Veri-R1: Toward Precise and Faithful Claim Verification via Online Reinforcement Learning

Qi He[1,2], **Cheng Qian**[1], **Xiusi Chen**[1], **Bingxiang He**[3], **Yi R. (May) Fung**[1,4], **Heng Ji**[1]
[1] University of Illinois Urbana-Champaign, [2] Fudan University
[3] Tsinghua University, [4] Hong Kong University of Science and Technology
qhe22@m.fudan.edu.cn, {xiusic,hengji}@illinois.edu

Abstract

Claim verification with large language models (LLMs) has recently attracted growing attention, due to their strong reasoning capabilities and transparent verification processes compared to traditional answer-only judgments. However, existing approaches to online claim verification, which requires iterative evidence retrieval and reasoning, still mainly rely on prompt engineering or pre-designed reasoning workflows, without unified training to improve necessary skills. Therefore, we introduce **Veri-R1**, an online reinforcement learning (RL) framework that enables an LLM to interact with a search engine and to receive reward signals that explicitly shape its planning, retrieval, and reasoning behaviors. The dynamic interaction between models and retrieval systems more accurately reflects real-world verification scenarios and fosters comprehensive verification skills. Empirical results show that Veri-R1 improves joint accuracy by up to 30% and doubles evidence score, often surpassing its larger-scale model counterparts. Ablation studies further reveal the impact of reward components, and the link between output logits and label accuracy. Our results highlight the effectiveness of online RL for precise and faithful claim verification, and provide a foundation for future research.

1 Introduction

Claim verification with LLMs has emerged as an increasingly important and complex challenge in natural language processing. With the society producing an enormous number of claims on the internet and increasing LLM generated contents, the proliferation of unverified claims on the web has accelerated at an unprecedented rate. Manual verification by humans is no longer a feasible solution given the overwhelming volume of information. Consequently, automatic and effective approaches to claim verification are becoming a more urgent need (Dmonte et al., 2024).

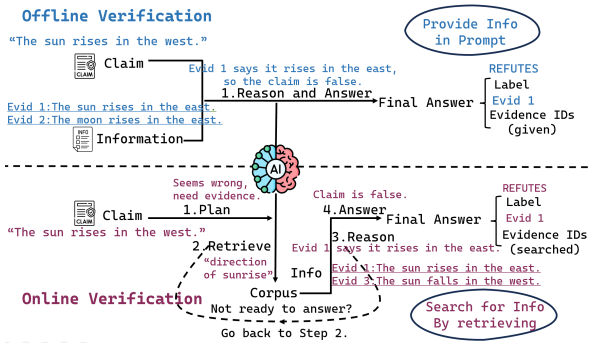


Figure 1: Conceptual comparison of **Offline Claim Verification** and **Online Claim Verification**. In the offline setting, models are provided with both the claim and relevant evidence, requiring only reasoning to produce the final answer. In the online setting, models must iteratively retrieve relevant information from a corpus before reasoning and producing the final answer.

Previous studies (Weng et al., 2022; Li et al., 2023) have primarily focused on enhancing a model's verification capability by providing both claims and corresponding evidence. This setting, referred to as Offline Claim Verification (OFFCV), as illustrated in Figure 1, requires the model to verify a given claim based solely on the provided evidence. However, in real-world scenarios, claim verification often lacks direct and concrete supporting evidence. Instead, the model must actively retrieve relevant information to assess the claim's validity. We define this process as Online Claim Verification (ONCV), where the model is given only the claim and access to a trusted corpus. The model is asked to actively acquire and utilize relevant evidence to perform verification, as shown in Figure 1. In this work, we focus on the more realistic ONCV setting and evaluate model performance in this context.

Online claim verification poses additional great challenges to the LLMs since it demands all-around capabilities, which combine information retrieval, reasoning, and judgment. Similar to a sophis-

Findings of the Association for Computational Linguistics: ACL 2026, pages 20372–20391

ticated human claim verifier, the LLM must iteratively reason over retrieved content and interact with a search engine across multiple turns before producing a final judgment. Traditional approaches (Hanselowski et al., 2018; Zhang et al., 2023; Pan et al., 2021) often adopt attribute-oriented methods (Gangi Reddy et al., 2022; Reddy et al., 2022; Li et al., 2022) to analyze claims, followed by a verifier that outputs the final decision without explanation (Soleimani et al., 2020). While LLMs have made the verification process more transparent (Lee et al., 2020; Guan et al., 2023), they still exhibit notable limitations in handling the full ONCV pipeline. Recent work (Trinh et al., 2025; Kim et al., 2024) has attempted to improve LLM performance on specific claim datasets by providing structured reasoning paths or introducing tailored analytical methods. However, these methods are typically specific to certain domains and challenges, whereas claim verification spans a wide range of domains (Augenstein et al., 2019; Wang et al., 2024) and encompasses diverse challenges such as multi-hop reasoning, entity disambiguation, numerical reasoning and more. Consequently, improving model performance in a more general claim verification setting requires addressing a central question:

How can we comprehensively enhance a model's ability to search, reason, and judge across varied claim verification scenarios and challenges?

In this work, we propose **Veri-R1**, a novel training framework designed to enhance the online claim verification capability of LLMs within a unified pipeline. Unlike common approaches such as supervised fine-tuning (SFT) (Zheng and Lee, 2025), we adopt a reinforcement learning (RL) paradigm, as it has demonstrated superior generalization ability in numerous prior studies and does not require explicit reasoning trajectories for training. Moreover, we employ online RL as our primary methodology due to its previous empirical success (Wang et al., 2025c) and its closer correspondence to real-world settings in which evidence should be retrieved and identified. During training rollouts, LLMs are required to iteratively search and reason across multiple turns before producing final answers. Since high-quality data is essential for effective training, and mislabeled samples can hinder the learning process, we filter and select samples from two high-quality datasets, FEVEROUS (Aly et al., 2021) and EX-FEVER (Ma et al., 2023), covering a wide range of claim verification challenges. To guide the training process, we design a task-specific reward function tailored to claim verification. Leveraging the ground-truth labels, the reward encourages models to learn robust judgment skills, while golden evidence is incorporated to ensure that models retrieve and identify evidence both completely and precisely. Furthermore, we conduct comprehensive comparisons among different training paradigms, including SFT and RL in both online and offline settings, and perform an ablation study to analyze the contribution of each reward component.

Empirically, online RL models from our Veri-R1 pipeline yielded the best performance in most cases compared to other training methods, yielding up to a 30% absolute gain in joint accuracy, a 23% improvement in verification accuracy, and a 22% improvement in label accuracy on evaluation data. It also enhanced the evidence-scoring metric by up to 150%. Our ablation study shows that the evidence reward effectively improves the model's ability to identify gold evidence, while the validity weight maintains training consistency by preventing reward hacking and encouraging correct label prediction with sufficient supporting evidence. Finally, we probe the relationship between model confidence and prediction correctness, finding that low confidence consistently correlates with low accuracy for SUPPORT/REFUTE labels. Moreover, larger models tend to avoid answering with NOT ENOUGH INFO, as they are more confident in their predictions.

In summary, our core contributions are as follows:

- **Veri-R1 framework**: We propose a unified RL-based pipeline tailored for Online Claim Verification. By contrasting Online RL and Offline RL against SFT, we highlight how interactive, feedback-driven learning better captures the dynamics of real-world verification tasks.
- **Data & Reward Design**: We construct a high-quality dataset from FEVEROUS and EX-FEVER and design a robust reward system that jointly emphasizes multi-level accuracy and precise evidence retrieval and identification. This alignment of supervision with verification objectives ensures that models learn to reason in ways directly tied to label and evidence precision, while maintaining faithful reasoning trajectories.
- **Confidence–Accuracy Analysis**: We conduct a

detailed investigation of model confidence, showing that low-confidence predictions frequently correspond to errors in both SUPPORT and REFUTE cases. Moreover, we find that larger models tend to exhibit overconfidence, favoring definitive judgments while underutilizing the NOT ENOUGH INFO category. These results highlight an alternative avenue for supervising model outputs, while also underscoring that the overconfidence of larger-scale models constitutes a potential drawback.

2 Related Work

2.1 LLM Empowered Claim Verification

Early investigations have primarily focused on offline claim verification, leveraging LLMs as the main reasoner and verifier. For example, pioneering work (Buchholz, 2023) demonstrated that LLMs can assess and validate factual assertions with promising accuracy. To further improve claim verification performance, several studies (Pisarevskaya and Zubiaga, 2025; Gong et al., 2025) have introduced structured reasoning prompts(Wei et al., 2022), guiding LLMs to decompose claims into sequences of analytic steps (Vladika et al., 2025; Hu et al., 2024), thereby enhancing transparency and fostering trust in their judgments. Other research has explored analytical tool integration, such as incorporating entity graphs to enforce systematic reasoning trajectories (Jeon and Lee, 2025; Huang et al., 2025).

More recent work has shifted toward online claim verification, where evidence must be retrieved by the model. Retrieval-augmented frameworks (Vykopal et al., 2025; Hagström et al., 2024) enable models to query external knowledge bases and rapidly incorporate emerging information. Given the additional challenges in the online setting, researchers have proposed new approaches, such as multi-agent systems (Hu et al., 2025b) and program-guided reasoning (Pan et al., 2023; Hu et al., 2025a), to help models adapt to diverse verification tasks.

Nevertheless, improving model performance for online claim verification across diverse challenges and domains remains an open problem. In this work, we introduce a framework designed to comprehensively enhance model verification capabilities in this complex setting.

2.2 Online Reinforcement Learning

Reinforcement learning (RL) (Kaelbling et al., 1996) has long been recognized as a promising paradigm for enabling agents to interact with their environment and learn from feedback (Watkins and Dayan, 1992; Rummery and Niranjan, 1994). Recently, RL techniques have emerged as a powerful mechanism for enhancing the inferential capabilities of large language models (LLMs) across a wide range of tasks, including mathematical problem solving (Shao et al., 2024), medical diagnosis (Lai et al., 2025; He et al., 2025), role-playing (Mou et al., 2024; Wang et al., 2025b), and dialogue generation (Chen et al., 2025).

As the complexity of interactions between LLMs and their counterparts, such as users or external tools, increases (Feng et al., 2025), there is a growing need for online RL training frameworks that support multi-turn interactions. Jin et al. (2025) proposed an RL framework enabling models to interactively search for information, demonstrating the effectiveness of online RL training. Subsequently, several studies (Mei et al., 2025; Xue et al., 2025) have sought to adapt the online RL pipeline to diverse application scenarios and improve training efficiency, for instance by refining rewards at each interaction turn (Wang et al., 2025d; Zeng et al., 2025).

However, the effectiveness of online RL remains largely unexplored in the context of claim verification, and no prior work has systematically compared online RL with offline RL (Levine et al., 2020). In this work, we implement both RL paradigms for claim verification and conduct an extensive comparison to evaluate their respective advantages and limitations.

3 Method

In the claim verification task, a model must not only determine the veracity of a claim but also provide evidence. Although supervised fine-tuning has been shown to enhance label prediction accuracy (Zheng and Lee, 2025), it typically requires high-quality reasoning trajectories and often falls short in terms of robustness and adaptability. Reinforcement learning (RL) has demonstrated substantial promise in optimizing model behavior across a wide range of tasks and scenarios. Empirical evidence (Qian et al., 2025) indicates that RL training can even surpass prevailing supervised fine-tuning in terms of robustness and adaptability. Inspired by

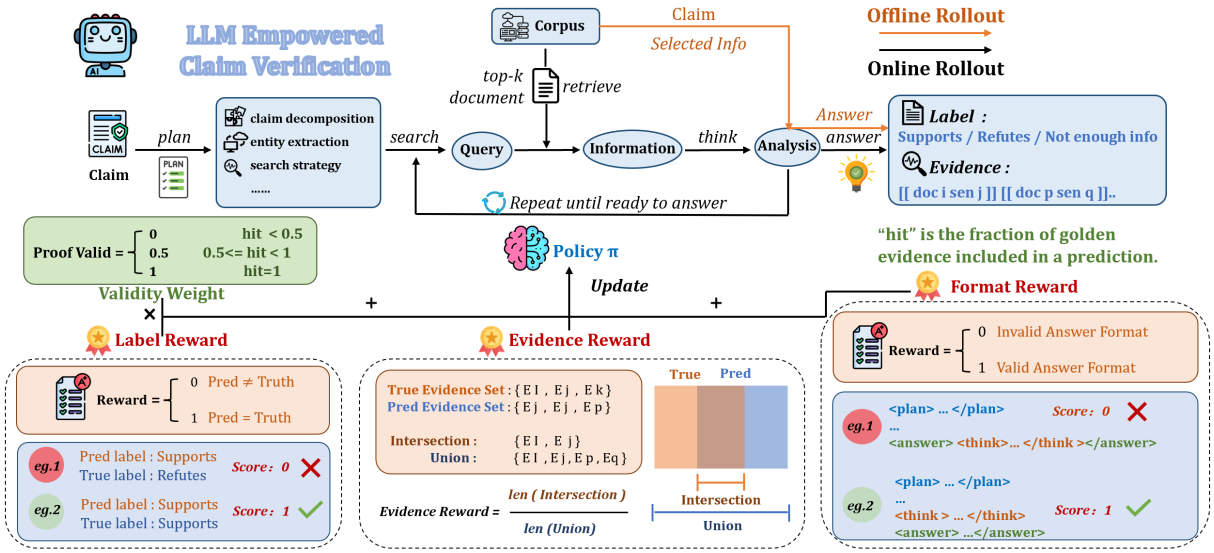


Figure 2: Comprehensive framework of Veri-R1, depicting the Online Claim Verification (ONCV) and Offline Claim Verification (OFFCV) workflows together with the calculation of label, evidence, and format rewards.

the power of RL training, we construct the **Veri-R1** training pipeline, built primarily on Online RL. Crucially, we devise a carefully calibrated reward function tailored to the claim-verification task, which ensures that the model improves in both label accuracy and evidence accuracy.

3.1 Task Definition

Online Claim Verification (ONCV) denotes an end-to-end process in which an LLM autonomously leverages external search engines to gather evidence, performs intermediate reasoning steps, and produces a final veracity label. A typical verification trajectory comprises an initial planning stage, multiple search–reasoning iterations, and a concluding judgment accompanied by selected evidence. Concretely, given a claim c, the model first generates a verification plan p. At each iteration $i \in \{1, \ldots, k\}$, it issues a search action s_i with query q_i, retrieves information i_i from the corpus, and executes an internal reasoning step t_i. Finally, the model produces an answer a, as detailed in Section G. We denote the full trajectory as

$$T = \big(p,\ (s_1, i_1, t_1),\ (s_2, i_2, t_2), \ldots, (s_k, i_k, t_k),\ a\big).$$

Under this framework, we optimize the model's policy π_θ by maximizing the expected reward

$$\max_{\theta}\ E_{T\sim\pi_\theta}\big[R(a)\big]$$

where $R(a)$, described in Section 3.3, is a scalar feedback signal computed from the final answer. In practice, we adopt *Group Relative Policy Optimization* (GRPO) (Shao et al., 2024), where rewards are normalized within each group before policy updates. The GRPO objective can be formulated as

$$L(\theta) = \mathbb{E}\Big[\min\big(r(a)\,\hat{R}_{\text{grp}}(a),\ \text{clip}(r(a), 1-\epsilon, 1+\epsilon)\,\hat{R}_{\text{grp}}(a)\big)\Big],$$

where

$$r(a) = \frac{\pi_\theta(a \mid s)}{\pi_{\theta_{\text{old}}}(a \mid s)}, \qquad \hat{R}_{\text{grp}}(a) = \frac{R(a) - \mu_g}{\sigma_g + \epsilon},$$

with μ_g, σ_g denoting the mean and standard deviation of rewards within the group, respectively.

3.2 Rollout Settings

Offline Rollout Offline rollout refers to the process in which, given an initial prompt containing a target claim, supporting or refuting evidence, and contextual sentences, a model internally reasons based on these inputs without any further interaction, and then generates its final verdict.

Online Rollout By contrast, online rollout mandates adherence to the specified algorithmic trajectory: the model may issue up to k search turns, interleaving retrieval and reasoning to support or refute the claim.

3.3 Reward Design

Rule-oriented reward strategies have demonstrated strong efficacy across a wide range of experiments and gain broad support (Icarte et al., 2022; Li et al., 2025; Wang et al., 2025a). Because an effective policy often needs to balance multiple objectives,

many studies combine several sub-reward functions using a linear weighted sum. In our framework, we define two primary components—**label reward** and **evidence reward**—to more directly steer the verification process. Additionally, we incorporate a **format reward** to enforce conformity to the desired output schema and to mitigate generation errors during reasoning. Finally, we introduce a validity weight to further steer and stabilize the training process.

Format Reward We enforce format compliance according to the following principles: (1) **Tag adherence**: all actions must strictly follow the prompt and be emitted within the prescribed tags; (2) **No extraneous tags**: the model must not introduce `<information>. . . </information>` (or any other undeclared tags) on its own. Invalid `<information>` tags are detected based on whether they immediately follow a `<search>` tag, since the system automatically identifies `<search>` tags and appends the corresponding `<information>` tags. Finally, if the verification process conforms to all of these rules, it is awarded a format reward of 1; otherwise, it receives 0:

$$R_{\text{format}} = \begin{cases} 1, & \text{if all format rules are satisfied,} \\ 0, & \text{otherwise.} \end{cases}$$

Evidence Reward Evidence is essential for validating a model's prediction and guarding against correct guesses made by chance. Accordingly, we define the evidence reward as the ratio of the intersection to the union between the predicted evidence set and the gold-standard evidence set. This reward is maximized only when the model retrieves all true evidence while avoiding irrelevant selections.

$$R_{\text{evidence}} = \frac{\left|E_{\text{pred}} \cap E_{\text{gold}}\right|}{\left|E_{\text{pred}} \cup E_{\text{gold}}\right|}, \quad R_{\text{evidence}} \in [0, 1].$$

E_{pred} denotes the set of evidence items selected by the model, and E_{gold} denotes the set of ground-truth evidence items.

Label Reward In claim verification, the label space is limited to `SUPPORT`, `REFUTE` or `NOT ENOUGH INFO`. Because label accuracy is the most critical metric for developing reliable verification models, we amplify the base label reward by giving it a higher value, thereby boosting its relative weight in the overall reward function.

$$R_{\text{label}} = \begin{cases} 2, & \hat{y} = y, \\ 0, & \text{otherwise,} \end{cases}$$

$\hat{y}$ is the model's predicted label and y is the ground-truth label.

Validity Weight There are cases in which a model arrives at the correct label via an unsound or "shortcut" reasoning path—for instance, predicting `SUPPORT` after verifying only a single subclaim. Such behavior undermines the model's capacity for genuine scrutiny and may even lead it astray. To guard against this, we introduce a validity constraint on the label reward. We define the hit rate as

$$h = \frac{|E_{\text{pred}} \cap E_{\text{gold}}|}{|E_{\text{gold}}|}.$$

A policy can earn the full label reward ($w_{\text{validity}} = 1$) for a `SUPPORT` or `REFUTE` decision only if it "hits" all of the gold evidence. To mitigate sparsity, we grant a half reward ($w_{\text{validity}} = 0.5$) whenever the hit rate exceeds 50%. We set the threshold at 50% since it is the natural midpoint of the evidence hit rate: retrieving more than half of the gold evidence indicates that the model has captured the majority of the reasoning path. For `NOT ENOUGH INFORMATION (NEI)` labels, however, we apply no validity weighting, since NEI cases often lack explicit evidence or involve only partially relevant facts.

$$w_{\text{validity}} = \begin{cases} 1, & y = \text{N} \vee \big(y \in \{\text{S}, \text{R}\} \wedge h = 1\big), \\ 0.5, & y \in \{\text{S}, \text{R}\} \wedge h > 0.5, \\ 0, & \text{otherwise.} \end{cases}$$

N, S, and R abbreviate `NEI`, `SUPPORT`, and `REFUTE`, respectively.

Final Reward The reward components are subsequently integrated to formulate the final reward for optimization:

$$R_{\text{final}} = R_{\text{label}} \cdot w_{\text{validity}} + R_{\text{evidence}} + R_{\text{format}}$$

4 Experiment

4.1 Dataset

To ensure robust and generalizable performance, we comprehensively trained and evaluated the base model on five claim verification datasets: *FEVEROUS* (Aly et al., 2021), *EX-FEVER* (Ma et al., 2023), *FEVER* (Thorne et al., 2018), *SciFACT* (Wadden et al., 2020), and *HOVER* (Jiang et al., 2020), which together present a wide spectrum of verification challenges. Recognizing the crucial role of data quality, we also implemented

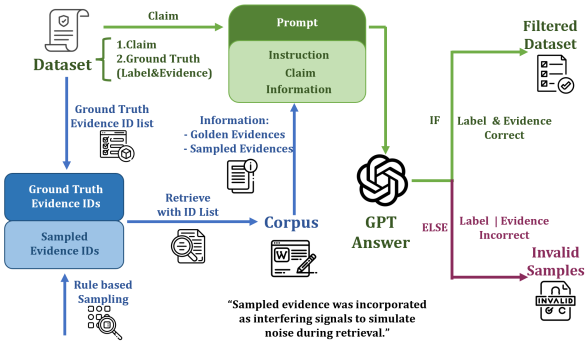


Figure 3: To mitigate annotation-related issues and ambiguities in the raw dataset, we developed a pipeline that simulates offline rollout, using GPT-4o to filter and preserve only high-quality data.

a filtering pipeline to retain only high-quality examples. Details of dataset and filtering process are described in Appendix E.

4.2 Experiment Setting

Training Our online claim verification system builds upon Verl (Sheng et al., 2025) and Search-R1 (Jin et al., 2025). During RL training, we employ the GRPO algorithm to preferentially learn from high-quality trajectories. We cap the number of search turns at three and require the model to produce its answer once this budget is exhausted.

Evaluation We evaluate our approach across five datasets: the FEVEROUS and EX-FEVER test sets, as well as selected subsets of FEVER, HOVER, and SCIFACT. Within each dataset, **the class labels are uniformly distributed** to eliminate potential biases arising from class imbalance. In addition to the evidence score (as described in our methodology), we report three distinct accuracy metrics:

- **Joint Accuracy (Joint Acc)**: the proportion of instances in which both the predicted label and the retrieved evidence are correct.
- **Verification Accuracy (Veri Acc)**: the proportion of instances in which the model predicts the correct label and retrieves all gold-standard evidence (allowing redundant evidence).
- **Label Accuracy (Label Acc)**: the proportion of instances in which the predicted class label is correct.

All evaluations are conducted in an **online setting**, where the model retrieves the necessary information from the corpus using queries. For instances classified as *NOT ENOUGH INFO (NEI)*, we relax the requirement of retrieving all evidence items, since such cases often lack sufficient evidence or contain inherently subjective evidence annotations.

4.3 Baseline

To assess the effectiveness of Online RL training, we compare our approach against several baselines, including the *Raw Instruct Model*, the *SFT Model*, the *Offline RL Model*, and the *Raw Instruct Model with Larger Scale*. Further details are provided in the Appendix H.

4.4 Result

Main Result

The comparisons between online training and baseline methods on the training dataset's test set (FEVEROUS and EX-FEVER) and the held-out datasets(FEVER, HOVER, and SciFACT) are presented in Table 1. Table 2 additionally reports a label-wise comparison of joint accuracy. We also record evidence scores of all models across the five datasets in Table 3. From these results, we draw several key observations:

Online claim verification remains a challenging task for current models. Even the state-of-the-art model, GPT-4o, achieves joint accuracies ranging from 26.64% to 59.30% across different datasets. This highlights the inherent difficulty of online claim verification, which requires iterative reasoning, information retrieval, and judgment. Furthermore, models face additional challenges stemming from multi-hop reasoning, the need for induction, and the ambiguous nature of many claims.

Online RL training substantially enhances models' verification capabilities. Models trained with Online RL achieve the highest scores across all metrics among models of the same parameter scale and, in many cases, even outperform larger-scale counterparts. This outcome aligns with our expectations, as online training mirrors the evaluation environment, enabling the model to better adapt to the system. These findings underscore the effectiveness of our Veri-R1 framework.

RL training outperforms SFT in the claim verification domain. In most cases, models trained with RL achieve higher scores than their SFT counterparts. A possible reason is that SFT relies on reasoning trajectories generated by GPT-4o, which

Model	FEVEROUS			EX-FEVER			FEVER			SciFACT			HOVER		
	Joint	**Veri**	**Label**	**Joint**	**Veri**	**Label**	**Joint**	**Veri**	**Label**	**Joint**	**Veri**	**Label**	**Joint**	**Veri**	**Label**
GPT-4o	26.64	40.70	64.96	24.71	26.48	54.75	41.11	60.44	74.33	38.26	54.43	74.96	59.30	61.90	73.60
Qwen-3B	16.89	24.15	49.55	17.62	19.90	42.71	19.00	47.22	62.78	22.78	37.27	54.43	46.30	50.50	62.30
Qwen-3B-SFT	16.33	23.81	47.39	17.62	20.28	42.71	19.89	47.56	62.00	23.77	36.85	55.56	44.30	49.10	60.90
Qwen-3B-OffRL	19.16	27.89	53.40	18.00	20.41	50.19	30.11	51.22	67.22	28.41	39.94	65.12	45.80	51.00	61.50
Qwen-7B	14.17	25.51	53.97	14.96	17.62	46.39	26.00	49.78	66.11	24.33	36.01	**65.96**	53.60	**57.10**	**65.70**
Qwen-3B-OnRL	**28.91**	**36.28**	**61.22**	**31.69**	**32.45**	**61.09**	**49.11**	**55.89**	**69.56**	**38.82**	**43.18**	63.43	**53.70**	55.10	63.70
Llama-3B	13.27	19.27	41.38	17.24	20.15	37.26	13.11	37.67	53.78	14.35	24.75	47.54	31.70	36.30	56.00
Llama-3B-SFT	12.24	18.48	43.65	15.08	18.76	37.39	16.00	41.67	56.00	15.61	26.44	50.49	32.50	38.40	57.40
Llama-3B-OffRL	16.10	21.88	44.10	17.49	20.79	41.83	17.89	37.78	57.67	18.85	24.61	51.76	35.90	40.60	57.20
Llama-8B	**19.73**	**27.32**	50.91	24.08	26.87	52.34	23.56	**53.33**	67.44	16.32	32.63	58.93	51.50	**55.60**	**67.70**
Llama-3B-OnRL	19.27	26.30	**53.17**	**28.52**	**30.29**	**59.32**	**40.11**	53.11	**68.44**	**31.65**	**44.30**	**66.53**	**52.60**	54.90	65.40

Table 1: Performance comparison across FEVEROUS, EX-FEVER, FEVER, SciFACT, and HOVER. Joint/Veri/Label denote joint, verification, and label accuracy. Abbreviated model names: Qwen-3B/7B = Qwen2.5-3B/7B-Instruct; Llama-3B/8B = Llama3.2-3B-Instruct/Llama3.1-8B-Instruct; OffRL/OnRL = offline/online RL. **Bold** and underline indicate the best and second best results within each model group.

Model	FEVEROUS			EX-FEVER			FEVER			SciFACT			HOVER	
	Sup.	**Ref.**	**NEI**	**Sup.**	**Ref.**	**NEI**	**Sup.**	**Ref.**	**NEI**	**Sup.**	**Contr.**	**NEI**	**Sup.**	**N.Sup.**
GPT-4o	32.65	4.08	43.19	16.73	13.31	44.11	44.00	38.00	41.33	16.46	18.99	79.32	50.60	68.00
Qwen-3B	10.20	0.00	40.48	6.46	4.56	41.83	5.00	1.67	50.33	2.11	2.53	63.71	27.00	65.60
Qwen-3B-SFT	7.82	0.34	40.82	8.75	3.80	40.30	7.00	3.00	49.67	1.27	2.11	67.93	24.00	64.60
Qwen-3B-OffRL	19.73	4.08	33.67	7.98	5.32	40.68	25.67	21.00	43.67	5.91	10.13	69.20	28.80	62.80
Qwen-7B	16.33	2.72	23.47	11.03	6.84	27.00	30.67	23.33	24.00	8.86	10.13	54.01	34.00	**73.20**
Qwen-3B-OnRL	**30.27**	**11.22**	**45.24**	**17.49**	**10.65**	**66.92**	**45.67**	**46.67**	**55.00**	**11.39**	**14.77**	**90.30**	**42.00**	65.40
Llama-3B	3.74	0.00	36.05	7.22	1.90	42.59	3.67	3.33	32.33	0.42	0.42	42.19	17.40	46.00
Llama-3B-SFT	5.10	0.68	30.95	7.60	1.90	35.74	4.33	4.67	**39.00**	0.42	0.42	45.99	17.80	47.20
Llama-3B-OffRL	11.22	2.04	35.03	8.75	9.13	34.60	13.33	13.33	27.00	1.27	1.27	54.01	21.40	50.40
Llama-8B	16.67	3.40	**39.12**	15.97	6.84	49.43	23.00	14.00	33.67	3.80	2.11	43.04	42.60	60.40
Llama-3B-OnRL	**21.77**	**4.08**	31.97	**20.91**	**12.55**	**52.09**	**41.33**	**44.33**	34.67	**8.02**	**12.66**	**74.26**	**43.80**	**61.40**

Table 2: Class-wise **Joint Accuracy** across five datasets. Values are percentages. Sup./Ref./Contr./N.Sup. denote support, refute, contradict, and not-supported labels. Model abbreviations follow Table 1. Within each model group, **bold** denotes the best performance and underline denotes the second best.

may lead the model to imitate the format of reasoning rather than genuinely learning the verification process. Moreover, these trajectories may differ from the reasoning logic required for other datasets, thereby limiting the model's generalizability. In contrast, RL enables the model to actively explore and learn from higher-quality trajectories, which substantially enhances its verification capability.

Other Findings

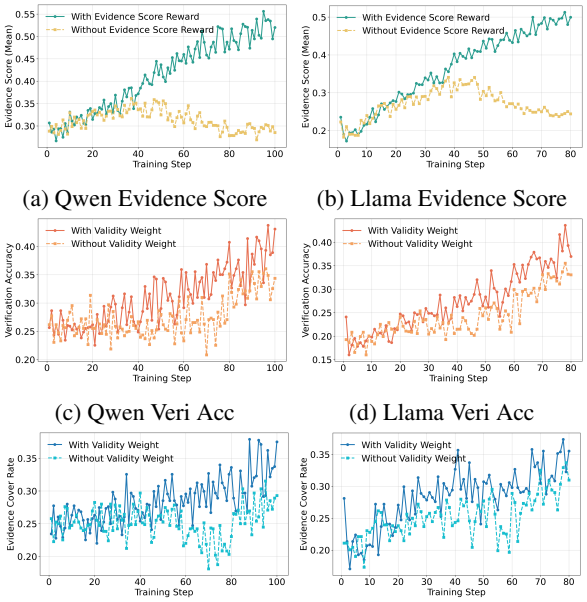
In addition to demonstrating the efficacy of online reinforcement learning, we have identified several other noteworthy findings, which are presented below.

Supervised Fine-Tuning does not guarantee improvement In many cases, SFT versions underperform their base counterparts, suggesting that pure supervised tuning can lead to overfitting on training data distributions. This overfitting may reduce the model's ability to generalize to out-of-distribution claims, especially when real-world fact verification requires handling noisier or more varied evidence than the training set. Additionally, SFT may bias the model toward producing confident but incorrect predictions if the supervision data lacks sufficient coverage of borderline or ambiguous cases.

Model size advantage is task-dependent As is shown in Table 2, larger models (e.g., LLaMA3.1-8B) tend to excel in SUPPORT/REFUTE judgments, likely because increased parameter capacity enhances their ability to absorb broad factual

Model	FEVEROUS	EX-FEVER	SciFEVER	SciFACT	HOVER
GPT-4o	0.4558	0.4185	0.4242	0.3187	0.6945
Q-3B	0.2735	0.2893	0.2327	0.1940	0.5022
Q-3B-SFT	0.2468	0.2960	0.2397	0.1972	0.4753
Q-3B-Off	0.3837	0.3361	0.3441	0.2459	0.5389
Q-7B	0.3397	0.3323	0.3383	0.1891	0.5646
Q-3B-On	**0.4769**	**0.4635**	**0.4630**	**0.2713**	**0.6562**
L-3B	0.1934	0.2444	0.1732	0.0963	0.3484
L-3B-SFT	0.1995	0.2433	0.1742	0.1004	0.3766
L-3B-Off	0.3097	0.3140	0.2601	0.0863	0.4587
L-8B	0.3763	0.3969	0.3368	0.2073	0.6209
L-3B-On	**0.4607**	**0.4717**	**0.4389**	**0.2410**	**0.6610**

Table 3: Evidence score comparison across five datasets. Q/L denote Qwen/Llama; Off/On denote offline/online RL. Within each model group, **bold** denotes the best performance and underline denotes the second best.



(a) Qwen Evidence Score (b) Llama Evidence Score

(c) Qwen Veri Acc (d) Llama Veri Acc

(e) Qwen Evid Cover Rate (f) Llama Evid Cover Rate

Figure 4: Training curves from the ablation study. Panels (a)–(b) report evidence score under the evidence score ablation, while panels (c)–(d) show verification accuracy and panels (e)–(f) show evidence cover rate under the validity weight ablation.

knowledge, retrieve relevant information, and perform multi-hop reasoning when sufficient evidence exists. In contrast, NEI detection requires a different skill set: careful uncertainty calibration, recognition of evidence gaps, and the ability to resist making unwarranted inferences. These abilities are less tied to sheer model size and more dependent on training signals that penalize overconfident answers in the absence of proof. Online RL appears to aid NEI performance by encouraging more cautious, evidence-driven predictions, as seen in cases like Qwen2.5-3B-Instruct-OnlineRL's high NEI accuracy in EX-FEVER.

REFUTE is generally the hardest label for models Across datasets, REFUTE (or CONTRADICT) accuracies are usually lower and more variable than SUPPORT or NEI. This difficulty likely stems from the need to retrieve not just relevant evidence, but specific counter-evidence that clearly contradicts the claim. Contradictions often require more precise reasoning about causal relationships, negation, or temporal mismatches, and errors in evidence retrieval or reasoning chains can easily lead the model to misclassify REFUTE as NEI or SUPPORT.

5 Analysis

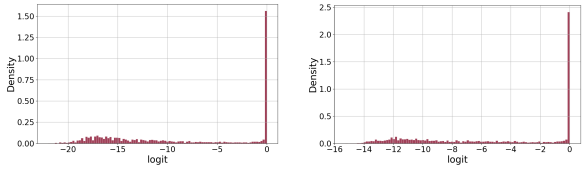
5.1 Effect of Evidence Reward

To directly incentivize precise evidence selection, we introduce an evidence-score based reward during the online reinforcement-learning phase. Figure 4a and 4b compare the evolution of this metric with and without the evidence reward on the Qwen2.5-3B-Instruct model. While the "no-reward" variant exhibits an initial uptick in evidence score over the first 40 training steps, its performance subsequently degrades. By contrast, the model trained with the evidence-score reward maintains consistent improvement, demonstrating that explicit supervision via the evidence score is essential for robust, high-precision evidence retrieval in an interactive verification setting.

5.2 Effect of Weight Validity

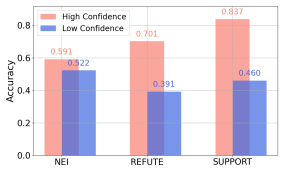
To encourage LLMs to learn from genuinely correct reasoning rather than shortcut paths that merely happen to yield the right answer, we introduce a **validity weight**. We evaluate two key metrics during training: **evidence cover rate** and **verification accuracy**. The evidence cover rate is defined as the proportion of responses in which the model covers all gold-standard evidences.

As illustrated in Figure 4c and 4d, both the Qwen and Llama models augmented with validity weight achieve higher verification accuracy throughout training. Because we incorporate an evidence-based reward into the total objective, models receive greater reward when they retrieve more gold evidences. As a result, they learn to produce an-

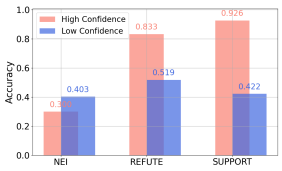


(a) Qwen Logit Dist

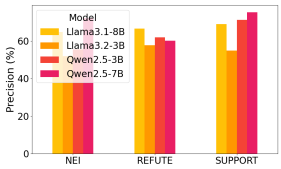
(b) Llama Logit Dist



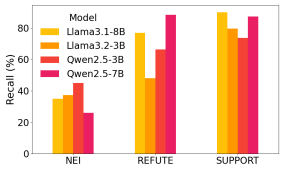
(c) Qwen Acc vs. Conf



(d) Llama Acc vs. Conf



(e) Precision Analysis



(f) Recall Analysis

Figure 5: Bar-based analysis of accuracy and logits. Panels (a)–(b) display the logit distributions of Qwen and Llama, panels (c)–(d) report accuracy versus confidence, and panels (e)–(f) present recall and precision across models and labels.

swers grounded in the gold evidence set; yet, even with this evidence reward, they still underperform the variants trained with validity weight.

We also observe notable differences in evidence cover rate in Figure 4e and 4f. For the Qwen model without validity weight, the evidence cover rate declines sharply after sixty training steps, implying that the model has discovered a shortcut to verify claims rather than performing the comprehensive verification we intended. Although this model gradually improves its ability to retrieve gold evidences, it remains inferior to the validity-weighted variant.

In conclusion, incorporating a validity weight guides the model toward reasoning trajectories that yield both correct answers and sound justifications, thereby stabilizing the training process and preventing the model from exploiting the reward function.

5.3 Relationship between Accuracy and Confidence

We conducted an experiment to quantify model confidence by extracting the output logit corresponding to each predicted label and recording these as triplets of label–logit pairs. Examination of the resulting logit distribution (Figure 5a and 5b) reveals that most values cluster near zero, suggesting that, in general, models exhibit high confidence in their predictions. We attribute this phenomenon, at least in part, to the fact that extensive intermediate reasoning processes help models resolve uncertainty before producing a final judgment.

To investigate the relationship between answer confidence and accuracy, we defined two confidence tiers: low confidence (logit < 0.85) and high confidence (logit > 0.95). As shown in Figure 5c and 5d, for SUPPORT and REFUTE labels, both Qwen and Llama models demonstrate higher confidence levels accompanied by correspondingly higher accuracy rates. In contrast, the NOT ENOUGH INFO label exhibits lower logits overall, and notably, higher confidence does not consistently translate into improved accuracy for this category—indeed, the trend reverses for Qwen and Llama in this case.

We further observe in Figure 5e and 5f that increasing model scale correlates with a decreased propensity to predict the "NOT ENOUGH INFO" category: in both architectures, the 7 and 8 billion-parameter variants exhibit lower NEI recall than their 3 billion-parameter counterparts. In particular, the notably poor NEI recall of Qwen2.5-7B-Instruct appears to result from a systematic confusion in which many genuine NEI instances are misclassified as REFUTE.

6 Conclusion

To address the critical challenge of enhancing verification capabilities across diverse scenarios, we introduce the **Veri-R1** framework, which leverages reinforcement learning to guide models in reasoning, evidence retrieval, and judgment under an online claim verification setting. Empirically, models trained with Online RL consistently outperform their counterparts of the same scale trained via SFT or Offline RL, and in many cases even surpass larger-scale models within the same series, demonstrating the effectiveness of the Veri-R1 paradigm. Our component-wise reward analysis elucidates the specific contributions of each reward signal to the training process, while logit probing reveals the relationship between output confidence and answer accuracy. We envision this work as a step toward more precise and faithful claim verification pipelines capable of handling diverse real-world challenges.

7 Limitations

Our training and evaluation are performed in an online setting with a fixed local corpus and retriever. In practice, claim verification often operates over far larger, dynamic corpora with continuously emerging information. Integrating a real-world–scale retriever into this framework would better reflect practical conditions and likely improve the reliability and robustness of both training and evaluation.

References

Rami Aly, Zhijiang Guo, Michael Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. Feverous: Fact extraction and verification over unstructured and structured information. *arXiv preprint arXiv:2106.05707*.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. Multifc: A real-world multi-domain dataset for evidence-based fact checking of claims. *arXiv preprint arXiv:1909.03242*.

Mars Gokturk Buchholz. 2023. Assessing the effectiveness of gpt-3 in detecting false political statements: A case study on the liar dataset. *arXiv preprint arXiv:2306.08190*.

Xiusi Chen, Gaotang Li, Ziqi Wang, Bowen Jin, Cheng Qian, Yu Wang, Hongru Wang, Yu Zhang, Denghui Zhang, Tong Zhang, and 1 others. 2025. Rm-r1: Reward modeling as reasoning. *arXiv preprint arXiv:2505.02387*.

Alphaeus Dmonte, Roland Oruche, Marcos Zampieri, Prasad Calyam, and Isabelle Augenstein. 2024. Claim verification in the age of large language models: A survey. *arXiv preprint arXiv:2408.14317*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Jiazhan Feng, Shijue Huang, Xingwei Qu, Ge Zhang, Yujia Qin, Baoquan Zhong, Chengquan Jiang, Jinxin Chi, and Wanjun Zhong. 2025. Retool: Reinforcement learning for strategic tool use in llms. *arXiv preprint arXiv:2504.11536*.

Revanth Gangi Reddy, Sai Chetan Chinthakindi, Zhenhailong Wang, Yi Fung, Kathryn Conger, Ahmed ELsayed, Martha Palmer, Preslav Nakov, Eduard Hovy, Kevin Small, and Heng Ji. 2022. NewsClaims: A new benchmark for claim detection from news with attribute knowledge. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6002–6018, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Haisong Gong, Jing Li, Junfei Wu, Qiang Liu, Shu Wu, and Liang Wang. 2025. Strive: Structured reasoning for self-improvement in claim verification. *arXiv preprint arXiv:2502.11959*.

Jian Guan, Jesse Dodge, David Wadden, Minlie Huang, and Hao Peng. 2023. Language models hallucinate, but may excel at fact verification. *arXiv preprint arXiv:2310.14564*.

Lovisa Hagström, Sara Vera Marjanović, Haeun Yu, Arnav Arora, Christina Lioma, Maria Maistro, Pepa Atanasova, and Isabelle Augenstein. 2024. A reality check on context utilisation for retrieval-augmented generation. *arXiv preprint arXiv:2412.17031*.

Andreas Hanselowski, Hao Zhang, Zile Li, Daniil Sorokin, Benjamin Schiller, Claudia Schulz, and Iryna Gurevych. 2018. Ukp-athene: Multi-sentence textual entailment for claim verification. *arXiv preprint arXiv:1809.01479*.

Zhitao He, Haolin Yang, Zeyu Qin, and Yi R Fung. 2025. Medtutor-r1: Socratic personalized medical teaching with multi-agent simulation. *arXiv preprint arXiv:2512.05671*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Qisheng Hu, Quanyu Long, and Wenya Wang. 2024. Decomposition dilemmas: Does claim decomposition boost or burden fact-checking performance? *arXiv preprint arXiv:2411.02400*.

Qisheng Hu, Quanyu Long, and Wenya Wang. 2025a. Boost: Bootstrapping strategy-driven reasoning programs for program-guided fact-checking. *arXiv preprint arXiv:2504.02467*.

Qisheng Hu, Quanyu Long, and Wenya Wang. 2025b. Coordinating search-informed reasoning and reasoning-guided search in claim verification. *arXiv preprint arXiv:2506.07528*.

Yani Huang, Richong Zhang, Zhijie Nie, Junfan Chen, and Xuefeng Zhang. 2025. A graph-based verification framework for fact-checking. *arXiv preprint arXiv:2503.07282*.

Rodrigo Toro Icarte, Toryn Q Klassen, Richard Valenzano, and Sheila A McIlraith. 2022. Reward machines: Exploiting reward function structure in reinforcement learning. *Journal of Artificial Intelligence Research*, 73:173–208.

Hyewon Jeon and Jay-Yoon Lee. 2025. Graphcheck: Multi-path fact-checking with entity-relationship graphs. *arXiv preprint arXiv:2502.20785*.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. Hover: A dataset for many-hop fact extraction and claim verification. *arXiv preprint arXiv:2011.03088*.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*.

Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. 1996. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285.

Kyungha Kim, Sangyun Lee, Kung-Hsiang Huang, Hou Pong Chan, Manling Li, and Heng Ji. 2024. Can llms produce faithful explanations for fact-checking? towards faithful explainable fact-checking via multi-agent debate. *arXiv preprint arXiv:2402.07401*.

Yuxiang Lai, Jike Zhong, Ming Li, Shitian Zhao, and Xiaofeng Yang. 2025. Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models. *arXiv preprint arXiv:2503.13939*.

Nayeon Lee, Belinda Z Li, Sinong Wang, Wen-tau Yih, Hao Ma, and Madian Khabsa. 2020. Language models as fact checkers? *arXiv preprint arXiv:2006.04102*.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. 2020. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*.

Manling Li, Revanth Gangi Reddy, Ziqi Wang, Yi-Shyuan Chiang, Tuan Lai, Pengfei Yu, Zixuan Zhang, and Heng Ji. 2022. Covid-19 claim radar: A structured claim extraction and tracking system. In *Proceedings of the 60th annual meeting of the association for computational linguistics: system demonstrations*, pages 135–144.

Miaoran Li, Baolin Peng, Michel Galley, Jianfeng Gao, and Zhu Zhang. 2023. Self-checker: Plug-and-play modules for fact-checking with large language models. *arXiv preprint arXiv:2305.14623*.

Xuefeng Li, Haoyang Zou, and Pengfei Liu. 2025. Torl: Scaling tool-integrated rl. *arXiv preprint arXiv:2503.23383*.

Yuxuan Liu, Hongda Sun, Wenya Guo, Xinyan Xiao, Cunli Mao, Zhengtao Yu, and Rui Yan. 2025. Bidev: Bilateral defusing verification for complex claim fact-checking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 541–549.

Huanhuan Ma, Weizhi Xu, Yifan Wei, Liuji Chen, Liang Wang, Qiang Liu, and Shu Wu. 2023. Ex-fever: A dataset for multi-hop explainable fact verification. *arXiv preprint arXiv:2310.09754*.

Jianbiao Mei, Tao Hu, Daocheng Fu, Licheng Wen, Xuemeng Yang, Rong Wu, Pinlong Cai, Xinyu Cai, Xing Gao, Yu Yang, and 1 others. 2025. O_2-searcher: A searching-based agent model for open-domain open-ended question answering. *arXiv preprint arXiv:2505.16582*.

Xinyi Mou, Xuanwen Ding, Qi He, Liang Wang, Jingcong Liang, Xinnong Zhang, Libo Sun, Jiayu Lin, Jie Zhou, Xuanjing Huang, and 1 others. 2024. From individual to society: A survey on social simulation driven by large language model-based agents. *arXiv preprint arXiv:2412.03563*.

Liangming Pan, Wenhu Chen, Wenhan Xiong, Min-Yen Kan, and William Yang Wang. 2021. Zero-shot fact verification by claim generation. *arXiv preprint arXiv:2105.14682*.

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. *arXiv preprint arXiv:2305.12744*.

Hoang Pham, Thanh-Do Nguyen, and Khac-Hoai Nam Bui. 2025. Verify-in-the-graph: Entity disambiguation enhancement for complex claim verification with interactive graph representation. *arXiv preprint arXiv:2505.22993*.

Dina Pisarevskaya and Arkaitz Zubiaga. 2025. Zero-shot and few-shot learning with instruction-following llms for claim matching in automated fact-checking. *arXiv preprint arXiv:2501.10860*.

Cheng Qian, Emre Can Acikgoz, Qi He, Hongru Wang, Xiusi Chen, Dilek Hakkani-Tür, Gokhan Tur, and Heng Ji. 2025. Toolrl: Reward is all tool learning needs. *arXiv preprint arXiv:2504.13958*.

Revanth Gangi Reddy, Sai Chetan Chinthakindi, Yi R Fung, Kevin Small, and Heng Ji. 2022. A zero-shot claim detection framework using question answering. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6927–6933.

Gavin A Rummery and Mahesan Niranjan. 1994. *Online Q-learning using connectionist systems*, volume 37. University of Cambridge, Department of Engineering Cambridge, UK.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2025. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pages 1279–1297.

Amir Soleimani, Christof Monz, and Marcel Worring. 2020. Bert for evidence retrieval and claim verification. In *European Conference on Information Retrieval*, pages 359–366. Springer.

Qwen Team. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*.

Tam Trinh, Manh Nguyen, and Truong-Son Hy. 2025. Towards robust fact-checking: A multi-agent system with advanced evidence retrieval. *arXiv preprint arXiv:2506.17878*.

Juraj Vladika, Ivana Hacajová, and Florian Matthes. 2025. Step-by-step fact verification system for medical claims with explainable reasoning. *arXiv preprint arXiv:2502.14765*.

Ivan Vykopal, Martin Hyben, Robert Moro, Michal Gregor, and Jakub Simko. 2025. A generative-ai-driven claim retrieval system capable of detecting and retrieving claims from social media platforms in multiple languages. *arXiv preprint arXiv:2504.20668*.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. *arXiv preprint arXiv:2004.14974*.

Haoran Wang, Aman Rangapur, Xiongxiao Xu, Yueqing Liang, Haroon Gharwi, Carl Yang, and Kai Shu. 2024. Piecing it all together: Verifying multi-hop multimodal claims. *arXiv preprint arXiv:2411.09547*.

Haoran Wang and Kai Shu. 2023. Explainable claim verification via knowledge-grounded reasoning with large language models. *arXiv preprint arXiv:2310.05253*.

Hongru Wang, Cheng Qian, Wanjun Zhong, Xiusi Chen, Jiahao Qiu, Shijue Huang, Bowen Jin, Mengdi Wang, Kam-Fai Wong, and Heng Ji. 2025a. Otc: Optimal tool calls via reinforcement learning. *arXiv e-prints*, pages arXiv–2504.

Yumeng Wang, Zhiyuan Fan, Jiayu Liu, Jen-tse Huang, and Yi R Fung. 2025b. Diversity-enhanced reasoning for subjective questions. *arXiv preprint arXiv:2507.20187*.

Zihan Wang, Kangrui Wang, Qineng Wang, Pingyue Zhang, Linjie Li, Zhengyuan Yang, Xing Jin, Kefan Yu, Minh Nhat Nguyen, Licheng Liu, and 1 others. 2025c. Ragen: Understanding self-evolution in llm agents via multi-turn reinforcement learning. *arXiv preprint arXiv:2504.20073*.

Ziliang Wang, Xuhui Zheng, Kang An, Cijun Ouyang, Jialu Cai, Yuhang Wang, and Yichao Wu. 2025d. Stepsearch: Igniting llms search ability via step-wise proximal policy optimization. *arXiv preprint arXiv:2505.15107*.

Christopher JCH Watkins and Peter Dayan. 1992. Q-learning. *Machine learning*, 8(3):279–292.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Shengping Liu, Bin Sun, Kang Liu, and Jun Zhao. 2022. Large language models are better reasoners with self-verification. *arXiv preprint arXiv:2212.09561*.

Zhenghai Xue, Longtao Zheng, Qian Liu, Yingru Li, Zejun Ma, and Bo An. 2025. Simpletir: End-to-end reinforcement learning for multi-turn tool-integrated reasoning. https://simpletir.notion.site/report. Notion Blog.

Siliang Zeng, Quan Wei, William Brown, Oana Frunza, Yuriy Nevmyvaka, and Mingyi Hong. 2025. Reinforcing multi-turn reasoning in llm agents via turn-level credit assignment. *arXiv preprint arXiv:2505.11821*.

Hengran Zhang, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. 2023. From relevance to utility: Evidence retrieval with feedback for fact verification. *arXiv preprint arXiv:2310.11675*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*.

Zhi Zheng and Wee Sun Lee. 2025. Reasoning-cv: Fine-tuning powerful reasoning llms for knowledge-assisted claim verification. *arXiv preprint arXiv:2505.12348*.

A Logit Distribution for Each Label

For the **SUPPORT** label, the answer logits produced by both models are primarily concentrated near 1, indicating high model confidence. Additionally, there is a secondary, low-density concentration in the lower logit regions — approximately between -20 and -10 for Qwen, and between -14 and -10 for LLaMA. As previously analyzed, answers falling within these low logit intervals are more likely to be incorrect.

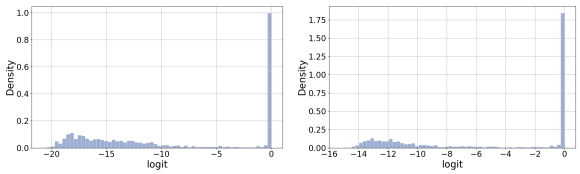


Figure 6: Logit Distribution for SUPPORT

For the **REFUTE** label, the logits are similarly centered near 1, reflecting high confidence for correct predictions. However, unlike the SUPPORT label, there is no notable secondary peak; instead, the remaining logits are relatively evenly distributed across the entire range, suggesting a more uniform uncertainty profile.

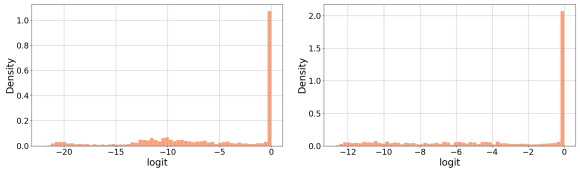


Figure 7: Logit Distribution for REFUTE

In contrast, for the **NOT ENOUGH INFO** label, the models exhibit substantially lower confidence. This is evident from the noticeably reduced density near logit = 1. Furthermore, the logits are more widely spread across the range, with elevated densities in non-peak regions compared to the SUPPORT and REFUTE labels. This indicates a broader distribution of model uncertainty when handling instances labeled as NOT ENOUGH INFO.

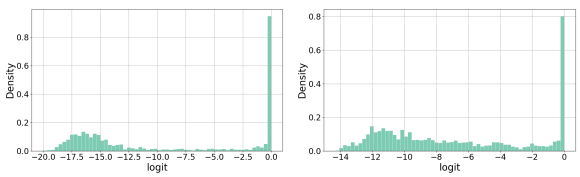


Figure 8: Logit Distribution for NEI

B Training Configuration

All experiments were conducted on a compute node with two NVIDIA A800-80G GPUs. Running Online RL training on this hardware takes approximately 15 hours to complete 100 steps. In addition, we group every three sentences into one entry, and in each turn we retrieve the most related three entries. For the maximum number of turns, it includes three search turns and one answer turn.

Setting	
Train Batch Size	256
Validation Batch Size	256
Max Prompt Length	4864
Max Response Length	512
Max Start Length	512
Max Observation Length	768
Actor Rollout	
Learning Rate	1e-6
LR Warmup Ratio	0.285
Use KL Loss	true
PPO Mini Batch Size	64
PPO Micro Batch Size	16
KL Loss Coefficient	0.001
KL Loss Type	low_var_kl
Actor Rollout Reference — Rollout	
Log-Prob Micro Batch Size	64
Tensor Model Parallel Size	1
Rollout Engine Name	vllm
GPU Memory Utilization	0.6
Number of Agents	3
Temperature	0.8
Actor Rollout Reference — Reference Model	
Ref Log-Prob Micro Batch Size	64
Ref FSDP Param Offload	True
Trainer Setting	
GPUs per Node	2
Save Frequency (epochs)	5
Test Frequency (epochs)	5
Total Epochs	10
Other	
Max Turns	4
Retriever Top-K	3

Table 4: Configuration For Training Process

C Comparison with current works

Comparison Results We further compare our models against classical and state-of-the-art baselines in Table 5. While BiDeV shows slight improvements on two-label datasets (e.g., HOVER), it fails to generalize to three-label claim verification. Across all three-label datasets, our Qwen 3B model consistently outperforms these baselines, highlighting the effectiveness of our proposed method.

Limitations in Baselines Current approaches such as ProgramFC (Pan et al., 2021) and BiDeV (Liu et al., 2025) are designed for binary classification (SUPPORT/REFUTE) and do not include the *Not Enough Info* label, which accounts for a substantial portion of real-world claims and is therefore indispensable in realistic claim verification scenarios. Moreover, our task formulation requires models not only to predict verdicts but also to identify the corresponding evidence IDs that support their decisions, which poses an additional and non-trivial challenge. As a result, prior works (Wang and Shu, 2023; Pham et al., 2025) exhibit limited generalization and practical credibility under this more realistic setting compared to our proposed training approach.

D Data Quality

Data quality plays a crucial role in maintaining a stable training process and ensuring the reliability of evaluation outcomes. During our data preparation stage, we identified several types of issues and ambiguities within the raw datasets. To illustrate the potential limitations inherent in the current data, we selected representative examples of these problems. Highlighting such issues not only reveals challenges in the existing datasets but also provides guidance for improving data creation practices in future work.

Table 6 illustrates three common data quality issues in claim verification: (1) **Overgeneralization** where the claim makes an overly absolute statement not fully supported by the evidence; (2) **Entity mismatch** where the subject in the claim is different from the evidence; and (3) **Incomplete context** where critical linking information is missing. Addressing these issues improves label accuracy and dataset reliability.

E Details of Dataset Processing

Training Dataset

FEVEROUS (Aly et al., 2021) is a comprehensive verification dataset presenting diverse challenges—such as entity disambiguation, numerical reasoning, and more. To counteract the imbalance in both task difficulty and label distribution, we include every sample from underrepresented challenge categories and uniformly subsample 3,000 instances for each label before adding them to our training pool. Besides, we evenly sampled data for each label from the development set as evaluation set. FEVEROUS dataset is under a CC BY-SA 3.0 license.

EX-FEVER (Ma et al., 2023) is designed to evaluate the explainability of the verification process, comprising thousands of multi-hop claims. In our experiments, we utilize only the gold labels and evidence (omitting the explanatory annotations), randomly selecting 3,000 examples per label for inclusion in the training pool. Besides, we evenly sampled data each label from the development set as evaluation set. All the data of EX-FEVER comes from the development set. EX-FEVER dataset is under a MIT license.

Final Training Dataset From the aggregated pool of FEVEROUS and EX-FEVER, we first apply our filtering pipeline to remove low-quality instances, then resample from the filtered data to assemble the final training set.

Held-out Dataset

FEVER (Thorne et al., 2018) is a canonical claim-verification benchmark composed largely of concise, single-sentence claims. We randomly select 900 data samples from the original FEVER dataset, comprising 300 instances for each label. As the gold-standard evidence in FEVER is annotated at the sentence level, we preprocess the source documents by splitting them into individual sentences and grouping every three consecutive sentences to construct compact retrieval units. FEVER dataset is under a CC BY-SA 3.0 license.

SciFACT (Wadden et al., 2020) is a scientific-domain fact-checking corpus containing expert-annotated claims from research papers. Due to the limited size of the SciFACT dataset, we select as many samples as available for the label with the fewest instances. To ensure label consistency

Model	FEVEROUS	EX-FEVER	FEVER	SciFACT	HOVER	Backbone
GPT-4o	**64.96**	54.75	**74.33**	**74.96**	73.60	GPT-4o
ProgramFC (Pan et al., 2023)	60.77	50.67	60.67	54.01	73.20	GPT-3.5-turbo
BiDeV (Liu et al., 2025)	58.73	48.04	58.78	54.71	**74.10**	GPT-3.5-turbo
Qwen-3B-OnRL (ours)	61.22	**61.09**	69.56	63.43	63.70	Qwen2.5-3B
Llama-3B-OnRL (ours)	53.17	59.32	68.44	66.53	65.40	Llama3.2-3B

Table 5: Label accuracy across five benchmarks. Values are percentages. Model abbreviations follow Table 1. **Bold** denotes the best performance and underline denotes the second best. Baseline methods are shown for reference, while our OnlineRL-based models demonstrate competitive performance across datasets.

Claim 1	It is illegal in Illinois to record a conversation.
Evidence 1	[[Illinois law]]: Illinois law prohibits recording private conversations without the consent of all parties.
Evidence 2	[[2014 Revision]]: In 2014, the Illinois Supreme Court struck down the old statute, and the revised law now bans only secret recordings of private conversations.
Label	~~REFUTE~~ NOT ENOUGH INFO
Note	**Overgeneralization:** The claim is too absolute. The evidence shows that the law only makes it illegal under specific conditions (i.e. for private conversations, lacking consent). It is not correct to say all conversations in Illinois may not be recorded.
Claim 2	Honda Racing is an annual event organized by Honda to publicize ongoing projects.
Evidence 1	[[Honda Winner_sentence_0]]: The Honda Winner is an underbone motorcycle from the Japanese manufacturer Honda.
Label	~~REFUTE~~ NOT ENOUGH INFO
Note	**Entity mismatch:** Evidence only describes Honda Winner as a motorcycle, not an event, leaving the claim unverified.
Claim 3	Jennifer Garner is an American actress raised in a city that is located at the confluence of the Elk and Kanawha rivers.
Evidence 1	[[Jennifer Garner]]: Jennifer Anne Garner (born April 17, 1972) is an American actress. Her breakthrough film debut was in the comedy *Dude, Where's My Car* (2000). Following a supporting role in *Pearl Harbor* (2001), Garner gained recognition for her performance as CIA officer Sydney Bristow in the ABC spy-action thriller *Alias*, which aired from 2001 to 2006. For her work on the series, she won a Golden Globe Award and a SAG Award and received four Emmy Award nominations.
Evidence 2	[[Charleston, West Virginia]]: Charleston is the capital and the largest city in the U.S. state of West Virginia, and the county seat of Kanawha County. It is located at the confluence of the Elk and Kanawha Rivers in Kanawha County.
Label	~~SUPPORT~~ NOT ENOUGH INFO
Note	**Incomplete context:** While the city's location is confirmed, there is no direct evidence that Garner was raised there.

Table 6: Examples of Data Quality Issues in Claim Verification

across datasets, we map the label CONTRADICT to REFUTE, aligning it with the labeling scheme used in FEVER. The document preprocessing procedure follows the same approach as in FEVER. SciFACT dataset is under Apache License 2.0.

HOVER (Jiang et al., 2020) defines a more challenging multi-hop verification task, demanding reasoning across 2 to 4 hops over Wikipedia entities. As a multi-hop fact verification dataset, HOVER contains a considerable proportion of ambiguous claims. To address this, we employ GPT-4o to filter and retain samples with clearer, more objective judgments. Additionally, given that multi-hop claims often require extensive evidence, which poses challenges for models in accurately retrieving all supporting sentences, we limit our evaluation to 2-hop claims. The evaluation set is curated to ensure a balanced distribution between the SUPPORTED and NOT SUPPORTED labels (448 samples each label). HOVER dataset is released under a CC BY-SA 4.0 license.

We do not directly use the test sets of each dataset for evaluation for the following reasons. First, the full test sets of these datasets are either inaccessible or incomplete, often containing only claims without corresponding ground-truth labels, and can only be evaluated through online submission systems. This makes it impossible to compute our evaluation metrics, including joint accuracy and verification accuracy. Second, different datasets adopt their own benchmark-specific evaluation protocols, and the resulting metrics are not aligned, preventing consistent and fair comparison across datasets. To ensure a unified and reproducible evaluation standard, we therefore conduct all evaluations on the development sets. Third, for transparency and reproducibility, we release all training and evaluation data along with our complete codebase and implementation details.

Data Filtering

Given the ambiguous nature of claims, there is often some ambiguity for a claim. We also find some toxic samples in the dataset may disturb a smooth training process. Considering the training effectiveness and evaluation credibility, we decided to filter the dataset with assistance of GPT-4o. We simulate the process of offline rollout shown in Figure 3, where we provide the model with claims and sufficient information. As is depicted in the figure, we only select the samples where GPT-4o is totally correct, which means it predicts the right label and evidence with no deviation. In general, approximately 70% of the samples are retained during the filtering process.

F Prompt

The system prompt for online claim verification and offline claim verification is respectively illustrated in Figure 9 and 10. Similar to the instruction in Figure 9 for online setting, prompt in Figure 10 is designed to leverage reinforcement learning (RL) for training LLMs in an offline setting. In this framework, the models are required solely to perform reasoning and provide answers.

G Rollout Details

Token Delimiters To enforce procedural rigor and facilitate answer parsing, each component of the workflow must be delimited by designated tokens:

`<plan>...</plan>`: Strategic planning

`<search>...</search>`: Retrieval queries

`<think>...</think>`: Reasoning Process

`<answer>...</answer>`: Final judgment

Answer Format The model's output must consist solely of two elements within the pre-defined tag:

`<answer>`

Label: one of `SUPPORT`, `REFUTE`, or `NOT ENOUGH INFO`

Evidence: a list of evidence id formatted as $[[\texttt{evid\_id}_1], [\texttt{evid\_id}_2]], \ldots$

`</answer>`

We adopt a strict answer format to simplify parsing, using `[[]]` to delimit each evidence ID and guarantee accurate extraction.

H Baseline

To evaluate the effectiveness of online RL training, we compare our approach against the following baselines:

1. **Raw Instruct Model**: the original instruction-tuned model without any additional training. For baselines, we adopt the widely used Qwen2.5-3B-Instruct(Team, 2024) and Llama3.2-3B-Instruct models(Dubey et al., 2024).

2. **SFT Model**: the instruct model supervised fine-tuned(SFT) on simulated reasoning paths generated by GPT-4o for samples in the training set. We prompt GPT-4o using the template in Figure 9 to guide it in producing high-quality verification reasoning for each claim. We perform LoRA adaptation(Hu et al., 2022) using the llama-factory toolkit(Zheng et al., 2024).

3. **Offline RL Model**: the instruct model trained using offline rollouts based on the prompt described in the appendix. The training follows the same reward function outlined in Section 3.3. We select the checkpoint with the highest validation accuracy for evaluation.

4. **Raw Instruct Model with Larger Scale**: we use the 7B variant of the Qwen Instruct model and the 8B variant of the Llama Instruct model to assess the effect of scale.

I Additional Results

To comprehensively evaluate models, we also report label-wise verification accuracy in Table 7, and label-wise label accuracy in Table 8. Online RL-trained models achieve the highest verification accuracy compared to other models of the same scale, and their performance is comparable to that of 7B and 8B counterparts. In terms of label accuracy, although Online RL does not exhibit overwhelmingly superior results, it consistently outperforms the baseline instruct models. Offline RL demonstrates particular effectiveness in improving label accuracy for the Qwen model. However, label accuracy may not always provide a fully reliable measure, as models can achieve high scores in specific labels while relying on incorrect reasoning, reflecting inherent biases or preferences.

J Case Analysis

In comparing the two RL trained models' outputs(Figure 11), two key dimensions emerge: **decomposition granularity** and **retrieval–evidence utilization**.

Decomposition Granularity

Offline RL - Coarse (word-level) decomposition The first model extracts isolated keywords — such as "Olympic medal", "Olympic Games", "first place", and "awarding criteria" — and issues a single undifferentiated query. This strategy overlooks the claim's internal logical structure and impedes precise verification of individual subclaims.

Online RL - Fine-grained (subclaim-level) decomposition By contrast, the second model partitions the overall assertion into two subclaims:

1. "An Olympic medal is awarded to successful competitors at one of the Olympic Games."

2. "First place receives a medal awarded for the highest achievement in a non-military field."

Separate and tailored searches are conducted for each subclaim, ensuring that each component of the original claim undergoes independent validation.

Retrieval and Evidence Utilization

Offline RL - Redundant retrieval and unfocused evidence The first model repeatedly issues the same generic query (e.g., What are the criteria for awarding Olympic medals ?), obtains relevant or background information, and ultimately produces a "NOT ENOUGH INFO" label with irrelevant evidence.

Online RL - Targeted retrieval and incremental confirmation The second model formulates distinct queries for each subclaim and immediately integrates the retrieved information into its reasoning. After confirming Subclaim 1, it proceeds to Subclaim 2, thereby constructing a coherent chain of thought. This workflow culminates in the correct "SUPPORT" judgment supported by precisely relevant evidence (e.g., the Olympic medal awarding criteria and the definition of a gold medal).

Therefore, a subclaim-based decomposition strategy, combined with targeted retrieval and stepwise evidence confirmation, substantially enhances the precision and reliability of fact-verification performance.

Claim Verification Assistant Prompt

You are a claim-verification assistant. You MUST follow this protocol exactly:

<plan>. . . </plan>
– Once at the start: sketch your high-level strategy, such as claim decomposition, entity recognition, etc.

<search>. . . </search>
– When you need a fact: emit exactly this tag with your query.
– To make the most of your search turns, don't repeat identical queries.
– You can search at most three times.

<information>
[[e_1]]: info1
[[e_2]]: info2
. . .
</information>
– You will be given claim related information in the format above.

<think>. . . </think>
– Use for every piece of reasoning; do not state your final verdict here.
– You must conduct reasoning inside <think> and </think> first every time you get new information.

<answer>
Label: SUPPORT / REFUTE / NOT ENOUGH INFO
Evidence: [[e_1]], [[e_3]], . . .
</answer>
– Emit exactly once at the end, no extra text or tags.
– Evidence ids such as e_1 will be replaced by real ids from the corpus. Include only those ids in your evidence list.
- Evidence outputs must strictly enforce the format [[e_i]], [[e_j]]. . .
- Answer Labels respectively stand for:
SUPPORT: The claim is consistent with the cited evidence and the evidence is sufficient to confirm the claim.
REFUTE: The claim contradicts the cited evidence and the evidence is sufficient to disprove the claim.
NOT ENOUGH INFO: The available evidence is insufficient to determine whether the claim is true or false.

– Process: plan → (search → information → think) repeat until conclusion → answer

Verify the claim: {claim}

Figure 9: System Prompt for Online Claim Verification.

Model	FEVEROUS			EX-FEVER			FEVER			SciFACT			HOVER	
	Sup.	Ref.	NEI	Sup.	Ref.	NEI	Sup.	Ref.	NEI	Sup.	Contr.	NEI	Sup.	N.Sup.
GPT-4o	40.13	38.77	43.19	19.77	15.59	44.11	70.67	69.33	41.33	40.93	43.04	79.32	55.80	68.00
Qwen-3B	18.37	13.61	40.48	12.55	5.32	41.83	48.33	43.00	50.33	26.58	2.53	63.71	35.40	65.60
Qwen-3B-SFT	17.01	13.61	40.82	15.97	4.56	40.30	52.33	40.67	49.67	22.78	2.11	67.93	33.60	64.60
Qwen-3B-OffRL	26.19	23.81	33.67	13.69	6.84	40.68	57.33	52.67	43.67	**28.69**	10.13	69.20	39.20	62.80
Qwen-7B	25.17	27.89	23.47	15.21	**10.65**	27.00	**61.67**	**63.67**	24.00	**28.69**	10.13	54.01	41.00	**73.20**
Qwen-3B-OnRL	**33.67**	**29.93**	**45.24**	**19.77**	**10.65**	**66.92**	58.00	54.67	**55.00**	19.41	**14.77**	**90.30**	**44.80**	65.40
Llama-3B	17.35	4.42	36.05	14.83	3.04	42.59	48.00	32.67	32.33	25.32	0.42	42.19	26.60	46.00
Llama-3B-SFT	17.35	7.14	30.95	16.73	3.80	35.74	49.67	36.33	**39.00**	26.58	0.42	45.99	29.60	47.20
Llama-3B-OffRL	14.63	15.99	35.03	16.35	11.41	34.60	43.33	43.00	27.00	12.66	1.27	54.01	30.80	50.40
Llama-8B	23.47	19.39	**39.12**	22.43	8.75	49.43	66.67	**59.00**	33.67	**37.13**	2.11	43.04	**50.80**	60.40
Llama-3B-OnRL	**26.53**	**20.41**	31.97	**25.10**	**13.69**	**52.09**	**67.33**	58.00	34.67	30.80	**12.66**	**74.26**	48.40	**61.40**

Table 7: Class-wise **Verification Accuracy** across five datasets. Values are percentages. Sup./Ref./Contr./N.Sup. denote support, refute, contradict, and not-supported labels. Model abbreviations follow Table 1. Within each model group, **bold** denotes the best performance and underline denotes the second best.

Claim Verification Assistant Prompt

You are a claim-verification assistant. You MUST follow this protocol exactly:

<information>
[[e_1]]: info1
[[e_2]]: info2
...
</information>
- You will be given claim related information above.

<think>. . .</think>
- Use for every piece of reasoning.
- During reasoning, you must verify the claim step by step based on the given information.

<answer>
Label: SUPPORT / REFUTE / NOT ENOUGH INFO
Evidence: [[e_1]], [[e_3]], ...
</answer>
- Emit exactly once at the end, no extra text or tags.
- Evidence id such as e_1 will be replaced by real ids from the corpus. You must include useful real ids when answering
- Evidence outputs must strictly enforce the format [[e_i]], [[e_j]]...
- Answer Labels respectively stand for:
SUPPORT: The claim is consistent with the cited evidence and the evidence is sufficient to confirm the claim.
REFUTE: The claim contradicts the cited evidence and the evidence is sufficient to disprove the claim.
NOT ENOUGH INFO: The available evidence is insufficient to determine whether the claim is true or false.

Verify the claim:
{claim}
<information>
{evidence}
</information>

Figure 10: System Prompt for Offline Claim Verification.

Model	FEVEROUS			EX-FEVER			FEVER			SciFACT			HOVER	
	Sup.	Ref.	NEI	Sup.	Ref.	NEI	Sup.	Ref.	NEI	Sup.	Contr.	NEI	Sup.	N.Sup.
GPT-4o	86.05	65.64	43.19	50.19	69.96	44.11	90.33	91.33	41.33	69.20	76.37	79.32	79.20	68.00
Qwen-3B	50.95	35.36	40.48	78.23	29.93	41.83	71.00	67.00	50.33	48.52	51.05	63.71	59.00	65.60
Qwen-3B-SFT	51.33	36.50	40.82	71.09	30.27	40.30	73.67	62.67	49.67	48.95	49.79	67.93	57.20	64.60
Qwen-3B-OffRL	50.57	59.32	33.67	75.85	50.68	40.68	80.00	78.00	43.67	65.82	60.34	69.20	60.20	62.80
Qwen-7B	44.87	**67.30**	23.47	78.91	**59.52**	27.00	78.33	75.33	24.00	**76.37**	**67.51**	54.01	58.20	**73.20**
Qwen-3B-OnRL	**52.85**	63.50	**45.24**	**79.25**	59.18	**66.92**	**86.67**	**87.67**	**55.00**	45.57	54.43	**90.30**	**62.00**	65.40
Llama-3B	47.91	21.29	36.05	68.37	19.73	42.59	77.33	51.67	32.33	78.06	22.36	42.19	66.00	46.00
Llama-3B-SFT	49.43	27.00	30.95	74.49	25.51	35.74	77.67	51.33	**39.00**	80.59	24.89	45.99	67.60	47.20
Llama-3B-OffRL	42.59	48.29	35.03	60.54	36.73	34.60	80.33	65.67	27.00	61.18	40.08	54.01	64.00	50.40
Llama-8B	**60.84**	46.77	**39.12**	75.85	37.76	49.43	**88.67**	**82.00**	33.67	**88.19**	45.57	43.04	**75.00**	60.40
Llama-3B-OnRL	60.08	**65.78**	31.97	**85.03**	**42.52**	**52.09**	88.00	80.67	34.67	60.34	**64.98**	**74.26**	69.40	**61.40**

Table 8: Class-wise **Label Accuracy** across five datasets. Values are percentages. Sup./Ref./Contr./N.Sup. denote support, refute, contradict, and not-supported labels. Model abbreviations follow Table 1. Within each model group, **bold** denotes the best performance and underline denotes the second best.

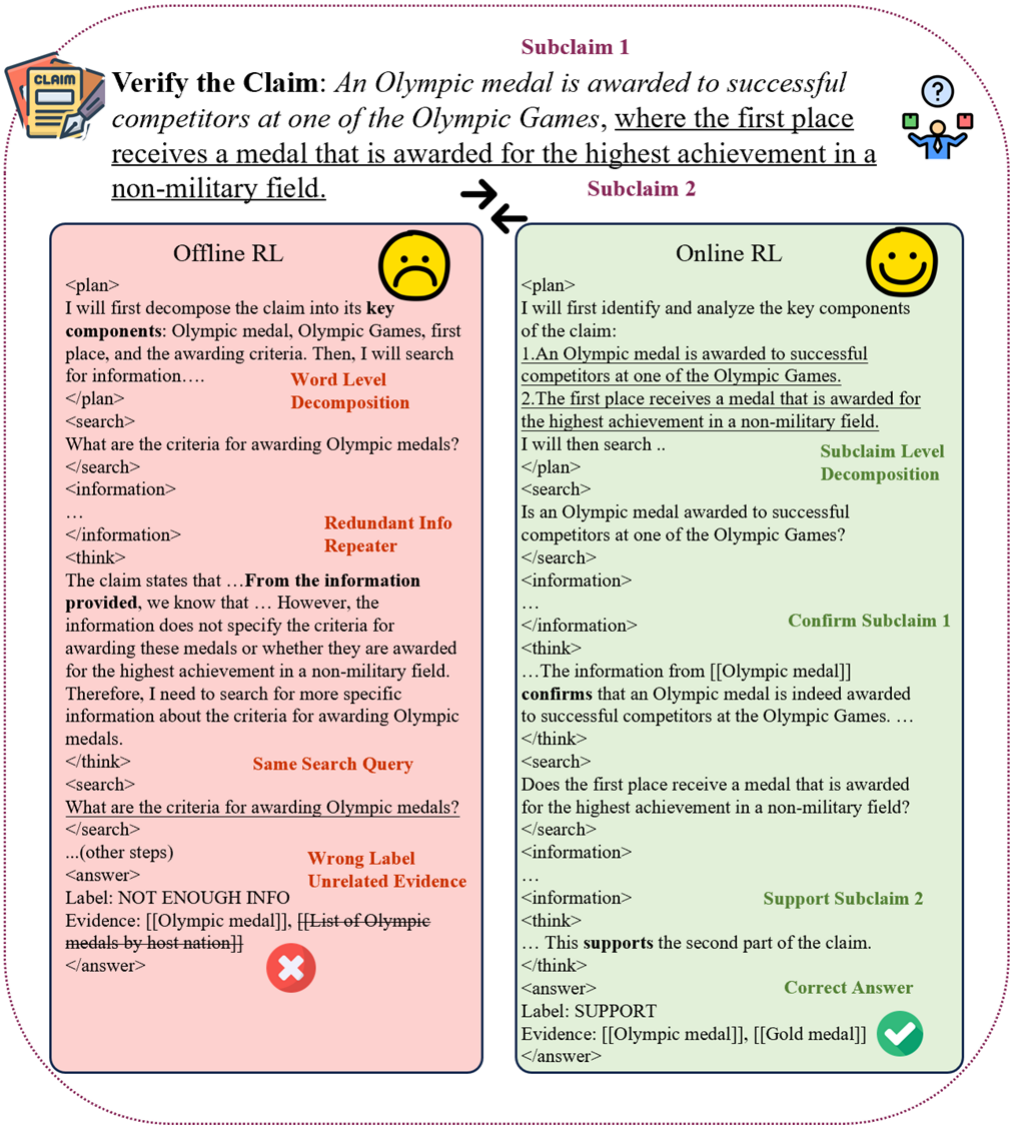


Figure 11: Case Analysis of Offline RL V.S. Online RL