ZoFia: Zero-Shot Fake News Detection with Entity-Guided Retrieval and Multi-LLM Interaction

Lvhua Wu[1,2]*, Xuefeng Jiang[1,2]*, Sheng Sun[1], Yan Lei[1,2], Tian Wen[1,2], Yuwei Wang[1], Min Liu[1,2]†

[1]Institute of Computing Technology, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
{wulvhua24s,jiangxuefeng21b,liumin}@ict.ac.cn

Abstract

The rapid spread of fake news threatens social stability and public trust, highlighting the urgent need for its effective detection. Although large language models (LLMs) show potential in fake news detection, they are limited by knowledge cutoff and easily generate factual hallucinations when handling time-sensitive news. Furthermore, the thinking of a single LLM easily falls into early stance locking and confirmation bias, making it hard to handle both content reasoning and fact checking simultaneously. To address these challenges, we propose ZoFia, a two-stage zero-shot fake news detection framework. In the first retrieval stage, we propose novel Hierarchical Salience and Salience-Calibrated Minimum Marginal Relevance (SC-MMR) algorithm to extract core entities accurately, which drive dual-source retrieval to overcome knowledge and evidence gaps. In the subsequent stage, a multi-agent system conducts multi-perspective reasoning and verification in parallel and achieves an explainable and robust result via adversarial debate. Comprehensive experiments on two public datasets show that ZoFia outperforms existing zero-shot baselines and even most few-shot methods. Our code has been open-sourced to facilitate the research community at https://github.com/SakiRinn/ZoFia.

1 Introduction

The rapid spread of fake news through social networks has become prevalent, posing severe threats to key domains such as politics (Fisher et al., 2016), economy (Bakir and McStay, 2018), and livelihood (Zhou and Zafarani, 2020). The swift advancement of generative models further exacerbates this concern (Chen and Shu, 2023; Nan et al., 2024; Li et al., 2025, 2024a). In this context, developing effective, efficient and interpretable fake news detection methods is indispensable.

Early studies mainly rely on supervised learning (Monti et al., 2019; Kaliyar et al., 2021) to train detection models, but their dependence on large-scale labeled data makes it hard to adapt to emerging news topics (Hoy and Koulouri, 2022). Large language models (LLMs), with broad pretraining knowledge and strong contextual understanding (Su et al., 2023), significantly advance few-shot and zero-shot methods, providing new opportunities to overcome this bottleneck. In few-shot methods, LLMs either serve as auxiliary tools to perform data augmentation for downstream classifiers (Hu et al., 2024a), or act directly as detectors through prompt learning (Jiang et al., 2022) and instruction tuning (Pavlyshenko, 2023), but they still do not fully escape reliance on training data. By contrast, zero-shot methods that do not require labeled samples are a more promising paradigm, which mainly guide LLM reasoning by context engineering (Zhang and Gao, 2023) and agentic architectures (Li et al., 2024b).

However, zero-shot methods face reliability challenges. On the one hand, the internal knowledge of LLMs is limited by knowledge cutoff (Cheng et al., 2024). Without the external information, they easily produce factual hallucinations (Ji et al., 2023) when handling dynamic news events. On the other hand, LLMs tend to lock their stance early (Echterhoff et al., 2024), where confirmation bias (Wan et al., 2025) drives subsequent reasoning to merely rationalize the initial judgment (Wang et al., 2025). Consequently, when external retrieval is introduced into a single LLM, it easily takes a cognitive shortcut that substitutes whether information can be retrieved for factual veracity, which seriously weakens content reasoning on the original news. This effect becomes much stronger when the retrieved evidence is irrelevant or contradictory (Tan et al., 2024). Therefore, we ob-

*Equal contribution.
†Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 21540–21556
July 2-7, 2026

serve that the inherent reasoning defects of a single LLM prevent it from performing both news content reasoning and external evidence verification well simultaneously.

We argue that content reasoning and fact checking should be decoupled and reliable judgment can be achieved through multi-agent interaction. Based on this motivation, we propose ZoFia, a two-stage zero-shot framework for fake news detection. In the first entity-guided retrieval stage, we introduce Hierarchical Salience to address semantic dilution (Hou et al., 2021) of news entities, which fully leverages global and local semantics to score entity salience, and design a novel Salience-Calibrated Minimum Marginal Relevance (SC-MMR) algorithm to accurately extract core entities. These entities then drive dual-source retrieval from Wikipedia and Open Web to mitigate knowledge cutoff and effectively suppress hallucinations. In the subsequent multi-LLM interaction stage, ZoFia assigns agents to perform content reasoning and fact checking in parallel and breaks stance locking of a single LLM via adversarial debate, ensuring the robustness and interpretability of the final verdict.

To sum up, our main contributions are outlined as follows:

- We propose ZoFia, a retrieval-augmented multi-agent zero-shot framework for fake news detection that effectively overcomes the inherent reasoning flaws of a single LLM.
- We introduce a novel granularity-aware metric Hierarchical Salience and design SC-MMR algorithm to extract news entities accurately for efficient retrieval.
- Comprehensive experiments on two public datasets demonstrate that ZoFia outperforms existing zero-shot baselines and even most few-shot methods.
- We open-source our code to facilitate the research community at `https://github.com/SakiRinn/ZoFia`.

2 Related Work

LLM in Fake News Detection. The application of Large Language Models (LLMs) has become a research frontier in fake news detection, primarily divided into few-shot and zero-shot paradigms. In few-shot methods, researchers often use LLMs as auxiliary tools for data augmentation (Hu et al., 2024a; Nan et al., 2024), or as detectors via post-training (Jiang et al., 2022; Pavlyshenko, 2023). However, these methods fail to completely eliminate dependence on labeled data. Zero-shot methods directly guide model reasoning through context engineering (Zhang and Gao, 2023) and agentic architectures (Li et al., 2024b; Liu et al., 2024), making judgments without labeled samples. Nevertheless, these methods fail to address the confirmation bias arising from a single reasoning chain. We decouple retrieval from reasoning via a multi-role multi-agent system and finally aggregate all evidence for judgment, achieving more comprehensive and robust zero-shot discrimination.

Multi-Agent System. Multi-Agent Systems (MAS) have emerged as an effective paradigm to enhance LLMs for complex tasks. The pioneering work Chateval (Chan et al., 2023) demonstrates that MAS improves both the robustness and accuracy of generation tasks. Subsequent studies introduce this paradigm to reasoning tasks. For instance, COLA (Lan et al., 2024) designs a collaboration framework for stance detection, but it remains limited to analyzing the original text. TruEDebate (Liu et al., 2025) applies structured debate to fake news detection, but its implementation tends to cause premature consensus convergence. Recent agentic systems (Zhang et al., 2026) also integrate reasoning more tightly with external knowledge access. Our ZoFia introduces external information retrieval and independent modular analysis. This design aims to fundamentally mitigate knowledge cutoff and ensure the diversity and independence of arguments and analyses.

3 Stage 1: Entity-Guided Retrieval

This stage aims to acquire reliable external knowledge and instant factual evidence for the subsequent multi-agent system. It consists of four sequential modules. The first three modules precisely extract a set of core entities from the original news. These entities serve as keywords and are concatenated into a query for the final retrieval module, which retrieves from Open Web and Wikipedia.

3.1 Entity Extractor

This module first uses a pre-trained BERT-NER (Tjong Kim Sang and De Meulder, 2003) model to perform named entity recognition (NER) on the news text. We adopt a lightweight BERT-NER model to obtain stable span-level confidence

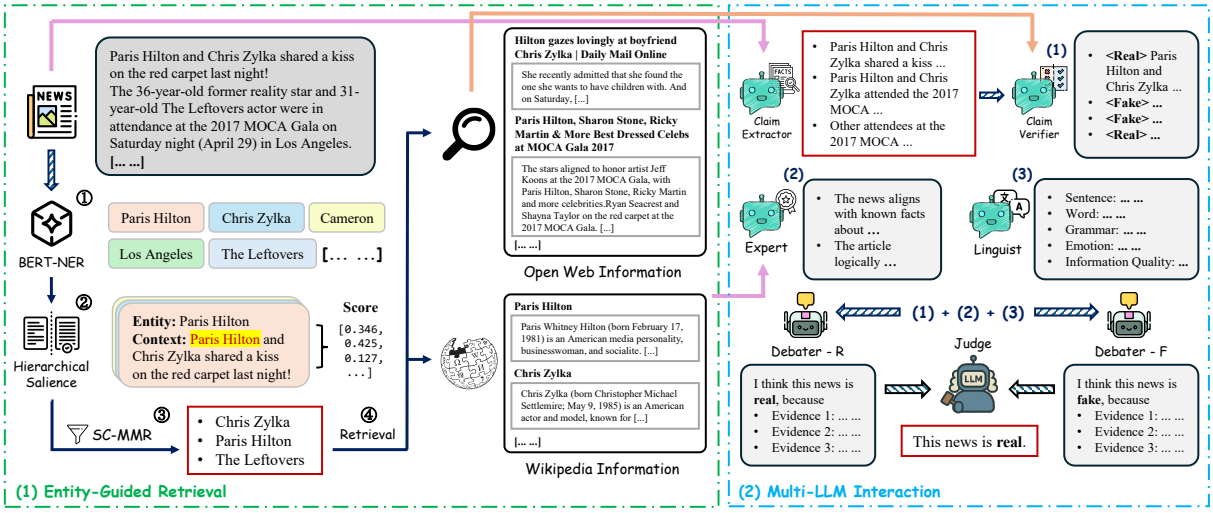

Figure 1: Overall architecture of our proposed ZoFia framework.

scores for dynamic filtering. This process can be expressed as:

$$\{(t_i, e_i, c_i)\}_{i=1}^{N} = \mathcal{M}_{\text{BERT-NER}}(T), \quad (1)$$

where $\mathcal{M}$ is the pre-trained model, T is the input news text. (t_i, e_i, c_i) denotes the recognized entity triplet, t_i is the entity token, e_i is the entity label, and c_i is the confidence score for the corresponding label, expressed as the conditional probability $c_i = P(e_i|t_i, T; \mathcal{M}_{\text{BERT-NER}})$.

Due to the large number of recognized entities, we aggregate consecutive entity tokens to form new entity units U_e. Its confidence score $c(U_e)$ is calculated by averaging the confidence scores of all tokens in U_e, expressed as:

$$c(U_e) = \frac{1}{|U_e|} \sum_{t_i \in U_e} c_i. \quad (2)$$

We establish a dynamic confidence threshold to select entities with high confidence scores. Starting from an initial confidence score λ_{init}, if the number of selected entities fails to meet the predetermined minimum n_{min}, the algorithm iteratively lowers the confidence score by $\Delta\lambda$ and repeats the selection, until at least n_{min} entities are obtained.

3.2 Salience Scorer

This module aims to accurately quantify the importance of news entities, namely Entity Salience (ES) (Dunietz and Gillick, 2014). Although the prior study (Bullough et al., 2024) has demonstrated that bi-encoder architectures can efficiently estimate entity salience, their performance is often limited by semantic dilution (Hou et al., 2021) problem, which leads to severe underestimation of the importance of key entities.

We propose a novel **Hierarchical Salience** that avoids a single, coarse evaluation between an entity and the whole text. It decomposes an entity's overall importance into two orthogonal and multiplicative components: Local Salience $\mathcal{S}_{\text{local}}$ and Global Salience $\mathcal{S}_{\text{global}}$. Local Salience measures the semantic alignment between the entity and its immediate context, and Global Salience measures how much the local context contributes to the text's main content.

Formally, for a news text T that is an ordered sequence of sentences, consider any entity U_i that appears in sentence s_j. We define its local context as $\mathcal{C}(U_i) = s_{j-1} \oplus s_j \oplus s_{j+1}$. Aligned with (Bullough et al., 2024), we use a unified Sentence-BERT (SBERT) (Reimers and Gurevych, 2019) encoder $\mathcal{M}_{\text{SBERT}}(\cdot)$ to embed the entity U_i, its local context $\mathcal{C}(U_i)$, and the full text T to vectors $\mathbf{v}_{U_i}$, $\mathbf{v}_{\mathcal{C}(U_i)}$, and $\mathbf{v}_T$. The hierarchical salience $\mathcal{S}_{\text{hier}}(U_i)$ is derived from the product of these two components as

$$\mathcal{S}_{\text{hier}}(U_i) = \underbrace{\frac{\mathbf{v}_{U_i} \cdot \mathbf{v}_{\mathcal{C}(U_i)}}{\|\mathbf{v}_{U_i}\| \cdot \|\mathbf{v}_{\mathcal{C}(U_i)}\|}}_{\mathcal{S}_{\text{local}}(U_i|\mathcal{C}(U_i))} \cdot \underbrace{\frac{\mathbf{v}_{\mathcal{C}(U_i)} \cdot \mathbf{v}_T}{\|\mathbf{v}_{\mathcal{C}(U_i)}\| \cdot \|\mathbf{v}_T\|}}_{\mathcal{S}_{\text{global}}(\mathcal{C}(U_i)|T)}. \quad (3)$$

Hierarchical Salience provides a finer and more robust estimate of each entity's importance to the news content. It serves as the key criterion for entity filtering in subsequent modules.

3.3 Keyword Selector

This module aims to select an informative subset of keywords from the candidate entity set $\mathcal{U}_{\text{selected}}$

to mitigate the query drift (Carpineto and Romano, 2012) problem. However, this process faces two practical challenges. First, the hierarchical salience score $\mathcal{S}_{\text{hier}}(U_i)$ is highly sensitive to context granularity, so a fixed screening threshold is not effective. Second, coreference in news introduces high-scoring entities with repeated semantics, which harms the diversity of the keyword set.

To optimize relevance and diversity under these constraints, we propose an improved MMR (Carbonell and Goldstein, 1998) algorithm, namely **Salience-Calibrated MMR (SC-MMR)**. At the k-th iteration, SC-MMR evaluates the score of a candidate entity U_i by

$$\text{MMR}(U_i) = \lambda_k \mathcal{S}_{\text{hier}}(U_i) - (1-\lambda_k) \max_{U_j \in \mathcal{U}_{\text{selected}}} \mathcal{S}(U_i, U_j), \tag{4a}$$

$$\mathcal{S}(U_i, U_j) = \frac{\mathbf{v}_{U_i} \cdot \mathbf{v}_{U_j}}{\|\mathbf{v}_{U_i}\| \, \|\mathbf{v}_{U_j}\|}. \tag{4b}$$

The key innovation of SC-MMR is to introduce a weight schedule λ_k that changes with the number of selected keywords k, so that the focus gradually shifts from relevance to diversity. We adopt an annealing schedule with a lower bound:

$$\lambda_k = \max(\lambda_{\min}, \lambda_{\max} - \exp(\alpha \cdot k - \beta)). \tag{5}$$

This form ensures that λ_k decreases monotonically with k while retaining a non-zero salience weight via $\lambda_{\min}$ to prevent the diversity term from fully dominating.

We further introduce a dynamic termination rule based on the relative change of the MMR score. The iteration continues only when the next best candidate $\mathcal{U}^*_{k+1}$ satisfies $\text{MMR}(\mathcal{U}^*_{k+1}) > \gamma \cdot \text{MMR}(\mathcal{U}^*_k)$, where γ is a decay factor. This design prevents the decline in entity quality caused by diminishing marginal utility.

3.4 Information Retrieval

This module uses the keyword set $\mathcal{K} = \{k_1, k_2, \ldots, k_N\}$ distilled in previous modules to build a comprehensive external knowledge base $\mathcal{E}$, serving as supplementary context for the subsequent detection stage. We implement a dual-source retrieval that gathers information from the Open Web and Wikipedia in parallel, ensuring both timeliness and authority.

Retrieval from Open Web. We compose all keywords into an aggregated query for the Open Web Q_{web} with the following logic:

$$Q_{\text{web}} = \left(\bigwedge_{i=1}^{N} k_i \right) \tag{6}$$

The operator $\wedge$ requires that results relate to all keywords, which enables strict cross-keyword verification. We explicitly exclude news and Wikipedia sources. The former avoids retrieving duplicate reports to prevent source contamination (Deng et al., 2023), and the latter prevents redundancy with the subsequent Wikipedia retrieval. For each returned page, we only extract its summary and search snippet as the raw corpus.

Retrieval from Wikipedia. The summary section of a Wikipedia entry usually provides the most precise definition for an entity. However, a single keyword k_i often corresponds to multiple Wikipedia entries, which introduces semantic ambiguity. We build a **context-aware disambiguation** mechanism that uses the original local context $\mathcal{C}(U_i)$ of keyword k_i in the news text to perform accurate matching.

When Wikipedia returns a list with M candidate senses $\mathcal{O}(k_i) = \{o_1, o_2, \ldots, o_M\}$, the mechanism examines each candidate o_m. It constructs a temporary modified context $\mathcal{C}(U_i \leftarrow o_m)$ by replacing the original entity U_i with the description text of o_m. The optimal sense o_i^* of U_i is defined as the option that maximizes the cosine similarity between the vector of the original context and the vector of the modified context:

$$o_i^* = \arg\max_{o_m \in \mathcal{O}(k_i)} \frac{\mathbf{v}_{\mathcal{C}(U_i)} \cdot \mathbf{v}_{\mathcal{C}(U_i \leftarrow o_m)}}{\|\mathbf{v}_{\mathcal{C}(U_i)}\| \|\mathbf{v}_{\mathcal{C}(U_i \leftarrow o_m)}\|}. \tag{7}$$

where $\mathbf{v}$ denotes the context vectors embedded by the pretrained SBERT $\mathcal{M}_{\text{SBERT}}$. After identifying the unique entry, we extract the first 3 sentences of its summary as supplementary material.

4 Stage 2: Multi-LLM Interaction

The external information provided by entity-guided retrieval and the prior knowledge of LLMs builds a **Multi-Source Information Matrix**. This stage employs a dual-state multi-LLM system that fully exploits this matrix to perform parallel and multi-perspective content reasoning and claim verification, which are finally aggregated to reach a robust and interpretable judgment.

This stage operates in two orthogonal interaction states. **LLM Collaboration** performs parallel analyses across multiple agents to reduce inferential variance. **LLM Debate and Judgment** introduces an adversarial debate to reduce systemic bias. These two states form a complete reasoning chain from mining divergent evidence to making a convergent judgment.

4.1 LLM Collaboration

LLM Collaboration state aims to reduce inferential variance. Through parallel analysis by multiple agents, it transforms the **Multi-Source Information Matrix** from the previous stage into a structured evidence pool, which provides a stable decision basis for the subsequent adversarial debate. As illustrated in Figure 2, the matrix systematically integrates 4 information sources that are orthogonal and complementary in evidence distribution:

- *In-news Information:* Comes from the original text of the target news and provides the basic core content.
- *Out-of-news Information:* Comes from Open Web retrieval and provides the most timely and broadest materials.
- *LLM internal Knowledge:* Comes from the model's prior knowledge and provides generalized common sense.
- *LLM external Knowledge:* Comes from Wikipedia retrieval and provides the most precise summary for the core entities.

The subsequent analysis decomposes the tasks and assigns them to agents with distinct roles, so that each agent focuses on a specific quadrant of the matrix for specialized processing.

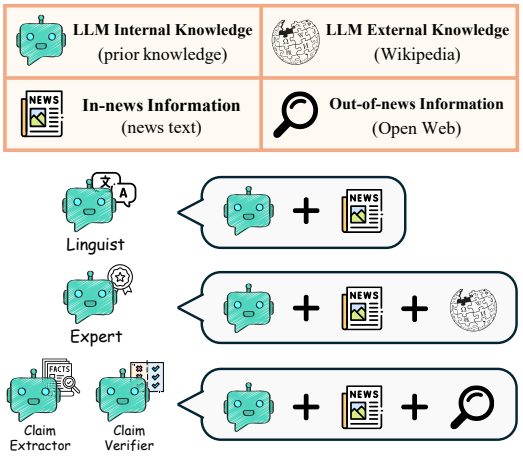

Figure 2: The diagram of Multi-Source Information Matrix and the quadrants used by LLM agents.

4.1.1 Linguist

Following prior studies (Shahid et al., 2022)(Zhou and Zafarani, 2018), the linguist agent is designed to systematically divide the text into 5 linguistic dimensions that are strongly associated with misinformation:

- *Sentence:* Lexical complexity, sentence length, and formality of tone.
- *Word:* Frequency of superlatives, affective language, and pronoun distribution.
- *Grammar:* Patterns of reported speech, passive voice, and negation.
- *Emotion:* Affective terms in the text and the headline, and the degree of incendiary tone.
- *Information Quality:* Presence of clickbait, information overload, or context mismatch.

To maintain objective independence, each dimension is evaluated in an isolated session. For each dimension, the LLM explicitly indicates whether it reflects the news is real or fake. It uses 2 quadrants of the Multi-Source Information Matrix: LLM internal knowledge and in-news information.

4.1.2 Domain-Specific Expert

It operates in a dynamic and adaptive manner. The system first identifies the most relevant domain from the news text and assigns the agent a precise expert role, such as "economist" or "journalist".

The expert with this role then analyzes along the following 2 dimensions:

- *Knowledge Concordance*: Examine all claims, viewpoints, and details for sound reasoning; identify departures from common sense.
- *Logical Integrity*: Examine argument-to-conclusion coherence with domain-specific common sense; identify logical errors or unsupported leaps.

It uses 3 quadrants of the Multi-Source Information Matrix: in-news information, LLM internal knowledge, and LLM external knowledge.

4.1.3 Claim Verification

Existing research (Niu et al., 2024) demonstrates that claim-based fact checking effectively serves as a reference for LLM-based detection. In our system, a serial pipeline composed of **Claim Extractor** and **Claim Verifier** deconstructs and verifies factual claims in the news text. These 2 agents use 3 quadrants of the Multi-Source Information Matrix: LLM internal knowledge, in-news information, and out-of-news information from Open Web.

Claim Extractor. The $\mathcal{M}_{\text{extractor}}$ agent converts unstructured news text T into a set of verifiable structured claims. Its function is formalized as follows:

$$\mathcal{M}_{\text{extractor}}(T) \rightarrow \{q_{\text{core}}, \{q_{\text{sub}_1}, \ldots, q_{\text{sub}_m}\}\} \quad (8)$$

where q_{core} is the core claim that determines the veracity of the news, and q_{sub} is a collection of supporting subclaims. All outputs are restated as concise and objective declarative sentences.

Claim Verifier. This agent verifies each claim q independently. To ensure precision and control costs (Purwar et al., 2024), a simple retrieval-augmented generation (RAG) (Lewis et al., 2020) is implemented to build a highly relevant context $\mathcal{C}_{\text{rel}}(q)$ for each q from the Open Web corpus $\mathcal{I}_{\text{web}}$.

The context consists of text chunks c_j whose cosine similarity with the claim representation $\mathbf{v}_q$ exceeds a threshold θ_{sim}:

$$\mathcal{C}_{\text{rel}}(q) = \{c_j \in \text{top-k}(\mathcal{I}_{\text{web}}, q) \mid \frac{\mathbf{v}_q \cdot \mathbf{v}_{c_j}}{\|\mathbf{v}_q\|\|\mathbf{v}_{c_j}\|} \geq \theta_{\text{sim}}\}. \quad (9)$$

All decisions are strictly based on $\mathcal{C}_{\text{rel}}(q)$. The final output includes a clear label (*"Supports"*, *"Refutes"*, or *"Not Enough Information"*), a brief reasoning, and evidence directly quoted from $\mathcal{C}_{\text{rel}}(q)$ to suppress hallucinations.

4.2 LLM Debate and Judgment

LLM Debate and Judgment state aims to suppress systemic bias. We introduce a multi-round adversarial debate framework that forces LLMs to explore both supporting and refuting views of the truthfulness of news equally. This design directly mitigates the thought degeneration (DoT) (Liang et al., 2023) phenomenon that often appears in a single linear reasoning chain.

Algorithm 1. Dynamic Adversarial Debate

Input: $\mathbb{E}$: The complete evidence pool
Output: D: The final decision $\{Real, Fake\}$
$\mathbb{H} \leftarrow \emptyset$; $A_{\text{con}} \leftarrow$ null; $D \leftarrow$ *Insufficient*
while $D =$ *Insufficient* **do**
  $A_{\text{pro}} \leftarrow \mathcal{M}_{\text{pro}}$.GenerateArgument($\mathbb{E}$, A_{con})
  $A_{\text{con}} \leftarrow \mathcal{M}_{\text{con}}$.GenerateArgument($\mathbb{E}$, A_{pro})
  $\mathbb{H} \leftarrow \mathbb{H} \cup \{(A_{\text{pro}}, A_{\text{con}})\}$
  $D \leftarrow$ Judge.Assess($\mathbb{H}$)
return D

As shown in Algorithm 1, a pair of debate agents $\mathcal{M}_{\text{pro}}$ and $\mathcal{M}_{\text{con}}$ act as opposing reasoners based on the evidence pool $\mathbb{E}$. In each round, the active debater first rebuts the opponent's last argument and then presents a new argument. After each exchange, a judge agent evaluates the debate history $\mathcal{H}$ and outputs a ternary judgment D.

This dynamic termination mechanism ensures that the debate stops once the information is sufficient for decision, effectively balancing the depth and efficiency of reasoning. The debate history $\mathbb{H}$ provides a traceable reasoning chain composed of pro. and con. arguments, making the final judgment highly interpretable.

5 Experiment

5.1 Experimental Setting

Datasets. Following previous state-of-the-art work (Liu et al., 2024), our experiments are conducted on two widely recognized fake news datasets: GossipCop and PolitiFact (Shu et al., 2020). GossipCop focuses on entertainment, mainly Hollywood celebrity news; PolitiFact focuses on politics, drawing on fact-checks of U.S. political figures. Considering that some links in the initial dataset have become invalid, we adopt the available version publicly re-released in (Su et al., 2023).

Metrics. We use accuracy and macro F1-Score as evaluation metrics. F1-Score is less affected by data imbalance, so it serves as the primary metric for assessment.

Baselines. We incorporate two groups of baselines. The first group includes advanced few-shot methods: PSM (Ni et al., 2020), MDFEND (Nan et al., 2021), ARG (Hu et al., 2024b), DKFND (Liu et al., 2024), and KPL (Jiang et al., 2022), with reliable metrics from the existing works (Jin et al., 2024; Hu et al., 2024c; Liu et al., 2024). The second group consists of representative zero-shot methods for comparison: Auto-CoT (Zhang et al., 2022), ReAct (Yao et al., 2023) equipped with a search API, HiSS (Zhang and Gao, 2023), Web Retrieval Agents (Tian et al., 2024), FactAgent (Li et al., 2024b), and CAPE-FND (Jin et al., 2025).

We implement ZoFia based on `DeepSeek-v3` (DeepSeekAI, 2024), `GPT-4o-mini` (OpenAI, 2024) and Qwen3-32B (Team, 2025), and compare it with only-LLM inference. All other LLM-based zero-shot methods are based on `DeepSeek-v3`.

Implementation details. We set the dynamic threshold of entity extraction as $\lambda_{\text{init}} = 0.8$ and $\Delta\lambda = 0.1$ and the decay factor of MMR as $\gamma = 0.5$. The maximum number of entries for Open Web retrieval is 10; for Wikipedia, we retrieve the first 3 sentences for each entry. The minimum similarity threshold for claim extraction is $\theta_{\text{sim}} = 0.1$. The NER model is selected as `dslim/bert-base-NER`

Table 1: Accuracy (Acc.) and F1-Score (F1) comparison of few-shot / zero-shot methods on PolitiFact and GossipCop. The **bold** and underlined denote the best and second-best performance.

Category	Method	LLM Usage	PolitiFact		GossipCop	
			Accuracy	F1-Score	Accuracy	F1-Score
Few-shot	PSM (Nan et al., 2021)	Non-LLM	70.00	49.15	77.44	41.73
	MDFEND (Nan et al., 2021)	Non-LLM	65.50	62.30	41.27	40.20
	KPL (Jiang et al., 2022)	Non-LLM	58.33	60.40	42.71	42.08
	ARG (Hu et al., 2024b)	LLM-assisted	74.00	67.16	61.41	42.32
	DKFND (Liu et al., 2024)	LLM-assisted	87.00	82.43	**82.37**	55.22
Zero-shot	Auto-CoT (Zhang et al., 2022)	LLM-based	89.65	73.67	60.01	48.15
	ReAct (Search API) (Yao et al., 2023)	LLM-based	74.73	67.64	74.03	47.30
	HiSS (Zhang and Gao, 2023)	LLM-based	64.82	56.80	68.81	40.40
	Web Retrieval Agents (Tian et al., 2024)	LLM-based	77.83	64.88	66.62	46.54
	FactAgent (Li et al., 2024b)	LLM-based	80.59	70.06	73.80	56.22
	CAPE-FND (Jin et al., 2025)	LLM-based	76.81	62.43	63.79	49.17
	DeepSeek-v3 (DeepSeekAI, 2024)	Only-LLM	78.99	40.24	61.03	28.17
	GPT-4o-mini (OpenAI, 2024)	Only-LLM	73.58	41.66	66.73	33.18
	Qwen3-32B (Team, 2025)	Only-LLM	76.57	43.35	63.82	32.45
	ZoFia (DeepSeek-v3)	LLM-based	**91.52**	**87.88**	79.04	**61.22**
	ZoFia (GPT-4o-mini)	LLM-based	75.28	75.19	68.90	56.20
	ZoFia (Qwen3-32B)	LLM-based	84.06	81.76	69.71	59.07

(Tjong Kim Sang and De Meulder, 2003), and the SBERT model is selected as `all-MiniLM-L6-v2` (Reimers and Gurevych, 2019). For all base LLMs, we set the temperature to 0.

Brave API[1] is selected as Open Web retrieval API. To strictly prevent label information leakage, we set the search cutoff date to the day before the publication date of each sample URL and exclude dataset-associated fact-checking domains from open-web retrieval (gossipcop.com, politifact.com, and snopes.com).

5.2 Main Experiment

ZoFia demonstrates exceptional performance on both datasets as shown in Table 1, consistently outperforming existing zero-shot and few-shot baselines. On the fact-intensive PolitiFact dataset, ZoFia's advantages are particularly pronounced, with its accuracy (91.52%) and F1-Score (87.88%) not only far exceeding other zero-shot methods but also surpassing all few-shot baselines by a substantial performance margin.

The detection task on the GossipCop dataset presents greater challenges, as its content often proves difficult to verify due to subjectivity and factual ambiguity. In this scenario, ZoFia achieves the highest F1-Score (61.22%), surpassing all baseline models. Although DKFND attains slightly higher accuracy, its F1-Score is notably lower. In contrast, ZoFia's superior F1-Score demonstrates more balanced and robust detection performance across both true and fake news categories while maintaining high precision.

The results clearly show that as a zero-shot method, ZoFia consistently outperforms all zero-shot baselines and even most few-shot methods. Compared to only-LLM inference, ZoFia provides stable and substantial performance gains, demonstrating the superiority of its architecture.

5.3 Ablation Study

To assess the contribution of each component in ZoFia, we conduct ablation studies on GossipCop with DeepSeek-v3 and Qwen3-32B. As shown in Table 2, removing Open Web retrieval causes clear degradation in both settings, which shows that timely external evidence is essential for both base models. Removing all retrieval sources further lowers F1-Score to 58.26% on DeepSeek-v3 and 50.85% on Qwen3-32B. In LLM collaboration, removing either the linguist or the expert weakens the framework, and the expert contributes more; removing both leads to a larger cumulative drop. Removing debate also consistently reduces F1-Score, which shows that it helps mitigate single-perspective bias. The F1-Score drops by 2.06% on DeepSeek-v3 and by 5.00% on Qwen3-32B, and this larger drop suggests that the smaller model is more sensitive to the loss of this mechanism.

[1] https://brave.com/search/api/

Table 2: Ablation study results on GossipCop. (ACC.: Accuracy, F1: F1-Score)

Component	DeepSeek-v3		Qwen3-32B	
	Acc.	F1	Acc.	F1
ZoFia	79.04	61.22	69.71	59.07
w/o Wikipedia retrieval	77.67	59.90	68.09	56.58
w/o Open Web retrieval	71.71	58.14	64.18	52.47
w/o All retrievals	71.37	58.26	62.65	50.85
w/o Linguist analysis	77.86	58.87	67.55	54.62
w/o Expert analysis	76.77	57.32	66.87	54.39
w/o All analyses	73.51	57.90	63.82	52.71
w/o Claim verification	79.23	60.50	67.48	55.17
w/o Debate	75.01	59.16	67.13	54.07

Table 3: Effectiveness of query construction variants in ZoFia on GossipCop.

Method	Accuracy	F1-Score
ZoFia	69.71	59.07
ZoFia (*w/o* All retrievals)	62.65	50.85
ZoFia (TF-IDF)	66.97	53.38
ZoFia (PromptNER)	63.08	55.13
ZoFia (Only-LLM NER)	61.69	52.74
ZoFia (Sentence Salience)	68.64	46.15
ZoFia (Core Claim)	66.65	48.81
ZoFia (OR Composition)	64.89	52.26

5.4 Effectiveness of Entity-Guided Retrieval

To verify the effectiveness of entity-guided retrieval in ZoFia, we fix the second stage and vary only how the query is constructed on GossipCop with Qwen3-32B. TF-IDF (Sparck Jones, 1972), PromptNER (Ashok and Lipton, 2023), and Only-LLM NER still build queries from keywords, but replace the entity acquisition mechanism, and their F1-Scores drop by 5.69%, 3.94%, and 6.33% relative to ZoFia. Sentence Salience (Bullough et al., 2024), Core Claim, and OR Composition instead use sentence-level queries or relax conjunctive matching from AND ($\wedge$) to OR ($\vee$), and their F1-Scores fall by 12.92%, 10.26%, and 6.81%. These results show that the gain does not come from arbitrary keyword lists, and it depends on accurate entity extraction and strict entity-level constraints, which make the original conjunctive entity-guided query the most effective design.

5.5 Sensitivity Analysis of SC-MMR

To motivate the design of the dynamic weight λ_k, we first conduct an experiment with the weight fixed at $\lambda = 0.5$ to observe the direct impact of the keyword count k on performance. As shown in Figure 3, the F1-Score remains stable when $k \leq 6$ but drops sharply beyond this point.

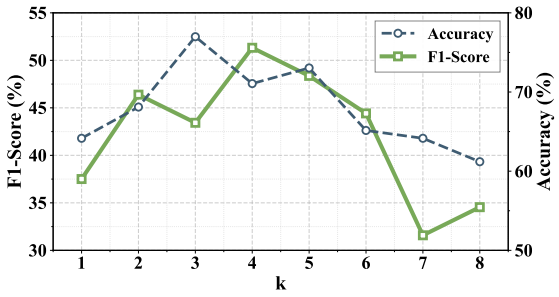

Figure 3: The effect of the number of keywords k on performance (F1-Score).

This phenomenon reveals an inflection point: when the number of keywords exceeds 6, the risk of introducing noise and redundancy begins to outweigh the benefits of new information. It suggests that the strategic focus must shift from relevance to diversity near the critical inflection point of $k \approx 6$. Consequently, we adopt the annealing schedule form defined in Equation 5 and determine its parameters as $\alpha = 0.3$ and $\beta = 2.5$ to fit this downward trend.

5.6 Efficiency of Token Utilization

To investigate ZoFia's reasoning efficiency, a controlled comparison experiment is conducted. We provide CoT (Wei et al., 2022), FactAgent (Li et al., 2024b), and ZoFia with the same materials and impose a unified limit on output tokens to evaluate their performance. Since ZoFia cannot complete full reasoning below 400 tokens, we set its evaluation range to 400 tokens and above.

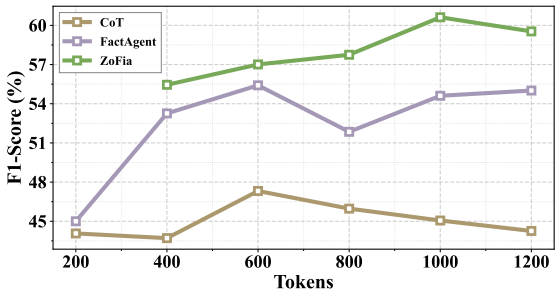

Figure 4: Performance (F1-Score) comparison of methods under maximum output token limits.

The results are shown in Figure 4. At all budget points above 400 tokens, ZoFia achieves a clearly higher F1-Score than other baseline meth-

ods, which demonstrates high reasoning efficiency. In contrast to CoT and FactAgent, whose performance saturates after around 600 tokens, ZoFia shows a stable upward trend even at 1200 tokens. This indicates that ZoFia can consistently convert tokens into performance gains within an acceptable budget.

Note that this experiment only constrains output tokens, because prompt caching (Gim et al., 2024) can extensively amortize input overhead, making output cost more critical for efficiency.

6 Conclusion

In this paper, we propose ZoFia, a retrieval-augmented multi-agent zero-shot framework for fake news detection, to address the cognitive conflict of a single LLM between content reasoning and fact checking. It first utilizes our novel Hierarchical Salience and Salience-Calibrated Minimum Marginal Relevance (SC-MMR) algorithm to accurately extract core entities from news text, which then guide dual-source retrieval from Open Web and Wikipedia. Next, the multi-agent system conducts analysis and verification in parallel and makes a final verdict via adversarial debate. This process effectively reduces confirmation bias from a single reasoning perspective and ensures robust and explainable results. Comprehensive experiments on two public datasets show that ZoFia outperforms existing zero-shot baselines in both performance and efficiency.

Limitations

Though ZoFia shows strong detection capabilities, its application and evaluation face multiple constraints. Due to copyright and privacy concerns, there has been a recent lack of high-quality, continuously updated public datasets. It prevents us from evaluating ZoFia on the most recent news. Building the next generation of benchmark datasets that meet ethical standards and reflect real-world information dynamics is a critical step in this field.

The efficiency of using external retrieval can also be improved. Since it is not our main focus, we integrate only a lightweight retrieval-augmented generation (RAG) module for claim verification. Future work that adopts more advanced RAG architectures, such as a re-ranking model and more advanced mechanisms, may further strengthen the exploitation of external knowledge. We plan to conduct more comprehensive benchmark evaluations of the detection capabilities of LLMs and multi-agent systems for fake news (Guo et al., 2025; Kuntur et al., 2024; Jiang et al., 2024).

ZoFia currently focuses on the text modality. Modern misinformation increasingly appears in multi-modal form that combines images and text. Extending ZoFia to the multimodal domain has strong potential. One direction is to introduce vision language models (VLMs) as a dedicated visual expert. Another is to study how different modalities interact during the debate process to achieve effective fusion.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62472410). We thank the reviewers and chairs for their constructive feedback.

References

Dhananjay Ashok and Zachary C Lipton. 2023. Promptner: Prompting for named entity recognition. *arXiv preprint arXiv:2305.15444*.

Vian Bakir and Andrew McStay. 2018. Fake news and the economy of emotions: Problems, causes, solutions. *Digital journalism*, 6(2):154–175.

Benjamin Bullough, Harrison Lundberg, Chen Hu, and Weihang Xiao. 2024. Predicting entity salience in extremely short documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 50–64.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336.

Claudio Carpineto and Giovanni Romano. 2012. A survey of automatic query expansion in information retrieval. *Acm Computing Surveys (CSUR)*, 44(1):1–50.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.

Canyu Chen and Kai Shu. 2023. Can llm-generated misinformation be detected? *arXiv preprint arXiv:2309.13788*.

Jeffrey Cheng, Marc Marone, Orion Weller, Dawn Lawrie, Daniel Khashabi, and Benjamin Van Durme. 2024. Dated data: Tracing knowledge cutoffs in large language models. *arXiv preprint arXiv:2403.12958*.

DeepSeekAI. 2024. Deepseekv3 technical report. *Preprint*, arXiv:2412.19437.

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. 2023. Investigating data contamination in modern benchmarks for large language models. *arXiv preprint arXiv:2311.09783*.

Jesse Dunietz and Dan Gillick. 2014. A new entity salience task with millions of training examples. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 205–209.

Jessica Maria Echterhoff, Yao Liu, Abeer Alessa, Julian McAuley, and Zexue He. 2024. Cognitive bias in decision-making with llms. In *Findings of the association for computational linguistics: EMNLP 2024*, pages 12640–12653.

Marc Fisher, John Woodrow Cox, and Peter Hermann. 2016. Pizzagate: From rumor, to hashtag, to gunfire in dc. *Washington Post*, 6:8410–8415.

In Gim, Guojun Chen, Seung-seob Lee, Nikhil Sarda, Anurag Khandelwal, and Lin Zhong. 2024. Prompt cache: Modular attention reuse for low-latency inference. *Proceedings of Machine Learning and Systems*, 6:325–338.

Hao Guo, Zihan Ma, Zhi Zeng, Minnan Luo, Weixin Zeng, Jiuyang Tang, and Xiang Zhao. 2025. Each fake news is fake in its own way: An attribution multi-granularity benchmark for multimodal fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 228–236.

Feng Hou, Ruili Wang, Jun He, and Yi Zhou. 2021. Improving entity linking through semantic reinforced entity embeddings. *arXiv preprint arXiv:2106.08495*.

Nathaniel Hoy and Theodora Koulouri. 2022. Exploring the generalisability of fake news detection models. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 5731–5740. IEEE.

Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 2024a. 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.

Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 2024b. 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.

Weiqi Hu, Ye Wang, Yan Jia, Qing Liao, and Bin Zhou. 2024c. A multi-modal prompt learning framework for early detection of fake news. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 651–662.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Gongyao Jiang, Shuang Liu, Yu Zhao, Yueheng Sun, and Meishan Zhang. 2022. Fake news detection via knowledgeable prompt learning. *Information Processing & Management*, 59(5):103029.

Xuefeng Jiang, Lvhua Wu, Sheng Sun, Jia Li, Jingjing Xue, Yuwei Wang, Tingting Wu, and Min Liu. 2024. Investigating large language models for code vulnerability detection: An experimental study. *arXiv preprint arXiv:2412.18260*.

Weiqiang Jin, Yang Gao, Tao Tao, Xiujun Wang, Ningwei Wang, Baohai Wu, and Biao Zhao. 2025. Veracity-oriented context-aware large language modelsbased prompting optimization for fake news detection. *International Journal of Intelligent Systems*, 2025(1):5920142.

Yiqiao Jin, Minje Choi, Gaurav Verma, Jindong Wang, and Srijan Kumar. 2024. Mm-soc: Benchmarking multimodal large language models in social media platforms. *arXiv preprint arXiv:2402.14154*.

Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. 2021. Fakebert: Fake news detection in social media with a bert-based deep learning approach. *Multimedia tools and applications*, 80(8):11765–11788.

Soveatin Kuntur, Anna WrĂłblewska, Marcin Paprzycki, and Maria Ganzha. 2024. Fake news detection: It's all in the data! *arXiv preprint arXiv:2407.02122*.

Xiaochong Lan, Chen Gao, Depeng Jin, and Yong Li. 2024. Stance detection with collaborative role-infused llm-based agents. In *Proceedings of the international AAAI conference on web and social media*, volume 18, pages 891–903.

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.

Jia Li, Lijie Hu, Zhixian He, Jingfeng Zhang, Tianhang Zheng, and Di Wang. 2024a. Text guided image editing with automatic concept locating and forgetting. *arXiv preprint arXiv:2405.19708*.

Jia Li, Lijie Hu, Jingfeng Zhang, Tianhang Zheng, Hua Zhang, and Di Wang. 2025. Fair text-to-image diffusion via fair mapping. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 26256–26264.

Xinyi Li, Yongfeng Zhang, and Edward C Malthouse. 2024b. Large language model agent for fake news detection. *arXiv preprint arXiv:2405.01593*.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*.

Ye Liu, Jiajun Zhu, Kai Zhang, Haoyu Tang, Yanghai Zhang, Xukai Liu, Qi Liu, and Enhong Chen. 2024. Detect, investigate, judge and determine: A novel llm-based framework for few-shot fake news detection. *arXiv preprint arXiv:2407.08952*.

Yuhan Liu, Yuxuan Liu, Xiaoqing Zhang, Xiuying Chen, and Rui Yan. 2025. The truth becomes clearer through debate! multi-agent systems with large language models unmask fake news. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 504–514.

Federico Monti, Fabrizio Frasca, Davide Eynard, Damon Mannion, and Michael M Bronstein. 2019. Fake news detection on social media using geometric deep learning. *arXiv preprint arXiv:1902.06673*.

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.

Qiong Nan, Qiang Sheng, Juan Cao, Beizhe Hu, Danding Wang, and Jintao Li. 2024. Let silence speak: Enhancing fake news detection with generated comments from large language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 1732–1742.

Bo Ni, Zhichun Guo, Jianing Li, and Meng Jiang. 2020. Improving generalizability of fake news detection methods using propensity score matching. *arXiv preprint arXiv:2002.00838*.

Cheng Niu, Yang Guan, Yuanhao Wu, Juno Zhu, Juntong Song, Randy Zhong, Kaihua Zhu, Siliang Xu, Shizhe Diao, and Tong Zhang. 2024. Veract scan: Retrieval-augmented fake news detection with justifiable reasoning. *arXiv preprint arXiv:2406.10289*.

OpenAI. 2024. Gpt-4o-mini. `https://platform.openai.com/docs/models`. Accessed: 2024-10.

Bohdan M Pavlyshenko. 2023. Analysis of disinformation and fake news detection using fine-tuned large language model. *arXiv preprint arXiv:2309.04704*.

Anupam Purwar and 1 others. 2024. Evaluating the efficacy of open-source llms in enterprise-specific rag systems: A comparative study of performance and scalability. *arXiv preprint arXiv:2406.11424*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Wajiha Shahid, Bahman Jamshidi, Saqib Hakak, Haruna Isah, Wazir Zada Khan, Muhammad Khurram Khan, and Kim-Kwang Raymond Choo. 2022. Detecting and mitigating the dissemination of fake news: Challenges and future research opportunities. *IEEE Transactions on Computational Social Systems*, 11(4):4649–4662.

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.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

Jinyan Su, Claire Cardie, and Preslav Nakov. 2023. Adapting fake news detection to the era of large language models. *arXiv preprint arXiv:2311.04917*.

Hexiang Tan, Fei Sun, Wanli Yang, Yuanzhuo Wang, Qi Cao, and Xueqi Cheng. 2024. Blinded by generated contexts: How language models merge generated and retrieved contexts when knowledge conflicts? *arXiv preprint arXiv:2401.11911*.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Jacob-Junqi Tian, Hao Yu, Yury Orlovskiy, Tyler Vergho, Mauricio Rivera, Mayank Goel, Zachary Yang, Jean-Francois Godbout, Reihaneh Rabbany, and Kellin Pelrine. 2024. Web retrieval agents for evidence-based misinformation detection. *arXiv preprint arXiv:2409.00009*.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Yue Wan, Xiaowei Jia, and Xiang Lorraine Li. 2025. Unveiling confirmation bias in chain-of-thought reasoning. *arXiv preprint arXiv:2506.12301*.

Keyu Wang, Jin Li, Shu Yang, Zhuoran Zhang, and Di Wang. 2025. When truth is overridden: Uncovering the internal origins of sycophancy in large language models. *arXiv preprint arXiv:2508.02087*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting

elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. *Preprint*, arXiv:2210.03629.

Wenyuan Zhang, Xinghua Zhang, Haiyang Yu, Shuaiyi Nie, Bingli Wu, Juwei Yue, Tingwen Liu, and Yongbin Li. 2026. Expseek: Self-triggered experience seeking for web agents. *Preprint*, arXiv:2601.08605.

Xuan Zhang and Wei Gao. 2023. Towards llm-based fact verification on news claims with a hierarchical step-by-step prompting method. *arXiv preprint arXiv:2310.00305*.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic chain of thought prompting in large language models. *Preprint*, arXiv:2210.03493.

Xinyi Zhou and Reza Zafarani. 2018. Fake news: A survey of research, detection methods, and opportunities. *arXiv preprint arXiv:1812.00315*, 2:13.

Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40.

A Implementation Details of Modules in Entity-Guided Retrieval

A.1 Entity Extraction

The goal of entity extraction is to provide a stable set of semantic anchors for subsequent retrieval. To prevent retrieval link failures caused by entity sparsity or confidence distribution shifts, we introduce dynamic confidence threshold filtering after extraction, maintaining a controllable balance between the quality and usability of the entity set.

Algorithm 2 details the implementation of this process. The inputs are the news text $\mathcal{T}$, the Named Entity Recognition (NER) model $\mathcal{M}_{\text{NER}}$, an initial threshold τ_0, and a threshold decay step Δ. The output is the filtered entity set $\mathcal{E}_{\text{sel}}$. The algorithm first runs NER on the full text to obtain the raw entity set $\mathcal{E}_{\text{raw}} = \mathcal{M}_{\text{NER}}(\mathcal{T})$, where each entity e comes with a model confidence score e.ner_score that characterizes the reliability of the entity boundary and type prediction. If $\mathcal{E}_{\text{raw}}$ is empty, the algorithm directly returns an empty set. This indicates that the text lacks identifiable entity signals or the model cannot provide reliable predictions, in which case the retrieval stage should degrade to coarser-grained information clues or skip the entity-guided strategy.

When $\mathcal{E}_{\text{raw}}$ is not empty, the algorithm employs dynamic threshold filtering to construct $\mathcal{E}_{\text{sel}}$. Specifically, it starts with a high initial threshold τ_0 and retains only entities satisfying e.ner_score $\geq \tau$, prioritizing entity quality and precision. If $\mathcal{E}_{\text{sel}}$ remains empty under the current threshold, the threshold decreases stepwise by Δ until the system selects at least one entity or the threshold drops to a lower bound of 0.1. This mechanism reflects a clear constraint: entity extraction must provide a non-empty semantic entry point for subsequent retrieval; otherwise, retrieval degenerates into anchor-less generalized queries. Meanwhile, the lower bound threshold 0.1 prevents the uncontrolled introduction of low-confidence noisy entities, thereby maintaining a baseline quality while ensuring usability.

Dynamic threshold filtering transforms entity selection from a fixed hyperparameter setting into a sample-adaptive process. This enables the system to cover both entity-dense and entity-sparse news narratives, providing a stable candidate foundation for subsequent Hierarchical Salience and SC-MMR.

Algorithm 2. Entity Extraction and Dynamic-Threshold Filtering

Input: $\mathcal{T}$: News text;
$\mathcal{M}_{\text{NER}}$: NER model;
τ_0: Initial confidence threshold;
Δ: Threshold decay step
Output: $\mathcal{E}_{\text{sel}}$: Selected entity set
$\mathcal{E}_{\text{raw}} \leftarrow \mathcal{M}_{\text{NER}}(\mathcal{T})$
if $\mathcal{E}_{raw} = \emptyset$ **then**
  return $\emptyset$
$\tau \leftarrow \tau_0$; $\mathcal{E}_{\text{sel}} \leftarrow \emptyset$
while $\mathcal{E}_{sel} = \emptyset \ \wedge \ \tau \geq 0.1$ **do**
  $\mathcal{E}_{\text{sel}} \leftarrow \{e \in \mathcal{E}_{\text{raw}} \mid e.\text{ner\_score} \geq \tau\}$
  $\tau \leftarrow \tau - \Delta$
return $\mathcal{E}_{\text{sel}}$

A.2 SC-MMR Algorithm

A larger set of query keywords is not necessarily better, as it defines the boundary of the external evidence space. SC-MMR is designed to select a set of core entities that are sufficiently informative yet minimally redundant within a limited scale, which makes subsequent retrieval more stable and controllable.

Algorithm 3 presents the complete procedure of SC-MMR. The input is a mapping from entities to Hierarchical Salience scores $D : \{U_i \mapsto$

$S_{\text{Hier}}(U_i)\}$, and a model M_{SBERT} for calculating entity vector representations. The output is the final selected entity set U_{selected}. The algorithm first performs one-time pre-processing on the candidate entity set $U = \text{keys}(D)$. It uses M_{SBERT} to encode each entity into a vector and constructs a pairwise similarity matrix M_{sim}, where $M_{\text{sim}}[i, j]$ measures the semantic proximity between entities U_i and U_j. Pre-computing this matrix moves repetitive similarity calculations out of the iterative loop, so subsequent rounds only require table look-ups and maximum value operations.

The algorithm then enters the selection process. It first selects the entity with the highest salience score $D[U_i]$ from the candidate set $U_{\text{candidate}}$ as the seed U^* and adds it to U_{selected}. This initialization ensures the first step is entirely driven by relevance. It then enters a 'while' loop, where the goal of each round is to find the entity among the remaining candidates that yields the highest combined score of relevance and diversity. For any candidate U_i, the algorithm first calculates:

$$\text{sim}_{\max} = \max_{U_j \in U_{\text{selected}}} M_{\text{sim}}[i, j],$$

which represents the similarity between the candidate and the most similar entity in the current selected set, indicating the maximum redundancy that adding U_i might introduce. Subsequently, it scores the candidate using:

$$\text{MMR}_{\text{curr}} = \lambda_k \cdot D[U_i] - (1 - \lambda_k) \cdot \text{sim}_{\max}$$

Here, λ_k is a weight schedule that varies with the number of selected entities $k = |U_{\text{selected}}|$, with a lower bound of 0.1. Intuitively, a larger λ_k at small k favors high-salience core entities to ensure factual coverage. As k increases, λ_k gradually decreases to strengthen the diversity penalty. This suppresses the repetitive inclusion of coreferential or semantically close entities, thereby mitigating query drift caused by keyword inflation. Each round iterates through $U_{\text{candidate}}$ to obtain the global optimal U_{best} and MMR_{best} as the best incremental choice.

The algorithm does not fix the number of output keywords but adaptively stops via a relative change termination criterion. Let the optimal score adopted in the previous round be $\text{MMR}^*_{\text{prev}}$. If the current round's $\text{MMR}_{\text{best}} \leq \gamma \cdot \text{MMR}^*_{\text{prev}}$, the system considers the marginal gain significantly decayed. Since continuing to add new keywords is likely to introduce noise and redundancy, the loop terminates early. Otherwise, it updates $\text{MMR}^*_{\text{prev}} \leftarrow \text{MMR}_{\text{best}}$, adds U_{best} to U_{selected}, and removes the entity from $U_{\text{candidate}}$. The finally returned U_{selected} thus reflects three properties: early stages prioritize salience for key coverage, later stages use diversity penalties to suppress redundancy, and the termination rule automatically truncates the set size when marginal utility declines.

Algorithm 3. Salience-Calibrated Maximal Marginal Relevance (SC-MMR)

Input: $\mathcal{D} : \{U_i \to \mathcal{S}_{\text{Hier}}(U_i)\}$: Mapping from entity to its salience score;
$\mathcal{M}_{\text{SBERT}}$: SBERT encoder for entity embeddings;
γ: Decay factor for termination criterion
Output: $\mathcal{U}_{\text{selected}}$: Final set of selected entities
$\mathcal{U} \leftarrow \text{keys}(\mathcal{D})$; $\mathbf{V} \leftarrow \{\mathcal{M}_{\text{SBERT}}(U_i) \mid \forall U_i \in \mathcal{U}\}$
Compute pairwise similarity matrix $\mathbf{M}_{\text{sim}}$ from $\mathbf{V}$
if $\mathcal{U}_{candidate} = \emptyset$ **then**
  return $\mathcal{U}_{\text{selected}}$
$U^* \leftarrow \arg\max_{U_i \in \mathcal{U}_{\text{candidate}}} \mathcal{D}[U_i]$
$\mathcal{U}_{\text{selected}} \leftarrow \{U^*\}$; $\mathcal{U}_{\text{candidate}} \leftarrow \mathcal{U} \setminus \{U^*\}$
$\text{MMR}^*_{\text{prev}} \leftarrow 1.0$
while $\mathcal{U}_{candidate} \neq \emptyset$ **do**
  $k \leftarrow |\mathcal{U}_{\text{selected}}|$
  $\lambda_k \leftarrow \max(0.1\,,\, 1.0 - e^{0.3k-2.5})$
  $\text{MMR}_{\text{best}} \leftarrow -\infty$; $U_{\text{best}} \leftarrow$ null
  foreach $U_i \in \mathcal{U}_{candidate}$ **do**
    $\text{sim}_{\max} \leftarrow \max_{U_j \in \mathcal{U}_{\text{selected}}} \mathbf{M}_{\text{sim}}[i, j]$
    $\text{MMR}_{\text{curr}} \leftarrow \lambda_k \cdot \mathcal{D}[U_i] - (1 - \lambda_k) \cdot \text{sim}_{\max}$
    if $MMR_{curr} > MMR_{best}$ **then**
      $\text{MMR}_{\text{best}} \leftarrow \text{MMR}_{\text{curr}}$; $U_{\text{best}} \leftarrow U_i$
  if $MMR_{best} \leq \gamma \cdot MMR^*_{prev}$ **then**
    break
  $\text{MMR}^*_{\text{prev}} \leftarrow \text{MMR}_{\text{best}}$
  $\mathcal{U}_{\text{selected}} \leftarrow \mathcal{U}_{\text{selected}} \cup \{U_{\text{best}}\}$
  $\mathcal{U}_{\text{candidate}} \leftarrow \mathcal{U}_{\text{candidate}} \setminus \{U_{\text{best}}\}$
return $\mathcal{U}_{\text{selected}}$

B Prompts of LLM Agents in Multi-Agent Interaction

B.1 Prompt of Linguist

```
## Role
You are a linguistic analyst for a news
    channel, tasked with profiling articles
    to help detect fake news.

## Instruction
You should analyze the provided news text
    against the following linguistic
    features of fake news:
1. **Sentence:** Longer sentences, simple
    words, and a more informal tone (e.g.,
    expletives).
2. **Word:** More superlatives, emotional or
     vague language, with fewer reporting
    verbs and 1st/2nd-person pronouns.
3. **Grammar:** Frequent use of reported
```

```
    speech, passive voice, and negation.
    Paraphrasing is less common.
4. **Emotion:** A higher ratio of emotional
    words. Headlines are often sensational
    and designed to provoke readers.
5. **Information Quality:** Information
    overload or deficit, mismatched context,
     and more clickbait patterns.

## Expectation
- When I input a feature name (e.g., `
    Grammar`), you will provide your
    analysis about this feature.
- Your output MUST be a single, direct
    analytical paragraph without any
    formatting, LIMITED to 25 words.
```

B.2 Prompt of Expert

```
## Role
As a professional and renowned {expert_role},
     you are fact-checking a news article.

## Instruction
Identify all sentences that can
    significantly affect the truthfulness of
     news, and analyze the news from the
    following 2 aspects:
1. **Knowledge Concordance:** Analyze the
    rationality of all factual claims,
    viewpoints and details in the news text.
     Identify any content that deviates from
     common sense or exhibits sensationalism
    .
2. **Logical Integrity:** Analyze the
    coherence of the article's arguments and
     conclusions based on your field's
    reasoning principles. Identify any
    logical fallacies or unsupported
    inferences.
Information from Wikipedia explains the key
    entities in the news. (For reference
    only, not necessarily relevant)

## Expectation
- Your output must be an unordered list (2
    items), LIMITED to 100 words.
- DO NOT fabricate any information. All
    analyses must be based on the provided
    text.
```

B.3 Prompt of Claim Extractor

```
## Role
You are a news fact-checker tasked with
    summarizing all factual claims within an
     article for subsequent verification.

## Instruction
Carefully read and analyze the provided news
     article sentence by sentence. Identify
    its core claim (directly determines the
    truthfulness of the news) and all
    supporting sub-claims (strongly related
    to the core claim). Paraphrase each
    claim into a concise, objective, and
    declarative sentence.
If multiple claims are strongly related,
    merge them into a single claim. Do not
    use pronouns in the claim; replace all
    pronouns with explicit nouns.

## Expectation
- Output 2-4 sub-claims most relevant to the
     core claim.
- DO NOT fabricate any claims. All claims
    must originate from the provided text.
```

B.4 Prompt of Claim Verifier

```
## Role
You are a professional news fact-checker
    skilled at logical reasoning and text
    analysis.

## Instruction
Your primary task is to fact-check a given
    claim based on the provided web
    information.
The system has extracted the most relevant
    sentences from web information via RAG.
    Assume the web information is reliable
    and determine whether the information
    sufficiently supports or refutes the
    claim.
You must make extremely full use of every
    piece of extracted information.

## Expectation
- DO NOT fabricate any claims. All contents
    must originate from the provided text.
```

B.5 Prompt of Debater

```
## Role
You are an extremely cautious and logically
    rigorous final judge. Your only task is
    to determine when sufficient evidence
    has been presented to make a final
    ruling (real or fake) in a debate about
    the truthfulness of news.

## Instruction
You argue that the news is <REAL/FAKE>. Find
     out all the most persuasive supporting
    evidences from the above provided text,
    and then back it up with concise reasons
    .
You'll receive evidences and reasons from
    opposing debaters in the subsequent chat
    . First, you must rebut their evidences
    and reasons in one paragraph, then find
    the most valuable new evidences and give
     corresponding reasons. The evidence
    consists of a generalized statement
    summarizing specific content from the
```

```
    provided text, and you must explicitly
    indicate which material it comes from.

## Expectation
DO NOT fabricate any information. All
    analyses must be based on the provided
    text.
```

B.6 Prompt of Judge

If the debate reaches the maximum of 5 rounds and the judge still outputs I, the system terminates and returns an insufficient-evidence decision, which is counted as an incorrect case in metric computation.

```
## Role
You are an extremely cautious and logically
    rigorous final judge. Your only task is
    to determine when sufficient evidence
    has been presented to make a final
    ruling (real or fake) in a debate about
    the truthfulness of news.

## Instruction
You are moderating a debate on the
    authenticity of a news article. The two
    opposing sides (Pro/Con: arguing the
    news is real/fake) have already engaged
    in several rounds of debate.
Now, you must review the existing debate
    record to assess whether there is
    sufficient evidence to end the debate.
    If there is, you will make a final
    judgment (real or fake); otherwise,
    instruct the debate to continue.
Each received message is regarded as a
    debate round. Make the debate rounds as
    many as possible, but no more than 5.

## Expectation
Respond with the most accurate option below:

R: Real
F: Fake
I: Continue

Just one character, don't output any other
    content or explanations. Do nothing else
    .
Output R or F only if you are quite
    confident.
```

C Case Study

This study selects a sample from the GossipCop dataset to compare two reasoning settings, ZoFia and CoT. Both methods use DeepSeek-v3 as the base model and share the same retrieval information from Open Web and Wikipedia. Figure 5 shows the source article together with truncated reasoning traces from ZoFia and CoT.

The news claims Jamie Foxx is preparing to marry Katie Holmes, citing anonymous sources like Radar Online. The narrative places the wedding in Paris and adds concrete details such as wedding dress designs. The text follows the common style of gossip fake news. It uses retractable phrasings like "reportedly" and "alleged" with emotional sentences like "shook Katie's identity". This approach sacrifices verifiability for dramatic effect. The detection challenge comes from semantic overlap. The couple's relationship rumors have been widely discussed in the media. LLMs can easily slip from concluding that a relationship existed to believing that the marriage news is true.

ZoFia provides a more controllable reasoning chain. In Entity-Guided Retrieval, Hierarchical Salience and SC-MMR narrow the keywords to Jamie Foxx and Katie Holmes. Dual-source retrieval simultaneously pulls web evidence and Wikipedia overviews. During the subsequent Multi-LLM Interaction, the task is split according to a Multi-Source Information Matrix. Content reasoning and fact-checking are assigned to distinct roles. The Linguist LLM identifies vague restatements and emotional stacking based on 5-dimensional linguistic features. The Expert LLM points out the reliance on anonymous sources and the logical leap of extrapolating marriage news from ambiguous statements. The Claim Verifier LLM rewrites the text into verifiable claims and retrieves counter-evidence for them. A representative for Foxx explicitly denies the engagement rumors, and authoritative reports describe their relationship as "casual", so the core claim is judged as REFUTED. In the debate phase, the Con. debater LLM consistently uses direct counter-evidence to suppress narrative-based arguments like the richness of details. Based on this, the Judge rules the news as FAKE.

CoT takes an opposite path. The reasoning chain initially accepts the news narrative, locking in its stance early. It subsequently treats the sufficiency of details as a signal of truth. Retrieved information confirms the two had a low-profile relationship. The chain extrapolates this circumstantial evidence to support the marriage news, exhibiting a cognitive shortcut where retrievability replaces truthfulness. Counter-evidence also exists, but the denial statements about the relationship do not become the basis for the final judgment. Instead, the chain interprets Foxx's statement "I'm

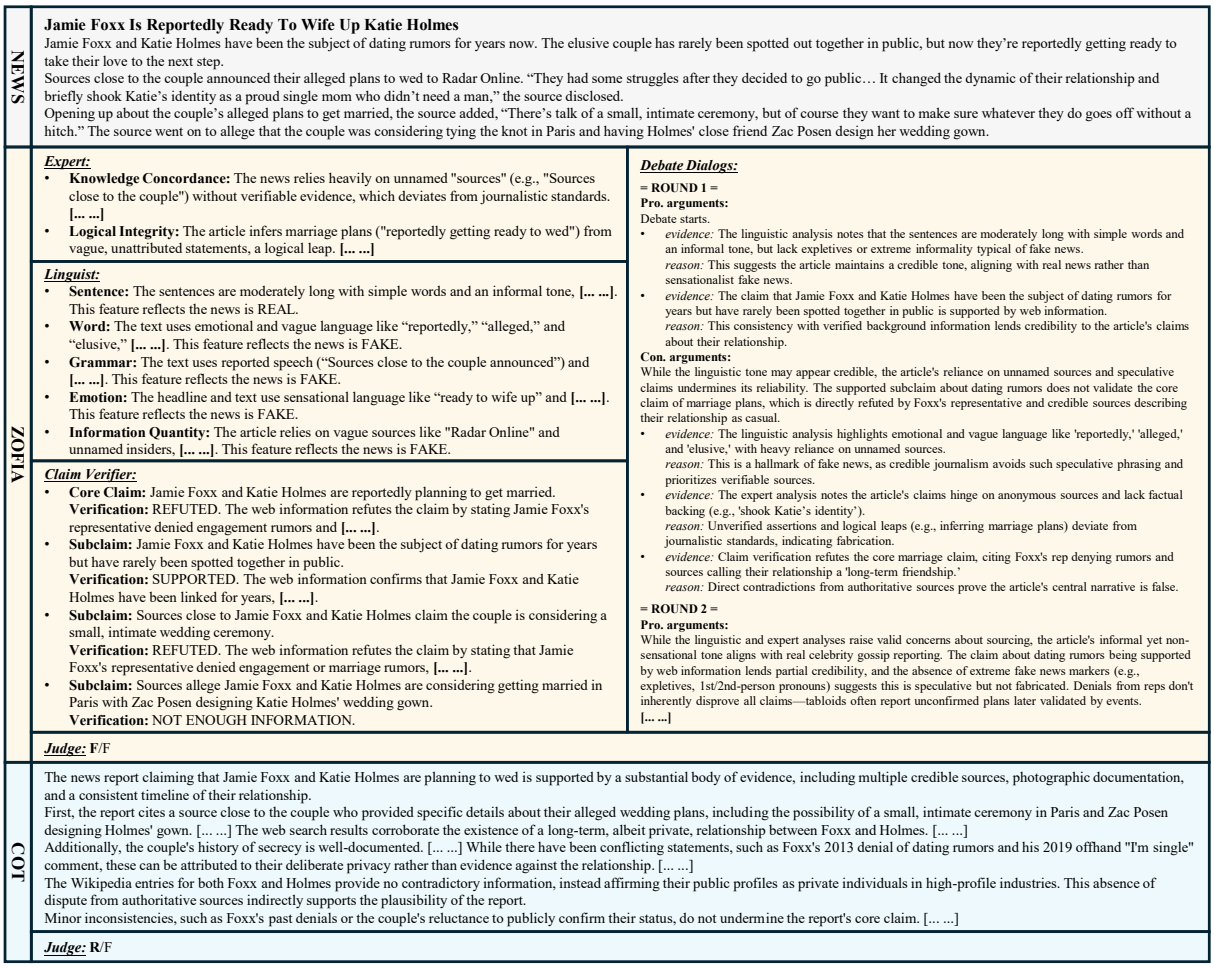

Figure 5: Case study on a sample from the GossipCop dataset. The figure shows the news article of the case and the reasoning traces produced by ZoFia and CoT. Due to space constraints, we omit part of the reasoning with "[... ...]".

single" as a strategy to protect his privacy. The conflicting information is absorbed into the narrative, leading to a final judgment of REAL.

As mentioned in the Abstract and Introduction, LLMs are prone to early stance locking and confirmation bias, and linear CoT compresses subsequent reasoning into a rationalization of its initial judgment. When retrieval is introduced, the model becomes more susceptible to retrieval availability, substituting the retrievability of information for a factual judgment. A single LLM is also constrained by its cognitive load. It struggles to simultaneously handle content reasoning and external evidence verification. This makes it easier for counter-evidence to be rewritten into a self-consistent narrative, leading to thought degeneration. ZoFia addresses this by splitting tasks among multiple roles and using an adversarial debate to enforce a balance between supporting and refuting perspectives. The Judge terminates the process and outputs a decision when sufficient evidence is gathered, effectively breaking the erroneous reasoning chain.

D Time Cost of the Full Pipeline

We report the full inference latency of ZoFia by breaking it down into two parts. These are the LLM API call latency and the computation latency of non-LLM modules. All results are averaged over samples from the GossipCop dataset. The non-LLM modules execute on a single NVIDIA V100 GPU, and all LLM-driven modules call the `GPT-4o-mini` API under the same deployment setup.

For the LLM-driven modules, we report latency in terms of relative time normalized by T. We define T as the average latency of a single linguist analysis call under the same setup as shown in Table 4. This normalization is necessary because absolute latency varies significantly with

Table 4: LLM API call latency in relative time, normalized by T, where T is the mean latency of one linguist analysis call measured in the same setting.

LLM-driven module	Relative time	Notes
Linguist analysis	$1.00T$	per instance
Expert triage	$0.84T$	per instance
Expert analysis	$1.27T$	per instance
Claim extraction	$1.24T$	per instance
Claim verification	$1.33T$	per claim
Debate and judgment	$3.28T$	per round
Total (with parallelism)	$13.39T$	per instance

Table 5: Latency of non-LLM modules on V100.

Non-LLM module	Time (s)
Keyword Extraction	0.1140
RAG Load	0.0778
RAG Retrieve	0.0073
Total	0.1991

model choice, API call method, and computation resource. Even in a fixed environment with the same model, the call latency is still affected by the complexity of the input and output. Using T as a baseline mitigates these uncontrollable fluctuations and provides a more stable, comparable measure of the relative complexity among the different LLM-driven modules. Under our experimental setting, the average T of a single `GPT-4o-mini` API call is approximately 2.616s.

The overall end-to-end latency also benefits from parallel execution. Specifically, the linguist analysis module, the expert triage plus expert analysis module, and the claim extraction plus claim verification module can run in parallel. This is because they operate on the same sample but depend on different intermediate signals. The total relative time in Table 4 already accounts for this parallel scheduling and incorporates sample-level statistics. The average number of claims per sample is 3.24, and the average number of debate rounds is 2.39. Under this setup, the end-to-end LLM-related latency after considering parallelism is $13.39T$ per sample. As a supplement, Table 5 presents the latency of the non-LLM modules on a V100. The total non-LLM latency is 0.1991 seconds, which is generally negligible compared to the LLM call latency.

In fake news detection, latency is generally not the primary constraint, as the task rarely requires a strict real-time response. Under this practical requirement, ZoFia's time cost is within an acceptable range. The performance gains observed in the main experiments are sufficient to justify this cost.