JustiLM: Few-shot Justification Generation for Explainable Fact-Checking of Real-world Claims

Fengzhu Zeng
Singapore Management University
80 Stamford Rd, Singapore 178902
fzzeng.2020@phdcs.smu.edu.sg

Wei Gao
Singapore Management University
80 Stamford Rd, Singapore 178902
weigao@smu.edu.sg

Abstract

Justification is an explanation that supports the veracity assigned to a claim in fact-checking. However, the task of justification generation has been previously oversimplified as summarization of a fact-check article authored by fact-checkers. Therefore, we propose a realistic approach to generate justification based on retrieved evidence. We present a new benchmark dataset called ExClaim (for Explainable fact-checking of real-world Claims), and introduce JustiLM, a novel few-shot Justification generation based on retrieval-augmented Language Model by using fact-check articles as an auxiliary resource during training only. Experiments show that JustiLM achieves promising performance in justification generation compared to strong baselines, and can also enhance veracity classification with a straightforward extension.[1]

1 Introduction

Automated fact-checking typically encompasses several stages: identify check-worthy claims, retrieve relevant evidence, determine the claim's veracity using the retrieved evidence, and generate justification for the verdict on the veracity (Guo et al., 2022). Despite a wealth of research focusing on the initial three stages, justification generation has remained under-explored in the past. Justifications present essential evidence and rationales used to arrive at a claim's veracity judgment, serving to convince readers and enhance the credibility of fact-checking systems. This explanatory process is of paramount importance in gaining the user's trust in automated fact-checking (Kotonya and Toni, 2020a; Atanasova et al., 2020).

Several methods have attempted to generate justification of verdict by summarizing fact-check articles that were previously authored by human fact-checkers (Kotonya and Toni, 2020b; Atanasova et al., 2020; Russo et al., 2023). Since a fact-check article per se is manually written to justify the verdict of a given claim with detailed presentation and reasoning over digested evidence, referring to reference documents collected from multiple sources, directly generating a summary from such a report as justification sidesteps the realistic challenges of evidence gathering and evidence-based reasoning for veracity assessment we essentially face in the fact-checking task. More importantly, these existing methods are impractical because fact-check articles are not available for new claims that are yet to check (Guo et al., 2022). Table 1 shows an example illustrating different types of information involved in the fact-checking practice and their relationship. To justify the veracity for a claim, the source of information that can be used practically ought to be the retrieved reference documents containing evidence rather than its fact-check article, which, as an outcome, has not been written during the checking process.

In this paper, we propose a more realistic approach for the task of justification generation based on a language model approach, which complies with the process of journalistic fact-checking by well-known fact-check organizations such as PolitiFact.[2] Our goal is to produce high-quality justifications, drawing upon evidence gathered from diverse sources. To this end, we construct a benchmark dataset for Explainable fact-checking of real-world Claims, named ExClaim, derived from a public dataset, WatClaimCheck (Khan et al., 2022), containing newsworthy claims along with their fact-check articles and reference documents. ExClaim provides a large searchable corpus by mixing the reference documents from all claims

[1] Code and dataset are released at https://github.com/znhy1024/JustiLM.

[2] https://www.politifact.com/.

Transactions of the Association for Computational Linguistics, vol. 12, pp. 334–354, 2024. https://doi.org/10.1162/tacl_a_00649
Action Editor: Fei Liu. Submission batch: 11/2023; Revision batch: 12/2023; Published 4/2024.

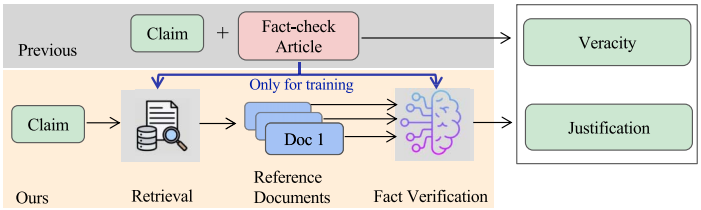
Claim	Biden: Gun manufacturers are “the only industry in the country” that have immunity from lawsuits.
Evidence Documents (References)	Doc1: *No, you can’t sue Pfizer or another manufacturer if you get a COVID-19 vaccine injury, but you can file for compensation.* The Pfizer-BioNTech COVID-19 vaccine received full approval from the Food and Drug ... Doc2: *Remarks by President Biden on Gun Violence Prevention.* THE PRESIDENT: Thank you, Kamala — Madam Vice President. Thank you very much. You know, we’re joined ... Doc3: *Clinton: Gun industry is ‘wholly protected’ from all lawsuits.* At the first Democratic debate of the 2016 presidential race, former Secretary of State Hillary Clinton criticized opponent ... Doc4: *Protection of Lawful Commerce in Arms Act.* The Protection of Lawful Commerce in Arms Act (PLCAA) is a United States law which protects firearms manufacturers and dealers from being held liable ... ...
Fact-check Article	... This isn’t the first time Biden has made this claim. He’s made it repeatedly, including April 2021 remarks about gun violence ... **But the claim is inaccurate. The gun industry is susceptible to some lawsuits, and there are federal laws restricting liability for a number of other types of businesses.** ... The law says gun dealers and manufacturers cannot be sued when their products are misused. But the law lists several situations that are not protected from liability. ... Other industries have exemptions in liability. ... until 2024, pharmaceutical companies that make the COVID-19 vaccines will have liability immunity ... **There’s also some liability protection in the medical devices and airline industries.** ...
Justification	Biden said that gun manufacturers represent the only industry in America that is exempt from being sued. This isn’t accurate. The gun industry is not entirely exempt from being sued and is susceptible to some lawsuits. Further, there are federal laws that restrict liability for a variety of other business sectors. We rate it False.
Veracity	FALSE
Procedure	 Previous methods just summarize fact-check articles as justification while our method follows a practical fact-checking process.

Table 1: An example claim along with the evidence documents, justification, and veracity. The *title* of each evidence document is italicized. The sentences in the fact-check article referring to evidence documents are marked in the same color as the corresponding documents, and the sentences that directly entail the justification are in **bold**.

in WatClaimCheck. Additionally, it curates the verdict justifications sourced from fact-check articles, typically located in a conclusive paragraph marked by cue phrases like ‘‘Our ruling’’ or ‘‘Our rating’’ for each claim. Furthermore, we develop a Justification Language Model called JustiLM for generating the rationales behind veracity judgement within the context of few-shot learning. Presumably, few-shot fine-tuning can mitigate the training resource requirements and its dependence on high-end hardware, often financially prohibitive, and also enables the model to achieve comparable effectiveness to state-of-the-art fully-trained models. JustiLM utilizes fact-check articles as auxiliary information in its training only via fine-tuning a pre-trained Retrieval-Augmented Generation (RAG) model on our curated justification dataset. Meanwhile, leveraging fact-check articles for training enhances the model’s proficiency in generating rationales based on evidence and articulating them in its generated content. Our contributions are threefold:

- We propose JustiLM, the first realistic justification generation method based on a retrieval-augmented language model that is trained end-to-end for explainable fact checking of real-world claims, leveraging fact-check articles as auxiliary information for model training only.

- We construct ExClaim, a new benchmark derived from the WatClaimCheck dataset (Khan et al., 2022) for explainable fact-checking, which contains 6,951 real-world claims and their corresponding veracity labels and human-written justifications, together with a large searchable corpus of 957,949 chunk-level documents for fine-grained evidence retrieval.

- JustiLM outperforms In-Context Learning (ICL) enabled language models, including Flan-T5, Llama2, and the state-of-the-art few-shot RAG model Atlas. JustiLM also shows promising performance compared to the latest GPT-4 model. A straightforward extension of JustiLM for joint veracity prediction and justification generation improves the veracity prediction task with large margins.

2 Related Work

2.1 Explanations for Fact-checking

Explanations for fact-checking claims have gained significant prominence in recent times, particularly due to the prevalent use of black-box models in automated fact-checking systems (Atanasova et al., 2020; Guo et al., 2022). Several methods have emerged to address this issue utilizing various techniques to provide human readable explanations. One stream of research leverages attention weights to highlight salient parts in the retrieved evidence as explanations (Popat et al., 2018; Ma et al., 2019; Yang et al., 2019; Shu et al., 2019; Lu and Li, 2020). Another stream of study is to adopt logic-based rules, such as knowledge graphs and natural logic relations designed by human experts (Ahmadi et al., 2019; Gad-Elrab et al., 2019; Vedula and Parthasarathy, 2021; Krishna et al., 2022a), where explanations are obtained by tracing the rules path to reach the veracity of the claim. However, these explanations are not presented in natural language, rendering them less accessible to general users. Furthermore, these rule-based systems encounter challenges when dealing with real-world claims that may not conform to predefined rules. In contrast, our work places a strong emphasis on generating textual justifications that are readily understandable for users, avoiding manual rule definitions.

A few studies have attempted to automatically generate textual justifications by summarizing fact-check articles (Kotonya and Toni, 2020b; Atanasova et al., 2020; Russo et al., 2023). Atanasova et al. (2020) employ DistilBERT (Sanh et al., 2019) to extract sentences from fact-check articles to form justifications. Kotonya and Toni (2020b) propose a two-step process, initially utilizing a Sentence-BERT (Reimers and Gurevych, 2019) to extract sentences from fact-check articles and subsequently using the BERTSUM model (Liu and Lapata, 2019) for abstractive justification generation based on the extracted sentences. Russo et al. (2023) explore several existing extractive summarization (Erkan and Radev, 2004; Reimers and Gurevych, 2019) and abstractive summarization (Raffel et al., 2020; Zhang et al., 2020; Shleifer and Rush, 2020) approaches for summarizing fact-check articles. These summarization methods come with inherent limitations practically, including complete reliance on fact-check articles (i.e., detailed human justification) as input, which is hardly available at the time of deployment, and complete omission of automatic evidence search and evidence-based reasoning. Different from these approaches, our method only assumes the availability of fact-check articles during model training and the key evidence exists within a large corpus which is searchable. Therefore, our approach generates justifications by harnessing the information from retrieved reference documents during inference, which is a more realistic solution for real-world scenarios. Similarly, Khan et al. (2022) infer claim veracity based on retrieved textual references, while Yao et al. (2023a) retrieve evidence for multi-modal fact-checking and generate explanations for predicted veracity labels using the BART model (Lewis et al., 2020a), both of which are stage-wise and full-dataset trained. In contrast, we base our approach on the latest RAG framework that is trained end-to-end and generates justifications by using fact-check articles to distill supervisory signals for training.

2.2 Few-shot Fact-checking

The need of few-shot learning is exacerbated by the continuous increase of computational and storage requirements for language model training. However, the specific application of few-shot learning techniques in the context of fact-checking has been relatively underexplored. Existing methods for few-shot fact-checking only focus on the so-called fact verification task (Lee et al., 2021; Zeng and Zubiaga, 2022; Zeng and Gao, 2023; Yue et al., 2023; Pan et al., 2023; Zhang and Gao, 2023) by feeding a few instances together with gold evidence into the model to predict the veracity of a claim. Different from these methods, our work primarily centers on generating justifications to substantiate the veracity of a claim based on the *retrieved* evidence. Importantly, we do not

assume the availability of annotated evidence. Instead, we necessitate the system to retrieve pertinent evidence, conforming to a more realistic and challenging scenario.

2.3 Retrieval-augmented Language Models

Equipping language models (LM) with external memory has shown to enhance their performance in knowledge intensive NLP tasks (Chen et al., 2017; Thorne et al., 2018; Guu et al., 2020; Lewis et al., 2020b; Sachan et al., 2021; Izacard and Grave, 2021b; Borgeaud et al., 2022; Izacard et al., 2023). Typically, a retriever is used to retrieve relevant documents from a large corpus, which enriches the input of a language model and contributes to the final output. However, due to the high cost of acquiring query-document annotations and training retrievers, many implementations rely on off-the-shelf retrievers, such as TF-IDF and BM25 (Jones, 2004; Robertson et al., 1994), which use term-matching techniques. In this setup, only the parameters of LMs are fine-tuned.

Recent research has demonstrated the advantages of jointly training the retriever and the LM in an end-to-end manner, which leverages the supervision signals from the LM to train the retriever (Guu et al., 2020; Lewis et al., 2020b; Sachan et al., 2021; Izacard and Grave, 2021b; Izacard et al., 2023). Moreover, considering the remarkable performance of large language models (LLMs) in various few-shot NLP tasks, some studies suggest enhancing LLMs with the retrievers or web search engines (Mallen et al., 2023; Si et al., 2023; Yu et al., 2023; Shi et al., 2023; Zhang and Gao, 2023). For example, REPLUG (Shi et al., 2023) optimizes the retriever by minimizing the KL divergence between the retrieval likelihood and the black-box LLM likelihood over retrieved documents. However, there exist inherent limitations in the interaction between retriever and black-box LLMs, such as their restricted ability to provide or access specific information. We refer readers to a comprehensive survey of retrieval-augmented LMs (Mialon et al., 2023).

3 Task Formulation

Let $\mathbf{C} = \{(\mathbf{x}, \mathbf{z}, \mathbf{y})\}$ be a fact-checking dataset of real-world news claims associated with a textual knowledge corpus $\mathcal{D}$. Each instance is composed of a claim $\mathbf{x}$ and its corresponding ground-truth justification $\mathbf{y}$ and fact-check article $\mathbf{z}$. $\mathbf{C}$ is divided as a training set and a test set, and only instances in the training set are associated with fact-check articles if available.

Given a claim $\mathbf{x}$ and the corpus $\mathcal{D}$, the goal of justification generation is to produce a sequence of tokens, denoted as $\hat{\mathbf{y}}$, that serves as an explanation for the veracity rendered on the claim using the evidence retrieved from the corpus. In the few-shot setting, we randomly select K instances from the training set, following the similar setup employed in previous studies for fact verification (Lee et al., 2021; Liu et al., 2022; Zeng and Gao, 2023), and we do not assume the availability of development set as this aligns to a more realistic scenario with limited data resources.

4 ExClaim Dataset

The existing fact-checking datasets based on real-world claims have limitations for justification generation. This is because the provided evidence sources might not cover the evidence documents that fact-checkers actually rely on when writing justifications. For example, some datasets (Vlachos and Riedel, 2014; Wang, 2017; Alhindi et al., 2018) only provide metadata like speaker, party, and date without a sizeable knowledge corpus for finding specific evidence. Some studies (Popat et al., 2016; Baly et al., 2018; Augenstein et al., 2019; Gupta and Srikumar, 2021; Yang et al., 2022; Hu et al., 2022) utilize web search to gather evidence documents, which result in retrieved information from non-authoritative sources or lead to the leak of ground truth by inadvertently including articles verifying the same claims by other organizations or sharing the fact-check information (Khan et al., 2022). More notably, certain studies (Hanselowski et al., 2019; Kotonya and Toni, 2020a; Atanasova et al., 2020; Ostrowski et al., 2021; Russo et al., 2023) regard fact-check articles as a primary source of evidence, a practice that may not align with realistic fact-checking procedures.

We use the WatClaimCheck (Khan et al., 2022) dataset that provides the real-world claims along with the text of reference documents cited by fact-check articles. However, WatClaimCheck is constructed for veracity classification and does not provide ground-truth justifications. For our task, we construct ExClaim based on WatClaimCheck, for which we additionally extract justifications

	Split	# Instance	Avg. # Tokens.
Claim	Train	5,964	25
	Test	987	25
Fact-check Article	Train	5,964	1,102
	Test†	987†	1,091†
Reference Documents	Train	40,089	2,656
	Test	6,647	2,404
Justification	Train	5,964	129
	Test	987	131

Table 2: Statistics of the ExClaim dataset. †: Note that fact-check articles in the test set are not used in our method, but exclusively utilized by baselines that rely on fact-check articles.

from fact-check articles based on the cue phrases such as ''Our ruling'' or ''Our rating'' in the reports following previous works (Alhindi et al., 2018; Augenstein et al., 2019; Kotonya and Toni, 2020a) and remove the instances that do not have such justification content. After extracting the justifications, we also remove them from fact-check articles.

Table 2 presents summary statistics of the ExClaim dataset with a total 6,951 real-world claims and justifications (i.e., 5,964 for training and 987 for testing). The data pose some challenges: 1) A single reference document is generally much longer than fact-check article, easily exceeding the context window of most text generation models (e.g., 512 tokens of T5 (Raffel et al., 2020) or 1,024 tokens of BART (Lewis et al., 2020a)). In particular, each claim may correspond to multiple reference documents from different sources, leading to excessively long text for evidence. 2) There is a lack of passage-/sentence-level annotation in reference documents and fact-check articles. Since fact-checkers generally refer to only several pieces of text in reference documents when writing justifications, most information in a reference document tend to be irrelevant for generating the justifications. To address these issues, we split each document into disjoint 100-word chunks following previous work (Lee et al., 2019; Karpukhin et al., 2020; Lewis et al., 2020b; Izacard et al., 2023), resulting in a large textual knowledge corpus $\mathcal{D}$ comprising a total of 957,949 chunk-level documents that systems can search fine-grained evidence text from. In the rest of the paper, we refer to these short text chunks as ''reference documents'' or simply ''documents''.

5 Methodology

We base our approach on the retrieval-augmented generation (RAG) framework (Lewis et al., 2020b; Sachan et al., 2021; Izacard and Grave, 2021b; Izacard et al., 2023), which contains a retriever for fine-grained evidence retrieval and a LM for textual justification generation. As shown in Figure 1, the retriever takes the claim text as input and retrieves the top-N chunk-level documents from the textual knowledge corpus, and the LM conditions on these documents together with the claim to generate justification. The retriever and LM can be jointly trained within a single RAG framework, which makes it possible to utilize fact-check articles as an auxiliary resource to provide supervisory signals during training, targeting to enhance the quality of generated justification. We employ Atlas (Izacard et al., 2023) as our backbone model considering two main reasons: 1) its strong few-shot learning ability in knowledge intensive tasks when its retriever and LM are jointly trained; 2) its flexibility for incorporating fact-check articles in the training process.

5.1 Retriever

Given a claim $\mathbf{x}$, the retriever should return the documents that help LM generate better justification. To enable the training of the retriever, Atlas utilizes a dense retriever named Contriever (Izacard et al., 2022), which is pre-trained using the MoCo contrastive loss (He et al., 2020). Contriever is a dual-encoder architecture that the pre-trained query encoder $\mathbf{E}_c$ and document encoder $\mathbf{E}_d$ encode the claim $\mathbf{x}$ and each document $\mathbf{d}_j \in \mathcal{D}$, respectively. The embeddings of documents can be pre-computed to build a collection of index using FAISS (Johnson et al., 2021) for fast retrieval. Documents are ranked by the similarity score $s(\mathbf{x}, \mathbf{d}_j) = \mathbf{E}_c(\mathbf{x})^\top \mathbf{E}_d(\mathbf{d}_j)$ that is calculated by taking the dot product of the embeddings of the claim $\mathbf{x}$ and document $\mathbf{d}_j$.

To mitigate the burden of re-computing embeddings for all documents when training the retriever, Atlas (Izacard et al., 2023) only updates the parameters corresponding to the query encoder while freezing the documents encoder, which still shows promising results in the few-shot setting. Therefore, we employ the document encoder for

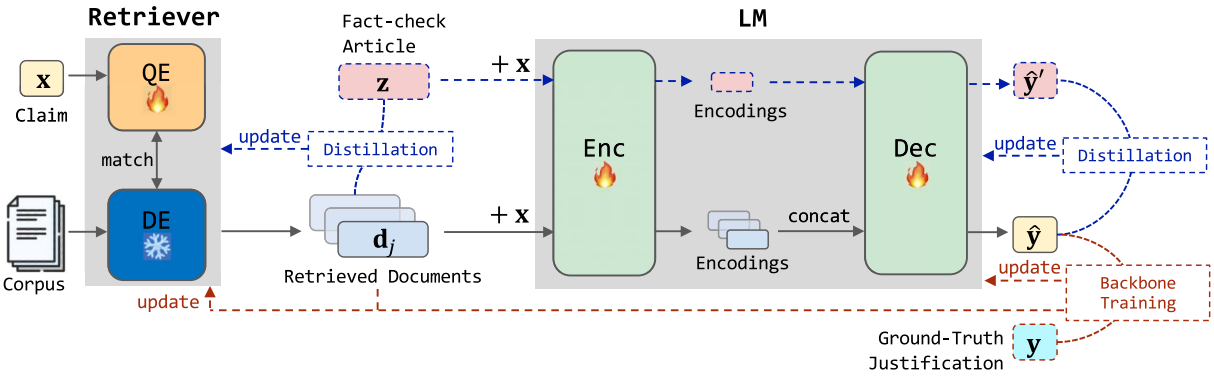


Figure 1: The architecture of JustiLM. Gray solid arrows present the inference process **without** fact-check article $\mathbf{z}$. Red dash arrows present the training process of backbone model, where the ground-truth justification provide supervisory signals to train both retriever and LM. Blue dashed arrows present the training process with the distillation of $\mathbf{z}$ as supervisory signals. The document encoder is fixed during training, while other modules are trainable. QE: Query Encoder; DE: Document Encoder; Enc: Encoder; Dec: Decoder.

encoding reference documents and the query encoder for encoding other inputs. Since there is no direct supervision available to train the retriever, Atlas proposes a Perplexity Distillation loss to leverage the supervisory signals from the LM. The intuition behind this is that documents contributing to the LM that help generate lower-perplexity outputs should be ranked higher (Izacard et al., 2023).

5.2 Language Model

The language model conditions on the top-N retrieved documents $D_N = \{\mathbf{d}_j\}_{j=1}^{N}$ by the retriever, together with the claim $\mathbf{x}$, to generate the justification. To aggregate evidence efficiently and effectively from multiple documents in LM, Atlas employs a T5 encoder-decoder model (Raffel et al., 2020) with the Fusion-in-Decoder (FiD) (Izacard and Grave, 2021b) modification. Each retrieved document $\mathbf{d}_j$ is encoded independently by the encoder, with the claim $\mathbf{x}$ prepended to it. All outputs of the encoder are then concatenated. The decoder takes as input this concatenation and performs cross-attention to fuse the evidence and generate outputs. The training objective is the standard language modeling loss that encourages the LM to assign higher probability to the target sequence $\mathbf{y}$ given the claim $\mathbf{x}$ and top-N retrieved documents.

5.3 Distillation Techniques

Although directly summarizing fact-check articles $\mathbf{z}$ can generate justifications with reasonable quality in previous work (Kotonya and Toni, 2020a; Atanasova et al., 2020), $\mathbf{z}$ is by no means available during inference for new claims in real-world deployment, as we discussed in §1, making the previous methods impractical. We propose a realistic approach to address this limitation: distilling information from $\mathbf{z}$ as auxiliary supervisory signals for training phase only. We introduce two types of techniques based on the granularity of distillation from fact-check articles. The first is article-level distillation, which utilizes aggregated information from the entire $\mathbf{z}$. The second is chunk-level distillation, where we split each article $\mathbf{z}$ as multiple disjoint 100-word chunks $\mathbf{z} = \{\mathbf{z}_i\}_{i=1}^{M}$, where $M = \lceil \frac{|\mathbf{z}|}{100} \rceil$. Chunk-level distillation utilizes individual information of each chunk $\mathbf{z}_i$. Both types of distillation techniques can be applied to train the retriever and LM.

5.3.1 Article-level Distillation

Article-level distillation is performed at the entirety of a fact-check article, aiming at utilizing the global-level alignment between fact-check article $\mathbf{z}$ and retrieved documents D_N as supervisory signals for model training. The basic idea is that the more similar D_N and $\mathbf{z}$ are, the easier it is for LM to generate justification based on D_N closely approximating that generated based on $\mathbf{z}$. This alignment serves two main purposes. Firstly, the similarity between D_N and $\mathbf{z}$ can act as a supervisory signal, guiding the retriever to prioritize the ranking of documents in D_N to resemble $\mathbf{z}$. Secondly, the justification generated by the LM

based on $\mathbf{z}$ can be used as a supervision signal to encourage the LM using D_N to generate justification as similar as those generated based on $\mathbf{z}$. Next, we will discuss two training losses that serve both purposes.

Retrieval Loss. The technique for training retriever is based on the similarity between the *entire* fact-check article $\mathbf{z}$ and retrieved documents D_N. However, the length of $\mathbf{z}$ is commonly larger than the maximum input length (i.e., 512 tokens) of query encoder. Therefore, we use the trainable query encoder $\mathbf{E}_c$ to represent $\mathbf{z}$ by aggregating the embeddings of all its chunks and obtain $\bar{\mathbf{E}}_c(\mathbf{z}) = \frac{1}{M}\sum_{i=1}^{M}\mathbf{E}_c(\mathbf{z}_i)$. The training objective is to minimize the mean-squared-error (MSE) loss between the embeddings of $\mathbf{z}$ and $\mathbf{d}_i$:

$$\mathcal{L}_g^{\text{ret}} = \frac{1}{N|\bar{\mathbf{E}}_c(\mathbf{z})|}\sum_{j=1}^{N}||\bar{\mathbf{E}}_c(\mathbf{z}) - \mathbf{E}_d(\mathbf{d}_j)||_2^2. \quad (1)$$

Generation Loss. The technique for training the LM generation is based on the distance between the generated justification using retrieved documents D_N and that directly using the fact-check article $\mathbf{z}$. During training, the generation $\hat{\mathbf{y}}$ of the LM using $\mathbf{z}$ as input is regarded as supervision signal to guide model's learning. Let $p_{\text{L}}(\mathbf{y} \mid \mathbf{x}, D_N) = \prod_{k=1}^{|\mathbf{y}|} p_{\text{L}}(t_k \mid \mathbf{x}, D_N, t_{<k})$ be the LM probability of generating the ground-truth justification $\mathbf{y}$ conditioned on $\mathbf{x}$ and D_N, where $p_{\text{L}}(t_k \mid \mathbf{x}, D_N, t_{<k})$ is the probability of each token t_k assigned by the LM and $t_{<k}$ denotes the tokens generated prior to t_k. Similarly, the LM probability of generating $\mathbf{y}$ conditioned on $\mathbf{z}$ is $p_{\text{L}}(\mathbf{y} \mid \mathbf{x}, \mathbf{z})$. The training objective is to minimize the MSE loss between these two distributions:

$$\mathcal{L}_g^{\text{lm}} = \frac{1}{|\mathbf{y}||\mathcal{V}|}\sum_{k=1}^{|\mathbf{y}|}\sum_{i=1}^{|\mathcal{V}|}(p_{\text{L}}(t_i \mid \mathbf{x}, D_N, t_{<k}) - p_{\text{L}}(t_i \mid \mathbf{x}, \mathbf{z}, t_{<k}))^2, \quad (2)$$

where $\mathcal{V}$ is the vocabulary of the LM.

5.3.2 Chunk-level Distillation

Chunk-level distillation is performed at the granularity of each chunk of fact-check article, leveraging the alignment between chunks $\{\mathbf{z}_i\}_{i=1}^{M}$ and documents $\{\mathbf{d}_j\}_{j=1}^{N}$ to provide supervisory signals for model training. The intuition is that different chunks of the fact-check article could be derived from rearranging or modifying specific text spans sourced from reference documents. Further, the chunks $\{\mathbf{z}_i\}_{i=1}^{M}$ may correspond to certain parts of the ground-truth justification $\mathbf{y}$. Thus, $\{\mathbf{z}_i\}_{i=1}^{M}$ can be seen as the ''connections'' between D_N and $\mathbf{y}$. Aligning $\{\mathbf{d}_j\}_{j=1}^{N}$ and $\{\mathbf{z}_i\}_{i=1}^{M}$ intuitively aids the model in learning the mapping from D_N to $\mathbf{y}$, hence improving its performance. However, there is no chunk-level annotation available, which poses an important challenge for training. We design two training techniques to address it for chunk-level distillation in both retriever and LM.

Retrieval Loss. The technique for training the retriever is based on the relation between similarity score and the LM perplexity, which is inspired by Izacard et al. (2023) and Shi et al. (2023). Intuitively, the more similar the text chunk $\mathbf{z}_i$ is to the document $\mathbf{d}_j$, the lower LM perplexity of generating $\mathbf{z}_i$ conditioned on $\mathbf{d}_j$:

$$s(\mathbf{z}_i, \mathbf{d}_j) \propto p_{\text{L}}(\mathbf{z}_i \mid \mathbf{x}, \mathbf{d}_j),$$

where $s(\mathbf{z}_i, \mathbf{d}_j) = \mathbf{E}_c(\mathbf{z}_i)^\top \mathbf{E}_d(\mathbf{d}_j)$. We train the retriever to learn the alignment between $\mathbf{d}_j$ and its most similar chunk $\mathbf{z}_{j^*}$, where $j^* = \arg\max_{i\in[1,M]} s(\mathbf{z}_i, \mathbf{d}_j)$. It involves minimizing the KL-divergence between the similarity score $s(\mathbf{z}_{j^*}, \mathbf{d}_j)$ and the corresponding LM probability of $\mathbf{z}_{j^*}$ conditioned on $\mathbf{d}_j$ and $\mathbf{x}$. Specifically, let the document distribution over D_N be $p_{\text{R}}(\mathbf{d}_j \mid \mathbf{z}_i) = \frac{\exp(s(\mathbf{z}_i, \mathbf{d}_j))}{\sum_{k=1}^{N}\exp(s(\mathbf{z}_i, \mathbf{d}_k))}$, and the document posterior distribution according to the LM be $q_{\text{L}}(\mathbf{z}_{j^*} \mid \mathbf{x}, \mathbf{d}_j) = \frac{\exp(\log p_{\text{L}}(\mathbf{z}_{j^*}|\mathbf{x}, \mathbf{d}_j))}{\sum_{k=1}^{N}\exp(\log p_{\text{L}}(\mathbf{z}_{j^*}|\mathbf{x}, \mathbf{d}_k))}$. Finally, the loss function for optimizing the retriever is given as:

$$\mathcal{L}_c^{\text{ret}} = \sum_{j=1}^{N} q_{\text{L}}(\mathbf{z}_{j^*} \mid \mathbf{x}, \mathbf{d}_j) \log \frac{q_{\text{L}}(\mathbf{z}_{j^*} \mid \mathbf{x}, \mathbf{d}_j)}{p_{\text{R}}(\mathbf{d}_j \mid \mathbf{z}_{j^*})}. \quad (3)$$

This loss is exclusively used to optimize the retriever's parameters, without affecting the LM.

Generation Loss. Our technique for training LM utilizes the attention scores of the LM to

train the LM itself, which is inspired by previous work of open-domain QA that trains a retriever by learning to approximate the attention scores of the reader (Izacard and Grave, 2021a; Izacard et al., 2023). The cross-attention scores between input and output can be used as a proxy of the usefulness of each input to the justification. We firstly average decoder cross-attention scores over all attention heads, layers, and tokens for each retrieved document $\mathbf{d}_j$, resulting an averaged attention score $a(\mathbf{x} \oplus \mathbf{d}_j)$, where $\oplus$ denotes concatenation. Then the score that indicates the usefulness of $\mathbf{d}_j$ is obtained by applying the softmax operator $p(\mathbf{d}_j) = \frac{\exp(a(\mathbf{x} \oplus \mathbf{d}_j))}{\sum_{k=1}^{N} \exp(a(\mathbf{x} \oplus \mathbf{d}_k))}$ following Izacard et al. (2023). Similarly, the score for each chunk $\mathbf{z}_i$ is $p(\mathbf{z}_i)$, while the score of the most similar chunk $\mathbf{z}_{j^*}$ to $\mathbf{d}_j$ is $p'(\mathbf{z}_{j^*}) = \frac{\exp(p(\mathbf{z}_{j^*}))}{\sum_{k=1}^{N} \exp(p(\mathbf{z}_{k^*}))}$. The objective is to encourage the score of $\mathbf{d}_j$ to approximate the score of its most similar chunk $\mathbf{z}_{j^*}$. We then minimize the KL-divergence between distributions of these two scores:

$$\mathcal{L}_c^{\text{lm}} = \sum_{j=1}^{N} p'(\mathbf{z}_{j^*}) \log \frac{p'(\mathbf{z}_{j^*})}{p(\mathbf{d}_j)}. \quad (4)$$

6 Experiments and Results

6.1 Evaluation Metrics

To assess the consistency of generated justifications with ground truth, we employ a spectrum of metrics to make our evaluation balance between factual accuracy and style diversity of verbal expressions: **ROUGE** (Lin, 2004) counts the number of overlapping units (e.g., *n*-gram and word sequences) between output justifications and ground truths. **MAUVE** (Pillutla et al., 2021) measures the divergence between output justifications and the ground truths, which could reflect whether the output is fluent and coherent to the ground (Xie et al., 2023; Krishna et al., 2022b; Gao et al., 2023; Xu et al., 2023). **SummaCC** expands the SummaC (Laban et al., 2022) to evaluate the coverage and factual consistency through checking entailment between the output justifications and ground truth. It sums the aggregating NLI scores over the pairs of the entire output justification and each sentence in the ground truth for coverage (Scialom et al., 2021; Gao et al., 2023), and reversely, the pairs of the entire ground truth justification and each sentence in the output justification for consistency (Laban et al., 2022).

6.2 Fallacy of Fact-Check Summarization

We investigate how the previous approach based on fact-check article summarization (Kotonya and Toni, 2020b; Atanasova et al., 2020) fails to generalize to the realistic setting given retrieved evidence rather than fact-check articles as input.

Experimental Setup. 1) Full training: We include two existing models, ExplainMT (Atanasova et al., 2020) and ExplainerFC (Kotonya and Toni, 2020b). ExplainMT is an extractive model while ExplainerFC is extractive-abstractive. We partition the training set of ExClaim into 5,000 instances for training and 964 for validation. We train the two models to summarize fact-check articles, and test them by inputting fact-check articles versus evidence documents retrieved with BM25 (Robertson et al., 1994). 2) Few-shot training: We train the RAG model Atlas (Izacard et al., 2023) under few shots with fact-check articles as input and test it using fact-check articles versus documents retrieved by its pre-trained retriever Contriever. In this setting, Contriever will be fixed during fine-tuning since the LM's input is fact-check articles. We use randomly sampled 30 shots from the training split, and report the results averaged over 3 trials based on different seeds.

Results. As shown in Table 3, for both settings, we observe that using retrieved documents as input dramatically declines the performance compared to inputting fact-check articles. This suggests that the fact-check article summarization approach struggles to generalize to the retrieved documents, especially in few-shot setting, indicating the impracticality of previous approaches and the importance of the more realistic framework outlined in §3. That is, models need to generate justifications based on retrieved evidence instead of fact-check articles which are not available for new claims during inference.

6.3 Few-shot Justification Generation

6.3.1 Baselines

1) **Lead-4** (Nallapati et al., 2017) selects as justification the first sentence from each document among the top-4 documents retrieved by BM25.

Method	#Para.	Test	ROUGE-1	ROUGE-2	ROUGE-L	SummaCC	MAUVE
ExplainMT$_{\text{Full-dataset}}$ (Atanasova et al., 2020)	132M	F.C. Article	$35.01_{(-)}$	$22.13_{(-)}$	$21.25_{(-)}$	$22.70_{(-)}$	$5.59_{(-)}$
		Retr. Docs	$19.33_{(-)}$	$9.55_{(-)}$	$17.59_{(-)}$	$9.34_{(-)}$	$5.27_{(-)}$
ExplainerFC$_{\text{Full-dataset}}$ (Kotonya and Toni, 2020b)	340M	F.C. Article	$62.10_{(-)}$	$38.03_{(-)}$	$54.25_{(-)}$	$50.67_{(-)}$	$14.63_{(-)}$
		Retr. Docs	$47.16_{(-)}$	$24.88_{(-)}$	$44.13_{(-)}$	$35.82_{(-)}$	$10.07_{(-)}$
Atlas$_{\text{Few-shot}}$ (Izacard et al., 2023)	∼3B	F.C. Article	$40.93_{(0.97)}$	$26.71_{(1.15)}$	$33.98_{(1.01)}$	$29.72_{(1.22)}$	$28.25_{(2.46)}$
		Retr. Docs	$28.14_{(0.87)}$	$13.91_{(1.31)}$	$21.87_{(1.12)}$	$12.64_{(0.87)}$	$25.37_{(0.69)}$

Table 3: Results of justification generation methods trained on Fact-check Article (F.C. Article) and tested on Fact-check Article / Retrieved Documents (Retr. Docs). Para.: Parameters. Standard deviation is in parentheses.

2) **Retriever + ICL-enabled LMs:** We use BM25 as the sparse retriever and Contriever (Izacard et al., 2022) as the dense retriever, and choose Flan-T5 (11B) (Chung et al., 2022), Llama2 (70B) (Touvron et al., 2023), and GPT-4 (OpenAI, 2023) as the ICL-enabled LMs. We prompt the model to generate justifications by concatenating few-shot training instances along with a test instance. 3) **Atlas** (Izacard et al., 2023) is the SoTA RAG model with strong few-shot ability, which consists of a trainable dense retriever Contriever and a LM-adapted variant of T5 (Lester et al., 2021) with FiD (Izacard and Grave, 2021b) modified to increase the number of retrieved documents. We also include a non-joint training setting by replacing the retriever with BM25.

6.3.2 Experimental Setup

For our method JustiLM, we randomly sample 30 instances from the training set for fine-tuning. We use the Atlas (Izacard et al., 2023) with its released pre-trained checkpoint[3] of 3B parameters as our backbone model. Following the Atlas paper, we retrieve top-20 documents for each instance. We set training steps as 100, batch size as 8, and learning rate as 4×10^{-5} with linear decay and 5 warmup steps for both the LM and the retriever.

For the distillation techniques to train the LM, we begin by fine-tuning the LM to take fact-check articles as auxiliary input and generate justification, which provides a warmup for LM. For BM25 + ICL-enabled LMs, we use the Pyserini[4] toolkit to build BM25 model. For Flan-T5, We use the code and pre-trained checkpoints from HuggingFace Transformers.[5] We use the original code and pre-trained checkpoints of Llama2.[6] We use the API service of GPT-4 from OpenAI.[7] Given different length constraints of these LMs, we intend to maximize the utilization of their specific input capabilities. We adjust the number of the shots and/or the number of retrieved documents to maximally utilize their input context windows. We prioritize to ensure that these models have access to as many of the top-20 retrieved documents as possible because effective generation requires an adequate amount of information, with the secondary goal to maximize the number of few-shot examples used. Specifically, we set 1-shot ICL with top-10 documents for Flan-T5, 2-shot ICL with top-20 documents for Llama2 and 3-shot with top-20 documents for GPT-4.

For fair and robust comparison, we perform experiments three times, with training instances sampled using different random seeds. We report the mean and standard deviation of each metric over the three runs in all experiments. The seeds and training instances are kept the same across different models. All the experiments use a server with 8 NVIDIA Tesla-V100 32GB GPUs.

6.3.3 Main Results

The results of few-shot justification generation methods are reported in Table 4a. Lead-4 that directly presents the retrieved documents as justification does not yield satisfactory results, due to simple evidence stacking without generating a clear explanation of the rationale.

[3] https://github.com/facebookresearch/atlas.
[4] https://github.com/castorini/pyserini.
[5] https://huggingface.co/google/flan-t5-xxl.
[6] https://github.com/facebookresearch/LLAMA.
[7] https://openai.com/gpt-4.

	#Parameters	ROUGE-1	ROUGE-2	ROUGE-L	SummaCC	MAUVE
Lead-4 (Nallapati et al., 2017)						
	–	$22.72_{(–)}$	$5.72_{(–)}$	$14.11_{(–)}$	$2.26_{(–)}$	$7.95_{(–)}$
Retriever + ICL-enabled LMs						
BM25 (Robertson et al., 1994)						
+ Flan-T5 (Chung et al., 2022)	11B	$27.99_{(2.39)}$	$14.14_{(1.06)}$	$20.74_{(1.66)}$	$14.55_{(0.90)}$	$12.42_{(1.22)}$
+ Llama2 (Touvron et al., 2023)	70B	$31.45_{(0.51)}$	$12.36_{(0.25)}$	$20.72_{(0.22)}$	$13.05_{(0.38)}$	$7.88_{(0.15)}$
+ GPT-4 (OpenAI, 2023)	Unkown	$\mathbf{39.72}_{(1.97)}$	$17.12_{(1.97)}$	$26.18_{(2.26)}$	$\mathbf{24.98}_{(2.49)}$	$14.73_{(2.86)}$
Contriever (Izacard et al., 2022)						
+ Flan-T5	∼11B	$23.75_{(1.91)}$	$11.34_{(1.17)}$	$18.11_{(1.48)}$	$9.93_{(0.29)}$	$12.07_{(0.90)}$
+ Llama2	∼70B	$31.28_{(0.51)}$	$11.52_{(0.82)}$	$20.42_{(0.70)}$	$11.06_{(0.14)}$	$7.91_{(0.09)}$
+ GPT-4	Unkown	$\underline{36.83}_{(1.37)}$	$14.10_{(1.66)}$	$23.36_{(1.75)}$	$\underline{20.07}_{(2.37)}$	$9.85_{(0.96)}$
Atlas (Izacard et al., 2023)						
No joint training	3B	$31.42_{(1.61)}$	$16.53_{(0.86)}$	$24.67_{(1.00)}$	$13.55_{(0.54)}$	$25.19_{(4.37)}$
Joint training	∼3B	$31.91_{(1.78)}$	$17.81_{(1.19)}$	$25.60_{(1.16)}$	$13.81_{(1.11)}$	$25.51_{(2.08)}$
JustiLM (Ours)						
$\mathcal{L}_g^{\text{ret}} + \mathcal{L}_g^{\text{lm}}$	∼3B	$33.48_{(1.33)}$	$18.59_{(0.79)}$	$27.12_{(0.81)}$	$15.04_{(1.27)}$	$20.29_{(2.00)}$
$\mathcal{L}_g^{\text{ret}} + \mathcal{L}_c^{\text{lm}}$	∼3B	$36.70_{(0.77)}$	$\mathbf{19.23}_{(0.84)}$	$\mathbf{28.39}_{(0.75)}$	$14.80_{(0.45)}$	$32.99_{(3.33)}$
$\mathcal{L}_c^{\text{ret}} + \mathcal{L}_g^{\text{lm}}$	∼3B	$36.51_{(1.01)}$	$18.67_{(1.00)}$	$27.94_{(0.96)}$	$14.77_{(0.19)}$	$\mathbf{37.08}_{(1.53)}$
$\mathcal{L}_c^{\text{ret}} + \mathcal{L}_c^{\text{lm}}$	∼3B	$36.30_{(0.91)}$	$\underline{18.68}_{(0.96)}$	$\underline{27.97}_{(0.99)}$	$14.69_{(0.48)}$	$\underline{35.30}_{(1.09)}$

(a) On the original test set with 987 claims indicated in Table 2.

	#Parameters	ROUGE-1	ROUGE-2	ROUGE-L	SummaCC	MAUVE
Lead-4 (Nallapati et al., 2017)						
	–	$21.87_{(–)}$	$3.95_{(–)}$	$12.61_{(–)}$	$1.70_{(–)}$	$6.62_{(–)}$
Retriever + ICL-enabled LMs						
BM25 (Robertson et al., 1994)						
+ Flan-T5 (Chung et al., 2022)	11B	$22.86_{(2.27)}$	$7.63_{(0.70)}$	$14.74_{(1.52)}$	$10.94_{(2.37)}$	$7.00_{(0.17)}$
+ Llama2 (Touvron et al., 2023)	70B	$31.01_{(0.29)}$	$9.64_{(0.32)}$	$18.73_{(0.17)}$	$11.49_{(0.69)}$	$6.99_{(0.60)}$
+ GPT-4 (OpenAI, 2023)	Unkown	$\mathbf{38.28}_{(1.44)}$	$13.74_{(1.75)}$	$23.36_{(2.20)}$	$\mathbf{25.10}_{(2.29)}$	$7.47_{(1.30)}$
Contriever (Izacard et al., 2022)						
+ Flan-T5	∼11B	$20.44_{(1.27)}$	$7.93_{(0.48)}$	$14.45_{(0.85)}$	$10.18_{(2.03)}$	$8.24_{(0.48)}$
+ Llama2	∼70B	$31.01_{(0.84)}$	$9.81_{(0.73)}$	$19.07_{(0.63)}$	$10.75_{(0.52)}$	$6.62_{(0.54)}$
+ GPT-4	Unkown	$\underline{35.93}_{(1.09)}$	$12.07_{(1.51)}$	$21.46_{(1.79)}$	$\underline{21.79}_{(2.22)}$	$6.25_{(0.37)}$
Atlas (Izacard et al., 2023)						
No joint training	3B	$29.76_{(0.98)}$	$13.40_{(0.34)}$	$22.16_{(0.32)}$	$10.78_{(0.55)}$	$12.56_{(1.56)}$
Joint training	∼3B	$30.78_{(1.95)}$	$15.75_{(1.72)}$	$23.84_{(1.48)}$	$12.20_{(0.45)}$	$14.09_{(2.34)}$
JustiLM (Ours)						
$\mathcal{L}_g^{\text{ret}} + \mathcal{L}_g^{\text{lm}}$	∼3B	$32.76_{(0.89)}$	$17.40_{(0.65)}$	$26.61_{(0.61)}$	$14.75_{(1.45)}$	$10.57_{(0.94)}$
$\mathcal{L}_g^{\text{ret}} + \mathcal{L}_c^{\text{lm}}$	∼3B	$35.55_{(0.31)}$	$\mathbf{17.84}_{(0.48)}$	$\mathbf{27.30}_{(0.21)}$	$14.11_{(1.40)}$	$16.78_{(4.64)}$
$\mathcal{L}_c^{\text{ret}} + \mathcal{L}_g^{\text{lm}}$	∼3B	$35.51_{(0.51)}$	$17.21_{(0.70)}$	$26.53_{(0.06)}$	$14.30_{(0.40)}$	$\mathbf{20.02}_{(7.39)}$
$\mathcal{L}_c^{\text{ret}} + \mathcal{L}_c^{\text{lm}}$	∼3B	$35.48_{(0.59)}$	$\underline{17.52}_{(0.86)}$	$\underline{26.92}_{(0.57)}$	$13.99_{(0.49)}$	$\underline{19.17}_{(7.04)}$

(b) On the new test set with 348 claims published later than the claims from the WatClaimCheck dataset used for training.

Table 4: Few-shot justification generation results on test set (a) and new test set (b). Standard deviation is in parentheses.

Both Flan-T5 and Llama2 outperform Lead-4, demonstrating the LM's ability to generate justifications based on retrieved evidence. Flan-T5 performs comparably with Llama2 in ROUGE and SummaCC scores and better in MAUVE, despite much fewer parameters. The reasons are likely two-fold: 1) Flan-T5's instruction fine-tuning on 1.8K tasks, which effectively enhances

the pre-trained language models (Sanh et al., 2022; Chung et al., 2022); 2) its fine-tuning on Chain-of-Thought (CoT) data (Wei et al., 2022), aligning with the common presentation of ground-truth justifications that provide rationales to conclude the veracity, as exemplified in Table 1.

Incorporating ICL-enabled LMs with the dense retriever Contriever does not exhibit improvement over using the sparse retriever BM25. Dense retrievers that trained on extensive in-domain training datasets like MS-MARCO (Nguyen et al., 2016), are often surpassed by sparse retrievers when applied to new domains without large annotated datasets (Thakur et al., 2021; Izacard et al., 2022). While Contriever is a strong unsupervised retriever for bridging this gap, BM25 still remains competitive (Izacard et al., 2022).

When training only the LM of Atlas, it demonstrates superior overall performance compared to Flan-T5 and Llama2, despite its much fewer parameters. This finding indicates that merely relying on the implicit knowledge of LMs without parameter updates is insufficient when the size of LM is not large enough. Joint training of the retriever and LM leads to further performance gains, implying its benefits in the few-shot setting.

Compared to Atlas, JustiLM makes improvements in different metrics, indicating that utilizing fact-check article as auxiliary training signals enhances justification quality. With our proposed distillation techniques, JustiLM considerably improves all ROUGE scores. Compared to Atlas, the combination of article-level distillation on retriever and chunk-level distillation on LM increases ROUGE-1, ROUGE-2, and ROUGE-L scores by 15.0%, 7.97%, and 10.9%, respectively, suggesting that JustiLM can generate justifications which are more similar to those written by fact-checkers. Furthermore, 3 out of 4 combinations of distillation techniques outperform Atlas in MAUVE scores, with the highest gain being 45%. This suggests that JustiLM's justifications are more fluent and coherent with ground truths. It can be attributed to our distillation method allowing the model to learn from fact-check articles that are much more informative and detailed than the explanatory justifications. Lastly, JustiLM effectively enhances the SummaCC score, indicating the improvements on the factual consistency of generated justifications.

GPT-4 demonstrates exceptionally strong ability in providing factually consistent responses and outperforms other ICL-enabled methods Flan-T5 and Llama2 across all metrics. In comparison, JustiLM falls relatively below GPT-4 in ROUGE-1 and SummaCC, but outperforms GPT-4 in ROUGE-2/L and MAUVE. This highlights its effectiveness, especially considering its small model size and independence from intensive compute and storage resources required by very large models. Also, its ease of fine-tuning with more and new training data provides significant flexibility in addressing the ever-changing landscape of misinformation.

6.3.4 Generalization on New Claims

To address the concern of pre-trained LMs having potentially seen the evaluation data during their pre-training, we investigate how different methods perform on a new test set with new emerging claims made after their training. Since the WatClaimCheck dataset exclusively encompasses claims prior to July 2021 (Khan et al., 2022) and the newest pre-training data of Llama2 are cut off by September 2022, we gather a new set of claims made between October 2022 and September 2023, yielding a new test set comprising 348 instances, each with their associated reference documents and justifications. Following the same steps detailed in §4, the newly collected reference documents are added into the corpus for model retrieval. As shown in Table 4b, all methods demonstrate performance drop on the new test set. Nonetheless, the findings obtained based on the original test set still hold true for the new test data. Additionally, compared to baseline methods, the relatively mild performance drop in JustiLM suggests stronger generalizability and robustness of our distillation techniques.

6.3.5 Ablation on Distillation Techniques

Table 5 reports the result of ablations on our distillation techniques. We observe that the distillation during LM training results in greater improvements compared to the retriever. This is expected, considering that the LM benefits from direct supervision from ground-truth justifications during training, while the retriever relies on the weak supervision from LM and the distillation of fact-check articles. Additionally, the LM has a larger number of parameters than the retriever, with 3 billion parameters for the LM compared to 110 million parameters for the retriever. As

Component	Loss	ROUGE-1	ROUGE-2	ROUGE-L	SummaCC	MAUVE
Retriever	$\mathcal{L}_g^{\text{ret}}$	$32.13_{(0.99)}$	$16.45_{(0.39)}$	$25.15_{(0.59)}$	$14.53_{(0.23)}$	$26.53_{(3.38)}$
	$\mathcal{L}_c^{\text{ret}}$	$31.29_{(1.53)}$	$17.26_{(0.94)}$	$25.19_{(1.15)}$	$13.77_{(1.41)}$	$19.17_{(1.57)}$
LM	$\mathcal{L}_g^{\text{lm}}$	$36.30_{(1.80)}$	$19.23_{(1.05)}$	$28.52_{(1.04)}$	$14.92_{(0.73)}$	$27.09_{(7.20)}$
	$\mathcal{L}_c^{\text{lm}}$	$37.03_{(0.80)}$	$18.89_{(0.90)}$	$28.29_{(0.79)}$	$14.56_{(0.69)}$	$34.16_{(3.73)}$

Table 5: Results of ablations on different distillation techniques. Parentheses enclose standard deviation.

Method	macro-F1	ROUGE-1	ROUGE-2	ROUGE-L	SummaCC	MAUVE
Majority	$23.34_{(-)}$	–	–	–	–	–
Atlas-CLS	$25.81_{(0.46)}$	–	–	–	–	–
JustiLM-$\mathcal{L}_g^{\text{lm}}$	$44.00_{(1.51)}$	$32.52_{(1.39)}$	$18.20_{(0.61)}$	$26.34_{(0.88)}$	$14.76_{(1.17)}$	$18.68_{(1.80)}$
JustiLM-$\mathcal{L}_c^{\text{lm}}$	$41.33_{(4.49)}$	$35.87_{(1.02)}$	$19.52_{(0.68)}$	$28.22_{(0.86)}$	$15.02_{(0.95)}$	$32.98_{(2.96)}$

Table 6: Results of joint veracity prediction and justification generation. Parentheses enclose standard deviation.

a result, the LM tends to capture more knowledge from fact-check article during the distillation process, leading to substantial improvements in performance.

6.4 Joint Veracity-Justification Performance

In this section, we demonstrate that JustiLM can be easily extended for joint veracity prediction and justification generation. We follow Khan et al. (2022) to map the original veracity labels assigned by fact-checking websites, resulting in 388, 532, and 67 instances for the false, mixture, and true classes in the test split, respectively. Such class imbalance is consistent with the report by Khan et al. (2022). To mitigate the impact of imbalanced class distribution, we balance the 30 training shots across the three classes by randomly sampling 10 instances per class from the training set.

We make the LM generate the justification and veracity label at the same time. For veracity label prediction, let $y_{\text{cls},i}$ be a veracity label, and its predicted score assigned by the LM conditioned on the claim and the retrieved documents is defined as $\beta(\mathbf{x}, y_{\text{cls},i}, D_N) = \frac{1}{|y_{\text{cls},i}|} \log p_{\text{L}}(y_{\text{cls},i} \mid \mathbf{x}, D_N)$ following Liu et al. (2022). In this way, we rank all classes by the predicted scores and select the top-ranked class. During training, we calculate the probability of prediction by applying softmax function on the predicted scores, and use cross-entropy as the loss function.

Table 6 presents the result. The Atlas-CLS, which directly predicts veracity label with Atlas, shows a limited improvement in macro-F1 score compared to the Majority method. This suggests that predicting the veracity of real-world claims remains challenging for this original RAG model in a few-shot setting. When performing joint veracity prediction and justification generation with the LM training, a substantial boost in verdict prediction is observed for our method. Specifically, we achieve absolute improvements of 18.19 and 15.52 in macro-F1 using article-level and chunk-level techniques, respectively. This indicates that justification generation can help veracity prediction by consolidating evidence from retrieved documents. We also find that jointly training JustiLM with the veracity prediction task does not improve the performance of justification generation, which is consistent with the findings by Atanasova et al. (2020). We conjecture that it remains challenging for the model to boost both tasks simultaneously with few-shot training instances. Potential solutions could consider either leveraging a larger multi-task training dataset, such as T0 (Sanh et al., 2022), or using an independent veracity classifier that can be jointly trained with the retriever and the LM. However, both options necessitate adding data and computational resources. We will leave this for future studies.

6.5 Case Study

Table 7 presents example justifications generated by JustiLM, the strong ICL baseline GPT-4, and the few-shot RAG model Atlas. Atlas's generated

Claim	Says Wisconsin women facing pay discrimination can't ''do something about it'' under bill passed by Republicans.
Ground Truth	Falk said women facing pay discrimination can't ''do something about it'' under a bill passed by Wisconsin Republicans. But of the four legal options available under the current law, only one would be eliminated by the bill. We rate Falk's statement False.
Atlas	Falk said that under a Republican-backed bill, women facing pay discrimination can't ''do something about it.'' Falk said that under the GOP bill, women facing pay discrimination can't ''do something about it.'' But the bill doesn't change the law. It simply makes it harder for women to file lawsuits. Falk said that under the GOP bill, women facing pay discrimination can't ''do something about it.''
GPT-4	A statement by Kathleen Falk, a former Dane County executive, claims that women in Wisconsin would no longer have ''a right to do something'' about pay discrimination on the job if a bill is signed by Governor Walker. The bill in question would take away the right of women who suffer pay discrimination to sue in state court. However, there are still other avenues for employment discrimination victims, including filing a complaint with the state Equal Rights Division, filing a complaint with the federal Equal Employment Opportunity Commission, and filing a lawsuit in federal court. Therefore, while the bill may limit one avenue for action, it does not completely prevent women from taking action against pay discrimination.
JustiLM	Falk said that women facing pay discrimination can't ''do something about it'' under a bill passed by Republicans. The measure would make it harder on victims of employment discrimination because it is easier to sue in state court than in federal court. But eliminating the state lawsuit option wouldn't affect the three other avenues of legal recourse available to workplace discrimination victims. And Falk's claim was women facing pay discrimination would have no options at all.

Table 7: An example of generated justifications by different methods compared to the ground-truth justification.

justification catches that the GOP bill does not change the law, but fails to highlight the key point that women still have viable avenues to address pay discrimination. Both GPT-4 and JustiLM successfully refute the claim by providing that crucial point.

More specifically, Atlas falls short in delivering convincing and comprehensive justification due to its tendency to provide incomplete and repetitive responses. In contrast, GPT-4, being the SoTA LLM, impresses with its ability to generate well-rounded justification, but appears to be lengthy and less focused. JustiLM, on the other hand, successfully highlights key points for fact-checking the claim with a precise and refined justification. Despite its relatively small model size, JustiLM may not always offer the same level of details as GPT-4, but it can produce concise and accurate justifications that closely resemble the ground truth, making JustiLM promising and valuable for users seeking quick and trustworthy fact-check explanations.

7 Discussion

There is no passage-/sentence-level annotation in the original long-form reference documents and fact-check articles, which are costly to obtain. We do not have ground truths for training and evaluating evidence retrieval model. Since these long documents bury specific evidence in them, directly using them for training will introduce a considerable amount of irrelevant text. While we mitigate this challenge by splitting each original reference document into disjoint 100-word chunks for retrieval, we believe that acquiring fine-grained evidence annotations will benefit the training and evaluation.

In our experimental setup, evidence retrieval is conducted under the assumption that the needed evidence for fact-checking a given claim exists in the retrieval corpus. However, in a real-world searching scenario where gold evidence may be absent from the retrieval corpus, it is valuable to investigate how justification generation methods

perform under this more challenging scenario by varying the ratio of gold reference documents in the retrieval corpus.

Additionally, while our experiments include the NLI-based metric SummaCC, providing automated evaluation on the factuality of generated justifications, we believe that a sound human evaluation should involve professional fact-checkers. Such evaluation, currently not conducted, necessitates close collaboration with fact-checking organizations and needs particular networking and setup, such as the integration with their existing workflow and the provision of motives for them to participate in evaluation, which could be warranted as a separate study by itself and is part of our future plan.

As the SoTA LLM, GPT-4 shows strong ability in generating factually consistent and informative justifications, therefore, developing justification methods based on those powerful API-based LLMs is beneficial. However, these blackbox LLMs have strict constraints on accessing their specific internal information, which poses important open challenges for being interacted with deeply and providing supervision signals to retriever.

In this work, we address the justification generation task with a realistic approach, which generates justifications based on the retrieved evidence using an end-to-end retrieval-augmented language model. Furthermore, incorporating our distillation techniques with the RAG model Atlas demonstrates a marked improvement in performance. This affirms that utilizing fact-check articles during training to provide supervision signals can strongly enhance justification generation.

8 Conclusion and Future Work

We propose a justification generation language model JustiLM for realistic fact-checking of real-world news claims, where justification generation is performed based on retrieved evidence from large textual corpus, and introduce a new benchmark dataset ExClaim for this task. JustiLM leverages fact-check articles as auxiliary resources during training to distill article-level and chunk-level training signals to guide justification writing. Experimental results show JustiLM outperforms ICL-enabled Flan-T5 and Llama2, as well as the SoTA few-shot RAG model Atlas. JustiLM also demonstrates comparable and promising performance when compared to GPT-4.

In the future, we will explore the adaptation of various LLM-based reasoning methods (e.g., CoT [Wei et al., 2022], ToT [Yao et al., 2023b], and GoT [Besta et al., 2023]) into JustiLM to enhance the reasoning ability for improving the task of justification generation, which aims to assist the LMs in providing better signals for guiding evidence retrieval and improving reasoning over retrieved evidence during justification generation. We also plan to develop a human evaluation scheme involving fact-checking experts to provide a more comprehensive and efficient assessment on machine-generated justifications.

Acknowledgments

We would like to thank the anonymous reviewers and action editor Fei Liu for their helpful suggestions. We are also grateful to Alessandro Moschitti for his valuable comments and discussion.

References

Naser Ahmadi, Joohyung Lee, Paolo Papotti, and Mohammed Saeed. 2019. Explainable fact checking with probabilistic answer set programming. In *Proceedings of the 2019 Truth and Trust Online Conference, TTO 2019*, London, UK. https://doi.org/10.36370/tto.2019.15

Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 85–90, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-5513

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

Ramy Baly, Mitra Mohtarami, James Glass, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. 2018. Integrating stance detection and fact checking in a unified corpus. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 21–27, New Orleans, Louisiana. Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-2004

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Michal Podstawski, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2023. Graph of thoughts: Solving elaborate problems with large language models. *CoRR*, arXiv:2308.09687v3.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. 2022. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning, ICML* 2022, volume 162 of *Proceedings of Machine Learning Research*, pages 2206–2240, Baltimore, Maryland, USA. PMLR.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-1171

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416v5.

Günes Erkan and Dragomir R. Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479. https://doi.org/10.1613/jair.1523

Mohamed H. Gad-Elrab, Daria Stepanova, Jacopo Urbani, and Gerhard Weikum. 2019. Exfakt: A framework for explaining facts over knowledge graphs and text. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, WSDM 2019*, pages 87–95, Melbourne, VIC, Australia. ACM. https://doi.org/10.1145/3289600.3290996

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.398

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

Ashim Gupta and Vivek Srikumar. 2021. X-fact: A new benchmark dataset for multilingual fact checking. 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 675–682,

Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-short.86

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20, pages 3929–3938. JMLR.org.

Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. 2019. A richly annotated corpus for different tasks in automated fact-checking. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 493–503, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/K19-1046

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020*, pages 9726–9735, Seattle, WA, USA. Computer Vision Foundation / IEEE. https://doi.org/10.1109/CVPR42600.2020.00975

Xuming Hu, Zhijiang Guo, GuanYu Wu, Aiwei Liu, Lijie Wen, and Philip Yu. 2022. CHEF: A pilot Chinese dataset for evidence-based fact-checking. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3362–3376, Seattle, United States. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.246

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*, 2022.

Gautier Izacard and Edouard Grave. 2021a. Distilling knowledge from reader to retriever for question answering. In *9th International Conference on Learning Representations, ICLR 2021*, Virtual Event, Austria. OpenReview.net.

Gautier Izacard and Edouard Grave. 2021b. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.74

Gautier Izacard, Patrick S. H. 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* , 24:251:1–251:43.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-scale similarity search with gpus. *IEEE Transations on Big Data*, 7(3):535–547. https://doi.org/10.1109/TBDATA.2019.2921572

Karen Spärck Jones. 2004. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 60(5):493–502. https://doi.org/10.1108/00220410410560573

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.550

Kashif Khan, Ruizhe Wang, and Pascal Poupart. 2022. WatClaimCheck: A new dataset for claim entailment and inference. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1293–1304, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.92

Neema Kotonya and Francesca Toni. 2020a. 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

Neema Kotonya and Francesca Toni. 2020b. 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. https://doi.org/10.18653/v1/2020.emnlp-main.623

Amrith Krishna, Sebastian Riedel, and Andreas Vlachos. 2022a. ProoFVer: Natural Logic Theorem Proving for Fact Verification. *Transactions of the Association for Computational Linguistics*, 10:1013–1030. https://doi.org/10.1162/tacl_a_00503

Kalpesh Krishna, Yapei Chang, John Wieting, and Mohit Iyyer. 2022b. RankGen: Improving text generation with large ranking models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 199–232, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.15

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

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1612

Nayeon Lee, Yejin Bang, Andrea Madotto, and Pascale Fung. 2021. Towards few-shot fact-checking via perplexity. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1971–1981, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.158

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.243

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.703

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020b. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*, pages 9459–9474, virtual.

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.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022*, New Orleans and virtual.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. 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 3730–3740, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1387

Yi-Ju Lu and Cheng-Te Li. 2020. GCAN: Graph-aware co-attention networks for explainable fake news detection on social media. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 505–514, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.48

Jing Ma, Wei Gao, Shafiq Joty, and Kam-Fai Wong. 2019. Sentence-level evidence embedding for claim verification with hierarchical attention networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2561–2571, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1244

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.546

Grégoire Mialon, Roberto Dessi, Maria Lomeli, Christoforos Nalmpantis, Ramakanth Pasunuru, Roberta Raileanu, Baptiste Roziere, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, Edouard Grave, Yann LeCun, and Thomas Scialom. 2023. Augmented language models: A survey. *Transactions on Machine Learning Research*. Survey Certification.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI 2017*, pages 3075–3081, San Francisco, California, USA. AAAI Press. https://doi.org/10.1609/aaai.v31i1.10958

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016)*, volume 1773 of *CEUR Workshop Proceedings*, Barcelona, Spain. CEUR-WS.org.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774v4.

Wojciech Ostrowski, Arnav Arora, Pepa Atanasova, and Isabelle Augenstein. 2021. Multi-hop fact checking of political claims. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021*, pages 3892–3898, Virtual Event / Montreal, Canada. ijcai.org. https://doi.org/10.24963/ijcai.2021/536

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

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaïd Harchaoui. 2021. MAUVE: measuring the gap between neural text and human text using divergence frontiers. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021*, pages 4816–4828, Virtual Event.

Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2016. Credibility assessment of textual claims on the web. In *Proceedings of the 25th ACM International Conference on Information and Knowledge*

Management, CIKM 2016, pages 2173–2178, Indianapolis, IN, USA. ACM. https://doi.org/10.1145/2983323.2983661

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research* , 21(1).

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at TREC-3. In *Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2–4, 1994*, volume 500–225 of *NIST Special Publication*, pages 109–126. National Institute of Standards and Technology (NIST).

Daniel Russo, Serra Sinem Tekiroglu, and Marco Guerini. 2023. Benchmarking the generation of fact checking explanations. *Transactions of the Association for Computational Linguistics*, 11:1250–1264. https://doi.org/10.1162/tacl_a_00601

Devendra Singh Sachan, Siva Reddy, William L. Hamilton, Chris Dyer, and Dani Yogatama. 2021. End-to-end training of multi-document reader and retriever for open-domain question answering. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021*, pages 25968–25981, Virtual Event.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: Smaller, faster, cheaper and lighter. In *NeurIPS EMC^2 Workshop*.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *The Tenth International Conference on Learning Representations, ICLR 2022*, Virtual Event.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.529

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. REPLUG: Retrieval-augmented black-box language models. *CoRR*, abs/2301.12652v4.

Sam Shleifer and Alexander M. Rush. 2020. Pre-trained summarization distillation. *CoRR*, abs/2010.13002v2.

Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining,*

KDD 2019, pages 395–405, Anchorage, AK, USA. ACM. https://doi.org/10.1145/3292500.3330935

Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan L. Boyd-Graber, and Lijuan Wang. 2023. Prompting GPT-3 to be reliable. In *The Eleventh International Conference on Learning Representations, ICLR 2023*, Kigali, Rwanda. OpenReview.net.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

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. https://doi.org/10.18653/v1/N18-1074

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288v2.

Nikhita Vedula and Srinivasan Parthasarathy. 2021. FACE-KEG: fact checking explained using knowledge graphs. In *The Fourteenth ACM International Conference on Web Search and Data Mining, WSDM '21*, pages 526–534, Virtual Event, Israel. ACM. https://doi.org/10.1145/3437963.3441828

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

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Zhouhang Xie, Sameer Singh, Julian J. McAuley, and Bodhisattwa Prasad Majumder. 2023. Factual and informative review generation for explainable recommendation. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023*, pages 13816–13824, Washington, DC, USA. AAAI Press. https://doi.org/10.1609/aaai.v37i11.26618

Jiacheng Xu, Caiming Xiong, Silvio Savarese, and Yingbo Zhou. 2023. Best-k search algorithm for neural text generation. In *Proceedings of*

the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 12385–12401, Toronto, Canada. Association for Computational Linguistics.

Fan Yang, Shiva K. Pentyala, Sina Mohseni, Mengnan Du, Hao Yuan, Rhema Linder, Eric D. Ragan, Shuiwang Ji, and Xia (Ben) Hu. 2019. Xfake: Explainable fake news detector with visualizations. In *The World Wide Web Conference, WWW 2019*, pages 3600–3604, San Francisco, CA, USA. ACM. https://doi.org/10.1145/3308558.3314119

Zhiwei Yang, Jing Ma, Hechang Chen, Hongzhan Lin, Ziyang Luo, and Yi Chang. 2022. A coarse-to-fine cascaded evidence-distillation neural network for explainable fake news detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2608–2621, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. 2023a. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2733–2743, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3539618.3591879

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik R. Narasimhan. 2023b. Tree of thoughts: Deliberate problem solving with large language models. In *Thirty-seventh Conference on Neural Information Processing Systems, NeurIPS 2023*.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2023. Generate rather than retrieve: Large language models are strong context generators. In *The Eleventh International Conference on Learning Representations, ICLR 2023*, Kigali, Rwanda. OpenReview.net.

Zhenrui Yue, Huimin Zeng, Yang Zhang, Lanyu Shang, and Dong Wang. 2023. MetaAdapt: Domain adaptive few-shot misinformation detection via meta learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5223–5239, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.286

Fengzhu Zeng and Wei Gao. 2023. Prompt to be consistent is better than self-consistent? Few-shot and zero-shot fact verification with pre-trained language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4555–4569, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.278

Xia Zeng and Arkaitz Zubiaga. 2022. Aggregating pairwise semantic differences for few-shot claim verification. *PeerJ Computer Science*, 8:e1137. https://doi.org/10.7717/peerj-cs.1137, PubMed: 36426249

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339, Virtual Event. PMLR.

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*, pages 996–1011, Nusa Dua, Bali. Association for Computational Linguistics.