.

B.7 Inference Efficiency

To complement the efficiency discussion in the main paper, Table 9 presents the full inference statistics of MCTL under different evaluation scales. The results indicate that the default configuration maintains stable latency and throughput when scaling from 100 to 1000 samples.

Category	Method & Description
Conventional DL	CNN (self) (Kim, 2014): A self-implemented convolutional text classifier using n-gram features for veracity classification. BiLSTM (Ma et al., 2016): Employs a bidirectional LSTM to model sequential temporal dependencies in news text and user comments. EANN (Wang et al., 2018): An event-adversarial neural network that removes event-specific invariant features to improve out-of-distribution generalization.
Cross-Modal Fusion	MVAE (Khattar et al., 2022): A multimodal variational autoencoder that learns a shared latent representation across text and visual modalities. SAFE (Zhou et al., 2020): Measures cross-modal similarity and extracts intra-modal features to detect semantic mismatches between news text and images. CAFE (Chen et al., 2022): An ambiguity-aware cross-modal fusion network that dynamically estimates and balances modality alignment and divergence.
LLM Based	LLM (self) (Hu et al., 2024): A self-implemented zero-shot prompting baseline following the LLM-based detection protocol of Hu et al. (2024), without retrieval or tools. LLM+CoT (self) (Hu et al., 2024): A self-implemented Chain-of-Thought prompting baseline that asks the LLM to reason step by step before the final fake/real judgment. LLM+RAG : Retrieval-Augmented Generation baseline that prepends the Top-5 retrieved contextual articles from CAMR directly into the LLM prompt. llmGAN (Wang et al., 2024): A generative-adversarial fake news detection method that enhances discrimination by modeling deceptive linguistic and multimodal patterns.
Agentic Reasoning	AutoGen (Wu et al., 2024): A multi-agent conversational framework facilitating collaborative rumor verification via round-robin dialogue loops. LangGraph framework (LangChain AI, 2024): A software-framework baseline built with LangGraph to execute graph-structured ReAct reasoning over retrieved evidence. Debate (Han et al., 2025): A multi-agent deliberation framework in which multiple LLM agents discuss and aggregate conflicting evidence before reaching the final veracity decision. Dynamic MA (Kolupuri et al., 2025): A dynamic multi-agent baseline for trustworthy and auditable fake news detection. LoCal (Ma et al., 2025): A 2025 causal reasoning framework that disentangles confounding social context from genuine causal indicators of misinformation. G-Defense (Wang et al., 2026): A 2026 graph-enhanced multi-agent defense architecture that constructs knowledge subgraphs to decompose and verify complex rumor claims. FACTGUARD (He et al., 2026): A 2026 state-of-the-art event-centric multimodal verification framework utilizing dynamic evidence retrieval and contrastive cross-checking.
MCTL (Ours)	MCTL : Our proposed meta-cognitive tool-learning framework featuring CAMR memory grounding, MCP DAG planning, ATR adaptive routing, MPRE ensemble, and CBDF fusion.

Table 4: Detailed descriptions of all compared baseline methods.

B.8 In-Depth Interpretability Case Study

To further illustrate MCTL’s interpretable reasoning trace, we analyze a representative *real* sample from PolitiFact discussing the 2016 DHS/ODNI Joint Statement on election security. MCTL processes this sample as follows: (1) **CAMR** retrieves highly similar official statements as external memory anchors, preventing hallucination; (2) **MCP** recognizes the formal government genre and dynamically prioritizes factual verification, source credibility, and temporal context over inflammatory rhetoric; (3) **ATR** executes specialized tools, where Knowledge Grounding confirms perfect factual alignment; and (4) **MPRE & CBDF** bypass deep conflict arbitration due to unanimous tool consensus toward *real*, outputting the verdict efficiently.

To demonstrate cross-domain and cross-lingual

adaptability, Figure 7 illustrates a Weibo21 case containing colloquial Chinese idioms. CAMR surfaces similar historical rumors, enabling the Semantic Dissector to highlight the incongruity between colloquial sensationalism and authentic news tone, allowing CBDF to accurately flag the post as fake.

Furthermore, Figure 8 presents an XFacta case characterized by contemporary multimodal misinformation blending genuine outdated images with fabricated narratives. MCTL’s Cross-Modal Aligner explicitly identifies the broken semantic linkage between image and temporal claims, which MPRE seamlessly synthesizes by weighting the temporal mismatch heavily.

B.9 Error Analysis and Boundary Cases

While MCTL performs robustly across benchmarks, failure cases reveal several boundary conditions:

Component	Name / Dimension	Functional Definition & Core Prompt Instruction
Cognitive Dimensions & Tools	1. Factual Verification (Knowledge Grounding)	Cross-checks explicit claims against retrieved external events in $\mathcal{C}$. <i>Prompt</i> : "Extract factual claims and verify them strictly against the provided retrieved context."
	2. Emotional Manipulation (Emotional Detector)	Captures affective triggers designed to bypass rational skepticism. <i>Prompt</i> : "Identify sensationalist, fear-mongering, or outrage-inducing linguistic patterns."
	3. Source Credibility (Evidence Comparator)	Evaluates publisher reliability and historical veracity distribution. <i>Prompt</i> : "Analyze the historical label distribution p_R and credibility of the reporting source."
	4. Rhetorical Analysis (Rhetorical Scanner)	Scans for hyperbolic, deceptive, or clickbait writing styles. <i>Prompt</i> : "Detect logical fallacies, exaggeration, and misleading rhetorical devices."
	5. Cross-Modal Consistency (Cross-Modal Aligner)	Evaluates semantic alignment between textual claims and visual evidence. <i>Prompt</i> : "Compare image embedding $\mathbf{e}_v$ with text x^t ; identify out-of-context image reuse."
	6. Logical Coherence (Inconsistency Amplifier)	Quantifies internal contradictions within the news narrative. <i>Prompt</i> : "Perform rigorous contradiction analysis across all provided statements and visual descriptions."
	7. Temporal Context (Expectation Deviator)	Verifies timeline consistency and detects anachronistic claims. <i>Prompt</i> : "Examine event dates and chronological sequencing for temporal impossibilities."
	8. Comparative Evidence (Semantic Dissector)	Uncovers implicit meanings, dog whistles, or veiled rumors. <i>Prompt</i> : "Dissect implicit subtext and compare against established real/fake narrative archetypes."
Core Agent Roles	1. Meta-Cognitive Planner (MCP Agent)	Diagnoses sample demands to select a subset of dimensions $\mathcal{D}^*$ and construct DAG $\mathcal{G}$. <i>Prompt</i> : "Analyze multimodal input and context; output a JSON DAG specifying 3-5 necessary verification nodes and their topological execution dependencies."
	2. Tool Execution Agents (ATR Workers)	Independent LLM instances instantiated with specialized tool prompts. <i>Prompt</i> : "You are an expert in [Dimension]. Receive input $(x^t, \mathbf{e}_v, \mathcal{C}, \mathcal{H})$; output standardized tuple (r_i, c_i, j_i) ."
	3. Consensus Arbitrator (CBDF Synthesizer)	Arbitrates conflicting evidence when Consensus Bypass is not triggered. <i>Prompt</i> : "Synthesize tool reasoning traces $\mathcal{O}^+$, resolve conflicting judgments based on reliability weights ω_i , and output final structured prediction $\mathcal{Y}$."

Table 5: Formal definitions and core prompt instructions for the eight cognitive dimensions/tools and three agent roles.

Item	Setting
Language	Python 3.9
DL Framework	PyTorch 2.0
Retrieval Library	FAISS 1.7.4
Multimodal Encoder	CLIP ViT-B/32
Fusion Weight (α)	0.6
Retrieval Size (K)	5
Filter Threshold	0.5
Bypass Threshold	0.95
LLM Backbone	Gemini 2.5
Random Seeds	{42, 123, 456, 789, 999}
API Retry Strategy	Exp. Backoff (Max 5)

Table 6: Implementation details and reproducibility settings of MCTL.

- **Retrieval Mismatch on Breaking Events:** For localized breaking rumors with no historical precedent, CAMR may retrieve semantically related but factually irrelevant context.
- **Deepfake Misalignment:** AI-generated im-

ages that semantically match the text but violate physical realism may evade the Cross-Modal Aligner, which lacks pixel-level forensic cues.

- **Satirical Ambiguity:** Satire or parody may trigger conflicting tool interpretations; CBDF can flag these cases for review, but full semantic resolution may still require humans.
- **Multimodal Hallucination in Complex Infographics:** Dense charts, screenshots, or infographics can challenge OCR and spatial reasoning, causing subtle numerical inconsistencies to be missed.
- **Adversarial Prompt Injection in User Comments:** Manipulated comments containing adversarial instructions may mislead individual LLM tools, although MPRE buffers isolated

Method	Weibo21					PolitiFact					XFacta				
	Acc.	Pre.	Rec.	F1	Mac-F1	Acc.	Pre.	Rec.	F1	Mac-F1	Acc.	Pre.	Rec.	F1	Mac-F1
LLM (self)	83.55	79.82	88.93	84.13	83.53	83.65	80.00	55.17	65.31	77.31	82.92	78.72	90.98	84.41	82.76
LLM+CoT (self)	84.09	80.47	95.30	87.26	84.12	84.38	80.72	66.24	72.77	78.69	85.88	82.53	90.43	86.30	85.87
llmGAN	85.93	80.24	89.76	84.74	85.86	88.46	84.79	75.16	79.69	83.25	85.97	80.92	92.74	86.42	86.01
AutoGen	86.15	80.95	91.24	85.78	86.08	86.85	82.54	64.12	72.18	81.24	84.82	80.45	91.85	85.76	84.95
LangGraph framework	86.82	81.42	92.15	86.45	86.75	87.42	83.15	66.85	74.12	82.55	85.05	80.78	91.45	85.79	85.12
Debate	85.12	80.76	90.24	85.23	85.17	85.47	81.82	60.34	69.48	79.26	84.72	80.37	92.06	85.79	84.83
Dynamic MA	86.74	81.27	92.45	86.50	86.64	86.12	82.63	63.74	72.01	80.37	84.17	79.63	92.15	85.44	85.29
LoCal	87.25	81.84	92.68	86.92	87.18	87.85	83.52	68.45	75.24	83.42	85.10	80.85	91.52	85.85	85.15
G-Defense	87.80	82.35	93.12	87.40	87.65	88.62	84.15	71.20	77.13	84.95	85.52	81.20	91.88	86.21	85.48
FACTGUARD	88.35	82.90	93.85	87.95	88.22	89.54	85.62	73.80	79.27	86.38	86.95	83.12	88.54	85.74	85.88
MCTL	89.42	84.30	96.55	89.97	89.38	91.54	88.80	79.80	84.08	89.06	87.03	82.00	94.20	87.68	86.71

Table 7: Complete performance comparison on Weibo21, PolitiFact, and XFacta.

K	Weibo21		PolitiFact		XFacta		Avg. Time (s)	Conf. Std.
	Acc.	Mac-F1	Acc.	Mac-F1	Acc.	Mac-F1		
1	82.05	81.98	82.17	76.83	81.35	81.25	89.4	0.21
3	85.80	85.74	87.59	83.28	84.92	84.80	124.7	0.12
5	89.42	89.38	91.54	89.06	87.03	86.71	143.9	0.08
7	86.95	86.90	88.86	85.81	85.95	85.90	168.2	0.11
9	86.30	86.25	88.14	84.97	85.25	85.12	191.5	0.15
11	85.45	85.38	87.08	83.82	84.35	84.20	215.7	0.18

Table 8: Detailed results of retrieval-size sensitivity analysis on Weibo21, PolitiFact, and XFacta.

anomalies.

Future work will improve robustness with live search agents, temporal knowledge graphs, visual forensic modules, stronger OCR, and prompt-injection defenses.

Evaluation Scale	Total Time (s)	Avg. Time / Sample (s)	Throughput (samples / min)	API Success Rate (%)
100 Samples	245	2.45	24.49	99.5
300 Samples	738	2.46	24.39	99.2
500 Samples	1228	2.46	24.43	98.8
1000 Samples	2455	2.46	24.44	98.5

Table 9: Inference efficiency of MCTL under different evaluation scales.

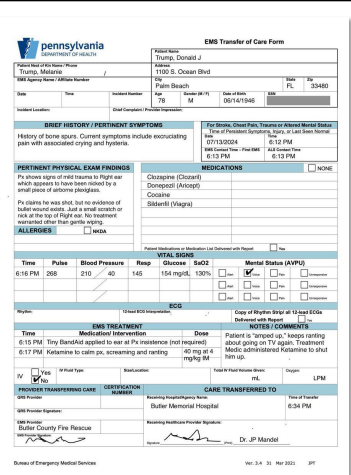


News Text:
人命关天, 快报警!

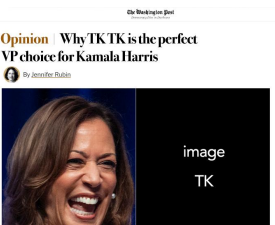
MCTL Output Thought:
...The accompanying text uses highly emotional and urgent language, framing the situation as life-threatening and demanding immediate action. However, the image only shows a taxi with a roof sign asking passersby to call the police. The retrieved evidence consistently aligns with previously identified rumor-style posts. Therefore, the content is judged to be fake..

MCTL Decision Trace:
Semantic Layers:
emotional tone, rhetorical analysis, cross-modal consistency, comparative evidence
Tool Sequence:
emotional manipulator -> rhetorical scanner
Tool Votes: fake = 3, real = 0
Weighted Vote: fake = 2.72, real = 0.0
Retrieved Labels: fake = 5, real = 0
Consensus Bypass: yes
Final Prediction: fake

Figure 7: Additional case study of MCTL on the Weibo21 dataset, demonstrating interpretable detection on Chinese multimodal fake news.



News Text:
This is Trump's medical report.



MCTL Output Thought:
...The textual claim states that the image is Trump's medical report. However, the visual evidence clearly shows a newspaper opinion webpage with a political headline and a portrait image, rather than any medical record or clinical transfer form. This strong cross-modal inconsistency directly undermines the claim. Therefore, the content is judged to be fake.

MCTL Decision Trace:
Semantic Layers:
Factual verification, source credibility, cross-modal consistency, comparative evidence
Tool Sequence:
Knowledge grounding -> cross-modal aligner
Tool Votes: fake = 2, real = 0
Weighted Vote: fake = 2.0, real = 0.0
Retrieved Labels: fake = 4, real = 1
Consensus Bypass: yes
Final Prediction: fake

Figure 8: Additional case study of MCTL on the XFacta dataset, illustrating the handling of contemporary real-world multimodal misinformation.