TELLER: A Trustworthy Framework For Explainable, Generalizable and Controllable Fake News Detection

Hui Liu[1,2] **Wenya Wang**[2] **Haoru Li**[3] **Haoliang Li**[1]
[1]City University of Hong Kong
[2]Nanyang Technological University
[3]University of Electronic Science and Technology of China
liuhui3-c@my.cityu.edu.hk, wangwy@ntu.edu.sg, lihaoru114@gmail.com, haoliang.li@cityu.edu.hk

Abstract

The proliferation of fake news has emerged as a severe societal problem, raising significant interest from industry and academia. While existing deep-learning based methods have made progress in detecting fake news accurately, their reliability may be compromised caused by the non-transparent reasoning processes, poor generalization abilities and inherent risks of integration with large language models (LLMs). To address this challenge, we propose TELLER, a novel framework for trustworthy fake news detection that prioritizes explainability, generalizability and controllability of models. This is achieved via a dual-system framework that integrates cognition and decision systems, adhering to the principles above. The cognition system harnesses human expertise to generate logical predicates, which guide LLMs in generating human-readable logic atoms. Meanwhile, the decision system deduces generalizable logic rules to aggregate these atoms, enabling the identification of the truthfulness of the input news across diverse domains and enhancing transparency in the decision-making process. Finally, we present comprehensive evaluation results on four datasets, demonstrating the feasibility and trustworthiness of our proposed framework. Our implementation is available at this link[1].

1 Introduction

Fake news has emerged as a prominent social problem due to the rampant dissemination facilitated by social media platforms (Zhou and Zafarani, 2021). Additionally, the swift progress of generative artificial intelligence has further amplified this issue (Cardenuto et al., 2023). While human fact-checking experts can accurately verify the authenticity of news, their efforts cannot scale with the overwhelming volume of online information. Consequently, researchers have turned to automatic fake news detection techniques.

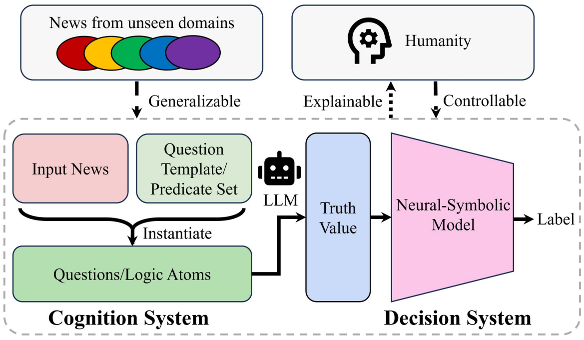

Figure 1: Three crucial aspects of trustworthy fake news detection algorithms and the correlation between these principles and our dual-sytem framework TELLER.

Despite the improved predictive accuracy achieved by current deep learning-based detection approaches (Ma et al., 2023; Qi et al., 2021; Mehta et al., 2022), these methods suffer from the lack of transparency because of the black-box nature of neural networks (Cui et al., 2019) and a limited ability to generalize to unseen data of which the distribution is different from training data, given the inherent diversity of online information (e.g., topics, styles and media platforms) (Liu et al., 2024). Moreover, the increasing integration with LLMs is prone to uncontrollable risks due to hallucinations and societal applications. Thus, a growing awareness emphasizes trustworthiness[2] of these systems (Liu et al., 2023; Sheng et al., 2022).

Unfortunately, the characteristics of a trustworthy fake news detector remain an open question. Hence, based on recent surveys of Trustworthy AI (Li et al., 2023; Jobin et al., 2019) and fake news detection (Shu, 2023), we identify three crucial aspects that go beyond accuracy for fake news detection technologies: explainability, generalizability, and controllability. These aspects work collectively

[1]https://github.com/less-and-less-bugs/Trust_TELLER

[2]In AI, trustworthiness refers to the extent to which an AI system can be trusted to operate ethically, responsibly, and reliably (Jobin et al., 2019).

Findings of the Association for Computational Linguistics ACL 2024, pages 15556–15583

to enhance system security and trustworthiness.

Firstly, explainability refers to understanding how an AI model performs decision (Miller, 2019). This mechanism serves as a fundamental requirement for establishing end-user trust in these tools, as it enables the disclosure of complex reasoning processes and the identification of potential flaws in neural networks. Secondly, generalizability represents the capability to acquire knowledge from limited training data to predict accurately in unseen situations (Wang et al., 2023a). Given the impracticality of exhaustively collecting and annotating vast amounts of data across various news domains, generalization ensures the affordable and sustainable deployment of data-driven fake news detection algorithms. Lastly, controllability encompasses the capacity for human guidance and intervention in the behavior of models (Ji et al., 2023a). This objective benefits models in understanding specific misinformation regulatory policies and rectifying deviations if necessary. While recent practices may satisfy the requirements of explainability (Xu et al., 2022; Liu et al., 2023) or generalization (Kochkina et al., 2018; Yue et al., 2023), they often fail to adhere to all three principles simultaneously.

To this end, we propose TELLER, a **T**rustworthy framework for **E**xplainable, genera**L**izable and contro**L**labe d**E**tecto**R**, drawing inspiration from the dual-system theory[3] (Daniel, 2017). This framework abstracts the existing pipeline of fake news detection into two components: the cognition and decision systems. As depicted in Fig. 1, the cognition system serves as the first step and is responsible for transforming meaningful human expertise from renowned journalism teams (Tsang, 2023; Sanders, 2023) into a set of Yes/No question templates that correspond to logic predicates. These decomposed questions are then answered using LLMs, which provide truth values for corresponding logic atoms.

On the other hand, the decision system, empowered by a differentiable neural-symbolic model (Cingillioglu and Russo, 2021), can integrate the output of the cognition system to deduce the final authenticity of input news by leveraging domain invariant logic rules learned from data automatically. This visible logic-based ensemble can mitigate the negative effects caused by inaccurate predictions of LLMs and allow for the correction of unreasonable rules through adjusting the weights in the model manually to align with human expertise.

Our framework ensures explainability by incorporating human-readable question templates (predicates) and a transparent decision-making process based on logic rules. This interpretability further enables the flexibility to adjust rules and enhances the model's robustness against false LLM predictions, thereby guaranteeing controllability. Moreover, our model exhibits generalizability, attributed to the generalizable performance of LLMs combined with reliable human experience as guidance and the utilization of the neural-symbolic model, which can learn domain-generalizable rules.

To summarize, the contributions of this work include: 1) We introduce a systematic framework comprising cognition and decision modules, aiming to uphold three crucial principles for establishing a trustworthy fake news detection system: explainability, generalizability, and controllability. 2) We validate the effectiveness of our framework by conducting comprehensive experiments using various LLMs on four benchmarks. The results demonstrate the feasibility and trustworthiness of TELLER across different scenarios.

2 Related Work

2.1 Trustworthy AI

Establishing comprehensive trustworthiness in AI is non-trivial due to its multi-objective nature, including robustness, security, transparency, fairness, safety, and ethical standards (Jobin et al., 2019). Achieving such trustworthiness necessitates considering the entire lifecycle of an AI system, spanning from data preparation and algorithm design, development, and deployment to management and governance (Li et al., 2023; Eykholt et al., 2018). Recent researchers have explored diverse approaches to enhance AI trustworthiness across various goals and stages to address this challenge. For example, regarding algorithm design, several topics, such as transfer learning, federated learning, and interpretable AI, have been proposed to improve models' robustness, security, and transparency. Moreover, the deployment of AI systems necessitates external government oversight, particularly for AGI (Bengio et al., 2023). Although our work focuses on enhancing the trustworthiness of detection systems from the algorithm design aspect, we acknowledge that there is still much room for improvement to achieve the ultimate goal.

[3]System 1 provides tools for intuitive, imprecise, and unconscious decisions akin to deep learning, while system 2 handles complex situations requiring logical and rational thinking akin to symbolic learning (Booch et al., 2021).

2.2 Trustworthy Fake News Detection

Recent fake news detection research has witnessed a notable paradigm shift from prioritizing accuracy to considering trustworthiness. In line with our work, we primarily examine studies that aim to enhance algorithms' explainability, generalizability, and controllability.

Regarding explainability, Cui et al. (2019); Xu et al. (2022); Liao et al. (2023) suggested obtaining key evidence for interpretation based on feature importance, while Liu et al. (2023) utilized logic clauses to illustrate the reasoning processing. However, these methods still need to be more transparent due to their probabilistic nature and complex architecture. Furthermore, another group of works (Huang and Sun, 2023; Hu et al., 2023; Yue et al., 2024; Qi et al., 2024), explored large generative language models (e.g., ChatGPT) and regarded the intermediate chain of thoughts as an explanation. Nevertheless, these explanations may not be reliable due to the hallucination phenomenon (Ji et al., 2023b) and the misalignment problem of AGI (Ji et al., 2023a). Moving on to generalizability, most methods, such as (Yue et al., 2023; Zhu et al., 2023; Qi et al., 2021), enhanced fake news detectors through transfer learning algorithms to learn domain-invariant features or domain-adaptive features. However, these methods inevitably introduce external costs of domain alignment, such as annotating domain labels. As for controllability, although some works (Silva et al., 2021; Mendes et al., 2023) incorporated the human-in-loop technique in data sampling and model evaluation, few works explore how to intervene and edit models to align with human expertise. More comparative discussion between TELLER and existing work can be found in Appendix E.

3 Methodology

Formally, given a piece of news T, the objective of the fake news detection task is to predict its label of truthfulness $y \in \mathcal{Y}$ where $\mathcal{Y}$ can fit in different levels of classification granularity. For example, in binary classification setting, $\mathcal{Y} = \{\text{true}, \text{false}\}$, and T is identified as real (fake) when y is true (false).

As depicted in Fig. 2, TELLER involves two main components: cognition and decision systems. The cognition system decomposes human expertise into Yes/No question templates corresponding to logic predicates. When presented with a new input T, the templates and predicates can be instantiated to form questions and logic atoms. By leveraging the parametric knowledge inside LLMs and gathering additional information from external tools (e.g., search engines), the cognition system can generate answers to these questions, represented as truth values of logic atoms. Then, the decision system takes these truth values as input and generates interpretable logic clauses to debunk misinformation by a neural-symbolic model, which can learn generic logic rules from data in an end-to-end manner.

3.1 Cognition System

To combat misleading information, existing deep learning-based algorithms fall short in gaining public trust, while fact-checking experts rigorously follow designated guidance and principles to facilitate transparent and fair evaluation. Our cognitive system aims to integrate the strengths of deep learning-based methods that can handle large-scale online information while maintaining the trustworthiness of manual checking.

3.1.1 Predicate Construction

To begin with, we describe the following symbol convention for clarity: calligraphic font $\mathcal{Q}$ and $\mathcal{P}$ for sets of question templates and predicates, capitalized letters Q, P, X for question templates, predicates, and variables, and corresponding lowercase letters q, p, x for instances of these entities (questions, logic atoms, values). The truth values of logic atoms are denoted by μ.

Inspired by the well-established fact-checking process in Table 5, we initially decompose it into a question template set, denoted as $\mathcal{Q}$, containing eight questions as detailed in Appendix A.1. Each template Q_i in $\mathcal{Q}$ consists of N_i variables and can be transformed into an N_i-ary logic predicate $\mathrm{P}_i(\mathrm{X}_{i,1}, \dots, \mathrm{X}_{i,N_i})$ in $\mathcal{P}$. The logic semantics of P_i is interpreted as the affirmative answer to Q_i and its truth value μ_i represents the probability that P_i holds. For instance, take Q_1 (i.e., "Background Information: $\mathrm{X}_{1,1}$. Statement: $\mathrm{X}_{1,2}$. Is the statement true?") in Fig. 2 as an example. The corresponding predicate $\mathrm{P}_1(\mathrm{X}_{1,1}, \mathrm{X}_{1,2})$ can be explained as "Given the background information $\mathrm{X}_{1,1}$, the statement $\mathrm{X}_{1,2}$ is true".

For each predicate $\mathrm{P}_i(\mathrm{X}_{i,1}, \dots, \mathrm{X}_{i,N_i})$, we can instantiate the variables $\mathrm{X}_{i,1}, \dots, \mathrm{X}_{i,N_i}$ with the actual contents taken from any input news to obtain logic atoms. Since an input piece of news may contain multiple background information and statements (instantiations), we use k to denote the kth

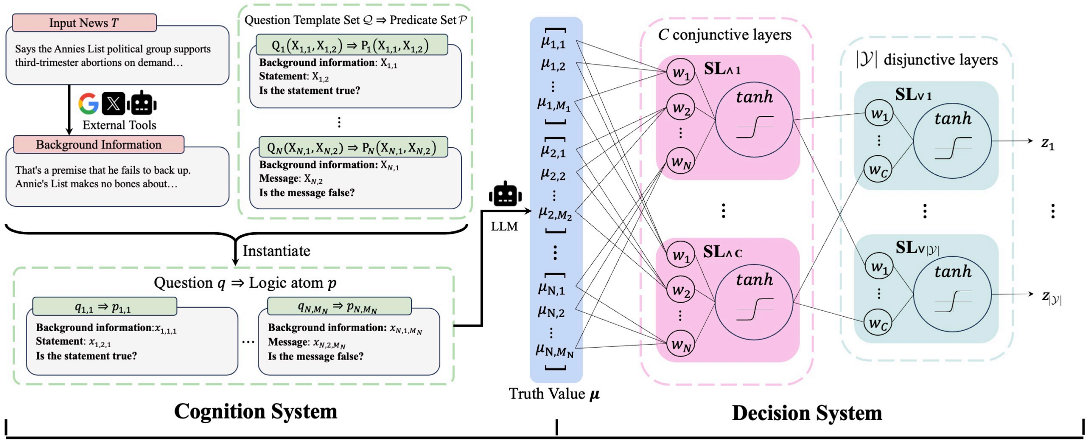

Figure 2: The architecture of the proposed framework TELLER. N represents the number of question templates (logic predicates), M_i denotes the number of logic atoms corresponding to the ith predicate, $\mathcal{Y}$ denotes the truthfulness label set. The semantics of question templates and logic predicates are described in Table 6.

instantiation where $1 \leq k \leq \prod_{j=1}^{N_i} |\mathrm{X}_{i,j}|$. Here $|\mathrm{X}_{i,j}|$ indicates the total number of possible instantiations for variable $\mathrm{X}_{i,j}$. Then we denote by $p_{i,k}$ the instantiated logic atom corresponding to the question $q_{i,k}$. Next, we introduce how to acquire the truth value of each logic atom.

3.1.2 Logic evaluation with LLMs

While decomposed questions can provide a comprehensive explanation of how the decision is made (Chen et al., 2022; Fan et al., 2020), directly answering these questions poses a challenge due to the impracticality of annotating enormous data to train multiple models for different questions. To address this issue, we resort to the more general-purpose LLMs (e.g., FLAN-T5 (Chung et al., 2022), Llama2 (Touvron et al., 2023b), and GPT-3.5) as the foundation for effectively answering these questions. Existing LLMs can be categorized into two groups: LLM$_{open}$, such as FLAN-T5 and Llama2, where the logits of output vocabulary can be obtained, and LLM$_{close}$, such as GPT-3.5, where the logits are not accessible.

To ensure compatibility with both categories of LLMs, we propose two strategies to obtain the final truth values of logic atoms. Concretely, we first input the question $q_{i,k}$ with a suffix (i.e., "Yes or No? Response:") to LLMs in order to measure their preference for the affirmative answer "Yes" versus the negative one "No". This preference is subsequently used to compute the truth value of the corresponding logic atom $p_{i,k}$.

For LLM$_{open}$, we follow (Gallego, 2023; Burns et al., 2023) to obtain pre-softmax logits of "Yes" and "No" tokens, denoted as v_{Yes} and v_{No} respectively. Compared with post-softmax logits, pre-softmax logits can mitigate the influence of other tokens in output vocabulary, particularly when LLMs tend to generate irrelevant tokens that may result in v_{Yes} or v_{No} becoming zero. Then the truth value μ for the logic atom p (here we omit the underscript i, k for ease of illustration) can be obtained as follows:

$$\mu = 2\frac{e^{v_{Yes}}}{e^{v_{No}} + e^{v_{Yes}}} - 1. \tag{1}$$

For LLM$_{close}$, we sample m times during decoding and count the frequency of "Yes" and "No" responses as m_{Yes} and m_{No}. Then we compute

$$\mu = 2\frac{m_{Yes}}{m_{No} + m_{Yes}} - 1. \tag{2}$$

In either case, μ is in the range of $[-1, 1]$. When $\mu \in [-1, 0)$, $\mu \in (0, 1]$, and $\mu = 0$, the corresponding logic atom p is evaluated as false, true, and unknown, respectively. Once the truth values of all logic atoms for a single predicate P_i (corresponding to a single question template) are obtained, we concatenate them as one vector, denoted as $\boldsymbol{\mu}_i$. Then we concatenate the value vectors for all predicates as the input for the final decision system.

In conclusion, our cognition system can generate diversified questions and logic atoms based on the input news T. These human-readable entities enhance explainability by showcasing potential intermediate reasoning steps and ensure controllability by allowing adjustments to $\mathcal{Q}$ and $\mathcal{P}$. Moreover, combining human expertise and LLMs provides the basis for the cognition system's satisfactory generalization performance in unseen domains.

3.2 Decision System

After acquiring responses to all questions, it is imperative to develop a decision system to effectively aggregate them to predict the label of the input news T while preserving trustworthiness in the reasoning process. However, prevalent heuristic strategies (e.g., majority voting) lack the flexibility to handle complex relationships among different questions and cannot tolerate false predictions, and deep-learning-based models cannot be comprehended literally by humans (Wang et al., 2023b).

Hence, we utilize a neural-symbolic model, named Disjunctive Normal Form (DNF) Layer (Cingillioglu and Russo, 2021; Baugh et al., 2023), as our decision system. This model includes conjunctive layers ($\mathrm{SL}_\wedge$) and disjunctive layers ($\mathrm{SL}_\vee$), which can progressively converge to symbolic semantics such as conjunction $\wedge$ and disjunction $\vee$ respectively during model training. Consequently, this model can automatically learn logic rules from data in an end-to-end manner, capturing generalizable relationships between logic predicates and the target label. As illustrated in Fig. 2, we stack C conjunctive layers $\mathrm{SL}_\wedge$ beneath $|\mathcal{Y}|$ disjunctive layers $\mathrm{SL}_\vee$ to construct the DNF Layer, where each $\mathrm{SL}_\vee$ corresponds to a truthfulness label $y \in \mathcal{Y}$.

However, the original DNF Layer proposed in (Cingillioglu and Russo, 2021) is not directly applicable to our work due to two issues. Firstly, the truth value of logic atoms μ ranges in $[-1, 1]$, while the original model can only handle values of -1 and 1. Secondly, each logic atom in the original DNF Layer is treated differently which loses logic semantics where atoms for the same logic predicate should share similar functionality. To address the aforementioned challenges, we propose a modified DNF layer which takes continuous values $\mu \in [-1, 1]$ as input and assigns the same weight for those atoms instantiated from the same logic predicate. The detailed description of our modified DNF layer can be found in Appendix G.

More concretely, in our proposed DNF Layer, every $\mathrm{SL}_\wedge$ takes truth values $\boldsymbol{\mu}$ of all logic atoms obtained in the cognition system as input, aiming to learn a conjunctive clause $\mathrm{conj} = \bigwedge_{p_{i,k} \in \mathcal{A}} p_{i,k}$ where $\mathcal{A} \subseteq \{p_{1,1}, \dots, p_{N,M_N}\}$, referring to a subset of the complete logic atoms, and outputs the truth value of this conjunctive clause. Subsequently, each $\mathrm{SL}_\vee$ receives the truth values of C conjunctive clauses to represent a disjunction of these conjunctions: $\bigvee_{c \in \mathcal{C}} \mathrm{conj}_c$ where $\mathcal{C} \subseteq \{1, \dots, C\}$, referring to a subset of all conjs. It then outputs the truth value of this disjunction formula, corresponding to the final probability that the input news T is identified as the label y. Hence, each label y will be associated with a DNF clause learned by the DNF layer. Intuitively, the conjunction simulates the idea that if the input news T gives affirmative answers to some questions simultaneously, it is highly probable that it should be assigned to label y. On the other hand, the disjunction provides more flexibility by considering different alternatives (the output is true if at least one of the conj is true) which makes the final decision less sensitive to incorrect atom values due to wrong predictions given by LLMs. For example, assume the learned rules are $\mathrm{conj}_1 \vee \mathrm{conj}_2$ where $\mathrm{conj}_1 = p_{1,1} \wedge p_{1,2}$ and $\mathrm{conj}_2 = p_{2,1} \wedge p_{3,1}$. Suppose conj_1 is true, then we can conclude that $\mathrm{conj}_1 \vee \mathrm{conj}_2$ is true even if conj_2 gives an incorrect value.

Last but not the least, we apply softmax function to the output of all disjunction layers $\mathrm{SL}_\vee$ to obtain the probability $\boldsymbol{z} \in \mathbb{R}^{|\mathcal{Y}|}$ for all possible labels. The entire decision system can be trained in an end-to-end fashion by minimizing the cross-entropy loss function as below:

$$\mathcal{L} = -\sum_{l=1}^{|\mathcal{Y}|} \mathbb{I}(y_l = y_T) \log z_l, \tag{3}$$

where y_T represents the ground truth label of T. During inference, we select the label corresponding to the highest value in $\boldsymbol{z}$ as the final result.

In summary, our decision system can extract interpretable symbolic rules from data that exhibit robustness across diverse domains and enable intervention by adjusting weights in the DNF Layer to align with prior knowledge (refer to Appendix C).

4 Experiments

In this section, we present the experiment setup and demonstrate the feasibility, explainability, generalizability and controllability of TELLER through extensive experiments.

4.1 Experimental Setting

Dataset. We conducted experiments using four challenging datasets, namely LIAR (Wang, 2017), Constraint (Patwa et al., 2021), PolitiFact, and GossipCop (Shu et al., 2020). LIAR comprises the binary classification and multi-classification setting with six fine-grained labels for truthfulness ratings. Moreover, Wang (2017); Alhindi et al.

(2018) curated relevant evidence (e.g., background information), serving as gold knowledge in an open setting. Constraint, PolitiFact and GossipCop are binary classification datasets related to COVID-19, politics, and entertainment domains, respectively.

LLMs. We select the open-source FLAN-T5 and Llama2 series, which encompass various parameter sizes, as large language models for constructing the cognition system. We also conduct experiments using GPT-3.5-turbo on the LIAR dataset to examine the versatility of our framework.

Baselines. We compare our model against *Direct*, *Few-shot Direct*, *Zero-shot COT*, *Few-shot COT*, *Few-shot Logic*. The baselines suffixed with *Direct* involve prompting large language models (LLMs) to predict the label of input news directly; those suffixed with *COT* utilize chain-of-thought techniques to enhance the performance of LLMs; those suffixed with *Logic* replace the thought process in COT with questions paired with their answers. We exclusively implement COT-related methods using GPT-3.5-turbo because they show no improvement over *Direct* on FLAN-T5 and Llama 2, as shown in Table 12. Additionally, we compare with small models, including BERT and RoBERTa, analyzed in Appendix E.

Implementation Detail. We evaluate the performance of our framework using the accuracy and Macro-F1, which accommodates class imbalance. For each dataset, we train our decision system using the training split; select the optimal model based on its performance on the validation split; and report the results on the test split. To assess the generalizability of our model, we consider each dataset as a separate domain and train our models using the train split from source domains; choose the best model on the validation split of source ones; and report results on the test split from the target domain. Moreover, to highlight the robustness of our framework, we keep all hyperparameters fixed in each setting. Details of the experiment setting, data leakage analysis, baselines, and model training are elaborated in Appendix B.

4.2 Feasibility Study

To validate the feasibility of our framework, we compare it against multiple baselines across a wide range of LLMs and scenarios (e.g., different classification granularities) in Table 1 and Table 2. These results uncover two crucial findings listed below:

Firstly, our framework demonstrates satisfactory performance in fake news detection tasks. Specifically, in the binary classification setting, TELLER achieves an accuracy of approximately 76% on the GossipCop dataset and over 80% on the other three datasets. Notably, when utilizing Llama 2 (13B) to drive the cognition system, TELLER outperforms all GPT-3.5-turbo based methods by a significant margin. These results highlight the effectiveness of TELLER in distinguishing between fake and genuine news. In the multi-classification setting on the LIAR dataset, our framework consistently outperforms *Direct* for FLAN-T5 and Llama2 series, even though these models may struggle to discriminate fine-grained labels. This observation underscores the capability of our decision system to mitigate the negative influences of noisy predictions in the cognition system, effectively unleashing the potential of LLMs through logic-based aggregation of answers to decomposed questions.

Secondly, our framework exhibits significant potential for the future. In the binary classification setting across four datasets, TELLER consistently outperforms *Direct* in terms of accuracy and macro-F1 scores by an average of 7% and 6%, respectively. Considering the swift improvement of LLM intelligence, these results imply that the performance of our framework is likely to scale with the evolution of LLMs. Additionally, due to the notable performance difference between closed and open settings on the LIAR dataset, it is promising to integrate external tools to acquire extensive evidence from credible sources, such as official government websites, to enhance the performance of our systems.

4.3 Explainability Verification

Explainability is a fundamental factor for establishing trust in AI technology. We demonstrate that our framework satisfies this aspect through its inherent mechanism and the visualization of rules.

Unlike approaches that rely heavily on LLMs, our cognition system incorporates expert knowledge to construct a more well-grounded worldview by generating well-defined question templates and logic predicates. Moreover, our decision system can learn interpretable rules from data to deduce logic clauses to debunk fake news by converging implicit parameters to conjunctive and disjunctive semantics. These symbolic units (e.g., questions and logic atoms) and the interpretable DNF Layer contribute to our framework's overall explainability and transparency.

Large Language Models	Method	Binary Classification				Multi-Classification			
		Closed		Open		Closed		Open	
		Acc(%)	Macro-F1(%)	Acc(%)	Macro-F1(%)	Acc(%)	Macro-F1(%)	Acc(%)	Macro-F1(%)
FLAN-T5-small (80M)	Direct	44.99	31.63	45.08	32.41	18.17	9.28	19.51	10.13
FLAN-T5-base (250M)	Direct	54.02	50.79	61.47	61.43	19.43	11.79	21.40	21.40
FLAN-T5-large (780M)	Direct	57.30	52.20	74.38	73.84	19.43	17.84	29.50	24.95
	TELLER	66.83$_{(9.53\uparrow)}$	**66.33**$_{(14.13\uparrow)}$	77.76$_{(3.38\uparrow)}$	77.32$_{(3.49\uparrow)}$	**26.99**$_{(7.55\uparrow)}$	18.04$_{(0.20\uparrow)}$	33.67$_{(4.17\uparrow)}$	27.50$_{(2.55\uparrow)}$
	w/ Intervention	65.64	65.12	77.46	77.14	26.28	18.49	35.25	30.05
FLAN-T5-xl (3B)	Direct	58.89	58.62	75.97	75.67	19.67	16.57	29.43	24.74
	TELLER	62.36$_{(3.48\uparrow)}$	60.18$_{(1.56\uparrow)}$	78.75$_{(2.78\uparrow)}$	78.55$_{(2.88\uparrow)}$	24.31$_{(4.64\uparrow)}$	17.40$_{(0.83\uparrow)}$	33.52$_{(4.09\uparrow)}$	27.22$_{(2.48\uparrow)}$
	w/ Intervention	63.65	61.82	79.34	79.07	25.57	19.62	34.46	33.59
FLAN-T5-xxl (11B)	Direct	56.41	56.08	75.17	75.15	22.42	18.31	32.18	28.12
	TELLER	66.63$_{(10.23\uparrow)}$	65.91$_{(9.82\uparrow)}$	80.24$_{(5.06\uparrow)}$	79.85$_{(4.70\uparrow)}$	26.83$_{(4.41\uparrow)}$	19.68$_{(1.36\uparrow)}$	35.48$_{(3.30\uparrow)}$	30.42$_{(2.30\uparrow)}$
	w/ Intervention	**67.03**	66.19	80.73	80.41	**26.91**	**21.30**	35.88	31.63
Llama2 (7B)	Direct	59.88	59.19	72.29	69.63	18.02	9.97	11.01	6.88
	TELLER	62.46$_{(2.58\uparrow)}$	62.45$_{(3.26\uparrow)}$	79.94$_{(7.65\uparrow)}$	79.80$_{(10.16\uparrow)}$	23.29$_{(5.27\uparrow)}$	15.51$_{(5.55\uparrow)}$	32.73$_{(21.72\uparrow)}$	25.55$_{(18.67\uparrow)}$
	w/ Intervention	64.15	62.77	81.93	81.84	23.92	15.14	34.30	27.58
Llama2 (13B)	Direct	56.90	56.90	69.31	63.77	7.32	2.85	10.86	8.25
	Ours	66.04$_{(9.14\uparrow)}$	66.03$_{(9.13\uparrow)}$	**82.52**$_{(13.21\uparrow)}$	**82.37**$_{(18.60\uparrow)}$	25.81$_{(18.49\uparrow)}$	17.71$_{(14.86\uparrow)}$	38.08$_{(27.22\uparrow)}$	29.27$_{(21.02\uparrow)}$
	w/ Intervention	**<u>67.73</u>**	**<u>66.97</u>**	**<u>84.21</u>**	**<u>84.03</u>**	25.10	16.78	38.63	30.60
GPT-3.5-turbo	Direct	42.40	51.48	76.27	74.21	20.46	20.34	26.20	25.12
	TELLER	-	-	79.15$_{(2.88\uparrow)}$	78.90$_{(4.69\uparrow)}$	-	-	31.94$_{(5.74\uparrow)}$	29.53$_{(4.41\uparrow)}$
	Zero-shot COT	30.88	41.87	72.49	70.83	7.16	9.20	39.81	**36.49**
	Few-shot	61.67	64.05	81.02	81.00	25.65	**<u>25.56</u>**	**<u>46.81</u>**	**<u>44.61</u>**
	Few-shot COT	52.04	56.15	74.48	76.21	20.69	17.20	**45.63**	36.36
	Few-shot Logic	49.26	48.85	61.67	60.92	16.37	13.98	20.54	19.22

Table 1: Results on LIAR dataset. "Closed" represents the cognitive system does not have access to any external knowledge source, while "Open" indicates that it can utilize gold evidence collected by human experts. The best results for each setting are highlighted with bold numbers and an underline, whereas sub-optimal results are only highlighted in bold. The number indicates that the performance of *w/ Intervention* is worse than TELLER. The number with ↑ indicates the performance gain of TELLER over *Direct*.

LLMs	Method	Constraint		PolitiFact		GossipCop	
		Acc(%)	Macro-F1(%)	Acc(%)	Macro-F1(%)	Acc(%)	Macro-F1(%)
FLAN-T5-large	Direct	78.06	77.97	56.62	54.84	67.43	58.76
	TELLER	80.32$_{(2.27\uparrow)}$	80.11$_{(2.14\uparrow)}$	67.65$_{(11.03\uparrow)}$	67.65$_{(12.81\uparrow)}$	69.53$_{(2.10\uparrow)}$	59.39$_{(0.63\uparrow)}$
	w/ Intervention	80.46	80.31	68.38	68.29	70.28	60.74
FLAN-T5-xl	Direct	75.32	74.79	55.88	50.72	67.73	52.80
	TELLER	83.77$_{(8.45\uparrow)}$	83.66$_{(8.88\uparrow)}$	68.82$_{(9.14\uparrow)}$	64.68$_{(13.95\uparrow)}$	69.58$_{(1.85\uparrow)}$	58.72$_{(5.91\uparrow)}$
	w/ Intervention	83.95	83.88	69.12	68.79	72.23	63.84
FLAN-T5-xxl	Direct	74.80	73.23	52.21	43.65	68.93	52.82
	TELLER	83.39$_{(8.59\uparrow)}$	83.24$_{(10.01\uparrow)}$	69.12$_{(16.91\uparrow)}$	68.57$_{(24.92\uparrow)}$	69.18$_{(0.25\uparrow)}$	57.21$_{(4.39\uparrow)}$
	w/ Intervention	83.62	83.54	69.12	68.95	71.48	62.12
Llama2 (7B)	Direct	81.83	81.73	77.21	77.00	66.78	52.23
	TELLER	83.72$_{(1.89\uparrow)}$	83.54$_{(1.81\uparrow)}$	83.82$_{(6.62\uparrow)}$	83.81$_{(6.81\uparrow)}$	70.68$_{(3.90\uparrow)}$	59.58$_{(7.35\uparrow)}$
	w/ Intervention	85.13	85.04	**83.82**	**83.82**	73.38	65.32
Llama2 (13B)	Direct	57.53	51.75	77.94	77.10	52.55	52.27
	TELLER	87.31$_{(29.78\uparrow)}$	87.29$_{(35.53\uparrow)}$	79.41$_{(1.47\uparrow)}$	79.41$_{(2.30\uparrow)}$	74.48$_{(21.93\uparrow)}$	66.32$_{(14.06\uparrow)}$
	w/ Intervention	**87.78**	**87.71**	78.68	78.65	**75.92**	**69.30**

Table 2: Results on Constraint, PolitiFact, and GossipCop datasets without access to retrieved background information. The best results for each setting are highlighted with bold numbers. The number and the number with ↑ have the same meaning as in Table. 1.

However, as the number of conjunctive and disjunctive layers grows, it is difficult for human beings to investigate logic rules derived from our decision system. To address this issue, we propose a strategy to prune unnecessary weights in the DNF Layer. For example, we present the rules extracted from the pruned model for GossipCop in Table 4, where each conjunctive clause identifies one candidate rule. The pruning algorithm and rules for other datasets are described in Appendix C.

Table 4 can be interpreted as learning DNF rules for both true and false labels of input news. Specifically, the true label is predicted if either $\neg\mathrm{conj}_{34}$ or $\neg\mathrm{conj}_{43}$ is true, i.e., either $\neg \mathrm{P}_2 \wedge \mathrm{P}_3 \wedge \mathrm{P}_6 \wedge \mathrm{P}_8$ or $\mathrm{P}_3 \wedge \mathrm{P}_6 \wedge \mathrm{P}_8$ is false when removing the negation. Given the semantics of these logic predicates shown in Table 6, we know that P_2, P_3 and P_8 check the consistency between the background information and a given message, whereas P_6 scrutinizes improper intention from the message alone. On the other hand, the news will be predicted as false if conj_{27} is true, i.e., P_4 is false which means that the background information in the message is neither accurate or objective according to Table 6.

4.4 Generalizability Verification

Ensuring the generalization ability of fake news decision systems is vital for their sustainable and practical deployment. As observed in Table 3, TELLER consistently outperforms *Direct* across all domains and LLMs without the assistance of any generalization algorithm, while only exhibiting a negligible

LLMs	Method	CP⟶G		GP⟶C		CG⟶P	
		Acc(%)	Macro-F1(%)	Acc(%)	Macro-F1(%)	Acc(%)	Macro-F1(%)
FLAN-T5-xl	Direct	67.73	52.80	75.32	74.79	55.88	50.72
	TELLER	$68.13_{(0.40\uparrow)}$	$56.54_{(3.74\uparrow)}$	$82.40_{(7.0\uparrow)}$	$82.09_{(7.31\uparrow)}$	$61.76_{(5.88\uparrow)}$	$60.92_{(10.19\uparrow)}$
FLAN-T5-xxl	Direct	68.93	52.82	74.80	73.23	52.21	43.65
	TELLER	$69.13_{(0.2\uparrow)}$	$53.15_{(0.34\uparrow)}$	$77.44_{(2.64\uparrow)}$	$76.21_{(2.98\uparrow)}$	$66.18_{(13.97\uparrow)}$	$66.17_{(22.52\uparrow)}$
Llama2 7B	Direct	66.78	52.23	81.83	81.73	77.21	77.00
	TELLER	$68.33_{(1.55\uparrow)}$	$59.33_{(7.10\uparrow)}$	$81.60_{(-0.24\downarrow)}$	$81.04_{(-0.69\downarrow)}$	$83.09_{(5.88\uparrow)}$	$82.82_{(5.82\uparrow)}$
Llama2 13B	Direct	52.55	52.27	57.53	51.75	77.94	77.10
	TELLER	$70.93_{(18.38\uparrow)}$	$60.90_{(8.63\uparrow)}$	$85.09_{(27.56\uparrow)}$	$84.87_{(33.1\uparrow)}$	$79.41_{(1.47\uparrow)}$	$79.41_{(2.30\uparrow)}$

Table 3: Results on cross-domain experiments. **C**, **P** and **G** represent Constraint, PolitiFact, and GossipCop datasets.

$$\begin{aligned}
&\mathrm{conj}_{34} = \neg \mathrm{P}_2 \wedge \mathrm{P}_3 \wedge \mathrm{P}_6 \wedge \mathrm{P}_8 \\
&\mathrm{conj}_{43} = \mathrm{P}_3 \wedge \mathrm{P}_6 \wedge \mathrm{P}_8 \\
&\mathrm{conj}_{27} = \neg \mathrm{P}_4 \\
&\mathrm{P}_{\mathrm{true}} = \neg \mathrm{conj}_{34} \vee \neg \mathrm{conj}_{43} \\
&\mathrm{P}_{\mathrm{false}} = \mathrm{conj}_{27}
\end{aligned}$$

Table 4: Extracted rules for the GossipCop dataset when using Llama2 (13B)

performance drop in the **GP⟶C** domain using Llama2 7B. This is attributed to the remarkable zero-shot ability of LLMs and the effectiveness of the DNF layer which further compensates for biased predictions made by LLMs through rule-based aggregation. Particularly, the performance gains of TELLER in cross-domain and in-domain experiments (refer to Table 2) are positively correlated, implying that the decision system manages to learn domain-agnostic rules. Moreover, the Pearson correlation coefficient between these two groups of performance gains shows a substantial improvement from 0.01 to 0.53 when transitioning from the FLAN-T5 series to the more powerful Llama2 series. This finding suggests that leveraging stronger LLMs to drive the cognition system enhances the generalization capability of our framework.

4.5 Controllability Verification

Controllability ensures that fake news detection systems are subject to effective human oversight and intervention. We demonstrate TELLER satisfies this attribute from two aspects. Firstly, we verify the feasibility of manually rectifying rules learned by our decision system that may exhibit irrational behavior. For instance, we observe that P_3 (i.e., "The message contains adequate background information") should have a positive logical relation with P_{true} instead of negation in Table 4. To correct this, we perform a manual adjustment by setting the corresponding weight to zero, effectively removing P_3 from the logic rule. However, this modification only leads to a negligible improvement in the test split. Further investigation reveals that the truth value of logic atoms pertaining to P_3 of most real samples is negative, possibly due to the preference of LLMs. This suggests the superiority of our logic-based decision system in reducing the negative effect of incorrect predictions made by LLMs automatically. Secondly, we simulate human experts by intervening in the actions of our cognition system. We achieve this by guiding LLMs to expand the question template set $\mathcal{Q}$ using Algorithm 1, referred to as *w/ intervention* in Tables 1 and 2. The new question template set for intervention is shown in Table 7. The results consistently indicate that *w/ intervention* outperforms TELLER, highlighting the potential of LLMs as an agency for automatically regulating the behaviors of the cognition system. Thus, our framework ensures a comprehensive control mechanism by simultaneously facilitating human and AI agents' oversight.

Furthermore, we conduct additional experiments to verify the effectiveness of the DNF Layer in logic formulation over other decision systems, namely decision trees, Naive Bayes classifiers and MLP. We replace the DNF Layers with these three algorithms to derive the final decisions. The results are shown in Tables 15 and 16 for in-domain and cross-domain settings, respectively in Appendix D.

5 Conclusion

In this work, we address the limitations of existing fake news detection methods, which struggle to establish reliability and end-user trust. To tackle this issue, we identify three crucial aspects for constructing trustworthy misinformation detection systems: explainability, generalizability, and controllability. By prioritizing these principles, we propose a dual-system framework TELLER that incorporates cognition and decision systems. To validate our framework's feasibility, explainability, generalizability, and controllability, we conduct extensive experiments on diverse datasets and LLMs. These results affirm the effectiveness and trustworthiness of our approach and highlight its significant potential through evolving both subsystems in the future. While we achieve trustworthiness from an algorithmic perspective, we emphasize the importance of further research to improve the trustworthiness of the entire lifecycle of fake news detection systems.

Limitations

We identify three main limitations of our work. Firstly, although our framework focuses on enhancing the trustworthiness of fake news detection algorithms, trustworthiness is also influenced by other stages of the AI system lifecycle, such as data collection and deployment. Given the advancements in AI techniques and the importance of online information security, we encourage future research to address the challenges of building trustworthy AI systems comprehensively.

Secondly, as shown in Table 1, integrating external tools to acquire high-quality background knowledge significantly improves the performance of fake news detection systems. However, collecting information that can effectively support detection tasks using such tools is non-trivial due to the complexities of open-domain information retrieval and the diversity of news content. For instance, we search for background information by inputting check-worthy claims of P_1 into a search engine and filter out as much useful information as possible using GPT-3.5-turbo. However, integrating this evidence led to a slight performance drop on Constraint, PolitiFact, and GossipCop datasets (Due to page limitations, we do not include this experiment in our paper). Therefore, we leave this for future research.

Thirdly, despite the excellent and robust performance of our decision system, especially in generalization ability, the expressiveness of the DNF Layer is still limited due to its simple architecture. For example, the DNF Layer learns rules from data without considering the semantics of logic predicates. It may be crucial to develop more powerful decision models to fully unleash the potential of large language models, such as incorporating the semantics of logic predicates. However, given the low-dimensional input and the need for trustworthiness, the DNF layer remains a prudent choice. Moreover, there also exists a trade-off between trustworthiness and the complexity of the decision system.

Ethics Statement

This paper adheres to the ACM Code of Ethics and Professional Conduct. Specifically, the datasets we utilize do not include sensitive private information and do not pose any harm to society. Furthermore, we will release our codes following the licenses of any utilized artifacts.

Of paramount importance, our proposed dual-system framework serves as an effective measure to combat fake news and safeguard individuals, particularly in the current era dominated by large generative models that facilitate the generation of deceptive content with increasing ease. Moreover, our approach fulfills explainability, generalizability, and controllability, thereby mitigating concerns regarding the security of AI products and enabling their deployment in real-world scenarios.

Acknowledgment

This research is supported by the Ministry of Education, Singapore, under its Academic Research Fund Tier 1 (023618-00001, RG99/23), Research Matching Grant Scheme, Hong Kong Government (9229106). We thank the SAC, PC and Editors of ACL 24/ARR February to read our appealing letter and coordinate the review.

References

Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 85–90, Brussels, Belgium. Association for Computational Linguistics.

Kexin Gu Baugh, Nuri Cingillioglu, and Alessandra Russo. 2023. Neuro-symbolic rule learning in real-world classification tasks. In *Proceedings of the AAAI 2023 Spring Symposium on Challenges Requiring the Combination of Machine Learning and Knowledge Engineering (AAAI-MAKE 2023), Hyatt Regency, San Francisco Airport, California, USA, March 27-29, 2023*.

Yoshua Bengio, Geoffrey E. Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, Gillian K. Hadfield, Jeff Clune, Tegan Maharaj, Frank Hutter, Atilim Günes Baydin, Sheila A. McIlraith, Qiqi Gao, Ashwin Acharya, David Krueger, Anca D. Dragan, Philip H. S. Torr, Stuart Russell, Daniel Kahneman, Jan Brauner, and Sören Mindermann. 2023. Managing AI risks in an era of rapid progress. *CoRR*, abs/2310.17688.

Grady Booch, Francesco Fabiano, Lior Horesh, Kiran Kate, Jonathan Lenchner, Nick Linck, Andrea Loreggia, Keerthiram Murugesan, Nicholas Mattei, Francesca Rossi, and Biplav Srivastava. 2021. Thinking fast and slow in AI. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 15042–15046.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2023. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*.

João Phillipe Cardenuto, Jing Yang, Rafael Padilha, Renjie Wan, Daniel Moreira, Haoliang Li, Shiqi Wang, Fernanda A. Andaló, Sébastien Marcel, and Anderson Rocha. 2023. The age of synthetic realities: Challenges and opportunities. *CoRR*, abs/2306.11503.

Canyu Chen and Kai Shu. 2023. Can llm-generated misinformation be detected? *CoRR*, abs/2309.13788.

Jifan Chen, Aniruddh Sriram, Eunsol Choi, and Greg Durrett. 2022. Generating literal and implied subquestions to fact-check complex claims. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 3495–3516. Association for Computational Linguistics.

Kewei Cheng, Nesreen K. Ahmed, and Yizhou Sun. 2023. Neural compositional rule learning for knowledge graph reasoning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

Nuri Cingillioglu and Alessandra Russo. 2021. pix2rule: End-to-end neuro-symbolic rule learning. In *Proceedings of the 15th International Workshop on Neural-Symbolic Learning and Reasoning as part of the 1st International Joint Conference on Learning & Reasoning (IJCLR 2021), Virtual conference, October 25-27, 2021*, pages 15–56.

Ganqu Cui, Shengding Hu, Ning Ding, Longtao Huang, and Zhiyuan Liu. 2022. Prototypical verbalizer for prompt-based few-shot tuning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 7014–7024.

Limeng Cui, Kai Shu, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. defend: A system for explainable fake news detection. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, pages 2961–2964.

Kahneman Daniel. 2017. *Thinking, fast and slow*.

Richard Evans and Edward Grefenstette. 2018. Learning explanatory rules from noisy data (extended abstract). In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 5598–5602.

Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. 2018. Robust physical-world attacks on deep learning visual classification. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1625–1634.

Angela Fan, Aleksandra Piktus, Fabio Petroni, Guillaume Wenzek, Marzieh Saeidi, Andreas Vlachos, Antoine Bordes, and Sebastian Riedel. 2020. Generating fact checking briefs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural*

Language Processing, EMNLP 2020, Online, November 16-20, 2020, pages 7147–7161. Association for Computational Linguistics.

Yi R. Fung, Christopher Thomas, Revanth Gangi Reddy, Sandeep Polisetty, Heng Ji, Shih-Fu Chang, Kathleen R. McKeown, Mohit Bansal, and Avi Sil. 2021. Infosurgeon: Cross-media fine-grained information consistency checking for fake news detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 1683–1698.

Víctor Gallego. 2023. ZYN: zero-shot reward models with yes-no questions. *CoRR*, abs/2308.06385.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3816–3830.

Claire Glanois, Zhaohui Jiang, Xuening Feng, Paul Weng, Matthieu Zimmer, Dong Li, Wulong Liu, and Jianye Hao. 2022. Neuro-symbolic hierarchical rule induction. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, pages 7583–7615.

Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 2023. Bad actor, good advisor: Exploring the role of large language models in fake news detection. *CoRR*, abs/2309.12247.

Yue Huang and Lichao Sun. 2023. Harnessing the power of chatgpt in fake news: An in-depth exploration in generation, detection and explanation. *CoRR*, abs/2310.05046.

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, Fanzhi Zeng, Kwan Yee Ng, Juntao Dai, Xuehai Pan, Aidan O'Gara, Yingshan Lei, Hua Xu, Brian Tse, Jie Fu, Stephen McAleer, Yaodong Yang, Yizhou Wang, Song-Chun Zhu, Yike Guo, and Wen Gao. 2023a. AI alignment: A comprehensive survey. *CoRR*, abs/2310.19852.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023b. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12):248:1–248:38.

Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nat. Mach. Intell.*, 1(9):389–399.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *CoRR*, abs/2001.08361.

Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. 2018. All-in-one: Multi-task learning for rumour verification. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 3402–3413.

Bo Li, Peng Qi, Bo Liu, Shuai Di, Jingen Liu, Jiquan Pei, Jinfeng Yi, and Bowen Zhou. 2023. Trustworthy AI: from principles to practices. *ACM Comput. Surv.*, 55(9):177:1–177:46.

Hao Liao, Jiahao Peng, Zhanyi Huang, Wei Zhang, Guanghua Li, Kai Shu, and Xing Xie. 2023. MUSER: A multi-step evidence retrieval enhancement framework for fake news detection. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2023, Long Beach, CA, USA, August 6-10, 2023*, pages 4461–4472.

Hui Liu, Wenya Wang, and Haoliang Li. 2023. Interpretable multimodal misinformation detection with logic reasoning. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 9781–9796.

Hui Liu, Wenya Wang, Hao Sun, Anderson Rocha, and Haoliang Li. 2024. Robust domain misinformation detection via multi-modal feature alignment. *IEEE Trans. Inf. Forensics Secur.*, 19:793–806.

Yibing Liu, Haoliang Li, Yangyang Guo, Chenqi Kong, Jing Li, and Shiqi Wang. 2022. Rethinking attention-model explainability through faithfulness violation test. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pages 13807–13824. PMLR.

Jing Ma, Jun Li, Wei Gao, Yang Yang, and Kam-Fai Wong. 2023. Improving rumor detection by promoting information campaigns with transformer-based generative adversarial learning. *IEEE Trans. Knowl. Data Eng.*, 35(3):2657–2670.

Nikhil Mehta, Maria Leonor Pacheco, and Dan Goldwasser. 2022. Tackling fake news detection by continually improving social context representations using graph neural networks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1363–1380.

Ethan Mendes, Yang Chen, Wei Xu, and Alan Ritter. 2023. Human-in-the-loop evaluation for early misinformation detection: A case study of COVID-19 treatments. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 15817–15835.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.*, 267:1–38.

Yonatan Oren, Nicole Meister, Niladri S. Chatterji, Faisal Ladhak, and Tatsunori B. Hashimoto. 2023. Proving test set contamination in black box language models. *CoRR*, abs/2310.17623.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023. Fact-checking complex claims with program-guided reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 6981–7004. Association for Computational Linguistics.

Parth Patwa, Shivam Sharma, Srinivas PYKL, Vineeth Guptha, Gitanjali Kumari, Md. Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. 2021. Fighting an infodemic: COVID-19 fake news dataset. In *Combating Online Hostile Posts in Regional Languages during Emergency Situation - First International Workshop, CONSTRAINT 2021, Collocated with AAAI 2021, Virtual Event, February 8, 2021, Revised Selected Papers*, pages 21–29.

Kellin Pelrine, Anne Imouza, Camille Thibault, Meilina Reksoprodjo, Caleb Gupta, Joel Christoph, Jean-François Godbout, and Reihaneh Rabbany. 2023. Towards reliable misinformation mitigation: Generalization, uncertainty, and GPT-4. In *EMNLP*, pages 6399–6429. Association for Computational Linguistics.

Peng Qi, Juan Cao, Xirong Li, Huan Liu, Qiang Sheng, Xiaoyue Mi, Qin He, Yongbiao Lv, Chenyang Guo, and Yingchao Yu. 2021. Improving fake news detection by using an entity-enhanced framework to fuse diverse multimodal clues. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, pages 1212–1220.

Peng Qi, Zehong Yan, Wynne Hsu, and Mong Li Lee. 2024. Sniffer: Multimodal large language model for explainable out-of-context misinformation detection. *arXiv preprint arXiv:2403.03170*.

Meng Qu, Junkun Chen, Louis-Pascal A. C. Xhonneux, Yoshua Bengio, and Jian Tang. 2021. Rnnlogic: Learning logic rules for reasoning on knowledge graphs. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *CoRR*, abs/2305.18290.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Katie Sanders. 2023. PolitiFact. `https://www.politifact.com/` [Accessed: (Accessed: December 5, 2023)].

Qiang Sheng, Juan Cao, H. Russell Bernard, Kai Shu, Jintao Li, and Huan Liu. 2022. Characterizing multi-domain false news and underlying user effects on chinese weibo. *Inf. Process. Manag.*, 59(4):102959.

Kai Shu. 2023. Combating disinformation on social media and its challenges: A computational perspective. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, page 15454.

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.

Amila Silva, Ling Luo, Shanika Karunasekera, and Christopher Leckie. 2021. Embracing domain differences in fake news: Cross-domain fake news detection using multi-modal data. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 557–565.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Stephanie Jean Tsang. 2023. HKBU Fact Check. `https://factcheck.hkbu.edu.hk/home/en/fact-check/our-process/` [Accessed: (Accessed: December 5, 2023)].

Haoran Wang and Kai Shu. 2023. Explainable claim verification via knowledge-grounded reasoning with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 6288–6304.

Jindong Wang, Haoliang Li, Haohan Wang, Sinno Jialin Pan, and Xing Xie. 2023a. Trustworthy machine learning: Robustness, generalization, and interpretability. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2023, Long Beach, CA, USA, August 6-10, 2023*, pages 5827–5828.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023b. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*, pages 422–426.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He. 2022. A survey of human-in-the-loop for machine learning. *Future Gener. Comput. Syst.*, 135:364–381.

Luwei Xiao, Xingjiao Wu, Junjie Xu, Weijie Li, Cheng Jin, and Liang He. 2024. Atlantis: Aesthetic-oriented multiple granularities fusion network for joint multimodal aspect-based sentiment analysis. *Information Fusion*, page 102304.

Luwei Xiao, Xingjiao Wu, Shuwen Yang, Junjie Xu, Jie Zhou, and Liang He. 2023. Cross-modal fine-grained alignment and fusion network for multimodal aspect-based sentiment analysis. *Information Processing & Management*, 60(6):103508.

Weizhi Xu, Junfei Wu, Qiang Liu, Shu Wu, and Liang Wang. 2022. Evidence-aware fake news detection with graph neural networks. In *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, pages 2501–2510.

Qichao Ying, Xiaoxiao Hu, Yangming Zhou, Zhenxing Qian, Dan Zeng, and Shiming Ge. 2023. Bootstrapping multi-view representations for fake news detection. In *AAAI*, pages 5384–5392. AAAI Press.

Zhenrui Yue, Huimin Zeng, Yimeng Lu, Lanyu Shang, Yang Zhang, and Dong Wang. 2024. Evidence-driven retrieval augmented response generation for online misinformation. *arXiv preprint arXiv:2403.14952*.

Zhenrui Yue, Huimin Zeng, Yang Zhang, Lanyu Shang, and Dong Wang. 2023. Metaadapt: Domain adaptive few-shot misinformation detection via meta learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 5223–5239.

Xinyi Zhou and Reza Zafarani. 2021. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Comput. Surv.*, 53(5):109:1–109:40.

Yongchun Zhu, Qiang Sheng, Juan Cao, Qiong Nan, Kai Shu, Minghui Wu, Jindong Wang, and Fuzhen Zhuang. 2023. Memory-guided multi-view multi-domain fake news detection. *IEEE Trans. Knowl. Data Eng.*, 35(7):7178–7191.

A Details of Cognition System

Deep learning has attracted increasing interest from academia and industry (Xiao et al., 2023, 2024). Unlike convolutional deep learning-based fake news detection frameworks that classify in a latent space, the cognition system of TELLER, aims to emulate human fact-checking experts by complying with specific policies to ensure transparency and controllability of the detection process. In this section, we describe the construction of the set of question templates $\mathcal{Q}$ and $\mathcal{Q}'$ for TELLER and $w/Intervention$ respectively in Appendix A.1. Furthermore, we introduce a trick for batch training by fixing the number of logic atoms for different inputs in Appendix A.2 and outline some potential techniques for further improvement of the cognition system in Appendix A.3.

A.1 Construction of Question Templates

To provide an overview, we present the referenced human-checking process in Table 5. In this table, Steps I, VI and VII are excluded from detection algorithms, as they either fall into the preliminary procedures or the post-processing stages of the fake news detection pipeline. These steps may involve data crawling, human-computer interaction, machine translation, etc. As a result, we concentrate on the other steps.

Subsequently, we decompose the process into a Yes/No question template set $\mathcal{Q}$, where each template Q_i in $\mathcal{Q}$ corresponds to a predicate P_i in the predicate set $\mathcal{P}$. All question templates and their corresponding predicates are listed in Table 6. Specifically, for Q_1, our objective is to determine the trustworthiness of statements in the input news. Here, statements represent crucial information in news articles, playing a vital role in debunking misinformation. Additionally, extracting statements from news is a challenging task. While previous studies like Liao et al. (2023); Fung et al. (2021) used pre-trained language models to generate summaries as statements, we choose to utilize GPT-3.5-turbo to generate statements for simplicity in implementation. The prompt used for this purpose is as follows:

> To verify the MESSAGE, what are the critical claims related to this message we need to verify? Please use the following format to answer. If there are no important claims, answer "not applicable".
>
> MESSAGE:
> CLAIM:
> CLAIM:
>
> MESSAGE: $MESSAGE$.

Then, we replace the "$MESSAGE$" with input news and take the generated claims as statements for Q_1 (P_1).

Additionally, when verifying the controllability of our framework, we propose adjusting the question template set to deal with the diversity of fake news. While this adjustment should be done by fact-checking experts to ensure the reasonableness of new questions, our empirical findings demonstrate the feasibility of guiding large language models, such as GPT-3.5-turbo, to generate new question templates. These templates are then manually filtered by us to create the final question template set $\mathcal{Q}'$, and the corresponding predicate set $\mathcal{P}'$ for intervention, as outlined in Algorithm 1. Such human verification is incorporated into our intervention method to ensure more controllability because the main point of controllability is to intervene via human knowledge instead of relying on models entirely. Moreover, such manual checking is not time-consuming, with only a few candidate questions being generated. Table 7 presents these newly added question templates and predicates. The prompt R used in this algorithm is as follows:

> Write some questions that can be used to determine whether a news report is misinformation. The questions should be answerable by large language models in a close-book situation without requiring additional information. Please format each question using the <s> and </s> tags, such as <s>A question</s>.

A.2 Trick for Batch Training

To enable batch training, we fix the number of logic atoms, denoted as M_i for each predicate P_i. Specifically, If $M_i < \prod_{j=1}^{N_i} |\mathrm{X}_{i,j}|$, we randomly select M_i atoms. Conversely, if $M_i > \prod_{j=1}^{N_i} |\mathrm{X}_{i,j}|$, we pad the vector by 0 accordingly. In the end, $\boldsymbol{\mu}$ can be represented as $[\mu_{1,1}, \ldots, \mu_{1,M_1}, \ldots, \mu_{N,1}, \ldots, \mu_{N,M_N}]$, where $\boldsymbol{\mu} \in \mathbb{R}^M$ and $M = \sum_{i}^{N} M_i$.

A.3 The Potential of Cognition System

It is noteworthy that specific techniques can be employed to improve the performance of our cognitive system. For instance, when obtaining the answers to questions as truth values for corresponding logic atoms in Sec. 3.1.2, we exclusively consider "Yes" and "No" tokens. However, considering the relationship between model outputs and final predictions, "Right" and "Wrong" tokens can also be suitable candidates. Therefore, drawing motivation from (Gao et al., 2021; Cui et al., 2022), existing manual or automatic verbalizer techniques that establish mappings between diverse model outputs and final labels can be leveraged to enhance performance. Additionally, the ensemble of prompts, similar to "Yes or No? The answer is: ", has proven effective for the "Yes" and "No" classification task in (Gallego, 2023). Consequently, our dual-system framework exhibits substantial potential for future improvements in the cognitive system.

Algorithm 1 Question Template Generation for Intervention Algorithm

Input: Prompt R, the original question template set $\mathcal{Q}$, and a copy of $\mathcal{Q}$ denoted as $\hat{\mathcal{Q}}$
Output: The question template set $\mathcal{Q}'$ for intervention

1: Set the number of iteration steps as T
2: **for** Iteration $t = 1, \ldots, T$ **do**
3:   Use R to guide GPT-3.5-turbo in generating a set of new question templates $\mathcal{Q}'$
4:   **for** each question template Q'_i in $\mathcal{Q}'$ **do**
5:     Compute the average similarity score between Q'_i and all templates in $\hat{\mathcal{Q}}$ using Sentence BERT.
6:   **end for**
7:   Add $\mathrm{Q}'_i \in \mathcal{Q}'$ with the lowest similarity score to $\hat{\mathcal{Q}}$.
8: **end for**
9: $\mathcal{Q}' = \hat{\mathcal{Q}} \setminus \mathcal{Q}$
10: Manually refine $\mathcal{Q}'$ by removing duplicate and impractical templates that are non-verifiable through LLMs, resulting in the final $\mathcal{Q}'$.

Step I: Selecting claims (1) To filter the information on news websites, social media, and online databases through manual selection and computer-assisted selection. (2) The public can submit suspicious claims. (3) Selecting suspicious claims based on their hotness in Hong Kong, considering factors such as the amount of likes, comments, and shares the message has received. A) Is the content checkable? B) Any misleading or false content? C) Does it meet public interest? D) Is it widespread?
Step II: Tracing the source (1) Determining the source of the information. (2) Identifying the publication date. (3) Investigating the publisher and their background and reputation. (4) Checking for similar information. (5) Capturing a screen record and attaching the URL link. (6) Providing two or more additional sources of information.
Step III: Fact-checking the suspicious information (1) Applying the Five Ws and an H: When, Where, Who, What, Why, How. (2) Searching for evidence to verify the information, such as official press releases, authoritative media reports, and research reports. (3) Attempting to engage the person or organization making the claim through email or telephone, if necessary. (4) Consulting experts in the relevant field, if necessary.
Step IV: Retrieving contextual information (1) Checking if the original claim contains adequate background information. (2) Assessing the accuracy and objectivity of the background information. (3) Identifying any intentionally eliminated content that distorts the meaning.
Step V: Evaluating improper intentions (1) Assessing if there is any improper intention (e.g., political motive, commercial purpose) in the information. (2) Investigating if the publisher has a history of publishing information with improper intentions.
Step VI: Self-checking (1) Fact-checkers signing a Declaration of Interest Form before joining the team. (2) Ensuring fact-checkers maintain objectivity and avoid biases during the process. (3) Upholding the principle of objectivity and avoiding emotional involvement.
Step VII: Publishing and reviewing reports (1) Completing a draft of the fact-check report, followed by editing and reviewing by professional editors and consultants. (2) Updating the report if any mistakes or defects are found, and providing clarification on correction reasons and date.

Table 5: Fake news detection policy of HKBU FACT CHECK Team (Tsang, 2023)

Question Template	Logic Predicate: Logic Semantics	Annotation
Q_1: Background Information: $X_{1,1}$. Statement: $X_{1,2}$. Is the statement true?	$P_1(X_{1,1}, X_{1,2})$: Given the background information $X_{1,1}$, the statement is true.	$X_{1,1}$: Background information for input news, $X_{1,2}$: Check-worthy statements in input news.
Q_2: Background Information: $X_{2,1}$. Message: $X_{2,2}$. Is the message true?	$P_2(X_{2,1}, X_{2,2})$: Given the background information $X_{2,1}$, the message is true.	$X_{2,1}$: Background information for input news, $X_{2,2}$: Input news.
Q_3: Message: $X_{3,1}$. Did the message contain adequate background information?	$P_3(X_{3,1})$: The message contains adequate background information.	$X_{3,1}$: Input news.
Q_4: Message: $X_{4,1}$. Is the background information in the message accurate and objective?	$P_4(X_{4,1})$: The background information in the message is accurate and objective.	$X_{4,1}$: Input news.
Q_5: Message: $X_{5,1}$. Is there any content in the message that has been intentionally eliminated with the meaning being distorted?	$P_5(X_{5,1})$: The content in the message has been intentionally eliminated with the meaning being distorted	$X_{5,1}$: Input news.
Q_6: Message: $X_{6,1}$. Is there an improper intention (political motive, commercial purpose, etc.) in the message?	$P_6(X_{6,1})$: The message has an improper intention.	$X_{6,1}$: Input news.
Q_7: Publisher Reputation: $X_{7,1}$. Does the publisher have a history of publishing information with an improper intention?	$P_7(X_{7,1})$: Given the publisher reputation $X_{7,1}$, the publisher has a history of publishing information with an improper intention.	$X_{7,1}$: Publishing history.
Q_8: Background Information: $X_{8,1}$. Message: $X_{8,2}$. Is the message false?	$P_8(X_{8,1}, X_{8,2})$: Given the background information $X_{8,1}$, the message is false.	$X_{8,1}$: Background information for input news, $X_{8,2}$: Input news.

Table 6: Question template set $\mathcal{Q}$ and logic predicate set $\mathcal{P}$

Question Template	Logic Predicate: Logic Semantics	Annotation
Q_9: News Report: $X_{9,1}$. Is the news report based on facts or does it primarily rely on speculation or opinion?	$P_9(X_{9,1})$: The news report is based on facts and relies on speculation or opinion.	$X_{9,1}$: Input news.
Q_{10}: News Report $X_{10,1}$: Are there any logical fallacies or misleading arguments present in the news report?	$P_{10}(X_{10,1})$: The news report has logical fallacies or misleading arguments.	$X_{10,1}$: Input news.
Q_{11}: Message: $X_{11,1}$. Does the message exhibit bias?	$P_{11}(X_{11,1})$: The message exhibits bias.	$X_{11,1}$: Input news.
Q_{12}: News report: $X_{12,1}$. Are there any grammatical or spelling errors in the news report that may indicate a lack of professional editing??	$P_{12}(X_{12,1})$: The news report has grammatical and spelling errors.	$X_{12,1}$: Input news.
Q_{13}: News report: $X_{13,1}$. Does the news report use inflammatory language or make personal attacks?	$P_{13}(X_{13,1})$: The news report uses inflammatory language and makes personal attacks.	$X_{13,1}$: Input news.

Table 7: Question template set $\mathcal{Q}'$ and logic predicate set $\mathcal{P}'$ generated by GPT-3.5-turbo for intervention

B Details of Experimental Setting

B.1 Datasets

LIAR is a publicly available dataset for fake news detection, sourced from POLITIFACT.COM. This dataset comprises six fine-grained labels for truthfulness ratings: true, mostlytrue, halftrue, barelytrue, false, and pantsfire. To align with the binary classification problem, we merge true, mostlytrue into true and merge barelytrue, false, and pantsfire into false, following (Liao et al., 2023). Moreover, Wang (2017); Alhindi et al. (2018) curated relevant evidence from fact-checking experts (e.g., publisher information, background information, etc.), which serve as gold knowledge in an open setting.

Constraint is a manually annotated dataset of real and fake news related to COVID-19. We adopt the data pre-processing procedures described in (Patwa et al., 2021), which involve removing all links, non-alphanumeric characters, and English stop words.

PolitiFact and **GossipCop** are two binary classification subsets extracted from FakeNewsNet (Shu et al., 2020). The PolitiFact subset comprises political news, while the GossipCop subset comprises entertainment stories. To optimize experimental costs and adhere to maximum context limitations, we exclude news samples longer than 3,000 words.

For dataset partitioning, we follow the default partition if specified; otherwise, we use a 7:1:2 ratio. Table 8 presents the statistics of each dataset.

Split	LIAR	Constraint	PolitiFact	GossipCop
Train	10202	6299	469	6999
Validation	1284	2139	66	999
Test	1271	2119	136	2002

Table 8: Statistics of four benchmarks

B.2 Data Leakage Analysis

In our work, we used four publicly available datasets to evaluate our proposed framework, TELLER. To begin with, following recent work (Oren et al., 2023), we refer to the problem of data leakage (data contamination) as the situation where the pretraining and finetuning dataset of LLMs contains the testing splits of datasets used in our work. To mitigate the risks associated with data leakage during our evaluation, we took three precautionary steps to ascertain that the probability of the occurrence of data leakage is particularly low:

Manual Check: For the open-public Flan-T5 and Llama 2 series, we double-checked the dataset cards of these two model families and did not find a data leaking problem. Concretely, we checked the finetuning data (i.e., Appendix F Finetuning Data Card of (Chung et al., 2022)) and pre-training data (i.e., C4 dataset in Sec. 3.4.1 of (Raffel et al., 2020)) for the family of Flan-T5 models and checked the pre-training data of Llama 1 (i.e., Sec. 2.1 of (Touvron et al., 2023a)) while the pre-training data of Llama 2 seems not publicly available yet.

Assumption Experiment: If data leakage were present, we would expect the detection accuracy of LLMs to scale with model size, given that the memorization ability of LLMs is positively correlated to the size of models empirically (Kaplan et al., 2020). However, our results in Tables 1 and 2 do not support this hypothesis, suggesting a low likelihood of data leakage.

Empirical Analysis: Some measurements for data leakage exist (Oren et al., 2023; Touvron et al., 2023b). We used the Sharded Rank Comparison Test, proposed by Oren et al. (2023) to analyze potential data leakage in our datasets on Llama2 (7B). We did not analyze the data leakage problem of the GPT series here due to the limited and expensive access, while Llama2 and FLAN-T5 are LLMs we mainly use. The results in Table 9 indicate no data leakage risk for Llama2 (i.e., when the p-value> 0.05 means there is no data leakage risk). However, these measurements of data leakage problems may compromise the accuracy of determining whether dataset contamination occurs and have contributed to evaluation performance sometimes because of many confounding factors (a detailed discussion in A.6 of (Touvron et al., 2023b)).

While TELLER has shown satisfactory accuracy on four open-public datasets, our main contribution is the systematic framework that adheres to explainability, generalizability, and controllability. As per our experimental results, TELLER's detection performance can scale by integrating more powerful LLMs and external techniques, demonstrating the effectiveness of our approach as LLMs and related techniques continue to evolve. Consequently, even if the possible data leakage problem may have a deceptively good influence on the detection accuracy, we argue that it will not decrease our work's contribution.

Table 9: The Sharded Rank Comparison Test for data leakage problem. We run this test on all testing splits of four datasets for Llama2 (7B).

Dataset	P-value
LIAR	0.8355
Constraint	0.7869
PolitiFact	0.7712
GossipCop	0.7802

B.3 Illustration of Different Baselines

We compare our model against *Direct*, *Few-shot Direct*, *Zero-shot COT*, *Few-shot COT*, *Few-shot Logic*. *Direct* utilizes LLMs to calculate the probability of each label using Eqs. 1-2 and then selects the label with the highest likelihood as the predicted label. Building upon *Direct*, *Few-shot Direct* incorporates demonstration samples with known labels as contextual information to enhance the model's performance. *Zero-shot COT* and *Few-shot COT* employ the chain-of-thought (COT) technique (Wei et al., 2022), enabling LLMs to engage in step-by-step reasoning. While *Zero-shot COT* immediately adds the prompt "Let us think step by step!", *Few-shot COT* provides multiple COT exemplars. For *Few-shot Logic*, we replace the thought process in COT with instantiated questions accompanied by corresponding answers generated by our cognition system. We omit comparisons with Few-shot and COT-based prompt methods for Llama 2 and FLAN-T5 because COT prompts have been found to yield performance gains basically when used with models of approximately 100B parameters (Wei et al., 2022), and both Few-shot and COT-based methods show no additional improvement over *Direct* as revealed by Table 12, we exclusively implement COT-related methods using GPT-3.5-turbo.

Below we show the templates for these five baselines for the fake news detection task in the closed setting without access to any external knowledge source.

Direct:

> Message: $MESSAGE$.
> Is the message $Label$?
> Yes or No? Response:

Then, we replace the "$MESSAGE$" with input news, "$Label$" with candidate truthfulness labels.

Few-shot Direct:

> Following given examples to answer Yes/No questions.
>
> Message: Says the Annies List political group supports third-trimester abortions on demand.
> Is the message true?
> Yes or No? Response: No
>
> Message: Says the Annies List political group supports third-trimester abortions on demand.
> Is the message false?
> Yes or No? Response: Yes
>
> (· · · more examples here · · ·)
>
> Message: $MESSAGE$.
> Is the message $Label$?
> Yes or No? Response:

Then, we replace the "$MESSAGE$" with input news, "$Label$" with candidate truthfulness labels. Furthermore, during the testing phase, the examples are randomly selected from the training set.

Zero-shot COT:

> You will be provided with a statement, and your task is to classify its truthfulness into one of two categories: true and false.
> Message: $MESSAGE$.
> Let's think step by step and give answer with the suffix "So the final answer is".

Then, we replace the "$MESSAGE$" with the input news.

Few-shot COT:

> You will be provided with a statement, and your task is to classify its truthfulness into one of two categories: true and false.
>
> Example One
> Message: Says the Annies List political group supports third-trimester abortions on demand.
> Let's think step by step and give answer with suffix "So the final answer is".
> Annie's List was comfortable with candidates who oppose more limits on late-term abortions

while he also supported candidates who voted for more limits this year. Both dose not mention of third-trimester abortions.
So the final answer is false.

(· · · more examples here · · ·)

Message: $MESSAGE$.
Let's think step by step and give answer with the suffix "So the final answer is".

Then, we replace the "$MESSAGE$" with the input news.

Few-shot Logic:

You will be provided with a statement, and your task is to classify its truthfulness into one of two categories: true and false.

Example One
Message: Says the Annies List political group supports third-trimester abortions on demand.
Decomposed Questions:
(1) Statement: The Annies List is a political group. Is the statement true?
Yes
(2) Statement: The Annies List supports third-trimester abortions. Is the statement true?
No
(3) Did the message contain adequate background information?
False

(· · · more examples here · · ·)

Message: $MESSAGE$.
Let's think step by step and give answer with the suffix "So the final answer is".

Then, we replace the "$MESSAGE$" with the input news.

Additionally, we conducted supplementary experiments comparing our framework with other non-LLM-based misinformation detectors (referred to as small models following convention), including BERT[4] and RoBERTa[5], presented in Tables 13 and 14 for in-domain and cross-domain settings, respectively. These small models are finetuned on misinformation detection datasets. Especially for the cross-domain setting, we consider each dataset as a separate domain and fine-tune these models using the train split from source domains, choose the model on the validation split of source ones, and report results on the test split from the target domain. Moreover, we do not compare our framework here with existing transfer learning algorithms because we assume the domain label and target domain data are unavailable in our work.

B.4 Model Training for Decision System

In the decision system of our framework, we employ the DNF Layer to learn human-readable rules from data differentially. To train this model, we utilize the Adam optimizer with a learning rate of 1e-3. Regarding the hyperparameters, we search the conjunction number C within the range [10, 20, 30, 40, 50], and the weight decay within the range [1e-3, 5e-4, 1e-4]. Furthermore, to showcase the superiority of our approach, we maintain consistent hyperparameters across different LLMs in each setting. For instance, all hyperparameters of TELLER in the closed setting for the binary classification task on the LIAR dataset remain unchanged. The batch size is set to 64, and the number of epochs is set to 30. Additionally, we progressively converge the model towards symbolic semantics by adjusting δ (refer to Appendix G for detail) to 1 or -1 before the first 15 epochs using exponential decay.

[4] https://huggingface.co/bert-base-uncased
[5] https://huggingface.co/FacebookAI/roberta-base

C Details of Explainability Study

To enhance the accessibility of the rules generated by the DNF Layer, we propose a pruning algorithm that extracts more concise logic clauses by eliminating insignificant weights. The algorithm is described in Algorithm 2. Furthermore, to demonstrate the explainability of our framework, we visualize the extracted rules obtained from the pruned model for Constraint, PolitiFact, and GossipCop datasets in Tables 10, 11 and 4, respectively. In these tables, P_{true} and $\mathrm{P}_{\text{false}}$ represent the proposition that the input news is identified as true or false, respectively. In our visualization experiments, we employ Llama2 (13B) as the LLM in the cognition system. We set the number of conjunctive layers C as 50, the performance drop threshold ϵ as 0.005, and b as 0.0001 to reduce the number of conjunction clauses. More details regarding these parameters can be found in Appendix G.

Similar to Symbolic AI, such as expert systems, our learned rules can be intuitively translated into natural language. For instance, consider the rules provided in Table 4, $\mathrm{conj}_{27} = \neg \mathrm{P}_4$, $\mathrm{P}_{\text{false}} = \mathrm{conj}_{27}$ and the semantics of P_4 in Table 6 is "The background information in the message is accurate and objective", $\mathrm{P}_{\text{false}}$ can be translated as "The input message (news) is **false** when the background information in the message is **not** accurate and objective".

Algorithm 2 Pruning Algorithm for the DNF Layer

Input: Trained DNF Layer Φ, performance drop threshold ϵ
Output: Pruned DNF Layer Φ' and extracted rule set $\mathcal{R}$

```
1:  Initialize R' as an empty set
2:  Initialize R by extracting rules from Φ
3:  Initialize Φ' using Φ
4:  while |R'| ≠ |R| do
5:      Initialize R by extracting rules from Φ'
6:      Prune disjunctions if the removal of a disjunction
        results in a performance drop smaller than ε
7:      Prune unused conjunctions that are not utilized by any
        disjunction
8:      Prune conjunctions if the removal of a conjunction
        results in a performance drop smaller than ε
9:      Prune disjunctions that use empty conjunctions
10:     Prune disjunctions again if the removal of a disjunc-
        tion results in a performance drop smaller than ε
11:     Update the pruned model as Φ' and extract rules from
        Φ' to obtain R';
12: end while
```

$\mathrm{conj}_{48} = \mathrm{P}_4 \wedge \neg \mathrm{P}_8$
$\mathrm{conj}_{25} = \neg \mathrm{P}_4 \wedge \neg \mathrm{P}_5 \wedge \mathrm{P}_8$
$\mathrm{conj}_{40} = \mathrm{P}_2 \wedge \mathrm{P}_4$
$\mathrm{P}_{\text{true}} = \mathrm{conj}_{48}$
$\mathrm{P}_{\text{false}} = \mathrm{conj}_{25} \vee \neg \mathrm{conj}_{40}$

Table 10: Extracted rules for the Constraint dataset when using Llama2 (13B).

$\mathrm{conj}_{36} = \mathrm{P}_3 \wedge \mathrm{P}_6 \wedge \mathrm{P}_8$
$\mathrm{conj}_{44} = \mathrm{P}_5 \wedge \mathrm{P}_1 \wedge \mathrm{P}_8$
$\mathrm{conj}_{0} = \mathrm{P}_1$
$\mathrm{conj}_{49} = \mathrm{P}_2 \wedge \mathrm{P}_3 \wedge \mathrm{P}_4$
$\mathrm{P}_{\text{true}} = \neg \mathrm{conj}_{36} \vee \neg \mathrm{conj}_{44}$
$\mathrm{P}_{\text{false}} = \neg \mathrm{conj}_{0} \vee \neg \mathrm{conj}_{49}$

Table 11: Extracted rules for the PolitiFact dataset when using Llama2 (13B).

LLMs	Method	Constraint		PolitiFact		GossipCop	
		Acc(%)	Macro-F1(%)	Acc(%)	Macro-F1(%)	Acc(%)	Macro-F1(%)
FLAN-T5-xl	Direct	75.32	74.79	55.88	50.72	67.73	52.80
	Few-shot	75.17	74.48	52.20	45.07	67.13	51.20
	Few-shot COT	52.67	45.76	58.08	56.62	46.65	46.50
FLAN-T5-xxl	Direct	74.80	73.23	52.21	43.65	68.93	52.82
	Few-shot	75.97	75.97	50.73	41.10	68.53	51.87
	Few-shot COT	52.66	45.33	50.61	41.43	65.98	47.15
Llama2 (7B)	Direct	81.83	81.73	77.21	77.00	66.78	52.23
	Few-shot	71.68	71.30	75.74	75.74	66.13	59.62
	Few-shot COT	52.10	34.77	55.14	42.89	47.95	47.43
Llama2 (13B)	Direct	57.53	51.75	77.94	77.10	52.55	52.27
	Few-shot	57.24	50.48	80.14	79.56	51.55	51.39
	Few-shot COT	53.79	44.98	50.01	33.33	65.28	50.92

Table 12: Comparison between different prompt methods on FLAN-T5 and Llama2 series.

Method	Constraint		PolitiFact		GossipCop		LIAR	
	Acc(%)	Macro-F1(%)	Acc(%)	Macro-F1(%)	Acc(%)	Macro-F1(%)	Acc(%)	Macro-F1(%)
BERT	96.98	97.11	85.29	85.71	81.97	86.45	63.06	62.42
RoBERTa	**97.07**	**97.21**	**88.97**	**89.36**	**82.72**	**87.07**	64.55	63.16
TELLER (best)	87.78	87.71	83.82	83.82	75.92	69.30	**67.73**	**66.97**

Table 13: Comparison between small models and TELLER on four datasets for **binary classification task** in an in-domain setting.

Method	**CP⟶G**		**GP⟶C**		**CG⟶P**	
	Acc(%)	Macro-F1(%)	Acc(%)	Macro-F1(%)	Acc(%)	Macro-F1(%)
BERT	46.97	31.12	65.69	70.65	48.53	46.97
RoBERTa	47.45	38.26	64.56	65.79	52.21	48.00
TELLER (best)	**70.93**	**60.90**	**85.09**	**84.87**	**83.09**	**82.82**

Table 14: Comparison between small models and TELLER for **binary classification task** in a cross-domain setting.

D Comparison with Different Decision Models

In our work, we utilize the DNF Layer to construct our decision system, guaranteeing explainability and controllability. However, there are also other alternatives, such as existing neural symbolic architectures and interpretable machine learning algorithms. By comparing the DNF Layer with these candidates, we demonstrate that our dual-system framework can achieve better performance by inventing a more effective decision model to unleash the ability of LLMs.

While existing neural symbolic architectures can extract useful rules from data (Booch et al., 2021), they indeed have certain limitations. Firstly, these architectures often require complex mechanisms to implement logical operations, which makes them unsuitable for immediate application in fake news detection tasks. For example, Qu et al. (2021); Cheng et al. (2023) developed neural-symbolic models for knowledge graph completion, but their reliance on well-defined graph structures makes them infeasible for our task. Secondly, these architectures often suffer from efficiency issues. For instance, δLP proposed in (Evans and Grefenstette, 2018) had high computational complexity, and HRI (Glanois et al., 2022) was incompatible with batch training, which externally required users to predefine rule templates to constrain the search space. Furthermore, to the best of our knowledge, there may be no neural-symbolic framework available that can simultaneously handle the challenges of missing values and multi-grounding problems (i.e., one predicate can be instantiated as multiple logic atoms), which are common in our tasks. Therefore, we acknowledge the need for future research to develop a more suitable and powerful neural-symbolic framework in the context of fake news detection.

Since each dimension in $\boldsymbol{\mu}$ is precisely bonded to a question template (logic predicate), we can employ traditional machine learning classification algorithms, including decision tree[6], naive Bayes Classifier[7] and multi-layer perceptron (MLP), to replace the DNF Layer to drive our decision system, while maintaining partial aspects of trustworthy AI. Therefore, we compare the DNF Layer with these three methods in both in-domain and cross-domain settings on three datasets, shown in Tables 15 and 16, respectively.

According to the results, we conclude that the decision tree and MLP perform better when the training and testing data are from the same domain. Meanwhile, the naive Bayes Classifier demonstrates more satisfactory generalization performance in cross-domain experiments across various LLMs. This implies that our proposed dual-system framework shows potential in developing a more powerful decision module, such as an ensemble of these algorithms. However, the DNF Layer still outperforms these three methods in most cases when using Llama2 (13B) as the driver of the cognition system, achieving a better trade-off between accuracy and generalization ability. Moreover, the DNF Layer also exhibits advantages over these methods in terms of its ability to handle missing values and multi-grounding problems, as well as its flexibility in efficiently searching logic rules in a large space, whereas the decision tree is constrained by depth and width.

[6] https://scikit-learn.org/stable/modules/tree.html

[7] https://scikit-learn.org/stable/modules/naive_bayes.html

LLMs	Method	Constraint		PolitiFact		GossipCop	
		Acc(%)	Macro-F1(%)	Acc(%)	Macro-F1(%)	Acc(%)	Macro-F1(%)
FLAN-T5-large	Decision Tree	78.53	78.30	67.65	67.19	70.88	62.76
	Bayes Classifier	80.93	80.86	66.18	66.15	68.33	61.04
	MLP	81.26	81.16	71.42	63.43	71.62	63.74
	TELLER	80.32	80.11	67.65	67.65	69.53	59.39
FLAN-T5-xl	Decision Tree	84.29	84.27	66.91	66.10	71.13	61.58
	Bayes Classifier	82.40	82.22	68.38	67.88	68.23	60.23
	MLP	84.52	84.44	70.28	60.74	70.78	62.76
	TELLER	83.77	83.66	68.82	64.68	69.58	58.72
FLAN-T5-xxl	Decision Tree	84.14	84.12	72.06	71.00	72.13	67.08
	Bayes Classifier	82.49	82.30	68.38	67.61	68.38	57.62
	MLP	83.29	83.15	72.78	65.82	72.52	65.98
	TELLER	83.39	83.24	69.12	68.57	69.18	57.21
Llama2 (7B)	Decision Tree	84.33	84.32	79.41	77.00	72.38	65.24
	Bayes Classifier	83.11	82.97	76.47	76.29	71.98	66.67
	MLP	84.99	84.94	74.68	68.80	74.83	68.86
	TELLER	83.72	83.54	**83.82**	**83.81**	70.68	59.58
Llama2 (13B)	Decision Tree	86.50	86.49	83.09	83.07	74.43	68.99
	Bayes Classifier	84.99	84.92	80.15	80.06	73.58	69.59
	MLP	87.31	87.31	77.37	72.72	**76.97**	**72.01**
	TELLER	**87.31**	**87.29**	79.41	79.41	74.48	66.32

Table 15: Results of different decision models on Constraint, PolitiFact, and GossipCop datasets without access to retrieved background information. The best results for each dataset are highlighted with bold numbers.

LLMs	Method	**CP**$\longrightarrow$**G**		**GP**$\longrightarrow$**C**		**CG**$\longrightarrow$**P**	
		Acc(%)	Macro-F1(%)	Acc(%)	Macro-F1(%)	Acc(%)	Macro-F1(%)
FLAN-T5-xl	Decision Tree	68.98	62.33	73.67	73.32	63.97	62.71
	Bayes Classifier	67.13	59.26	82.49	82.49	64.71	64.64
	MLP	67.63	55.67	74.80	74.78	64.71	63.76
	TELLER	68.13	56.54	82.40	82.09	61.76	60.92
FLAN-T5-xxl	Decision Tree	68.33	55.53	70.60	70.35	61.03	60.98
	Bayes Classifier	68.33	54.71	82.63	82.51	62.50	62.50
	MLP	67.58	53.96	74.23	74.22	66.18	65.81
	TELLER	69.13	53.15	77.44	76.21	66.18	66.17
Llama2 7B	Decision Tree	52.20	52.05	76.40	75.02	66.91	64.84
	Bayes Classifier	65.98	62.46	82.82	82.60	67.65	65.49
	MLP	65.73	64.87	81.50	80.82	75.00	74.65
	TELLER	68.33	59.33	81.60	81.04	**83.09**	**82.82**
Llama2 13B	Decision Tree	61.59	61.14	71.54	68.21	71.32	71.32
	Bayes Classifier	**71.53**	**69.09**	82.59	82.25	78.68	78.25
	MLP	71.33	68.48	78.76	77.62	80.15	79.96
	TELLER	70.93	60.90	**85.09**	**84.87**	79.41	79.41

Table 16: Results of different decision models on cross-domain experiments. **C**, **P** and **G** represent Constraint, PolitiFact, and GossipCop datasets, respectively. The best results for each dataset are highlighted with bold numbers.

E Comparison with Existing Work on Three Principles

It is imperative to compare our LLM-based framework with prevailing misinformation detection methods across dimensions of explainability, generalizability, and controllability. We conduct additional experiments to compare with small models to demonstrate the strength of TELLER in generalizability. However, quantitatively measuring explainability and controllability in deep learning is presently challenging (Li et al., 2023), necessitating substantial research endeavors.

Generalizability: We conduct additional experiments comparing with small models (BERT and RoBERTa) in Tables 13 and 14 for in-domain (with finetuning) and cross-domain settings, respectively. These results illustrate that small models only outperform TELLER in an in-domain setting, but TELLER excels in zero-shot generalization (around 30% improvement in terms of Accuracy and F1-Score) and can handle more complex misinformation detection tasks, exemplified by superior performance on the LIAR dataset. This advantage aligns with many real scenarios, characterized by the absence of training data and the presence of sophisticated misinformation (Pelrine et al., 2023). Consequently, TELLER proves significantly advantageous in such contexts.

Moreover, the feasibility and adaptability of TELLER are underscored by the resource-intensive nature of gathering adequate data for small models. Additionally, our framework, as a general and systematic framework, can achieve better in-domain accuracy by integrating small fine-tuned models into our cognition system, treating their binary classification outputs as truth values

Explainability: Current interpretative methods using feature importance, attention visualization, and multiview learning (Cui et al., 2019; Xu et al., 2022; Liao et al., 2023; Ying et al., 2023) may be unreliable and possess limited explanatory power, as indicated by (Liu et al., 2022). Another approach (Liu et al., 2023), employing neural-symbolic learning for multimodal misinformation detection, falls short of clause length and readability caused by its unexplainable predicates. Unlike small-model-based misinformation detectors, our cognition system incorporates expert knowledge to construct a more well-grounded worldview, which is unrealistic for small models to achieve. Furthermore, another group of work (Huang and Sun, 2023; Hu et al., 2023; Yue et al., 2024; Qi et al., 2024) explored large generative language models (e.g., ChatGPT) and regarded the intermediate chain of thoughts as an explanation. Nevertheless, these explanations may not be reliable due to the hallucination phenomenon and the misalignment problem of AGI (Chen and Shu, 2023). Compared with them, our decision system can learn interpretable rules to explicitly aggregate generated logic atoms for further double-checking instead of relying on the implicit aggregation of LLMs.

Controllability: As shown in Sec. 2.2, some studies integrated human-in-loop techniques (Wu et al., 2022) for data sampling and model evaluation, whereas our framework prioritizes algorithm design. Moreover, while recent RLHF techniques (Rafailov et al., 2023) can incorporate human guidance in model behaviors based on reinforcement learning, they indeed require external high-quality fine-tuning data and sophisticated finetuning. In contrast, our framework achieves controllability through natural manipulation of the question set and logic rules in our cognition and decision systems

In summary, TELLER effectively addresses challenges in explainability, generalizability, and controllability. We also emphasize TELLER is a general framework and does not sacrifice performance for explainability, generalizability, and controllability, considering its potential to integrate fine-tuned small models to improve the in–domain performance.

F Cost Analysis

One crucial consideration of TELLER is the expense associated with the N queries to LLMs. The specific costs, including inference time and token cost, will be discussed below.

Inference Time: Due to the limited access times of GPT-3.5-turbo in minutes, it is time-consuming to perform N queries for our framework. However, it is worthwhile that it may also require multiple queries for GPT-3.5-turbo to adopt self-consistency and least-to-most prompt techniques to achieve the comparable performance as our framework, given there is a performance gap between *Direct* and TELLER in Table 1.

Furthermore, our experiments indicate that utilizing smaller LLMs, like FLAN-T5 (XL and XXL) and Llama 2 (7B and 13B), suffices for effective misinformation detection. In this case, our framework stands out from COT-based methods (Pan et al., 2023; Pelrine et al., 2023; Wang and Shu, 2023) as it eliminates the necessity of generating numerous immediate reasoning steps sequentially. Specifically, our cognition system only requires decoding the first token (i.e., "yes"/"no") to compute truth values. Since the primary bottleneck in the inference time of LLMs arises from subsequential decoding, the cost of TELLER is lower than COT-based methods. For instance, consider a COT-based model that generates 100 tokens for input news. The theoretical inference time of our framework is thus $\frac{1}{100}$ of COT-based methods, assuming parallel decoding of the first token of N questions.

Token Cost: Assuming our framework needs N queries and other LLM-based methods requires one with input query length L and output length M, c_{in} is the price of input tokens, c_{out} is the price of output tokens, and the token cost ratio between our framework and LLM-based methods is $\frac{N\times(L\times c_{in}+1\times c_{out})}{L\times c_{in}+M\times c_{out}}$. In general, c_{out} is higher than c_{in}. Then if M is significantly high when the output of other LLM-based methods contains lots of tokens such as COT, the total cost does not give much difference.

Additionally, we conduct experiments on various language models to verify the versatility of our framework, especially for FLAN-T5-large (780M) in Table 1. That is to say, our framework can build on smaller models (780M) while the size of bert-large has been 334M. While there have been more and more distillation techniques for LLMs to obtain lightweight models, many engineering efforts can be made to reduce the running cost, which is not the focus of our work. Consequently, we conclude that the cost of my framework is acceptable.

G Formal Description of DNF Layer

In this section, we introduce modified Disjunctive Normal Form (DNF) Layer employed in our framework. The DNF Layer is built from semi-symbolic layers (SL), which can progressively converge to symbolic semantics such as conjunction $\wedge$ and disjunction $\vee$.

Specifically, for the truth value vector $\boldsymbol{\mu} \in \mathbb{R}^M$ mentioned in Sec. 3.1.2, SL can be formulated as follows:

$$\mu_o = \tanh\left(\sum_j^M w_j \mu_j + \beta\right), \tag{4}$$

$$\beta = \delta\left(b - \sum_j |w_j \mu_j|\right), \tag{5}$$

where w_j represents learnable parameters, $b = \max_j |w_j \mu_j|$ and $\delta \in [-1, 1]$ represents the semantic gate selector. μ_j is the truth value for the jth logic atom obtained from the cognitive system. The sign of the learned weight w_j indicates whether μ_j (if w_j is positive) or its negation (if w_j is negative) contributes to μ_o. Thus, logical negation (e.g., $\neg p_j$) can be computed as the multiplicative inverse of the input: $-\mu_j$.

Eq. 4 resembles a standard feed-forward layer, aiming to compute a single truth value from a collection of values μ_j corresponding to different instantiations of a single predicate/question. β serves as the bias term. As shown by (Cingillioglu and Russo, 2021), by adjusting δ from 0 to 1 during training, SL tends to converge to conjunctive semantics as $\mathrm{SL}_\wedge$ (e.g., $p_1 \wedge p_2, \ldots, \wedge p_M$), indicating that if at least one $w_j \mu_j$ is false, the output μ_o will be false; otherwise, μ_o will be true. Conversely, by gradually adjusting δ from 0 to -1, SL can attain disjunctive semantics as $\mathrm{SL}_\vee$ (e.g., $p_1 \vee p_2, \ldots, \vee p_M$), where if at least one $w_j \mu_j$ is true, μ_o will be true; otherwise, μ_o will be false. Additionally, b can guarantee μ_o being true (false) when all $w_j \mu_j$ are true (false) for $\mathrm{SL}_\wedge$ ($\mathrm{SL}_\vee$).

Since each dimension in $\boldsymbol{\mu}$ corresponds to the same predicate for different inputs, SL effectively represents the relationship among different instantiations and the target output μ_o, enabling the learning of generic rules for various inputs. Moreover, by employing rule-based aggregation, our framework exhibits noise tolerance against incorrect predictions of LLMs in the cognition system, particularly owing to the $\mathrm{SL}_\vee$.

Notably, one predicate can be instantiated by multiple assignments, i.e., P_i pertains to M_i logic atoms in Appendix A.2. Thus, the parameters bound to these M_i logic atoms should naturally share the logical semantics of P_i. Instead of gathering all possible combinations of M_i logic atoms for training ($\prod_{j=1}^{M_i} j$), we let these logic atoms share the same w. In this scenario, SL can be represented as follows:

$$\mu_o = \tanh(\sum_i^N \sum_j^{M_i} w_i \mu_{i,j} + \beta), \tag{6}$$

$$\beta = \delta(b - \sum_i^N \sum_j^{M_i} |w_i \mu_{i,j}|), \tag{7}$$

where N is the number of predicates.

Model	Method	Acc	Macro-F1	F1(T)	P(T)	R(T)	F1(F)	P(F)	R(F)
FLAN-T5-xl	Direct	75.55	75.12	71.82	79.90	65.22	78.42	72.78	85.00
	TELLER	84.10	84.02	82.95	84.97	81.03	85.10	83.36	86.90
FLAN-T5-xxl	Direct	75.13	73.59	67.21	90.76	53.36	79.97	69.03	95.03
	TELLER	82.73	82.59	81.06	85.11	77.37	84.13	80.90	87.62
Llama2 (7B)	Direct	71.68	71.31	74.60	65.26	87.06	68.02	82.96	57.63
	TELLER	85.18	85.11	84.35	0.8511	83.60	85.93	85.24	86.63
Llama2 (13B)	Direct	57.24	50.48	68.78	52.80	0.9862	32.19	93.89	19.42
	TELLER	87.49	87.47	0.8687	87.09	86.66	88.06	87.86	88.26

Table 17: Results on Constraint dataset, reporting F1, Precision, Recall metrics on real and fake news, separately. T and F represent fake or real news, respectively.