SYNTHVERIFY: Enhancing Zero-Shot Claim Verification through Step-by-Step Synthetic Data Generation

Rongwen Zhao and **Jeffrey Flanigan**
University of California, Santa Cruz
{rzhao17,jmflanig}@ucsc.edu

Abstract

Claim verification is a fundamental task in natural language processing (NLP), involving the assessment of whether available evidence supports or refutes a given claim. While large language models (LLMs) have shown promise in this area, they continue to struggle with domain-specific knowledge. Synthetic data generation has emerged as an effective solution to this challenge. However, existing methods are often either inefficient to scale across multiple domains or overly reliant on external documents. We introduce SYNTHVERIFY, a novel step-by-step prompting-based synthetic data generation framework designed to enhance zero-shot claim verification. Our core insight is that guiding generation with domain-specific claim patterns and structured evidence plans can bridge LLMs' knowledge gaps in specialized domains without requiring access to external corpora or sacrificing generalizability. Using SYNTHVERIFY, we construct a diverse synthetic dataset for zero-shot verification, enabling instruction fine-tuning tailored to the verification task. Empirical results across multiple specialized domains demonstrate significant accuracy improvements, including a 20.1-point gain on the Llama-3-8B model. Our results highlight the effectiveness of structured synthetic data generation in addressing the limitations of verification systems, particularly in domain-specific tasks.

1 Introduction

Claim verification, the task of determining whether a claim is supported by given evidence, has emerged as a crucial component in combating misinformation and ensuring information integrity (Litou et al., 2017; Hassan et al., 2017; Shu et al., 2017). While LLMs have shown impressive capabilities in various NLP tasks, LLMs face several critical challenges in claim verification (Zhang and Gao, 2023; Guan et al., 2024; Augenstein et al., 2024; Quelle and Bovet, 2024).

A primary concern is the lack of domain-specific knowledge during verification. When evaluating claims in specialized fields such as medicine, law, or scientific research, LLMs often fail to understand the relationships between these domain-specific terminologies (Vladika and Matthes, 2024). This limitation affects both the evidence analysis step, where models must identify relevant domain-specific information, and the logical reasoning step, where they need to apply domain-specific constraints.

Fine-tuning on domain-specific datasets can help LLMs mitigate these issues. While existing datasets like FEVER (Thorne et al., 2018), and VitaminC (Schuster et al., 2021) have driven progress in this field, they often focus on narrow domains (e.g., Wikipedia articles) or specific types of claims. This specialization leads to poor generalization when systems encounter claims from previously unseen domains (Zhu et al., 2022; Nan et al., 2022; Gu et al., 2023). Furthermore, the manual creation of claim verification datasets is both time-consuming and expensive, particularly when aiming to cover multiple knowledge domains with sufficient depth and diversity.

Synthetic data generation has emerged as a promising solution to address this challenge (Pan et al., 2021; Wright et al., 2022; Bussotti et al., 2024). Despite its potential, existing approaches face notable limitations. Some methods lack scalability across diverse domains, making them impractical for broad deployment (Pan et al., 2021; Wright et al., 2022), while others rely heavily on external documents to guide data synthesis, introducing dependencies that may not always be available or reliable (Pan et al., 2021; Bussotti et al., 2024). As a result, there is a growing need for more efficient and self-contained strategies that can generalize across multiple domains without compromising data quality or diversity.

To address these challenges, we propose SYN-

Findings of the Association for Computational Linguistics: ACL 2025, pages 3257–3274

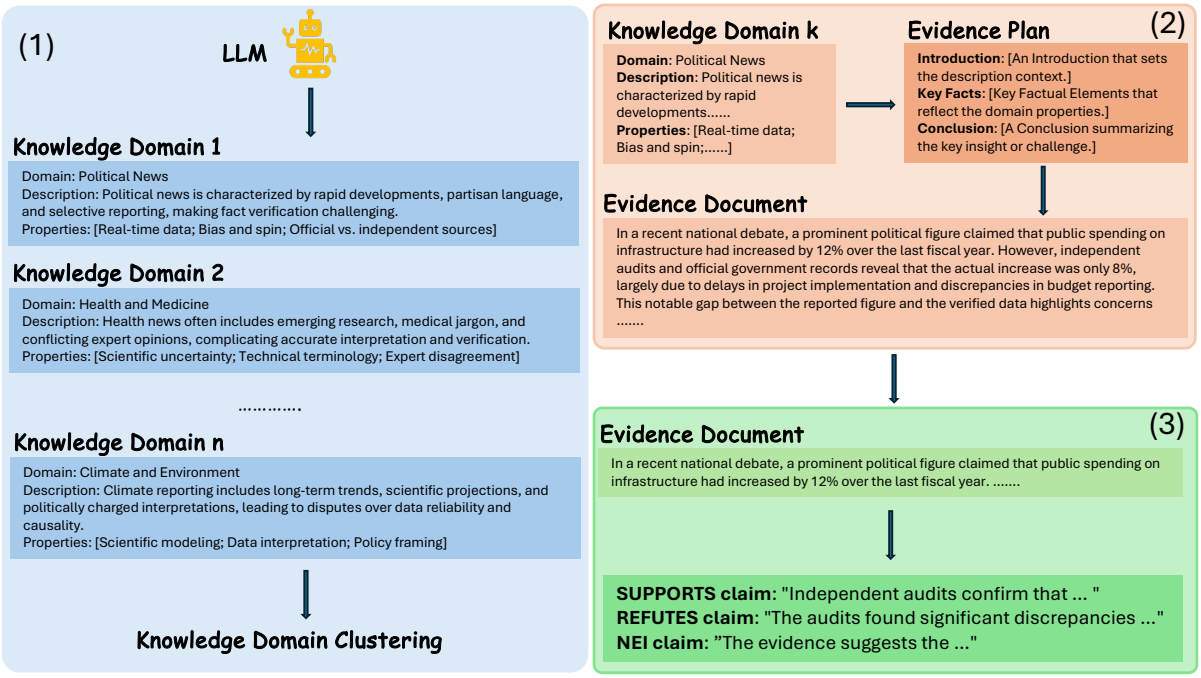

Figure 1: Overview of our synthetic data generation framework for claim verification. The framework consists of three main stages: (1) Knowledge Domain Generation (§3.2), (2) Plan-guided Evidence Synthesis (§3.3), and (3) Aspect-Conditioned Claim Generation (§3.4).

THVERIFY, a novel step-by-step prompting-based synthetic data generation framework. Our framework carefully controls the generation process to ensure data quality and verifiability. The core insight is that guiding generation with domain-specific claim patterns and structured evidence plans, without requiring access to external corpora or sacrificing generalizability.

First, we develop a structured domain specification stage that captures the nature of domain knowledge and verification requirements, providing a foundation for domain-aware claim verification. We then employ a simple but effective evidence synthesis method that provides clear verification signals, addressing the challenge of generating realistic and verifiable evidence. Building upon this stage, we implement an aspect-based claim generation approach that ensures both diversity in claim types and verifiability of generated claims.

Through extensive evaluation, we demonstrate that our synthetic data generation approach leads to prominent improvements on multiple claim verification tasks. Models trained on our synthetic dataset show improved zero-shot transfer capabilities, better handling of complex claims, and more robust verification across diverse knowledge domains. Our further analysis reveals that the structured nature of our generation process helps models learn generalizable verification strategies rather than superficial patterns specific to particular domains. Overall, our main contributions are:

- We propose a step-by-step prompting-based framework for generating synthetic claim verification data that enables zero-shot transfer across domains (§3.1).
- We introduce a structured framework for specifying knowledge domains and claim patterns that ensures both coverage and diversity (§3.2).
- We develop a simple but effective method for evidence synthesis (§3.3) and claim generation (§3.4) tailored to complex verification scenarios.
- We conduct extensive evaluations demonstrating that our approach significantly outperforms existing synthetic data baselines across multiple specialized domains (§5).

2 Background and Problem Setup

Problem Setup Following the prior work (Thorne et al., 2018; Eisenschlos et al., 2021; Schuster et al., 2021; Wadden et al., 2020), given a sentence-level claim c and a set of sentence-level

evidences [1] $E = \{e_1, ..., e_{|E|}\}$, the task of claim verification is to determine whether E supports or refutes c, or if there is insufficient information to make a determination. Here, we assume that each given claim c is *check-worthy* and can be fully verified based on the evidence set E, without relying on the external context. Formally, given an LLM M, we define a verification function $f_M : \mathcal{C} \times \mathcal{E} \rightarrow \mathcal{Y}$, where $\mathcal{C}$ represents the space of possible claims, $\mathcal{E}$ denotes the space of evidence sets and $\mathcal{Y} = \{\texttt{SUPPORTS}, \texttt{REFUTES}, \texttt{NOT\_ENOUGH\_INFO}\}$ is the set of verification labels.

For a given claim-evidence pair (c, E), the LLM M performs the verification through three main steps. First, M performs evidence analysis by processing each evidence $e_i \in E$ to extract and understand the information relevant to the claim. Second, M performs logical reasoning to determine the relationship between the extracted evidence and the claim, applying its understanding of the domain knowledge. Finally, M assigns a verification label $y \in \mathcal{Y}$ based on the aggregated evidence and reasoning outcome.

Challenges of LLM-based Claim Verification LLM-based claim verification systems face two significant challenges in real-world applications. First, although LLMs acquire extensive domain knowledge during pre-training, they often demonstrate limited generalization capabilities in unseen domains (Pan et al., 2023b). Human-annotated, domain-specific datasets can substantially enhance performance in such contexts. However, the high costs associated with annotation, both in terms of time and resources, constrain the coverage, diversity, and scalability of these datasets. Second, LLMs are susceptible to reasoning errors and over-reliance on spurious correlations (Tang et al., 2023), which can lead to incorrect veracity assessments. This is particularly problematic when claims require complex reasoning or contextual understanding that spans multiple knowledge types. In this work, we introduce an innovative synthetic data generation approach designed to bolster the capabilities of LLMs in addressing these challenges effectively.

[1]We acknowledge that retrieval is a critical component of real-world fact-checking pipelines. However, our evaluation setup assumes that claims are accompanied by evidences, focusing specifically on the claim verification stage rather than evidence retrieval.

Knowledge Domain

Domain: Climate Science
Description: A dynamic domain involving long-term alteration of temperature and typical weather patterns. Claims are challenging to verify due to reliance on complex climate models and long-term observational data.
Properties: [Temperature data, Weather patterns, Climate models]

Figure 2: An example of the generated knowledge domains. Our **Knowledge Domain Generation** module (§3.2) generates a diverse set of knowledge domains, coupled with a clustering-based diversity enforcement.

3 Methodology: Zero-Shot Claim Verification Data Generation

In this section, we describe the details on how we design a structured approach to generate synthetic claim verification data that enables zero-shot transfer across diverse domains. Our goal is to construct a dataset $\mathcal{D} = \{(D_i, C_i, E_i, y_i)\}_{i=1}^{N}$ of N instances, where D_i represents a specification of knowledge domain, C_i is a claim within that domain, E_i is the corresponding evidence, and $y_i \in \{\texttt{SUPPORTS}, \texttt{REFUTES}, \texttt{NOT\_ENOUGH\_INFO}\}$ is the verification label.

3.1 Overview

As shown in Figure 1 , our synthetic data generation framework consists of three stages. First, the **Knowledge Domain Generation** stage defines the scope and constraints of the target domain, organizing knowledge hierarchically to ensure generated content remains focused and relevant. Second, the **Plan-guided Evidence Synthesis** stage synthesizes informative evidence documents by leveraging the knowledge domains from the previous step. Third, the **Aspect-Conditioned Claim Generation** stage generate three different types of claims (Supported/Refuted/Not Enough Info) conditioned on each generated evidence document.

3.2 Knowledge Domain Generation

First, to collect a diverse set of domains where evidence and claims are sampled, a set of knowledge domains are randomly sampled from an LLM as structured triplets. Each knowledge domain O_i is formally characterized by: $O_i = (D_i, S_i, P_i)$, where D_i is the domain name, S_i is a basic description of the domain explaining why claims in this domain are challenging to verify. $P_i = \{p_1, ..., p_k\}$

is a set of domain properties. The knowledge domain O_i will be used to guide the generation of the evidence documents. The additional information S_i and P_i can lead to more non-trivial and domain-diverse generations compared to using D_i alone.

We prompt an LLM to generate a list of knowledge domains that require claim verification in the real-world. The detailed prompt and a sample of generated knowledge domain is shown in Appendix. However, naive LLM prompting often results in redundant or overlapped domains when generating full set of knowledge domains O_i, leading to low coverage (Ding et al., 2023).

To ensure broad coverage across knowledge spaces, we employ an iterative clustering-based generation and filtering process. Let $f(\cdot)$ be an embedding function that maps each domain D_i to a d-dimensional semantic space. For a batch of k generated knowledge domains, we compute their embeddings: $\mathbf{E} = \{\mathbf{e}_i = f(D_i)\}_{i=1}^{k}$. To identify and remove near-duplicate domains, we apply density-based clustering DBSCAN (Ester et al., 1996):

$$\text{clusters} = \text{DBSCAN}(\mathbf{E}, \epsilon, \text{min\_samples}) \quad (1)$$

where ϵ and min_samples are empirically determined parameters. This process continues iteratively until we obtain n diverse domains.

3.3 Plan-guided Evidence Synthesis

In this stage, we generate an informative evidence document by leveraging the knowledge domain from the previous step. However, in a preliminary experiment, prompting LLMs to get the evidence document based on the knowledge domain resulted in more generic content that lack sufficient details to generate meaningful and check-worthy claims in the future stage. To address this issue, inspired by some prior studies in the story generation (Yang et al., 2022, 2023), we generate the evidence document from the knowledge domain in two steps.

The key idea is to leverage a knowledge domain to generate a structured plan, which is subsequently expanded into a coherent evidence document. This two-step process ensures that the final output adheres not only to the factual but also to the stylistic nuances of the target domain but also maintains a consistent academic structure.

Step 1: Evidence Plan Generation. We first prompt the LLM to generate a concise plan that serves as a blueprint for the evidence document. The plan is designed to capture a brief domain-dependent statement and several specific details drawn from the domain properties.

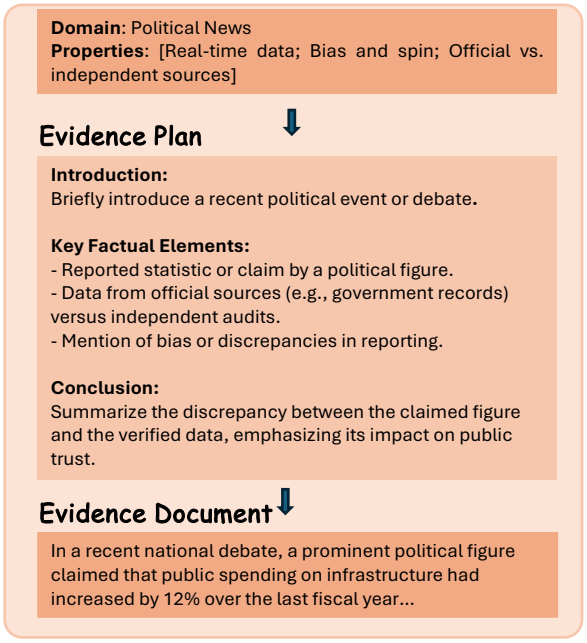

Figure 3: Given a knowledge domain, our **Plan-guided Evidence Synthesis** module (§3.3) first generates an evidence plan (upper), then it outputs a concise evidence document.

Step 2: Evidence Document Synthesis. Once the plan is generated, the next step is to expand it into a full evidence document. The model is instructed to integrate all plan components into a polished paragraph written in an information-preserving style. This expansion ensures that the evidence document is comprehensive, coherent, and reflective of the domain-specific characteristics.

Figure 3 provides an overview of the plan-guided evidence synthesis process. By grounding the generation process in a structured plan derived from the knowledge domain, our method achieves enhanced control over content structure, ensuring both consistency and domain-specific nuance. This approach has demonstrated improved quality in synthetic evidence, thereby contributing to more robust downstream claim verification across diverse domains.

3.4 Aspect-Conditioned Claim Generation

As shown in Figure 4, we generate three different types of claims (Supported/Refuted/Not Enough Info) conditioned on the generated evidence docu-

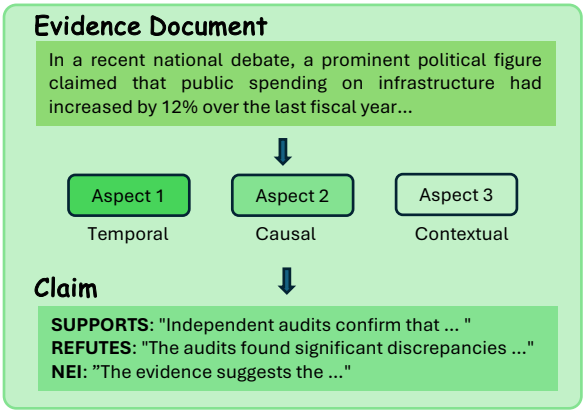

Figure 4: Our **Aspect-Conditioned Claim Generation** module (§3.4) takes in the generated evidence. It first generates three aspect features (e.g., temporal, causal and contextual) (middle), then it outputs the three claim (lower) based on the generated key aspects.

ment. Our approach increases the diversity and complexity of the generated claims through a aspect-based generation method that ensures both structural diversity and verifiability while maintaining semantic coherence within each source document. We first generate the supported claim based on the evidence document, then generate the other two claims conditioned on the supported claims.

Supported Claim Generation. We first extract several *key aspects* from the evidence document. These key aspects are concise descriptors that capture different dimensions or viewpoints present in the evidence (e.g., contextual background, numerical details, source credibility, or timing). By subsequently conditioning the claim generation on each identified aspect, the model produces multiple supported claims that, while all factually correct with respect to the evidence, emphasize different details or angles.

Given an evidence document and the associated knowledge domain, the model is prompted to generate a list of key aspects. Each key aspect is a short description that highlights a particular facet of the evidence. For example, one aspect might focus on the statistical discrepancy mentioned in the text, while another might highlight the reliability of the sources. This step ensures that the extracted aspects are aligned with the unique properties defined in the knowledge domain.

For each key aspect, a secondary prompt is created that conditions the model on both the original evidence and the specified aspect. The prompt instructs the LLM to generate a supported claim that is not only consistent with the evidence but also emphasizes the conditioned aspect. This encourages the model to explore multiple views of the evidence, thereby producing a diverse set of claim formulations.

Refuted Claim Generation. Starting from a supported claim, we generate refuted claims by applying a series of controlled perturbations designed to introduce discrepancies between the claim and the underlying evidence. These perturbations include: (1) **Entity Substitution:** Replacing critical entities with alternative, contextually plausible candidates. (2) **Temporal Modification:** Adjusting time-related references while preserving overall historical coherence. (3) **Relationship Reversal:** Inverting the relational dynamics between entities to alter the claim's meaning. (4) **Attribute Modification:** Modifying specific properties or characteristics to create a divergence from the original details.

Not-Enough-Info (NEI) Claim Generation. Given a supported claim, the model is instructed to modify it by removing or replacing the identified key elements with more vague, qualitative, or uncertain language. For example, a claim stating "The study shows a 70% reduction in hospitalization rates" may be transformed into "The study suggests a significant reduction in hospitalization rates," where the exact figure is omitted.

3.5 Human Evaluation

In order to investigate the accuracy and quality of the labels on our generated (evidence, claim) pairs. We randomly sample 50 examples from the generated dataset. Each example was annotated with a label by three NLP researchers. The examples are unedited outputs from the models, with no revisions made during annotation. The average Cohen's κ score between annotators is 65.83%, indicating substantial agreement. Furthermore, a majority label (agreed upon by 2 out of 3 annotators) was obtained for 94% of examples, while all three annotators reached unanimous agreement on 72% of examples. The model achieved a high accuracy of 79.6% against the majority label and 88.5% against the unanimous label. The Cohen's κ scores between the model's labels and both the majority and unanimous labels are 69.73% and 82.46%, respectively.

Instruction: You are given a claim and evidence. Your task is to verify the claim based solely on the evidence provided. Analyze the claim and evaluate the evidence. Based on your evaluation, return one of the following labels: "supports", "refutes" or "not enough info".

Evidence: {EVIDENCE}
Claim: {CLAIM}
Label: {LABEL}

Figure 5: The instruction template that we use for compiling the final instruction tuning dataset to instruct tuning LLMs for the claim verification.

3.6 The Final Instruction Tuning Dataset

Using our proposed framework SYNTHVERIFY, we generate a domain-diverse synthetic claim verification dataset for training zero-shot verification models. We use the generated dataset to fine-tune the various LLMs (e.g., Llama-3.1-8B). To achieve this, we combine the predefined instruction template in Figure 5 and instance input into a single prompt and train the LLMs in a standard supervised learning setup to generate the corresponding instance output.

4 Experimental Setup

Evaluation Benchmarks. In this work, following the prior work (Pan et al., 2023b), we use a wide range of existing claim verification benchmarks from various domains: (1) Wikipedia domain: We use FEVER (Thorne et al., 2018), VitaminC (Schuster et al., 2021) and FoolMeTwice (Eisenschlos et al., 2021). These datasets were created based on Wikipedia documents; (2) Climate domain: We use Climate-FEVER (Diggelmann et al., 2020), which consists of real-world claims related to climate change; (3) Science domain: We use SciFact (Wadden et al., 2020); (4) Health domain: We use PubHealth (Kotonya and Toni, 2020); (5) Other domains: we also use Covid-Fact (Saakyan et al., 2021) from the forum domain and FAVIQ (Park et al., 2022) from the question domain in our evaluations. The statistics of these benchmarks are shown in Table 5. Please refer to Appendix A for more details.

Evaluation Metric. Following the prior work (Laban et al., 2022), we use the Balanced Accuracy (BAcc), the average of recall obtained on each label, as our evaluation metric. Here, $\text{BAcc} = \frac{1}{|\mathcal{Y}|}\sum_{y\in\mathcal{Y}}\frac{\text{TP}_y}{\text{TP}_y+\text{FN}_y}$, where $\mathcal{Y}$ is the set of all class labels (e.g., supports, refutes, not_enough_info), TP_y is the number of true positives for class y, and FN_y is the number of false negatives for class y.

Baselines. (1) Zero-Shot: By default, we compare the model fine-tuned by our synthetic dataset to the zero-shot setting. We use the same prompt as shown in Figure 5. (2) Zero-Shot Chain-of-Thought (Kojima et al., 2022): We also compare to the proposed zero-shot CoT prompting method. We append the prompt "Let's think step by step" to the instruction prompt in Figure 5. (3) We also compare our method to QACG (Pan et al., 2021), which generates synthetic claims conditioned on the Wikipidia data. We take their released dataset and fine-tune Llama-3 and InternLM2.5 models.

Implementation Details. We compare against various state-of-the-art instruction-tuned LLMs, including: Qwen 2.5 (Yang et al., 2024), Gemma 2 (Team et al., 2024), InternLM (Cai et al., 2024), Mistral (Jiang et al., 2023), and Llama-3 family (Dubey et al., 2024).

For the synthetic data generation, we use GPT-4o-mini with temperature 0.9 for all prompting generation. For domain similarity computation, we employ the all-MiniLM-L6-v2 sentence transformer, which provides 384-dimensional embeddings. The clustering parameters ($\epsilon = 0.3$, min_samples $= 2$). The all prompts, dataset statistics and fine-tuning hyperparameters are available in Appendix.

5 Main Results

Table 1 presents a comprehensive evaluation of various language models on eight claim verification benchmarks. Several notable findings emerge from our experimental results.

First, our synthetic data framework demonstrates consistent and substantial improvements across all variants of the model. Most notably, InternLM2.5-7B-Chat with our method achieves the best overall performance with an average balanced accuracy of 62.8, surpassing its base version by a large margin of 13.2 absolute points (from 49.6 to 62.8). This improvement is particularly pronounced in other datasets like FEVER (+19.7 points) and VitaminC (+25.4 points).

Second, we observe that larger model size does not necessarily translate to better claim verification performance. For example, Gemma-2-2B-it, despite its relatively small size, achieves a com-

Model	FEVER	FM2	VitaminC	C-FEVER	PubHealth	SciFact	CovidFact	FAVIQ	Avg
Qwen2.5-0.5B-Instruct	35.7	50.0	37.0	32.5	31.3	38.1	52.0	50.0	40.8
Qwen2.5-1.5B-Instruct	34.0	50.0	33.3	33.3	33.3	34.8	50.0	49.0	39.7
gemma-2-2b-it	55.7	78.5	47.7	49.6	38.7	52.9	66.5	57.5	55.9
Qwen2.5-3B-Instruct	49.7	61.0	38.0	39.0	32.7	51.9	60.5	53.0	48.2
gemma-2-9b-it	37.7	51.0	33.0	38.2	34.7	34.8	63.0	55.5	43.5
Llama-2-13b-hf	33.3	50.0	33.3	33.3	33.7	33.3	50.0	49.5	39.6
InternLM2-Chat-20B	59.3	69.5	47.3	43.1	29.3	54.3	64.0	54.0	52.6
InternLM2.5-20B-Chat	46.0	56.0	47.3	39.0	36.3	36.7	57.0	52.0	46.3
Llama-3.2-1B-Instruct	43.7	53.0	31.3	35.8	35.7	37.1	52.5	56.0	43.1
Llama-3.2-1B-Instruct (Ours)	56.7	64.0	58.0	53.7	44.0	48.1	61.5	60.0	**55.8**
Llama-3.2-3B-Instruct	43.7	63.0	36.7	48.8	38.7	54.8	64.5	57.0	50.9
Llama-3.2-3B-Instruct (Ours)	61.0	75.5	56.0	49.6	51.2	53.3	64.5	60.5	**59.0**
Mistral-7B-Instruct-v0.3	46.0	54.0	36.0	34.1	36.3	36.2	50.5	48.0	42.6
Mistral-7B-Instruct-v0.3 (Ours)	59.0	74.5	54.0	53.7	52.3	52.9	67.0	55.5	**58.6**
Llama-3.1-8B-Instruct	37.0	60.5	32.7	38.2	36.3	38.1	54.5	56.0	44.2
Llama-3.1-8B-Instruct (Ours)	60.3	76.0	56.7	50.4	45.3	59.5	68.0	59.5	**59.5**
Llama-3-8B-Instruct	33.3	50.0	33.3	33.3	33.7	33.3	50.5	51.0	39.8
+ Zero-Shot CoT	43.7	53.0	47.3	39.0	36.3	36.7	52.5	53.0	42.7
+ QACG (Pan et al., 2021)	37.7	53.0	38.0	38.2	35.7	34.8	52.5	52.0	42.7
Llama-3-8B-Instruct (Ours)	**59.3**	**80.5**	**57.0**	**49.6**	**49.3**	**55.7**	**65.5**	**62.5**	**59.9**
InternLM2.5-7B-Chat	44.3	76.5	33.3	35.0	34.0	46.7	67.0	60.0	49.6
+ Zero-Shot CoT	46.0	76.5	36.7	35.8	36.3	48.1	67.0	60.5	50.9
+ QACG (Pan et al., 2021)	46.0	78.5	37.7	42.3	32.7	51.0	67.0	60.0	51.9
InternLM2.5-7B-Chat (Ours)	**64.0**	**83.5**	**58.7**	**53.7**	**49.3**	**57.1**	**71.0**	**65.0**	**62.8**

Table 1: Zero-shot performance across different datasets, using Balanced Accuracy (BAcc). For Llama-3 and InternLM2.5 models, we bold the best model for each dataset. Otherwise, we bold the best average score.

petitive average accuracy of 55.9, outperforming several larger models such as InternLM2-Chat-20B (52.6) and InternLM2.5-20B-Chat (46.3). This suggests that verification performance is more closely tied to a model's capabilities than its size.

Third, our method shows remarkable consistency in improving performance across different model architectures and scales. The enhancement is particularly evident in the Llama family: Llama-3.2-1B-Instruct improves from 43.1 to 55.8 (+12.7 points), Llama-3.2-3B-Instruct from 50.9 to 59.0 (+8.1 points), and Llama-3.1-8B-Instruct from 44.2 to 59.5 (+15.3 points). This consistent improvement pattern suggests the robustness and generalizability of our approach.

We also observe that our method is better than zero-shot chain-of-thought (CoT) prompting. Zero-shot CoT can usually get some improvements, but fine-tuning models on domain-relevant data would be more beneficial. Furthermore, our method is also better than QACG (Pan et al., 2021). The major reason is that our method benefits from LLMs. LLMs store a lot of domain knowledge during the pre-training stage.

Synthesis Framework	VitaminC	SciFact	FAVIQ
No Training	33.3	46.7	60.0
Direct Prompting	35.0	34.8	58.5
R1	57.0	**56.2**	62.5
R2	54.0	51.0	60.5
R3	56.0	54.3	62.5
Full Framework	**57.3**	**56.2**	**64.0**

Table 2: The results of our ablation studies. **R1** means removing knowledge domain speficiation. **R2** represents that removing the evidence plan generation. **R3** is removing the key aspect generation.

6 Further Analysis

In this section, we conduct more analysis studies, comparing our framework to a direct prompting method, analyzing the effect of training data size and the impact of domain knowledge.

6.1 Ablation Study

To understand the contributions of each component of our pipeline, we systematically evaluate our framework's three main components. We use the same domains as our framework to generate 20k samples for each setting. We use InternLM2.5-7B-Chat as our base model and evaluate on Vitaminc (Schuster et al., 2021), SciFact (Wadden et al.,

2020), and FAVIQ (Park et al., 2022) datasets. The full prompt is shown in Appendix.

Direct Prompting The most straightforward way is the direct prompting without any intermediate step. This method directly prompts the LLM to generate the claim-evidence pairs when given a domain label. Table 24 shows the prompt we use to perform direct prompting.

Removing knowledge domain Specification (R1) When removed, we replace the structured knowledge domain with simple domain labels (e.g., climate science), eliminating the domain-specific constraints. This tests the impact of structured knowledge domain representation in our framework. The prompt is shown in Table 21.

Removing Evidence Plan Generation (R2) We replace our evidence synthesis module with basic prompting generation, removing the evidence plan generation. This tests the value of the evidence plan in guiding the evidence document generation. The prompt is shown in Table 22.

Removing Key aspect Generation (R3) Without this component, we use basic prompt-based generation without key aspects to generate the supported claim. This evaluates the benefit of our diversity-oriented claim generation approach over simple prompting. The prompt is shown in Table 23.

Table 2 shows the results of our ablation studies. The results demonstrate the effectiveness of our proposed framework. We can observe that performance decreases slightly when removing a specific component. The results also show consistent improvements over both the no-training baseline and direct prompting approaches. Specifically, our full framework achieves the highest performance across all datasets. The improvement is particularly pronounced for VitaminC, where our framework outperforms the no-training baseline by 24 points. Interestingly, direct prompting shows mixed results compared to no training, performing slightly better on VitaminC (+1.7 points) but worse on SciFact (-11.9 points). These findings strongly suggest that our designed framework contributes significantly to the framework's overall effectiveness.

6.2 Impact of the Budget Control

Since our data synthesis framework uses more API calls compared to the direct prompting method.

Synthesis Framework	VitaminC	SciFact	FAVIQ
No Training	33.3	46.7	60.0
Direct Prompting (30k)	38.0	48.1	60.0
Full Framework (5k)	**56.0**	**55.7**	**62.5**

Table 3: Ablation study by considering the budget constraints.

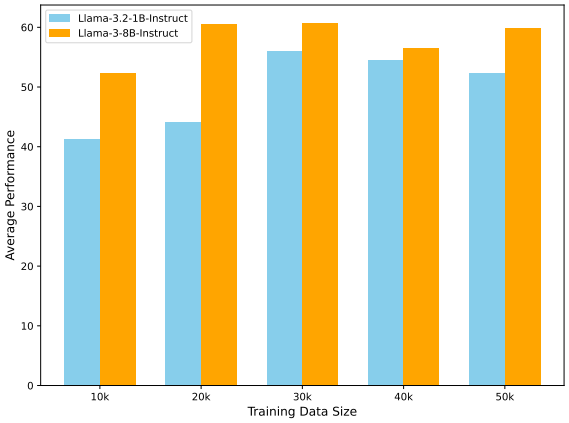

Figure 6: The results of how the performance of Llama-3.2-1B-Instruct and Llama-3-8B-Instruct models varies when trained on different sizes of synthetic instruction tuning data.

In order to investigate the benefit of the structure-aware data synthesis, we also compare our framework to the direct prompting method incorporating budget constraints. Since it is hard to control the synthesis budget directly, we conduct an additional experiment under the rough estimation. The table Table 3 compares a model fine-tuned with 30k samples from direct prompting against a model fine-tuned with only 5k samples generated using our framework. As we can see, our framework is still significantly better than the direct prompting even under the similar budget.

6.3 Effect of the Training Dataset Size

Here we study the effect of the size of the training dataset. In particular, we study how the performance of the Llama-3.2-1B-Instruct and Llama-3-8B-Instruct models varies when trained on different sizes of synthetic instruction tuning data. Figure 6 shows that training on more steps typically improves performance. We find that fine-tuned LLMs on all datasets reach the peak performance generally after 30,000 steps. It means training on more of the synthetic data may not be necessary.

Model	C-FEVER	SciFact	CovidFact
Llama-3.2-1B-Instruct	35.8	37.1	52.5
w/ FEVER	38.2	43.8	54.5
w/ (SynthVerify & FEVER)	42.3	47.7	57.0
Llama-3-8B-Instruct	33.3	33.3	50.5
w/ FEVER	40.3	43.8	61.0
w/ (SynthVerify & FEVER)	50.4	58.6	67.0

Table 4: Performance comparison across C-FEVER, SciFact, and CovidFact datasets.

6.4 Impact of Domain Knowledge

Fine-tuning models on unseen domains can improve the verification performance (Pan et al., 2023a). We hypothesize that the synthetic data generated by SYNTHVERIFY can benefit models further from the general verification training. To investigate the effectiveness of SYNTHVERIFY in domain adaptation setting, we conduct experiments comparing the performance of two Llama models (3.2-1B and 3-8B) across three distinct verification datasets. We examine three training configurations for each model: the base instruction-tuned model, fine-tuning with FEVER data only, and fine-tuning with both FEVER and our synthetic data generated by SYNTHVERIFY.

The results in Table 4 demonstrate that SynthVerify consistently enhances verification performance across all domains. The improvement is particularly pronounced in the Llama-3-8B model, where the combination of SYNTHVERIFY and FEVER training leads to substantial gains, improving from 33.3 to 58.6 on SciFact. Notably, even the smaller 3.2-1B model shows consistent improvements with SynthVerify, suggesting that our approach effectively improves domain adaptation across model sizes.

7 Related Work

7.1 LLM-based Claim Verification

Numerous datasets have been compiled for fact verification across various domains, such as politics (Vlachos and Riedel, 2014), encyclopedias (Thorne et al., 2018, 2021; Eisenschlos et al., 2021), news (Pérez-Rosas et al., 2018), climate (Diggelmann et al., 2020), science (Wadden et al., 2020), and healthcare (Kotonya and Toni, 2020). Honovich et al. (2022) consolidated several datasets to evaluate the ability to measure input-output consistency. The statements in these datasets are typically single sentences, generated either by crawling specific websites (Vlachos and Riedel, 2014), manually altering factual sentences (Thorne et al., 2018), or reformulating QA pairs (Thorne et al., 2021). In this work, we study the problem of LLM-based zero-shot claim verification across multiple domains.

7.2 Synthetic Data Generation for Fact Verification

Recent advancements in synthetic data generation have significantly enhanced fact verification systems by reducing reliance on costly human annotations. Pan et al. (2021) introduced QACG, a framework that generates claims from Wikipedia by transforming question–answer pairs into supported, refuted, or unverifiable statements. Building upon this, Wright et al. (2022) proposed CLAIMGEN-BART and KBIN to automatically produce atomic scientific claims and their negations in the biomedical domain, enabling zero-shot fact checking with performance reaching up to 90% of fully supervised models. Extending to multimodal data, Bussotti et al. (2024) developed UNOWN, a system that synthesizes training examples from both textual and tabular sources. Collectively, these works underscore the effectiveness of synthetic data in scaling fact verification across diverse domains and data modalities. In contrast to these works, we propose a step-by-step prompting-based generation framework that does not rely on external input documents and remains scalable across diverse domains.

8 Conclusion

We present SYNTHVERIFY, a novel synthetic data generation framework for synthesizing domain-diverse claim verification datasets. Our approach addresses the critical challenge of limited domain coverage in existing datasets through automated generation of claim-evidence pairs across novel domains. Experimental results demonstrate that models trained on SYNTHVERIFY-generated data significantly outperform baselines on multiple benchmarks. This success highlights that the step-by-step nature of our generation process helps models learn generalizable verification strategies.

Limitations

LLM Prompting Since our method is prompting an LLM to generate synthetic data. Some less powerfull LLMs may not be able to do this task.

We will explore how to self-improve using small LLMs.

Semantic Redundancy in Generated Data While SYNTHVERIFY successfully generates valuable training datasets for zero-shot claim verification, it does not ensure that each synthesized claim is semantically unique. For zero-shot applications, this limitation is less problematic, as such models are designed to adapt to any provided claims and evidence to determine veracity. Addressing this redundancy would improve the applicability of SYNTHVERIFY for broader training scenarios.

Label Quality Given the fully automated nature of our data synthesis process, some noise in the veracity labels is unavoidable. This noise likely underestimates the true impact of training data domain diversity on zero-shot claim verification, as models trained on SYNTHVERIFY are inadvertently exposed to, and learn to predict, this noise. Ideally, a dataset with the same level of domain diversity as SYNTHVERIFY, but with gold-standard veracity labels, would serve as the basis for our experiments. The lack of such a dataset underscores the motivation for our work.

Acknowledgments

We are thankful for the computing resources provided by the Pacific Research Platform's Nautilus cluster, supported in part by National Science Foundation (NSF) awards CNS-1730158, ACI-1540112, ACI-1541349, OAC-1826967, OAC-2112167, CNS-2100237, CNS-2120019, the University of California Office of the President, and the University of California San Diego's California Institute for Telecommunications and Information Technology/Qualcomm Institute. Thanks to CENIC for the 100Gbps networks. We also thank the reviewers for their valuable feedback.

References

Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, et al. 2024. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nature Machine Intelligence*, 6(8):852–863.

Jean-Flavien Bussotti, Luca Ragazzi, Giacomo Frisoni, Gianluca Moro, and Paolo Papotti. 2024. Unknown claims: Generation of fact-checking training examples from unstructured and structured data. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12105–12122.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. 2024. Internlm2 technical report. *ArXiv preprint*, abs/2403.17297.

Tri Dao. 2024. Flashattention-2: Faster attention with better parallelism and work partitioning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. Climate-fever: A dataset for verification of real-world climate claims. *ArXiv preprint*, abs/2012.00614.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3029–3051, Singapore. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *ArXiv preprint*, abs/2407.21783.

Julian Eisenschlos, Bhuwan Dhingra, Jannis Bulian, Benjamin Börschinger, and Jordan Boyd-Graber. 2021. Fool me twice: Entailment from Wikipedia gamification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 352–365, Online. Association for Computational Linguistics.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231.

Jiawei Gu, Xuan Qian, Qian Zhang, Hongliang Zhang, and Fang Wu. 2023. Unsupervised domain adaptation for covid-19 classification based on balanced slice wasserstein distance. *Computers in Biology and Medicine*, page 107207.

Jian Guan, Jesse Dodge, David Wadden, Minlie Huang, and Hao Peng. 2024. Language models hallucinate, but may excel at fact verification. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1090–1111, Mexico City, Mexico. Association for Computational Linguistics.

Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017*, pages 1803–1812. ACM.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 161–175, Dublin, Ireland. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *ArXiv preprint*, abs/2310.06825.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754, Online. Association for Computational Linguistics.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Iouliana Litou, Vana Kalogeraki, Ioannis Katakis, and Dimitrios Gunopulos. 2017. Efficient and timely misinformation blocking under varying cost constraints. *Online Social Networks and Media*, 2:19–31.

Qiong Nan, Danding Wang, Yongchun Zhu, Qiang Sheng, Yuhui Shi, Juan Cao, and Jintao Li. 2022. Improving fake news detection of influential domain via domain- and instance-level transfer. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2834–2848, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Liangming Pan, Wenhu Chen, Wenhan Xiong, Min-Yen Kan, and William Yang Wang. 2021. Zero-shot fact verification by claim generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 476–483, Online. Association for Computational Linguistics.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023a. 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)*, pages 6981–7004, Toronto, Canada. Association for Computational Linguistics.

Liangming Pan, Yunxiang Zhang, and Min-Yen Kan. 2023b. Investigating zero- and few-shot generalization in fact verification. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 511–524, Nusa Dua, Bali. Association for Computational Linguistics.

Jungsoo Park, Sewon Min, Jaewoo Kang, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022. FaVIQ: FAct verification from information-seeking questions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5154–5166, Dublin, Ireland. Association for Computational Linguistics.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Dorian Quelle and Alexandre Bovet. 2024. The perils and promises of fact-checking with large language models. *Frontiers in Artificial Intelligence*, 7:1341697.

Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. COVID-fact: Fact extraction and verification of real-world claims on COVID-19 pandemic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2116–2129, Online. Association for Computational Linguistics.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin C! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643, Online. Association for Computational Linguistics.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36.

Ruixiang Tang, Dehan Kong, Longtao Huang, and Hui Xue. 2023. Large language models can be lazy learners: Analyze shortcuts in in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4645–4657, Toronto, Canada. Association for Computational Linguistics.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *ArXiv preprint*, abs/2408.00118.

James Thorne, Max Glockner, Gisela Vallejo, Andreas Vlachos, and Iryna Gurevych. 2021. Evidence-based verification for real world information needs. *ArXiv preprint*, abs/2104.00640.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22, Baltimore, MD, USA. Association for Computational Linguistics.

Juraj Vladika and Florian Matthes. 2024. Comparing knowledge sources for open-domain scientific claim verification. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2103–2114, St. Julian's, Malta. Association for Computational Linguistics.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Dustin Wright, David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Isabelle Augenstein, and Lucy Lu Wang. 2022. Generating scientific claims for zero-shot scientific fact checking. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2448–2460, Dublin, Ireland. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *ArXiv preprint*, abs/2407.10671.

Kevin Yang, Dan Klein, Nanyun Peng, and Yuandong Tian. 2023. DOC: Improving long story coherence with detailed outline control. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3378–3465, Toronto, Canada. Association for Computational Linguistics.

Kevin Yang, Yuandong Tian, Nanyun Peng, and Dan Klein. 2022. Re3: Generating longer stories with recursive reprompting and revision. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4393–4479, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xuan Zhang and Wei Gao. 2023. Towards LLM-based fact verification on news claims with a hierarchical step-by-step prompting method. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 996–1011, Nusa Dua, Bali. Association for Computational Linguistics.

Yongchun Zhu, Qiang Sheng, Juan Cao, Qiong Nan, Kai Shu, Minghui Wu, Jindong Wang, and Fuzhen Zhuang. 2022. Memory-guided multi-view multi-domain fake news detection. *IEEE Transactions on Knowledge and Data Engineering*, 35(7):7178–7191.

A Appendix: Evaluation Benchmarks and Fine-Tuning Details

A.1 Dataset Statistics of the Evaluation Benchmarks

In this study, we used the following pre-processed datasets in Pan et al. (2023b). The statistics of these dataset are shown in Table 5. Claims are typically sentence-level statements, which can be either real-world natural claims sourced from websites, textbooks, and forums, or artificially gener-

	Dataset	Domain	Claim	Evidence	Label	# Claims		Avg. # tokens	
						Train	Test	Claim	Evid.
I	FEVER	Wikipedia	artificial	sent-level	S (52%), R (22%), N (26%)	145,327	19,972	9.4	35.9
	VitaminC	Wikipedia	artificial	sent-level	S (50%), R (35%), N (15%)	370,653	63,054	12.6	29.5
	FoolMeTwice	Wikipedia	artificial	sent-level	S (49%), R (51%)	10,419	1,169	15.3	37.0
II	ClimateFEVER	Climate	natural	sent-level	S (25%), R (11%), N (64%)	6,140	1,535	22.8	33.8
	Sci-Fact	Science	natural	sent-level	S (43%), R (22%), N (35%)	868	321	13.8	61.9
	PubHealth	Health	natural	sent-level	S (60%), R (36%), N (4%)	8,370	1,050	15.7	137.6
	COVID-Fact	Forum	natural	sent-level	S (32%), R (68%)	3,268	818	12.4	82.5
	FAVIQ	Question	natural	doc-level	S (50%), R (50%)	17,008	4,260	15.2	304.9

Table 5: List of the 8 fact verification datasets for our study and their characteristics.

ated by crowd-workers. Evidence, the key information used to validate a claim, is often drawn from textual sources such as news articles, academic papers, and Wikipedia documents. Claim labels vary, with common formats being binary (supports/refutes) or three-class labels (supports/refutes/not enough info).

FEVER (Thorne et al., 2018) It treats Wikipedia as the primary evidence source and instructs crowd-workers to alter sentences from Wikipedia articles to create claims, which are then classified as supporting, refuting, or lacking sufficient information.

VitaminC (Schuster et al., 2021) It generates contrastive evidence pairs for each claim, where the pairs are nearly identical in wording and content, except that one supports the claim while the other does not.

FoolMeTwice (Eisenschlos et al., 2021) develops a multi-player game that encourages diverse strategies for constructing claims (*e.g.*, temporal inference) based on Wikipedia, leading to more complex claims with reduced lexical overlap with the evidence.

Climate-FEVER (Diggelmann et al., 2020) It consists of 1,535 real-world claims related to climate change, collected from the Internet. The five most relevant sentences from Wikipedia are retrieved as evidence, and human annotators label each sentence as supporting, refuting, or providing insufficient information for verification.

Sci-Fact (Wadden et al., 2020) It includes 1.4K expert-generated scientific claims, each linked to abstracts containing supporting evidence, annotated with labels and sentence-level rationales.

PubHealth (Kotonya and Toni, 2020) It contains 11.8K claims accompanied by gold standard judgments crafted by journalists to support or refute the claims. The claims come from five fact checking platforms, news headlines, and news reviews. We pair each claim with its corresponding judgment text as evidence.

COVID-Fact (Saakyan et al., 2021) It consists of 4,086 claims related to the COVID-19 pandemic, collected from the *r/COVID19* subreddit. We use the sentence-level evidence annotated by crowd-workers as the supporting evidence.

FAVIQ (Park et al., 2022) It contains 26K claims derived from naturally occurring ambiguous questions posed by real users. Each claim is paired with an answer-containing Wikipedia paragraph as its document-level evidence.

Since many of the original datasets do not provide a publicly available test set, we use their original dev splits as our evaluation sets. Furthermore, we standardize label naming conventions to `supports`, `refutes`, and `not enough info`.

A.2 Fine-Tuning Details

We use LoRA (Hu et al., 2022) and FlashAttention-2 (Dao, 2024) with a cosine learning rate schedule and an initial learning rate of $1e{-}4$ to fine-tune LLMs. The maximum sequence length is 1024 and the batch size is 4. The LoRA rank is 16 for all models. We use the Huggingface Transformers (Wolf et al., 2020) library to train all models.

B Appendix: Prompts

This appendix provides the complete set of prompts used in our synthetic data generation pipeline, along with example outputs for each stage.

Prompt	Table Reference
Synthetic data generation	
knowledge domain generation	Table 7
Evidence plan generation	Table 9
Evidence document synthesis	Table 11
Key aspect extraction	Table 13
Supported claim generation	Table 15
Refuted claim generation	Table 17
Not-Enough-Info claim generation	Table 18
Model evaluation	
Zero-shot evaluation	Table 19
Zero-shot Chain-of-Thought evaluation	Table 20
Ablation studies	
Removing knowledge domain Specification	Table 21
Removing Evidence Plan Generation	Table 22
Removing Key Aspect Generation	Table 23
Direct prompting	Table 24

Table 6: A collection of all prompts used in synthetic data generation, model evaluation and ablation studies.

B.1 Knowledge Domain Generation Prompt

B.2 Prompts for Plan-guided Evidence Synthesis

B.3 Prompts for Aspect-Conditioned Claim Generation

B.4 Prompts for Model Evaluation

B.5 Prompts for Ablation Studies

Please generate a list of diverse domains that is highly relevant for claim verification tasks. For each domain, provide:

1. The domain name. Do not include "&" or "and" in the generated domain name. Do not combine two or more words together to form a single domain name.
2. A concise description (1–2 sentences) highlighting its unique linguistic and factual characteristics and explaining why claims in this domain are challenging to verify.
3. A list of several domain properties.

Format your response exactly as follows:
Domain: [Domain Name]
Description: [Your description here]
Properties: [Property 1;Property 2;Property 3]

Table 7: The prompt for knowledge domain generation.

Example Output:
Domain: Climate Science
Description: A dynamic domain involving long-term alteration of temperature and typical weather patterns. Claims are challenging to verify due to reliance on complex climate models and long-term observational data.
Properties: ["Temperature data", "Weather patterns", "Climate models"]

Table 8: The example output for knowledge domain generation.

You are an expert academic writer in the field of claim verification. Given the following knowledge domain:

Domain: [Domain Name]
Description: [A concise description (1–2 sentences) highlighting its unique linguistic and factual characteristics and explaining why claims in this domain are challenging to verify.]
Properties: [Property 1; Property 2; Property 3]

Please generate a concise bullet-point plan for a synthetic evidence document in this domain. Your plan should include:
• An Introduction that sets the description context.
• Key Factual Elements that reflect the domain properties.
• A Conclusion summarizing the key insight or challenge.

Table 9: The prompt for evidence plan generation.

Example Output:
Domain: Political News
Description: Political news is characterized by rapid developments, partisan language, and selective reporting, making fact verification challenging.
Properties: Real-time data; Bias and spin; Official vs. independent sources

Evidence plan:
• Introduction: Briefly introduce a recent political event or debate.
• Key Factual Elements:
  - Reported statistic or claim by a political figure.
  - Data from official sources (e.g., government records) versus independent audits.
  - Mention of bias or discrepancies in reporting.
• Conclusion: Summarize the discrepancy between the claimed figure and the verified data, emphasizing its impact on public trust.

Table 10: The example output for evidence plan generation.

You are a domain expert in claim verification. Expand the following evidence plan into a full evidence document suitable for a claim verification dataset. Ensure that the final document is coherent, detailed, and written in a formal academic style.

Evidence plan:
Insert the plan generated here.

Please generate the complete evidence document following this structure.

Table 11: The prompt for evidence document synthesis.

Example Output:
Evidence Document:
In a recent national debate, a prominent political figure claimed that public spending on infrastructure had increased by 12% over the last fiscal year. However, independent audits and official government records reveal that the actual increase was only 8%, largely due to delays in project implementation and discrepancies in budget reporting. This notable gap between the reported figure and the verified data highlights concerns over potential bias and spin in political reporting, which may ultimately undermine public trust in government accountability.

Table 12: The example output for evidence document synthesis.

You are a domain expert in claim verification. Given the following evidence document and its associated knowledge domain, please extract and list three key Aspects that capture different facets of the evidence. Each key aspect should be a concise description that highlights a distinct dimension (e.g., numerical details, source credibility, or temporal context) and aligns with the domain properties.

Evidence:
In a recent national debate, a prominent political figure claimed that public spending on infrastructure increased by 12% last fiscal year, while independent audits reported an actual increase of only 8%.

knowledge domain:
Domain: Political News
Description: Political news often features conflicting reports and varying interpretations due to partisan Aspects.
Properties: Real-time data; Bias and spin; Official vs. independent sources

Output Format:
1. [Key Aspect 1]
2. [Key Aspect 2]
3. [Key Aspect 3]

Table 13: The prompt for the key aspect extraction from the evidence document.

Example Output:
Statistical Aspect: Focuses on the numerical difference between the 12% claimed and the 8% audited increase.
Source Reliability Aspect: Emphasizes the contrast between the official claim and the independent audits.
Contextual Aspect: Highlights the influence of political debate and potential bias in reporting.

Table 14: The example output for the key aspect extraction from the evidence document.

You are a domain expert in claim verification. Based on the evidence document provided below and a specified key aspect, generate a supported claim that is factually consistent with the evidence while emphasizing the given aspect.

Evidence:
In a recent national debate, a prominent political figure claimed that public spending on infrastructure increased by 12% last fiscal year, while independent audits reported an actual increase of only 8%.

Key Aspect:
Statistical Aspect: Focuses on the numerical difference between the 12% claimed and the 8% audited increase.

Output Format:
Supported Claim: [Your claim here]

Table 15: The prompt for generating supported claims based on key aspects.

Example Output:
Supported Claim: Independent audits confirm that public spending on infrastructure actually increased by only 8% last fiscal year, highlighting a significant statistical discrepancy compared to the 12% claimed by the political figure.

Table 16: The example output for generating supported claims based on key aspects.

You are a highly skilled claim verification expert. Your task is to generate a refuted claim from a given supported claim by applying controlled perturbations that introduce discrepancies between the claim and the underlying evidence. In doing so, consider the following perturbation strategies:

1. Entity Substitution: Replace critical entities in the claim with alternative, contextually plausible candidates.
2. Temporal Modification: Adjust any time-related references, while maintaining overall historical coherence.
3. Relationship Reversal: Invert the relational dynamics between entities, altering the claim's meaning.
4. Attribute Modification: Modify specific properties or characteristics to create a divergence from the original details.

Now, given the following information, generate a refuted claim.

Supported Claim: [Insert Supported Claim Here]
Evidence: [Insert Relevant Evidence Here]

Output your response in the following format:
Refuted Claim: [Your refuted claim here]

Table 17: The prompt for generating the refuted claim based on the supported claim.

You are a domain expert in claim verification. Given a supported claim that is fully detailed and verifiable based on the evidence, your task is to generate a "Not-Enough-Info" (NEI) claim by intentionally omitting or generalizing the key verifiable elements. Replace precise numerical figures and explicit qualifiers with vague, qualitative, or uncertain language, so that the claim no longer contains enough concrete information for conclusive verification.

For example:
Supported Claim: The study shows a 70% reduction in hospitalization rates.
Not-Enough-Info Claim: The study suggests a significant reduction in hospitalization rates.

Now, please generate a Not-Enough-Info claim for the following input:

Supported Claim: [Insert Supported Claim Here]
Evidence: [Insert Evidence Here]

Output your response in the following format:
Not-Enough-Info Claim: [Your generated Not-Enough-Info claim here]

Table 18: The prompt for generating the Not-Enough-Info claim.

You are a highly skilled claim verification expert. You are given a claim and evidence. Your task is to verify the claim based solely on the evidence provided. Analyze the claim and evaluate the evidence. Based on your evaluation, return one of the following labels: "supports", "refutes" or "not enough info".

Evidence: {EVIDENCE}
Claim: {CLAIM}
Label: {LABEL}

Table 19: The prompt used for zero-shot evaluation.

You are a highly skilled claim verification expert. You are given a claim and evidence. Your task is to verify the claim based solely on the evidence provided. Analyze the claim and evaluate the evidence. Based on your evaluation, return one of the following labels: "supports", "refutes" or "not enough info". Let's think step by step.

Evidence: {EVIDENCE}
Claim: {CLAIM}
Label: {LABEL}

Table 20: The prompt used for zero-shot chain-of-thought evaluation.

You are an expert academic writer in the field of claim verification. Given the following domain:

Domain: [Domain Name]
Please generate a concise bullet-point plan for a synthetic evidence document in this domain.

Table 21: The prompt for the ablation study of removing knowledge domain Specification.

You are an expert academic writer. Generate an evidence document based on the following knowledge domain specification. Ensure that the final document is coherent, detailed, and written in a formal academic style.

Domain: [Domain Name]
Description: [A concise description (1–2 sentences) highlighting its unique linguistic and factual characteristics and explaining why claims in this domain are challenging to verify.]
Properties: [Property 1; Property 2; Property 3]
Please generate the complete evidence document following this structure.

Evidence Document:

Table 22: The prompt for the ablation study of removing evidence plan generation.

You are a domain expert in claim verification. Based on the evidence document provided below, generate a supported claim that is factually consistent with the evidence.

Evidence:[Insert the evidence document here.]

Output Format:
Supported Claim: [Your claim here]

Table 23: The prompt for the ablation study of removing key aspect extraction.

You are a domain expert in claim verification. Given a predefine domain, generate one evidence document and three claims about that domain so that their labels are “supports”, “refutes” and “not enough info”, respectively.

Domain: [Domain name]

Output Format:
Evidence Document: [Evidence Document]
Supported Claim: [Supported Claim]
Refuted Claim: [Refuted Claim]
Not-Enough-Info Claim: [Not-Enough-Info Claim]

Table 24: The prompt for the ablation study of the direct prompting method.