Can LLMs Automate Fact-Checking Article Writing?

Dhruv Sahnan[1] **David Corney**[2] **Irene Larraz**[3] **Giovanni Zagni**[4]
Ruben Miguez[3] **Zhuohan Xie**[1] **Iryna Gurevych**[1,5] **Elizabeth Churchill**[1]
Tanmoy Chakraborty[6] **Preslav Nakov**[1]

[1] MBZUAI, UAE [2] Full Fact, UK [3] Newtral, Spain [4] Pagella Politica, Italy
[5] TU Darmstadt, Germany [6] IIT Delhi, India

dhruv.sahnan@mbzuai.ac.ae david.corney@fullfact.org
irene.larraz@newtral.es ruben.miguez@newtral.es
zhuohan.xie@mbzuai.ac.ae churchill@acm.org
tanchak@iitd.ac.in preslav.nakov@mbzuai.ac.ae

Abstract

Automatic fact-checking aims to support professional fact-checkers by offering tools that can help speed up manual fact-checking. Yet, existing frameworks fail to address the key step of producing output suitable for broader dissemination to the general public: While human fact-checkers communicate their findings through fact-checking articles, automated systems typically produce little or no justification for their assessments. Here, we aim to bridge this gap. In particular, we argue for the need to extend the typical automatic fact-checking pipeline with *automatic generation of full fact-checking articles*. We first identify key desiderata for such articles through a series of interviews with experts from leading fact-checking organizations. We then develop QRAFT, an LLM-based agentic framework that mimics the writing workflow of human fact-checkers. Finally, we assess the practical usefulness of QRAFT through human evaluations with professional fact-checkers. Our evaluation shows that while QRAFT outperforms several previously proposed text-generation approaches, it lags considerably behind expert-written articles. We hope that our work will enable further research in this new and important direction. The code for our implementation is available at https://github.com/mbzuai-nlp/qraft.git.

1 Introduction

According to a March 2024 survey (Pew Research Center, 2024), nearly 80% of American adults on major social media platforms regularly encounter news-related content. While these platforms allow the rapid spread of *breaking news*, they have also been heavily misused to circulate dubious claims.

In response, the role of fact-checkers has grown to be increasingly vital. However, manual fact-checking efforts cannot match the scale of global dis/misinformation, prompting a push towards automating the process (Vlachos and Riedel, 2014). As shown in Figure 1, most existing studies frame automatic fact-checking to comprise five main steps: (*i*) check-worthy claim detection, (*ii*) verified claim retrieval, (*iii*) evidence retrieval, (*iv*) claim verification, and (*v*) brief explanation generation (Vlachos and Riedel, 2014; Nakov et al., 2021a,b; Guo et al., 2022). Note that some studies skip (*i*), (*ii*), and (*v*).

In practice, fact-checkers do more than just verifying claims; they also publish detailed *fact-checking articles* that guide readers toward a clear understanding of a claim by presenting factual arguments and explaining how they lead to the final verdict (Graves, 2017). Yet existing automatic fact-checking pipelines fail to account for this key time-consuming step of the manual fact-checking workflow. One may argue that *brief explanation generation* serves as the automated counterpart to fact-checking article writing; however, expert requirements for high-quality articles transcend the traditional definition of fact-checking explanations (Warren et al., 2025). Such explanations are typically concise and aim to justify the factuality decision (Kotonya and Toni, 2020). In contrast, fact-checking articles usually aim not just to provide the evidence leading to the verdict, but also to give the readers the necessary background and contextual information, to explain the origins of the claim, to clarify all likely interpretations of the claim, to address counter-arguments, etc. We thus argue that the typical automatic fact-checking pipeline needs to be extended with an extra step, *automatic*

Transactions of the Association for Computational Linguistics, vol. 14, pp. 489–509, 2026. https://doi.org/10.1162/TACL.a.644
Action Editor: Dan Goldwasser. Submission batch: 3/2025; Revision batch: 7/2025; Published 4/2026.

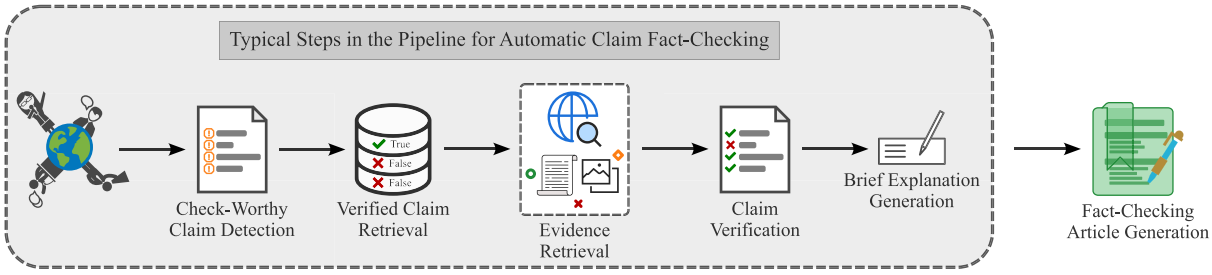

Figure 1: **Our proposed pipeline for automatic fact-checking.** We extend the typical steps in the automatic claim fact-checking pipeline to include a new task: *fact-checking article generation*.

generation of a fact-checking article, as shown in Figure 1.

Moreover, it has repeatedly been highlighted that automatic fact-checking overlooks insights from professional fact-checkers (Nakov et al., 2021a; Juneja and Mitra, 2022; Das et al., 2023). Thus, we study the task of generating fact-checking articles in close collaboration with experts from several world-leading fact-checking organizations, aiming to understand the key elements of a good fact-checking article. We further provide a brief analysis of why professional fact-checkers generally do not trust Artificial Intelligence (AI) to fully automate this task. Finally, based on what we have learned from this study, we develop **QRAFT**, a framework using multiple AI agents that collaborate to write and iteratively refine a fact-checking article, designed as the first attempt at a comprehensive solution to cater to the requirements highlighted by the experts.

Our framework leverages large language models (LLMs) as its foundation, as they can generate fluent, long-form text (Yang et al., 2023; Xie et al., 2023; Wang et al., 2025). Recent studies have highlighted that agentic workflows can further enhance performance, particularly for tasks involving the generation of documents grounded in a set of facts (Huot et al., 2025; Shao et al., 2024; Wang et al., 2024a). Thus, we design QRAFT as a multi-agent collaboration that mimics the fact-checking article writing process of human experts. QRAFT breaks the writing process down into two main stages. In the first stage, QRAFT gathers evidence nuggets relevant to the claim, formulates an outline, and then populates it to produce an initial draft. In the second stage, QRAFT simulates an editorial review that uses conversational question-answering interactions between LLM agents to formulate a list of edits to refine the draft and to ensure professional standards of writing.

We further benchmark QRAFT against several state-of-the-art text generation frameworks using automatic evaluation strategies.

In addition, we conduct evaluations with professional fact-checkers focusing on assessing QRAFT's real-world usefulness. We find that QRAFT shows better performance on all automatic metrics; yet, the human evaluations reveal key limitations, indicating the need for constant expert supervision when using LLM-based frameworks for this task.

To sum up, our contributions are as follows:

- We introduce the novel task of automatically generating full fact-checking articles, working in close collaboration with experts from leading fact-checking organizations to enrich the task with their insights.
- We propose **QRAFT**, a unique framework designed to generate a fact-checking article given (*i*) a claim, (*ii*) its veracity, and (*iii*) a set of evidence documents; we further show that QRAFT outperforms several generic text generation approaches.
- We conduct expert evaluations to assess the practical usefulness of **QRAFT** and find that it falls short of expert-written articles. We also discuss the key limitations as highlighted by experts to facilitate future research.

2 What Expert Fact-Checkers Want from a Fact-Checking Article

> *''A fact-checking article for a claim is a form of communication from a fact-checker to the public that provides the necessary context around the claim, argues its veracity, and explains why it may or may not be exactly as claimed.''*
> – An experienced fact-checker

It is crucial for fact-checkers to maintain public trust in the fact-checking process, which means that they need to be extremely careful when reporting data towards fact-checking a claim.

To support this, the International Fact-Checking Network has developed a *fact-checkers' code of principles*[1] to streamline the fact-checking process across organizations. However, fact-checking article writing guidelines still vary between organizations, and individual fact-checkers have their own unique writing styles. In order to further refine our perception of how fact-checking articles must be composed, we interviewed four experts from world-leading fact-checking organizations. To maintain the experts' anonymity, we refer to them as P1, P2, P3, and P4.

2.1 Expert Desiderata for the Articles

We focused our interviews on gathering insights into expert's expectations for a fact-checking article, details that must be present, and preferences about the article's structure and writing style. Subsequently, we extracted key insights from the responses and categorized them into distinct *characteristics* of a fact-checking article.

Consistent with Graves (2017), we found that the foremost expectation of a fact-checking article is that it *accurately clarifies* all nuances of the claim, with arguments on how the evidence supports or refutes it, and how they lead to the verdict. The experts also highlighted that the article must present details on the *origins* of the claim, which includes the necessary background information, details on the context in which it was made, and its implied meaning(s). Moreover, they emphasized that the article must be written in a *transparent* manner. P3 elaborated that all arguments made in the article must be supported by publicly available sources, so that readers can verify by themselves that no argument misrepresents information from the original evidence. These insights echo those by Warren et al. (2025); however, their work focuses on how fact-checkers construct explanations for a claim's veracity in a fact-checking article. We further expand these findings to cover the composition of the full fact-checking article.

Adding details toward the *structure* of the fact-checking articles, P2 explained that it varied across organizations, but an article usually began with some background on the claim, including key information regarding the context in which it was spreading. This is followed by justifications for all likely interpretations of the claim and its proposed veracity assessment, before concluding the article with a summary of the findings.

P1 further revealed that some fact-checking organizations require fact-checkers to specify why the claim was *important* to fact-check. Table 1 presents a list of the identified characteristics along with direct quotes from the experts, offering a concise overview of their expectations for a fact-checking article. The questions we used for the interviews are given in Appendix A.

Characteristic	Comments
Accurately clarifies the claim	**P3** →''*We must have clarity regarding the veracity of the claim and understand the reasons behind it after reading the article.*'' **P2** →''*The article must accurately convey why the evidence contradicts the claim, providing all the necessary context.*''
Origin of the claim	**P4** →''*The article must present the context in which the claim was made, what it meant, and how it affects world events.*''
Transparent writing style	**P3** →''*We must be able to verify everything by consulting the listed sources by ourselves.*''
Structure	**P2** →''*It varies, but generally it begins by introducing the claim, followed by arguments towards its veracity, and ends with a conclusion of our findings.*''
Importance of the fact-check	**P1** →''*Why is the claim worth fact-checking?*'' **P1** →''*The article gives context about where the claim was spreading, its harm potential and why it is important to fact-check.*''

Table 1: Summary of the important characteristics of fact-checking articles, according to human experts, along with direct quotes from our interviews with them, explaining what they expected from an article.

2.2 Can AI Do It?

We also asked the experts whether they had used AI to assist them in writing fact-checking articles. Most were unaware of any suitable tools for that. Those who had used general-purpose AI systems said that they had certain limitations, which required human intervention. Therefore, we collected expert opinions about the limitations they foresaw. We summarized the key points from their responses and categorized them into anticipated

[1] ifcncodeofprinciples.poynter.org.

Issue	Comments
Hallucinations	**P1** →''*LLMs make up arguments and non-existent URLs just to satisfy the veracity of the claim.*''
Lack of world-knowledge	**P4** →''*AI does not have access to world-knowledge like humans do, and thus understanding the relevance of some evidence can be tough.*''
Unable to capture context	**P3** →''*AI may not be able to put the claim into context because it does not distinguish between what was said and what was implicitly meant.*''
Evidence presentation	**P3** →''*It is hard for AI to have all the context to understand the data and knead it together to argue the claim; instead, it would just give a summary of the evidence as an explanation of the claim's veracity.*''

Table 2: Reasons why experts do not trust an LLM-based system for generating fact-checking articles, together with some direct quotes from our interviews to further detail these issues.

''issues'' concerning an AI system's capability to perform this task.

Notably, many of these issues align with well-known challenges associated with LLMs, e.g., all experts stated that LLMs were known to *hallucinate* content to make the generated text sound convincing (Huang et al., 2025b). P1 mentioned that in his experience, LLMs would make up baseless arguments to satisfy their assessment of the claim's veracity, even though the evidence clearly stated otherwise. P3 echoed the same concern stating that LLMs cited non-existing URLs as the source for their made up arguments (Zuccon et al., 2023).

Experts also highlighted issues such as LLMs *lacking world knowledge* and likely not being able to *capture the context* in which a claim was made, despite exhaustive evidence being provided (Ko et al., 2024). Additionally, P3 mentioned a key shortcoming: He expected AI systems to be unable to construct correct arguments towards the claim's veracity. In his experience, AI systems had equated justifying the claim's veracity to a summary of the evidence sources. However, a fact-checking article must communicate why a claim was assigned a given veracity label, instead of merely summarizing all of the evidence that consulted to arrive with this veracity label.

Table 2 lists all identified issues, along with direct quotes from fact-checkers providing a brief explanation for their concerns.

3 Generating Fact-Checking Articles

3.1 Task Definition

Given a datapoint $X = \{C, V, E_{C,V}\}$, where C represents a claim, V is its veracity, and $E_{C,V}$ is the evidence set consulted for fact-checking the claim, we aim to generate a fact-checking article, $D_{C,V}$, corresponding to C and V.

We assume that the datapoint X is explicitly provided by the fact-checker, and we do not address automatic evidence retrieval or claim verification here. For evaluating the quality of the generated article, we further have a ground-truth expert-written reference article $D^{gt}_{C,V}$.

3.2 Methodology

Experienced writers generally approach any long-form expository writing task by first gathering relevant pieces of evidence (evidence nuggets) from various sources. These nuggets allow the writers to create an outline that defines the overall structure of their draft and specifies the data to be included in each section (Rohman, 1965). This is followed by populating the outline with content, forming the initial draft. Subsequently, the draft is subjected to proofreading and copy-editing, which are critical steps to refine the writing quality before publication for public consumption (Wates and Campbell, 2007).

Here we have a similar approach. We propose QRAFT, an agentic pipeline that replicates the human process of organizing and structuring information for the writing task. QRAFT uses conversational interactions between three AI agents as the backbone of the pipeline:

- *Planner* $\boldsymbol{\mathcal{P}}$, which plays the role of an *AI assistant*, responsible for gathering evidence nuggets from a given evidence set and planning an outline for the draft.
- *Writer* $\boldsymbol{\mathcal{W}}$, which plays the role of a *human fact-checker*, responsible for composing a draft for the fact-checking article.
- *Editor* $\boldsymbol{\mathcal{E}}$, which plays the role of a *human expert editor*, responsible for reviewing and helping to refine the drafts composed by $\boldsymbol{\mathcal{W}}$.

We illustrate the workflow of QRAFT in Figure 2 and Figure 3, which show the two main stages:

(a) Planning and Compiling the First Draft, where the article's structure is defined and an initial draft is composed, and *(b) Simulating Editorial Review*, where the draft is iteratively refined

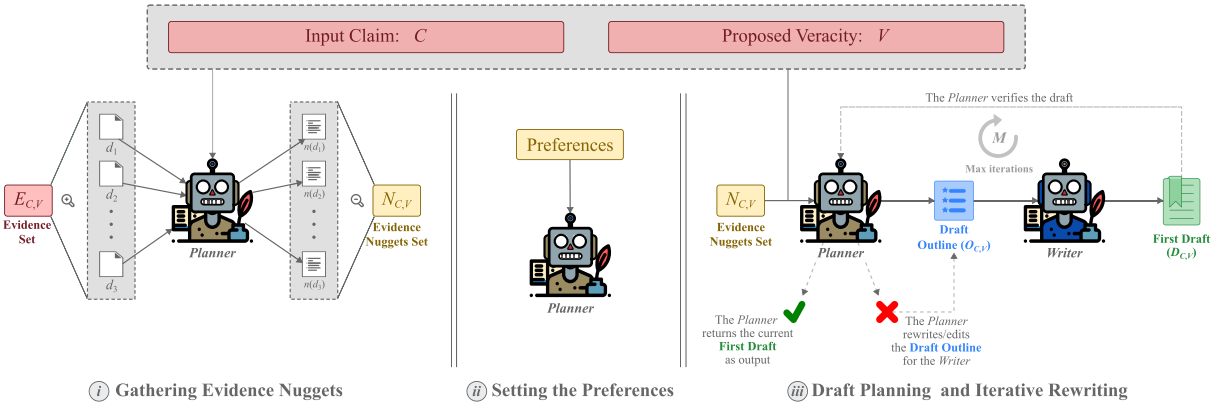

Figure 2: **Workflow of QRAFT: stage (a) – *planning and compiling the first draft*.** We use two agents, *Planner* and *Writer*, and we split this stage into three steps: *(i) Gathering Evidence Nuggets*, *(ii) Setting the Preferences*, and *(iii) Draft Planning and Iterative Rewriting*. See §3.2 for more detail.

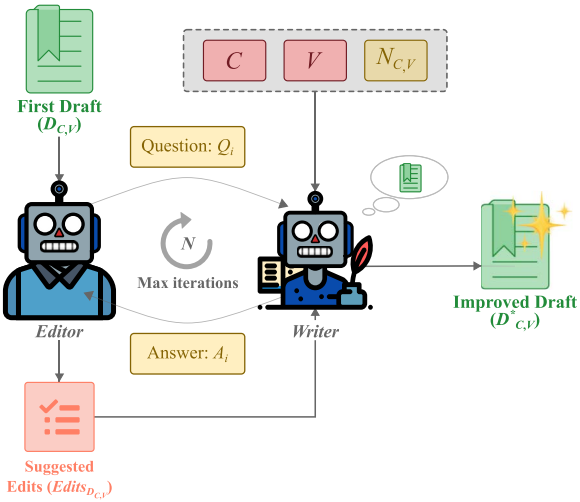

Figure 3: **Workflow of QRAFT: stage (b) – *simulating an editorial review*.** We simulate conversational question-answering between the *Editor* and the *Writer* in order to generate feedback on how to improve the first draft from QRAFT's stage (a).

using feedback from simulated writer ↔ editor interactions. We present pseudocode for the implementation of QRAFT in Appendix B.

(a) Planning and Compiling the First Draft

We further break this stage into three broad steps as shown in Figure 2, which we detail below.

(i) Gathering Evidence Nuggets. In this step, the *Planner* $\mathcal{P}$ extracts key evidence relevant to the claim C and its veracity V from each document $d \in E_{C,V}$ and summarizes it into concise bullet points, denoted by $n(d)$. This results in the formation of a set of evidence nuggets $N_{C,V}$:

$$N_{C,V} = \{n(d) | \forall\, d \in E_{C,V}\}$$

(ii) Setting Preferences. In order to encourage QRAFT to generate articles that align with the *characteristics* of a fact-check article outlined in §2.1, we design a list of preferences comprising high-level guidelines about the expected structure and content of the draft. We provide these preferences to $\mathcal{P}$ in the form of an instruction prompt.

(iii) Draft Planning and Writing. Here, the *Planner* $\mathcal{P}$ first proposes an outline, $O_{C,V}$, aligning with the preferences specified earlier, using $N_{C,V}$ as the source of evidence nuggets relevant to the claim. The *Writer* $\mathcal{W}$ then uses C, V, and $N_{C,V}$ to populate $O_{C,V}$, resulting in the first draft, $D_{C,V}$. Note that $\mathcal{W}$ does not access the preferences directly; instead, $\mathcal{P}$ implicitly encodes them into the outline $O_{C,V}$. This design simplifies the *Writer*'s task, allowing it to focus solely on expanding the outline into an article, without needing to redundantly process the preferences again.

Afterwards, the *Planner* $\mathcal{P}$ verifies whether the compiled draft $D_{C,V}$ aligns with the preferences set in the previous step. Based on this assessment, $\mathcal{P}$ may either approve $D_{C,V}$ for the second stage of QRAFT's workflow or it may self-reflect and refine $O_{C,V}$ to ensure that a draft constructed over a modified outline adheres more closely to our preferences. The interactions in this step are repeated until $\mathcal{P}$ approves the draft $D_{C,V}$ or for a maximum of $M = 5$ iterations.

(b) Simulating an Editorial Review

As illustrated in Figure 3, in this stage, we simulate conversational interactions between the *Editor* $\mathcal{E}$ and the *Writer* $\mathcal{W}$, to derive feedback for

improving the first draft $D_{C,V}$. $\mathcal{E}$ begins by identifying unclear parts in the first draft $D_{C,V}$ and formulating questions for $\mathcal{W}$ seeking clarification.

This initiates a conversational question-answering based interaction session between the two LLM agents. As part of each interaction i, $\mathcal{E}$ generates a question Q_i based on $D_{C,V}$ and the interaction history $\{Q_1, A_1, \ldots, Q_{i-1}, A_{i-1}\}$. $\mathcal{W}$ then responds with an answer A_i for Q_i. The session is conducted for a total of $N = 10$ such interactions, after which $\mathcal{E}$ generates a list of suggested edits $Edits_{D_{C,V}}$ to improve the draft. Once the interaction session concludes, $\mathcal{W}$ receives $Edits_{D_{C,V}}$ and applies it on $D_{C,V}$ to produce the improved draft $D^*_{C,V}$. The entire workflow of this stage is repeated on the improved draft $D^*_{C,V}$ returned at the end of each cycle. The iterative process continues until the editor determines that no further edits are required, or until a maximum of $K = 5$ improvement cycles have been completed.

For the purpose of our experiments, we use `GPT-4o-mini` for Planner $\mathcal{P}$ and Writer $\mathcal{W}$. The Editor $\mathcal{E}$, who is responsible for guiding the Writer to improve its draft, is instantiated using `GPT-4o`. This configuration balances efficiency and performance by using a lightweight low-cost model for the resource-intensive task of long-form draft generation, and reserving a more capable LLM for the editorial phase, where high-quality critique is crucial.

4 Experimental Setup

4.1 Datasets

We used two datasets, which we merged into one:

ExClaimCheck (Zeng and Gao, 2024) contains claims from publicly accessible fact-checking websites (published between November 2008 and March 2021), together with a veracity label, a set of webpages used as evidence, and the full expert-written fact-checking article. We used 987 examples from its test set, for which the fulltext of the evidence articles was available.

AmbiguousSnopes is a fresh and hard dataset that we curated. It comprises 240 claims, along with their veracity labels, evidence documents, and the expert-written fact-checking article from *Snopes*[2] (published between October 2022 and August 2024). We only kept hard claims that could not be judged outright as completely True or completely False. This was suggested by the experts: that a real-world system must also excel at claims that are challenging even for expert fact-checkers.

[2] `https://www.snopes.com`.

Dataset	# Examples	Veracity Labels	# Labels
ExClaimCheck	987	*True* (67), *Mostly True* (125), *Half True* (137), *Barely True* (132), *False* (380), *Pants on Fire* (64), *Mostly False* (1), *Partly False* (8), *Mixture* (1), *Misleading* (49), *Mixed* (1), *Missing Context* (10), *Faux* (1), *Distorts the Facts* (2), *Out of Context* (2), *False Attribution* (1), *No evidence* (2), *Satire* (4).	18
AmbiguousSnopes	240	*Mostly True* (16), *Mixture* (118), *Mostly False* (24), *Misattributed* (28), *Scam* (14), *Correct Attribution* (34), *Legit* (6).	7
Total	**1,227**	***True* (67), *Mostly True* (141), *Half True* (137), *Barely True* (132), *False* (380), *Pants on Fire* (64), *Mostly False* (25), *Partly False* (8), *Mixture* (119), *Misleading* (49), *Mixed* (1), *Missing Context* (10), *Misattributed* (28), *Faux* (1), *Correct Attribution* (34), *False Attribution* (1), *Scam* (14), *Legit* (6), *Out of Context* (2), *No evidence* (2),Satire (4), *Distorts the Facts* (2).**	**22**

Table 3: **Statistics about the *test* dataset we used to evaluate QRAFT:** number of examples and veracity labels (with frequency of each such label, shown in parentheses).

Table 3 gives statistics about the datasets and the inventory of labels they used. Below, we report evaluation results on the union of the two datasets: a total of 1,227 examples using 22 veracity labels.

4.2 Baselines

We compare QRAFT to several text-generation approaches. To the best of our knowledge, there is no prior work on generating entire fact-checking articles. Therefore, we compare to several prior methods and baselines as described below.

Naïve. We collected the top-k semantically similar sentences (to the claim C) from each document in the evidence set, then we concatenated them to form a candidate fact-checking ‘‘article.’’ We set $k = 3$ for the purpose of this experiment.

Summarization. Several previous studies modeled the generation of fact-checking *explanations* as a text summarization task (Atanasova et al., 2020; Xing et al., 2022), and thus wanted to try this as a baseline for generating fact-checking *articles*. In particular, we fine-tuned *PRIMERA* (Xiao et al., 2022) using the evidence webpages as the source

and the corresponding fact-checking article as the target (for this, we used additional 5,964 examples from the training split of *ExClaimCheck*).

Justification. We used *JustiLM* (Zeng and Gao, 2024), a preexisting method from the literature for justification generation, which uses Atlas (Izacard et al., 2023) as the base large language model. We modified it to generate full fact-checking articles (it uses a mini-fine-tuning on 30 examples randomly sampled from training) to align with the requirements of our task.

LLM-Based Approaches. We used the following two LLM-based approaches: (*i*) *Vanilla-GPT*, where we prompted GPT-4o-mini (OpenAI, 2023) to generate a full fact-checking article given the input in a zero-shot setting, and (*ii*) *Storm* (Shao et al., 2024), which is an agentic long-form text generation pipeline, designed to write Wikipedia-like articles from scratch. We retained the workflow of Storm, but we modified it to generate fact-checking articles by customizing the prompts in the underlying components, and restricting the retrieval component to use the evidence documents from our dataset.

5 Automatic Evaluation

Below, we report the evaluation results for our QRAFT framework in comparison to the methods described in the previous section; we also perform ablation studies.

Our evaluation focuses on assessing whether the generated text aligns with the characteristics of the fact-checking articles outlined in §2.1 and whether it exhibits traits that exacerbate the experts' distrust in AI, as identified in §2.2.

5.1 Evaluation Measures

We used several automatic evaluation measures, namely, *ROUGE*[3] (Lin, 2004) and *BERTScore* (Zhang et al., 2020) to measure the lexical and the semantic similarity of the generated full fact-checking articles with respect to the references, and *FactScore*[4] (Min et al., 2023) for evaluating their factuality.

We further used the percentage of *Hallucinated Citations*, which calculates the proportion of URLs cited in the text that do not exist in the claim's evidence set, and *Entailment Score*, which measures the consistency and the coverage of the generated fact-checking article $D_{C,V}$ with respect to the ground-truth $D^{gt}_{C,V}$ (Zeng and Gao, 2024). The latter measure calculates the mean of SummaC scores (Laban et al., 2022) over the ordered pairs $(D_{C,V}, D^{gt}_{C,V})$ and $(D^{gt}_{C,V}, D_{C,V})$.

Moreover, we performed LLM-as-a-judge evaluations (Zheng et al., 2023) on the generated articles, assessing them with respect to *Relevance*, *Comprehensibility*, *Importance*, and *Evidence presentation*. For this, we collaborated with the fact-checking experts to design five-point rubrics to score the generated articles on each of these aspects (see Appendix C for rubric details). Since we evaluated on custom criteria, we selected Prometheus-2 (Kim et al., 2024), an LLM tuned for such assessments, as the judge.

5.2 Performance Comparison

Automatic Evaluation. Table 4 presents a performance comparison of various frameworks across multiple evaluation measures. We can see that LLM-based approaches consistently achieve better performance.

QRAFT outperforms the other LLM-based approaches, surpassing them on 6 out of 7 evaluation measures. Notably, QRAFT achieves the highest FactScore, demonstrating 11 points of improvement over the next-best method, while also having the least amount of hallucinated citations. *Storm* emerges as the best baseline with scores within 3 points of the highest on each measure, except for FactScore. *Vanilla-GPT* follows closely behind, maintaining competitive scores; however, it exhibits a relatively higher number of hallucinated citations, and underperforms both on BERTScore and ROUGE. On deeper analysis, we also find that it cites only about 30% of the available evidence sources on average, whereas QRAFT cites more than 90% of all evidence sources in the article it generates. Moreover, from Table 5, we can see that models maintain similar performance across both datasets. *Vanilla-GPT* produces a relatively much higher proportion of hallucinated citations on *AmbiguousSnopes* at 3.93%—compared to 1.89% on *ExClaimCheck*—while *Storm* shows degradation in the Entailment Score on *AmbiguousSnopes*. QRAFT largely remains stable across both datasets, with slight degradations in performance on *AmbiguousSnopes*.

[3] We report ROUGE-1, ROUGE-L, and ROUGE-Lsum.

[4] For FactScore, we restricted the knowledge base to the expert-written articles from our evaluation datasets.

Method	ROUGE ↑			BERTScore ↑	Entailment Score ↑	FactScore ↑	Hallucinated Citations (%) ↓
	R_1	R_L	R_{Lsum}				
Naïve Top-k	0.20	0.09	0.09	0.83	0.20	***NA***	***NA***
PRIMERA	0.26	0.11	0.11	0.83	0.14	0.20	***NA***
JustiLM	0.07	0.05	0.05	0.80	0.23	0.55	***NA***
Vanilla-GPT	0.29	0.13	0.16	0.79	**0.34**	0.74	2.30
Storm	0.35	0.13	0.20	0.82	0.32	0.72	1.63
QRAFT(a)	0.36	0.13	0.19	0.84	0.29	0.81	3.51
– w/o evidence nuggets	0.32	0.13	0.18	0.83	0.28	0.80	3.62
– w/o draft verification	0.31	0.13	0.18	0.84	0.29	0.81	†
– w/o outline	0.31	0.12	0.18	0.84	0.27	0.79	3.33
– w/o preferences	0.34	0.13	0.18	**0.85**	0.27	0.81	†
QRAFT(a) + (b)	**0.38**	**0.14**	**0.21**	**0.85**	0.30	**0.83**	**1.29**
– w/o question-asking	0.36	**0.14**	0.20	**0.85**	0.29	0.81	1.94

Table 4: **Automatic evaluation.** ↑ and ↓ indicate that a higher or a lower score is better, respectively. ***NA*** denotes that the measure is not applicable for that method, and † signals the absence of in-text citations. A **bold** score represents the best performance for an evaluation measure, while an **underlined bold** is a tie for the best results.

Method	ExClaimCheck			AmbiguousSnopes		
	ES ↑	FS ↑	HC (%) ↓	ES ↑	FS ↑	HC(%) ↓
Naïve Top-k	0.21	***NA***	***NA***	0.17	***NA***	***NA***
PRIMERA	0.12	0.19	***NA***	0.23	0.22	***NA***
JustiLM	0.24	0.56	***NA***	0.19	0.51	***NA***
Vanilla-GPT	**0.34**	0.74	1.89	**0.33**	0.72	3.93
Storm	**0.34**	0.73	1.58	0.26	0.71	1.83
QRAFT(a)	0.30	0.81	3.86	0.27	**0.81**	2.09
QRAFT(a) + (b)	0.31	**0.85**	**1.25**	0.29	**0.81**	**1.46**

Table 5: **Automatic evaluation on individual datasets.** ***NA*** denotes that the measure is not applicable for that method. A **bold** score represents the best performance for an evaluation measure.

Method	Rel	Com	Imp	Evi
Naïve Top-k	2.31	1.77	1.96	1.60
PRIMERA	1.13	1.09	1.21	1.15
JustiLM	1.21	1.14	1.07	1.05
Vanilla-GPT	4.31	4.33	4.21	4.11
Storm	4.65	4.11	4.13	4.32
QRAFT(a)	4.52	4.15	4.49	4.25
QRAFT(a) + (b)	**4.70**	**4.77**	**4.54**	**4.56**

Table 6: **LLM-as-a-judge evaluation.** We report scores on a five-point scale for four aspects: *Relevance* (Rel), *Comprehensibility* (Com), *Importance* (Imp), and *Evidence presentation* (Evi).

LLM-as-a-Judge Evaluation. Table 6 presents results for LLM-as-a-judge evaluations on the four aspects we discussed above. These experiments reinforce the trend that LLM-based frameworks outperform other baselines on this task.

Storm outperforms *Vanilla-GPT*, benefiting from its multi-perspective question-asking approach. This lets *Storm* conduct a more thorough analysis, leading to a clearer explanation of the claim and detailed, but sometimes overly elaborate, evidence presentation.

Regardless of these strengths, QRAFT still emerges as the preferred method, demonstrating the best scores across all four aspects. The editorial review simulated in the second stage of our framework enables it to focus on clarifying the claim by arguing its truthfulness with evidence without presenting unnecessary details.

5.3 Ablation Studies

Considering that QRAFT decomposes the process of generating fact-checking articles into multiple steps (§3), it is natural to question the need for each step. Tables 4 to 7 include some ablation experiments, which we discuss below.[5]

QRAFT's Stages. We observed that dropping stage (b) from our QRAFT framework led to degradation across all automatic measures, including aspects of the articles assessed through LLM-as-a-judge evaluations. Most notably, there is a 2.7 times increase in the proportion of hallucinated citations and 0.62 point absolute drop in *Comprehensibility*, which is likely due to the absence of an *Editor* to help the *Writer* clarify and refine its arguments. However, it is important to highlight that our QRAFT framework (a) is still on par with the best baseline on most aspects, and even outperforms it on FactScore by 0.09.

Steps Within Each Stage. We assessed four variations of QRAFT(a): (*i*) *w/o evidence nuggets*,

[5]In Tables 5 and 7, **ES** denotes Entailment Score; **FS** denotes FactScore; and **HC** denotes Hallucinated Citations.

Method	Avg. Cost	ES ↑	FS ↑	HC (%) ↓
QRAFT(a)				
– draft verification				
$M = 0$	\$0.0036	0.29	0.81	†
$M = 1$	\$0.0050	**0.30**	0.80	4.17
$M = 3$	\$0.0064	0.28	0.81	7.03
$M = 5$	\$0.0096	0.29	0.81	3.51
$M = 7$	\$0.0125	0.29	0.81	4.04
QRAFT(a) + (b)				
– question-asking				
$N = 0$	\$0.049	0.29	0.81	1.94
$N = 3$	\$0.054	0.27	0.82	1.93
$N = 7$	\$0.058	0.29	**0.83**	1.97
$N = 10$	\$0.063	**0.30**	**0.83**	1.29
$N = 12$	\$0.067	**0.30**	**0.83**	1.40
$N = 15$	\$0.070	0.29	**0.83**	**1.23**
– editorial review				
$K = 1$	\$0.012	0.27	0.81	1.72
$K = 3$	\$0.037	0.28	0.82	1.52
$K = 5$	\$0.063	**0.30**	**0.83**	1.29
$K = 7$	\$0.076	**0.30**	**0.83**	1.36

Table 7: **Cost and efficiency analysis.** † signals the absence of in-text citations. A **bold** score represents the best performance for an evaluation measure.

where we skip the compression of evidence into concise nuggets, (*ii*) *w/o draft verification*, where the *Planner* does not verify the draft compiled by the *Writer*, (*iii*) *w/o outline*, where the *Writer* generates the draft without an outline, and (*iv*) *w/o preferences*, where no preferences are provided.

We observe that each of these variations showed slight degradations across all evaluation measures. Interestingly, dropping preferences or draft verification resulted in the complete absence of any in-text citations using URLs. On closer examination, we found that the evidence was instead cited using webpage titles without URLs, leaving no way for the reader to verify whether the information was correctly presented. Moreover, with respect to QRAFT(b), we found that using it *w/o question-asking* yields to slight degradation across all evaluation measures with a notable increase in the hallucinated citations.

Cost and Efficiency. Since QRAFT comprises multiple cycles of LLM invocations with long-form articles, we assessed the average cost of the framework against the quality of the generated articles across three variables: the number of *draft verification* cycles M, the number of *question-asking* interactions N, and the number of *editorial review* cycles K.[6]

[6]We fixed $M = 5$ and $K = 5$ while varying N, and we set $M = 5$ and $N = 10$ while varying K.

We can see that for QRAFT(a), increasing M raises the cost per generated article gradually from \$0.003 to \$0.012, offering a 1.2–2× reduction in the proportion of hallucinated citations, while the performance on other metrics remains stable. Introducing QRAFT(b) incurs higher costs—up to \$0.076—but leads to substantial performance gains, especially, a much lower percentage of hallucinated citations (a best of 1.23%).

We note that varying the number of question-asking interactions (N) is relatively inexpensive, and provides improvements in the form of higher factual accuracy and a reduced amount of hallucinated citations. Increasing the number of editorial review cycles (K) is costlier, but yields consistent performance gains across all metrics. We choose $M = 5$, $N = 10$, and $K = 5$ as the default configuration for QRAFT because this setup produces high-quality articles while maintaining cost-efficiency.

6 Expert Evaluation

While automatic evaluations indicate that QRAFT outperforms existing methods, they are insufficient to assess the real-world usefulness of the generated fact-checking articles for expert fact-checkers. Several studies have demonstrated ROUGE and BERTScore to be less reliable in single-reference settings (Sheng et al., 2024). In particular, due to their reliance on lexical or semantic similarity-based text overlap, their ability to capture the diversity of expression in text is limited. Perfectly acceptable fact-checking articles can be written for the same claim in multiple different ways. However, because there is only one human-written reference per claim in the dataset, these metrics could unfairly treat generated articles as of lower quality, solely due to stylistic differences. FactScore also exhibits several limitations, such as overestimating factuality in certain cases (Chiang and Lee, 2024), or assuming all claims in the generated text are verifiable (Song et al., 2024). These limitations, combined with the fact that FactScore is specifically optimized to assess the factuality of generated biographies using Wikipedia as evidence, render the metric insufficient to evaluate the performance of LLM-based frameworks on our task. Moreover, LLM-based evaluation is not infallible: LLMs are known to be biased towards LLM-generated text assigning them higher scores (Ye et al., 2020).

Method	Rel	Com	Imp	Evi	Pub
Vanilla-GPT	4.41	3.91	2.83	3.62	3.16
Storm	3.75	3.50	3.41	3.29	2.00
QRAFT(a) + (b)	4.47	4.16	**3.75**	3.45	3.25
Expert	**5.00**	**4.66**	3.25	**4.79**	**4.83**

Table 8: **Expert evaluation.** We report average scores on a five-point scale for five aspects: *Relevance* (Rel), *Comprehensibility* (Com), *Importance* (Imp), *Evidence presentation* (Evi), and *Publishability* (Pub).

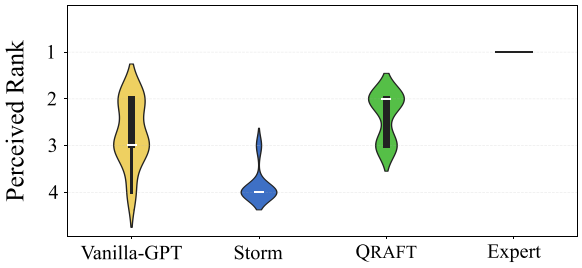

Figure 4: Distribution of the rankings received by the four fact-check articles based on relative publishability as perceived by the experts.

Furthermore, LLM-as-a-judge may not reflect expert preferences due to the specialized nature of the task and our need for specifically tailored evaluation criteria (Szymanski et al., 2025; Huang et al., 2025a). The usefulness of the generated text extends beyond these automatic evaluation strategies, requiring expert judgment to assess whether the fact-check articles align with professional standards.

We conducted human evaluations with the help of our expert fact-checkers. We designed a questionnaire to rate the fact-checking articles using a 5-point Likert scale on the aspects listed in §5.1, along with *Publishability*, which served as a proxy to measure the quality and the usefulness of the generated articles (see Appendix D for questionnaire details). We manually collected 12 claims related to climate change, which were fact-checked by *different* fact-checking organizations, and generated automatic fact-checking articles for these claims using *Vanilla-GPT*, *Storm*, and QRAFT. We then presented each of our experts with a claim and asked them to blindly rate the corresponding fact-checking articles: the three automatic ones and the original article. Finally, we asked the expert fact-checkers to rank these articles by relative *Publishability*.

Table 8 presents the scores for the five evaluation aspects, while Figure 4 illustrates the distribution of the rankings for the four articles. We can see that all LLM-based frameworks achieve considerably lower scores than the expert-written articles, which are consistently recognized as the best. All three generative frameworks are rated within the 3–4 point range, indicating expert uncertainty or considerable room for improvement in quality across most aspects. In terms of rankings, expert-written articles were clearly found to be of the highest quality, while *Storm*'s generations were ranked the lowest.

This is further reinforced by the *Publishability* scores, where the experts indicated that they would rather write an article from scratch than use *Storm*'s version. QRAFT was ranked second by most experts, while *Vanilla-GPT* received inconsistent rankings, showing a median rank of third. This is again reflected in the *Publishability* scores, where the experts expressed uncertainty regarding the potential usefulness of these articles indicating the need for considerable manual editorial effort to raise them up to professional writing standards.

We further asked the experts for open-ended qualitative feedback explaining their scores:

Articles generated by LLM frameworks are too elaborative. The experts noted that while *Storm* and QRAFT thoroughly discussed the evidence, they often included extraneous details that did not help explain why the claim might or might not be true. These details made the articles harder to read and the arguments less comprehensible. This concern relates to all the issues relayed by experts in the initial interviews: hallucinations, lack of world knowledge, inability to capture context, and poor presentation of evidence. In addition, when multiple evidence sources offered similar information, these articles tended to be overly repetitive. Expert-written articles generally consolidate such evidence into a single concise argument, citing all sources to strengthen their case. *Storm* exhibits this limitation the most, due to its overly elaborate articles, while QRAFT is less concerning.

QRAFT presents factual information, but makes errors in using it to clarify the claim. The experts highlighted that QRAFT presented data that was relevant to the claim; however, it failed to construct arguments using this data to explain the claim's veracity.

For instance, in an article, QRAFT reported the amount of CO_2 emissions from fossil fuel burning, but did not mention how this related to a significant percentage of atmospheric CO_2 being anthropogenic, and not natural as claimed.

Moreover, the articles occasionally presented information from certain evidence sources as facts, while, in reality, this might be only true within a specific context, and does not necessarily constitute *the whole truth*.

This indicates that the articles lack a thorough and accurate analysis, failing to provide all the necessary context around a claim to explain its veracity.

Organizations exhibit considerable diversity in their guidelines for article composition. In §2.1, we mentioned that the article writing guidelines varied between fact-checking organizations, as different fact-checking organizations have distinct writing guidelines for their articles, and this point was further reinforced by our evaluation. We asked the experts whether the articles explicitly specified why the claim was ''important'' to fact-check, and we observed low scores for all four fact-checking articles. The experts clarified that, while even the expert-written articles did not always explain the importance of the fact-check, these would still be considered acceptable by some organizations, although their own organization might have written it differently. In addition, the experts also voiced their disagreement on the choice of evidence, noting that their organizations would have preferred to use official documents as sources, while others would be more flexible and could use third-party news reportings about the same.

Experts do not trust fully AI-generated articles. Fact-checking articles are written in a way that they transparently justify the verdict of a claim by guiding readers through the investigative process (Graves, 2017). Consequently, experts carefully analyze any fact-checking article they encounter and verify for themselves, using the cited evidence, whether all arguments are presented accurately. They stressed that an AI-generated article requires an even higher level of scrutiny and cannot be trusted as-is, emphasizing that a complete draft generated by QRAFT does not save them much time, but offers ideas on possible structures for their own fact-checking article.

Yet, the experts highlighted the potential of an AI framework that generates full fact-checking articles under human supervision. In such a framework, fact-checkers could specify their preferences and constantly guide the underlying models at the intermediate steps to align with their requirements and preferences. This would enable them to quickly prototype initial drafts for their fact-checking article, which they could easily edit and polish for actual publication.

7 Related Work

Automatic Fact-Checking. Previous work has focused on automating fact-checking, and not so much on assisting expert fact-checkers. Yet, they decomposed the fact-checking process into useful subtasks, which could help experts (Vlachos and Riedel, 2014; Kotonya and Toni, 2020; Nakov et al., 2021a; Guo et al., 2022): claim check-worthiness detection (Hassan et al., 2015; Wright and Augenstein, 2020; Konstantinovskiy et al., 2021; Barrón-Cedeño et al., 2023), detecting fact-checked claims (Shaar et al., 2020), evidence retrieval (Zhou et al., 2019; Soleimani et al., 2020; Zou et al., 2023), claim verification (Wang, 2017; Augenstein et al., 2019; Xie et al., 2025), and explanation generation (Popat et al., 2018; Atanasova et al., 2020; Zeng and Gao, 2024).

However, these tasks overlook the important step of generating fact-checking articles, which is a crucial step of the manual fact-checking workflow (Graves, 2017). It is a time-consuming process, and recent work has advocated for the development of NLP tools to assist experts in this task (Liu et al., 2024). Warren et al. (2025) recently presented insights into how fact-checkers compose explanations for a claim's veracity in fact-checking articles, highlighting features consistent with those that we outline in this work (see §2). Yet, the emphasis of their work is on improving the alignment of automatic fact-checking explanations with the needs of professional fact-checking. Explanations traditionally serve as a rationale of the underlying model's decision-making process (Kotonya and Toni, 2020), whereas experts write fact-checking articles to communicate their findings and clear the air regarding a claim by guiding readers through the exact investigation process (Graves, 2017). Thus, here we have argued for the need to introduce the *generation of fact-checking articles* as an additional task.

Long-Form Text Generation. Long-form text generation is challenging, as it requires ensuring that the generated text remains on topic, follows a natural narrative arc, and preserves the intended meaning. To address these issues, previous studies have broken down the process into stages inspired by human writing processes, including: planning, drafting, rewriting, and editing (Yao et al., 2019; Yang et al., 2022; Hu et al., 2022; Yang et al., 2023; Liang et al., 2024). Moreover, expository writing requires the text to be grounded on external evidence (Weaver and Kintsch, 1991), which demands a thorough sense-making process over the evidence, along with an ability to collate information into a cohesive narrative (Shen et al., 2023). While some methods have been proposed (Balepur et al., 2023), recent research has highlighted the effectiveness of agentic frameworks, which emphasize the pre-writing stage (Shao et al., 2024; Wang et al., 2024a,b; Liu and Chang, 2025). Fact-checking article writing is one such task, and our approach uses LLM agents to structure it into the above four stages of long-form writing. This ensures that QRAFT maintains topical consistency, and clarifies the claim with strong support from the evidence.

8 Conclusion and Future Work

Previous automatic fact-checking research has focused on tasks that can potentially assist expert fact-checkers to perform claim verification more efficiently. However, fact-checkers also perform another largely overlooked, yet essential task: communicating their findings regarding the claim through detailed fact-checking articles. In this paper, we argued that the typical fact-checking pipeline must be extended to include a new task: *the automatic generation of fact-checking articles*. We defined this task in close collaboration with experts from leading fact-checking organizations, deepening our understanding with insights into what constitutes a good fact-checking article. Based on these insights, we proposed QRAFT, a multi-agent collaboration framework that mimics the fact-checking article writing process of human experts. Moreover, through comprehensive evaluation, using automated evaluation measures, LLM as a judge, as well as expert judgments and qualitative analysis, we showed that QRAFT outperforms several preexisting approaches from the literature.

As human evaluation revealed that QRAFT still falls short compared to expert-written articles, particularly in terms of practical usefulness, we aim to compile a more detailed set of requirements for such articles. Moreover, as existing automatic evaluation measures are insufficient to capture the real-world utility of the generated articles, we aim to develop a more robust suite of evaluation measures for this task. We also aim to benchmark stronger, more recent LLMs for this purpose, especially such tuned for complex reasoning—commonly known as *large reasoning models*.

In our experiments, we further observed that QRAFT failed to construct arguments using the evidence to explain the claim's veracity in some cases. To alleviate this gap, in future work, we aim to introduce an intermediate reasoning step, which would focus on deeper sense-making of the claim and the broader narrative using context provided by the evidence. Finally, we plan to explore the potential for human–AI cooperation (Dutta et al., 2025), which would enable for higher-quality fact-checking articles thanks to human oversight and iterative feedback.

Limitations

Our work introduced a new task into the typical automatic fact-checking pipeline, namely, *the automatic generation of full fact-checking articles*. We also proposed QRAFT, a framework for the end-to-end generation of such articles given a claim and the supporting evidence. Here, we acknowledge some limitations of our work, which remain open research questions for future work.

First, we relied on the inherent capabilities of the underlying LLMs in our framework to preserve relevant context when extracting concise evidence nuggets from source evidence documents. However, this strategy may not ensure the retention of all context—particularly for complex claims that often depend on indirectly related or contextually distant information. We have left a thorough analysis of this limitation for future work.

Second, QRAFT consists of three LLM agents, which collaborate to generate and iteratively refine full fact-checking articles. In our experiments, we exclusively used OpenAI's `GPT-4o` and `GPT-4o-mini` as the underlying LLMs. We have left a systematic performance comparison of

alternative model combinations as agents in such a multi-agent framework to future research.

Finally, we designed QRAFT with a focus on *assisting professional fact-checkers* in the time-consuming and labor-intensive fact-checking article writing process, instead of generating drafts intended for public dissemination as-is without human oversight. We expect the users of our framework to be professional fact-checkers, and thus, we only consulted fact-checking experts during the human evaluation phase of our work. However, an in-depth investigation of how non-experts perceive the generated fact-checking articles is also important and would be interesting to study in future work.

Acknowledgments

The authors would like to acknowledge the Multi-Institutional Faculty Interdisciplinary Research Project (MFIRP) between IIT Delhi and MBZUAI and Anusandhan National Research Foundation (DST/INT/USA/NSF-DST/Tanmoy/P-2/2024) for financial support. T. C. further acknowledges the support of the Rajiv Khemani Young Faculty Chair Professorship in Artificial Intelligence.

References

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating fact checking explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7352–7364, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.656

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. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4685–4697, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1475

Nishant Balepur, Jie Huang, and Kevin Chang. 2023. Expository text generation: Imitate, retrieve, paraphrase. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11896–11919, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.729

Alberto Barrón-Cedeño, Firoj Alam, Tommaso Caselli, Giovanni Da San Martino, Tamer Elsayed, Andrea Galassi, Fatima Haouari, Federico Ruggeri, Julia Maria Struß, Rabindra Nath Nandi, Gullal S. Cheema, Dilshod Azizov, and Preslav Nakov. 2023. The CLEF-2023 CheckThat! lab: Checkworthiness, subjectivity, political bias, factuality, and authority. In *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part III*, pages 506–517, Berlin, Heidelberg. Springer-Verlag. https://doi.org/10.1007/978-3-031-28241-6_59

Cheng-Han Chiang and Hung-yi Lee. 2024. Merging facts, crafting fallacies: Evaluating the contradictory nature of aggregated factual claims in long-form generations. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2734–2751, Bangkok, Thailand. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.160

Anubrata Das, Houjiang Liu, Venelin Kovatchev, and Matthew Lease. 2023. The state of human-centered NLP technology for fact-checking. *Information Processing & Management*, 60(2): 103219. https://doi.org/10.1016/j.ipm.2022.103219

Subhabrata Dutta, Timo Kaufmann, Goran Glavaš, Ivan Habernal, Kristian Kersting, Frauke Kreuter, Mira Mezini, Iryna Gurevych, Eyke Hüllermeier, and Hinrich Schütze. 2025. Problem solving through Human-AI preference-based cooperation. *Computational Linguistics*, pages 1–35. https://doi.org/10.1162/coli.a.19

Lucas Graves. 2017. Anatomy of a fact check: Objective practice and the contested epistemology of fact checking. *Communication, Culture*

and Critique, 10(3):518–537. https://doi.org/10.1111/cccr.12163

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206. https://doi.org/10.1162/tacl_a_00454

Naeemul Hassan, Chengkai Li, and Mark Tremayne. 2015. Detecting check-worthy factual claims in presidential debates. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management, CIKM 2015, Melbourne, VIC, Australia, October 19–23, 2015*, pages 1835–1838. ACM. https://doi.org/10.1145/2806416.2806652

Zhe Hu, Hou Pong Chan, Jiachen Liu, Xinyan Xiao, Hua Wu, and Lifu Huang. 2022. PLANET: Dynamic content planning in autoregressive transformers for long-form text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2288–2305, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.163

Hui Huang, Xingyuan Bu, Hongli Zhou, Yingqi Qu, Jing Liu, Muyun Yang, Bing Xu, and Tiejun Zhao. 2025a. An empirical study of LLM-as-a-judge for LLM evaluation: Fine-tuned judge model is not a general substitute for GPT-4. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5880–5895, Vienna, Austria. Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.findings-acl.306

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025b. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2). https://doi.org/10.1145/3703155

Fantine Huot, Reinald Kim Amplayo, Jennimaria Palomaki, Alice Shoshana Jakobovits, Elizabeth Clark, and Mirella Lapata. 2025. Agents' Room: Narrative generation through multi-step collaboration. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24–28, 2025*. OpenReview.net.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(1).

Prerna Juneja and Tanushree Mitra. 2022. Human and technological infrastructures of fact-checking. *Proceedings of the ACM on Human Computer Interaction*, 6(CSCW2). https://doi.org/10.1145/3555143

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4334–4353, Miami, Florida, USA. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.248

Dayoon Ko, Jinyoung Kim, Hahyeon Choi, and Gunhee Kim. 2024. GrowOVER: How can LLMs adapt to growing real-world knowledge? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3282–3308, Bangkok, Thailand. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.181

Lev Konstantinovskiy, Oliver Price, Mevan Babakar, and Arkaitz Zubiaga. 2021. Toward automated factchecking: Developing an annotation schema and benchmark for consistent automated claim detection. *Digital Threats*, 2(2). https://doi.org/10.1145/3412869

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking: A survey. In *Proceedings of the 28th International*

Conference on Computational Linguistics, pages 5430–5443, Barcelona, Spain (Online). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.474

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. https://doi.org/10.1162/tacl_a_00453

Yi Liang, You Wu, Honglei Zhuang, Li Chen, Jiaming Shen, Yiling Jia, Zhen Qin, Sumit Sanghai, Xuanhui Wang, Carl Yang, and Michael Bendersky. 2024. Integrating planning into single-turn long-form text generation. *ArXiv preprint*, abs/2410.06203.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Houjiang Liu, Anubrata Das, Alexander Boltz, Didi Zhou, Daisy Pinaroc, Matthew Lease, and Min Kyung Lee. 2024. Human-centered NLP fact-checking: Co-designing with fact-checkers using Matchmaking for AI. *Proceedings of the ACM on Human Computer Interaction*, 8(CSCW2). https://doi.org/10.1145/3686962

Yuxiang Liu and Kevin Chen-Chuan Chang. 2025. Writing like the best: Exemplar-based expository text generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25739–25764, Vienna, Austria. Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-long.1250

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.741

Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021a. Automated fact-checking for assisting human fact-checkers. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4551–4558, International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2021/619

Preslav Nakov, Giovanni Da San Martino, Tamer Elsayed, Alberto Barrón-Cedeño, Rubén Míguez, Shaden Shaar, Firoj Alam, Fatima Haouari, Maram Hasanain, Nikolay Babulkov, Alex Nikolov, Gautam Kishore Shahi, Julia Maria Struß, and Thomas Mandl. 2021b. The CLEF-2021 CheckThat! Lab on detecting check-worthy claims, previously fact-checked claims, and fake news. In *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part II*, pages 639–649, Berlin, Heidelberg. Springer-Verlag. https://doi.org/10.1007/978-3-030-72240-1_75

Team OpenAI. 2023. GPT-4 technical report.

Pew Research Center. 2024. How Americans get news on TikTok, X, Facebook and Instagram. Accessed: 2025-02-25.

Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. DeClarE: Debunking fake news and false claims using evidence-aware deep learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 22–32, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1003

D. Gordon Rohman. 1965. Pre-writing the stage of discovery in the writing process. *College Composition and Communication*, 16(2):106–112. https://doi.org/10.2307/354885

Shaden Shaar, Nikolay Babulkov, Giovanni Da San Martino, and Preslav Nakov. 2020. That is a known lie: Detecting previously

fact-checked claims. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3607–3618, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.332

Yijia Shao, Yucheng Jiang, Theodore Kanell, Peter Xu, Omar Khattab, and Monica Lam. 2024. Assisting in writing Wikipedia-like articles from scratch with large language models. 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 6252–6278, Mexico City, Mexico. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.347

Zejiang Shen, Tal August, Pao Siangliulue, Kyle Lo, Jonathan Bragg, Jeff Hammerbacher, Doug Downey, and Joseph Chee Chang. 2023. Beyond summarization: Designing AI support for real-world expository writing tasks. *ArXiv*, abs/2304.02623.

Shuqian Sheng, Yi Xu, Luoyi Fu, Jiaxin Ding, Lei Zhou, Xinbing Wang, and Chenghu Zhou. 2024. Is reference necessary in the evaluation of NLG systems? When and where? 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 8580–8596, Mexico City, Mexico. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.474

Amir Soleimani, Christof Monz, and Marcel Worring. 2020. BERT for evidence retrieval and claim verification. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II*, pages 359–366, Berlin, Heidelberg. Springer-Verlag. https://doi.org/10.1007/978-3-030-45442-5_45

Yixiao Song, Yekyung Kim, and Mohit Iyyer. 2024. VeriScore: Evaluating the factuality of verifiable claims in long-form text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9447–9474, Miami, Florida, USA. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.552

Annalisa Szymanski, Noah Ziems, Heather A. Eicher-Miller, Toby Jia-Jun Li, Meng Jiang, and Ronald A. Metoyer. 2025. Limitations of the LLM-as-a-judge approach for evaluating LLM outputs in expert knowledge tasks.In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, IUI '25, pages 952–966, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3708359.3712091

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. https://doi.org/10.3115/v1/W14-2508

Qianyue Wang, Jinwu Hu, Zhengping Li, Yufeng Wang, Daiyuan Li, Yu Hu, and Mingkui Tan. 2025. Generating long-form story using dynamic hierarchical outlining with memory-enhancement. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1352–1391, Albuquerque, New Mexico. Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.naacl-long.63

Qiyao Wang, Shiwen Ni, Huaren Liu, Shule Lu, Guhong Chen, Xi Feng, Chi Wei, Qiang Qu, Hamid Alinejad-Rokny, Yuan Lin, and Min Yang. 2024a. AutoPatent: A multi-agent framework for automatic patent generation. *ArXiv preprint*, abs/2412.09796.

William Yang Wang. 2017. ''Liar, liar pants on fire'': A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-2067

Yidong Wang, Qi Guo, Wenjin Yao, Hongbo Zhang, Xin Zhang, Zhen Wu, Meishan Zhang, Xinyu Dai, Min zhang, Qingsong Wen, Wei Ye, Shikun Zhang, and Yue Zhang. 2024b. AutoSurvey: Large language models can automatically write surveys. In *Advances in Neural Information Processing Systems*, volume 37, pages 115119–115145. Curran Associates, Inc. https://doi.org/10.52202/079017-3655

Greta Warren, Irina Shklovski, and Isabelle Augenstein. 2025. Show me the work: Fact-checkers' requirements for explainable automated fact-checking. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3706598.3713277

Edward Wates and Robert Campbell. 2007. Author's version vs. publisher's version: An analysis of the copy-editing function. *Learned Publishing*, 20(2):121–129. https://doi.org/10.1087/174148507X185090

Constance A. Weaver and Walter Kintsch. 1991. Expository text. In *Handbook of Reading Research*, volume 2, pages 230–245. Lawrence Erlbaum Associates, Mahwah, NJ.

Dustin Wright and Isabelle Augenstein. 2020. Claim check-worthiness detection as positive unlabelled learning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 476–488, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.43

Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2022. PRIMERA: Pyramid-based masked sentence pre-training for multi-document summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5245–5263, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.360

Zhuohan Xie, Trevor Cohn, and Jey Han Lau. 2023. The next chapter: A study of large language models in storytelling. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 323–351, Prague, Czechia. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.inlg-main.23

Zhuohan Xie, Rui Xing, Yuxia Wang, Jiahui Geng, Hasan Iqbal, Dhruv Sahnan, Iryna Gurevych, and Preslav Nakov. 2025. FIRE: Fact-checking with iterative retrieval and verification. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2901–2914, Albuquerque, New Mexico. Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.findings-naacl.158

Rui Xing, Shraey Bhatia, Timothy Baldwin, and Jey Han Lau. 2022. Automatic explanation generation for climate science claims. In *Proceedings of the 20th Annual Workshop of the Australasian Language Technology Association*, pages 122–129, Adelaide, Australia. Australasian Language Technology Association.

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. https://doi.org/10.18653/v1/2023.acl-long.190

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. https://doi.org/10.18653/v1/2022.emnlp-main.296

Lili Yao, Nanyun Peng, Ralph M. Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-Write: Towards better automatic storytelling. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence,*

EAAI 2019, Honolulu, Hawaii, USA, January 27–February 1, 2019, pages 7378–7385. AAAI Press. https://doi.org/10.1609/aaai.v33i01.33017378

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V. Chawla, and Xiangliang Zhang. 2020. Justice or prejudice? Quantifying biases in LLM-as-a-judge. In *The Thirteenth International Conference on Learning Representations, ICLR 2025*.

Fengzhu Zeng and Wei Gao. 2024. JustiLM: Few-shot justification generation for explainable fact-checking of real-world claims. *Transactions of the Association for Computational Linguistics*, 12:334–354. https://doi.org/10.1162/tacl_a_00649

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26–30, 2020*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10–16, 2023*.

Jie Zhou, Xu Han, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2019. GEAR: Graph-based evidence aggregating and reasoning for fact verification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 892–901, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1085

Anni Zou, Zhuosheng Zhang, and Hai Zhao. 2023. Decker: Double check with heterogeneous knowledge for commonsense fact verification. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11891–11904, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.752

Guido Zuccon, Bevan Koopman, and Razia Shaik. 2023. ChatGPT hallucinates when attributing answers. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, SIGIR-AP '23, pages 46–51, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3624918.3625329

Section	Content
I.	**General Question Round**
	Q. As a fact-checker, what do you expect from a fact-check article?
	Q. Can you comment on any challenges for AI that you may foresee in the process of writing a fact-check article?
II.	**Sample Data Presentation**
	Claim: ''In December 2023, conspiracy theorist <person_1>'s Twitter/X account, @<username_p1>, was reinstated by <person_2> after having been banned for abusive behavior since 2018.''
	Veracity: True
	Evidence: {Website 1, Website 2, . . . }
III.	**Focused Question Round 1**
	Q. What do you expect from an article written for the given data?
	Q. What do you think an AI-based system might miss if it generated a fact-check article from the given data?
IV.	**Presenting Generated Article**
	Article: – Redacted –
V.	**Focused Question Round 2**
	Q. What are your opinions on the article based on your expectations? Do you think your expectations were met?
	Q. What do you think of the utility of this article for a fact-checker?
VI.	**Concluding Question Round**
	Q. Do you have any other points or any comments on something we might have missed?

Table 9: **Structure of our interview along with the questions asked in each round.** Each interview took approximately one hour and was held one-on-one with a fact-checking expert. The article in Section IV was generated using few-shot prompting on GPT-4o-mini.

Algorithm 1 Pseudocode for QRAFT.

Require: $X = \{C, V, E_{C,V}\}$
where $E_{C,V} = \{d \mid d \text{ is an evidence document}\}$

$\mathcal{P}$: the Planner agent
$\mathcal{W}$: the Writer agent
$\mathcal{E}$: the Editor agent
$Prefs$: Guidelines for the draft's structure

M: Max iterations for draft planning and writing
N: Max iterations for Question-Asking interactions between $\mathcal{W}$ and $\mathcal{E}$

procedure QRAFT (A)
- ▷ (i) *Evidence nugget from the input by* $\mathcal{P}$ ◁
- Initialize $N_{C,V} \leftarrow \{\}$
- **for all** $d \in E_{C,V}$ **do**
  - $n(d) \leftarrow \mathcal{P}$.gatherEvidenceNuggets$(d, C, V)$
  - $N_{C,V}$.add$(d : n(d))$
- ▷ (ii) *Telling* $\mathcal{P}$ *the preferences for the draft's structure* ◁
- $\mathcal{P}$.setPreferences$(Prefs)$
- ▷ (iii) *Planning and writing the first draft by collaboration between* $\mathcal{P}$ *and* $\mathcal{W}$ ◁
- Initialize $i \leftarrow 0$
- Initialize $O_{C,V} \leftarrow$ " "
- Initialize $D_{C,V} \leftarrow$ " "
- **while** $i < M$ **do**
  - $O_{C,V} \leftarrow \mathcal{P}$.proposeOutline$(N_{C,V}, C, V, O_{C,V}, D_{C,V})$
  - $D_{C,V} \leftarrow \mathcal{W}$.writeDraft$(O_{C,V}, C, V, N_{C,V})$
  - **if** $\mathcal{P}$.approveDraft$(D_{C,V})$ **then**
    - break
  - $i \leftarrow i + 1$
- **return** $D_{C,V}$

procedure QRAFT (B)
- Initialize $history \leftarrow \{\}$
- Initialize $i \leftarrow 0$
- $\mathcal{E}$.reviewDraft$(D_{C,V})$
- **while** $i < N$ **do**
  - ▷ *Question-Asking based interactions between* $\mathcal{E}$ *and* $\mathcal{W}$ ◁
  - $Q_i \leftarrow \mathcal{E}$.makeQuestion$(history)$
  - $A_i \leftarrow \mathcal{W}$.answer$(Q_i, D_{C,V})$
  - $history$.add$(\{Q_i, A_i\})$
  - $i \leftarrow i + 1$
- $Edits_{D_{C,V}} \leftarrow \mathcal{E}$.suggestEdits$(history)$
- $D^*_{C,V} \leftarrow \mathcal{W}$.improveDraft$(D_{C,V}, Edits_{D_{C,V}}, C, V, N_{C,V})$
- **return** $D^*_{C,V}$

A Details about the Interviews

In §2, we presented the key findings from our interviews with experts from leading fact-checking organizations. Here, we present the structure of our interviews along with the questions we asked in Table 9. We kept the questions open-ended in order to keep the process of gathering insights unbiased.

B Details about the QRAFT Implementation

In §3, we introduced QRAFT, our framework designed to rely on conversational question-asking between multiple LLM agents to draft fact-checking articles. Here, we give some specifics on QRAFT's workflow through a pseudocode in Algoritm 5.

C Details about the LLM-as-a-Judge Evaluation

In §5, we have already explained that we performed LLM-as-a-judge evaluations using Prometheus-2 as the evaluator LLM, and that we collaborated with fact-checking experts to design five-point rubrics to score the generated articles on several aspects. Here, we present Table 10, which shows the particular rubrics we used in our evaluation on the set of four specific aspects we addressed: *Relevance*, *Comprehensibility*, *Importance*, and *Evidence Presentation*.

Aspect	Question & Rubrics
Relevance	**Q. Is the content of the article relevant to the claim and its proposed veracity?**
	Score 1: The content is irrelevant to the claim and/or its proposed veracity.
	Score 2: Most of the content is inconsistent with the proposed veracity of the claim.
	Score 3: Some of the content is consistent with the proposed veracity of the claim, but some is not.
	Score 4: Most of the content is consistent with the proposed veracity of the claim.
	Score 5: The content in the article is consistent with the proposed veracity of the claim.
Comprehensibility	**Q. Is the article easy to understand and follow for readers without background knowledge on the claim?**
	Score 1: Clearly not, the reader would definitely need additional background knowledge on the claim to understand the article.
	Score 2: Probably not, the claim is hard to understand without background knowledge, which is not present in the article.
	Score 3: Unsure, readers might need some additional background knowledge to understand the article.
	Score 4: Mostly yes, the article contains most of the necessary context about the claim, but some is missing.
	Score 5: Definitely, the article contains all the necessary context about the claim.
Importance	**Q. Does the article explain why the claim is being fact-checked?**
	Score 1: The article puts no effort into explaining why the claim is being fact-checked.
	Score 2: Mostly not, it requires some effort while reading the article to guess why the claim is being fact-checked.
	Score 3: The reader could infer why the claim is being fact-checked, but it is not explicitly stated.
	Score 4: The article gives some justification for why the claim is being fact-checked, but it could do better.
	Score 5: The article clearly explains why the claim is being fact-checked.
Evidence Presentation	**Q. Does the article construct arguments using evidence to explain the claim's veracity?**
	Score 1: The article does not discuss the evidence at all.
	Score 2: The article does not create arguments, mostly just summarizes the evidence.
	Score 3: The article mostly creates arguments, but some evidence is summarized with no focus on how it helps explain the claim's veracity.
	Score 4: The article constructs arguments using evidence to explain the claim's veracity, but there are some gaps in logic.
	Score 5: The article constructs accurate arguments using the evidence to explain the claim's veracity.

Table 10: **The questions and rubrics used to judge the generated articles in our LLM-as-a-judge evaluations.** We used Prometheus-2 (7B) as the evaluator LLM.

D Details about the Expert Evaluation

In §6, we presented the expert evaluations that we conducted to judge the practical usefulness of the AI-generated articles, along with the expert-written fact-checking article. As we explained there, we evaluated these articles across 5 aspects: *Relevance*, *Comprehensibility*, *Importance*, *Evidence Presentation*, and *Publishability*. Here, in Table 11, we present the actual questionnaire we used to rate the articles qualitatively and quantitatively.

Aspect	Question & Rubrics
Relevance	**Q 1. Does the article clearly state the claim that is being fact-checked?**
	Score 1: Definitely not, the claim is not mentioned at all.
	Score 2: Maybe not, the claim is not articulated clearly, and it is hard to tell what is being fact-checked.
	Score 3: Unsure, the claim is not stated clearly.
	Score 4: Mostly yes, the claim is mentioned, but it is hard to find (e.g., maybe because it is not stated clearly in the introductory part of the article).
	Score 5: Definitely, the claim is clearly stated.
	Q 2. Does the article clearly state the claim's veracity?
	Score 1: Definitely not, no mention of the claim's veracity in the article at all, or the article states a wrong veracity label
	Score 2: Not really, the veracity of the claim is not explicitly stated, but one might be able to guess it with some effort by reading the entire article carefully.
	Score 3: Maybe, the veracity of the claim can be inferred from the contents of the full article, but it is not clearly stated.
	Score 4: Somewhat, the article only states the veracity of the claim in the introduction and/or in the conclusion, but does not discuss it otherwise.
	Score 5: Definitely yes, the article clearly states the veracity of the claim.
	Q 3. Do the contents of the article support the proposed veracity assessment of the claim?
	Score 1: Definitely not, the presented content is irrelevant to the claim and/or its proposed veracity.
	Score 2: Mostly not, most of the content is inconsistent with the proposed veracity of the claim.
	Score 3: Unsure, some of the content is consistent with the proposed veracity of the claim, and some is not.
	Score 4: Mostly yes, most of the content is consistent with the proposed veracity of the claim.
	Score 5: Definitely yes, the content in the article is consistent with the proposed veracity of the claim.
Comprehensibility	**Q. Do you think the article is easy to understand for readers without background knowledge about the claim?**
	Score 1: Clearly not, the reader would definitely need additional background knowledge on the claim to understand the article.
	Score 2: Probably not, the claim is hard to understand without background knowledge, which is not present in the article.
	Score 3: Unsure, readers might need some additional background knowledge to understand the article.
	Score 4: Mostly yes, the article contains most of the necessary context about the claim, but some is missing.
	Score 5: Definitely, the article contains all the necessary context about the claim.
Importance	**Q. Does the article explain why the claim is being fact-checked?**
	Score 1: Not at all, the article puts no effort to justify why the claim is being fact-checked.
	Score 2: Mostly not, it requires some effort while reading the article to guess why the claim is being fact-checked.
	Score 3: Unsure, one could infer from the article why the claim needs to be fact-checked, but this is not clearly articulated.
	Score 4: Mostly yes, the article gives some justification about why the claim is being fact-checked, but it could do better.
	Score 5: Definitely yes, the article gives enough justification about why the claim is being fact-checked.
Evidence Presentation	**Q. How does the article discuss the evidence?**
	Score 1: Does not discuss the evidence at all.
	Score 2: Mostly just summarizes the contents of the evidence sources.
	Score 3: Mostly constructs arguments, but contents of some evidence sources are just summarized with no focus on how they help the claim's veracity assessment.
	Score 4: Constructs arguments, but there are some gaps in their logic.
	Score 5: Constructs accurate arguments towards the claim's veracity based on the contents of each evidence source.
Publishability	**Q. Could this article be published as-is or does it need extra work?**
	Score 1: The article is of no use.
	Score 2: The article gives me some ideas about the structure, but I would rather write one from scratch.
	Score 3: The article could serve as a first draft and could save time for me compared to writing a review-ready article from scratch.
	Score 4: The article could be published after some edits.
	Score 5: The article could be published as is.

Table 11: **The questionnaire used to judge the four (three AI-generated and one expert-written) articles through expert evaluations.** We presented each expert with a claim and asked them to rate the four corresponding fact-checking articles using this questionnaire. For *Relevance*, we calculated the the mean score across the three questions.