Universidade Estadual de Campinas
Instituto de Computação

Jing Yang

Combating Disinformation with Explainable and Efficient Fact-checking

Combatendo desinformação com verificação factual eficiente e explicável

CAMPINAS
2024

Jing Yang

Combating Disinformation with Explainable and Efficient Fact-checking

Combatendo desinformação com verificação factual eficiente e explicável

Tese apresentada ao Instituto de Computação da Universidade Estadual de Campinas como parte dos requisitos para a obtenção do título de Doutora em Ciência da Computação.

Thesis presented to the Institute of Computing of the University of Campinas in partial fulfillment of the requirements for the degree of Doctor in Computer Science.

Supervisor/Orientador: Prof. Dr. Anderson de Rezende Rocha

Este exemplar corresponde à versão final da Tese defendida por Jing Yang e orientada pelo Prof. Dr. Anderson de Rezende Rocha.

CAMPINAS
2024

Ficha catalográfica
Universidade Estadual de Campinas (UNICAMP)
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Silvania Renata de Jesus Ribeiro - CRB 8/6592

Yang, Jing, 1993-
Y16c Combating disinformation with explainable and efficient fact-checking / Jing Yang. – Campinas, SP : [s.n.], 2024.

Orientador(es): Anderson de Rezende Rocha.
Tese (doutorado) – Universidade Estadual de Campinas (UNICAMP), Instituto de Computação.

1. Verificação de fatos. 2. Inteligência artificial explicável. 3. Eficiência de dados. I. Rocha, Anderson de Rezende, 1980-. II. Universidade Estadual de Campinas (UNICAMP). Instituto de Computação. III. Título.

Informações complementares

Título em outro idioma: Combatendo desinformação com verificação factual eficiente e explicável
Palavras-chave em inglês:
Fact-checking
Explainable artificial intelligence
Data efficiency
Área de concentração: Ciência da Computação
Titulação: Doutora em Ciência da Computação
Banca examinadora:
Anderson de Rezende Rocha [Orientador]
Agma Juci Machado Traina
Roberto Marcondes Cesar Junior
Paula Dornhofer Paro Costa
Marcelo da Silva Reis
Data de defesa: 25-11-2024
Programa de Pós-Graduação: Ciência da Computação

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0002-0035-3960
- Currículo Lattes do autor: http://lattes.cnpq.br/5918926421966556

Universidade Estadual de Campinas
Instituto de Computação

Jing Yang

Combating Disinformation with Explainable and Efficient Fact-checking

Combatendo desinformação com verificação factual eficiente e explicável

Banca Examinadora:

- Prof. Dr. Anderson de Rezende Rocha
  Instituto de Computação/Universidade Estadual de Campinas
- Prof. Dr. Agma Juci Machado Traina
  Instituto de Ciências Matemáticas e de Computação/Universidade de São Paulo
- Dr. Roberto Marcondes Cesar Junior
  Instituto de Matemática e Estatística/Universidade de São Paulo
- Prof. Dr. Paula Dornhofer Paro Costa
  Faculdade de Engenharia Elétrica e de Computação/Universidade Estadual de Campinas
- Prof. Dr. Marcelo da Silva Reis
  Instituto de Computação/Universidade Estadual de Campinas

A ata da defesa, assinada pelos membros da Comissão Examinadora, consta no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria do Programa da Unidade.

Campinas, 25 de novembro de 2024

Agradecimentos

I express my greatest gratitude to my parents, who provided unconditional love and support for my decision to study a PhD in Brazil, being so far way. I am also grateful to my brother, who took the responsibility of taking care of our parents while I was absent.

I would like to express my sincere gratitude to my advisor Prof. Anderson. His support is essential for this PhD. Personally, Prof. Anderson is approachable and kind, always giving me advice as a friend and mentor. Professionally, he is always there to help tracking my progress and guide me to the correct direction. In time where I was struggling to continue my research project, he supported me and guided me in changing the research direction, which led to this important topic of fact-checking.

I would like to thank the Ubiquitous Knowledge Processing (UKP) lab for hosting me during my internship in Germany. I thank Prof. Iryna Gurevych for her valuable advices and guidance, and Max Glockner for the weekly meetings and patient discussions. I would also like to thank all the colleagues and friends I met in UKP, for the collaborations, friendships (Qianqian, Kexin, Haishuo, Fengyu, Xu, Haau-Sing, Thy, Yongxin, Bob, Furkan...), and a great time in Germany.

Without the help of all the friends I met in Brazil, this PhD journey would not be possible. I am grateful to Padilha for helping me applying the FAPESP scholarship, and Antônio for picking me up when I first arrived in Brazil. I am grateful to Aurea and Ma Lin for taking me to the supermarket and shopping for the first time. I am also grateful to Victor who helped me go through the visa procedure while I had no knowledge of Portuguese. I am grateful for spending the Chinese New Year with Liu Si and Jiabin, and all other times we spent together. I am grateful to all RECOD.ai members (João, José, Juan, Darwin, Werneck, Ramon, Elian, Gabriel, Padilha and many others...) for providing such a friendly and welcoming lab environment. Although being far from home, I was so welcomed by everyone that cured my home-sickness, and helped me adapting to the PhD life without too much struggle.

I am lucky to have met Alceu, who has always been there for my happiness and sorrow, for helping me dealing with all aspects of life, in all moments. And thank you for bringing Lathifa and Yoda to me, they provided so much joy and comfort to us.

Finally, I thank IC and UNICAMP for providing an excellent environment for students and researchers. This study was financed in part by the São Paulo Research Foundation (FAPESP), process number #2019/04053-8 and #2022/05002-0. I also would like to thank the Dejavu (#2017/12646-3) and Horus project (#2023/12865-8) for supporting this PhD research.

Resumo

A Internet tornou-se uma parte integrante de nossas vidas, conectando quase todos globalmente. Essa vasta conectividade permite que informações falsas se espalhem rapidamente, potencialmente prejudicando a reputação de indivíduos, empresas e até mesmo nações. Dado o impacto significativo da desinformação, a verificação automatizada de fatos tornou-se uma área crítica de pesquisa. No entanto, as soluções de verificação automatizada de fatos têm gerado preocupações entre verificadores humanos, que geralmente não confiam nas decisões baseadas em máquinas. Além disso, muitos algoritmos apresentaram problemas comuns à inteligência artificial: eles não são explicáveis, podem aprender características espúrias e, às vezes, não conseguem generalizar. Apesar desses desafios, os verificadores humanos de fatos, sozinhos, não conseguem acompanhar o volume cada vez maior de dados gerados diariamente, tornando a assistência da máquina essencial. Portanto, esta tese visa preencher a lacuna entre verificadores humanos de fatos e algoritmos automatizados, para que possam se complementar e alcançar eficiência e explicabilidade. Em termos de eficiência, desenvolvemos métodos para a detecção de desinformação com agrupamento e sumarização de textos, reduzindo drasticamente a redundância de dados brutos de redes sociais. Também exploramos o uso eficiente de dados selecionando poucos dados anotados ou gerando dados sintéticos com poucas amostras para ajuste fino do modelo. No campo da explicabilidade, introduzimos a resposta a perguntas no processo de verificação de fatos, o que pode identificar o local do erro de uma alegação. Além disso, para lidar com a falta de dados anotados de explicação, realizamos um estudo em larga escala sobre auto-racionalização (a tarefa de gerar uma previsão de rótulo e uma explicação em texto livre juntas) em cenários fora de distribuição (OOD). Para a falta de explicações de referência, propusemos uma métrica independente de referência para a avaliação da explicação. No estudo, mostramos que os modelos podem aprender a partir de um subconjunto muito pequeno de dados e generalizar de forma comparável aos modelos ajustados em todos os dados de treinamento. Também mostramos que, para a geração de explicações, a qualidade dos dados é um fator-chave para obter melhores explicações fora de distribuição. Avançando ainda mais na explicabilidade, investigamos a melhoria da auto-racionalização para conjuntos de dados de verificação de fatos do mundo real. Ao encontrar um conjunto de dados com mais de três classes, a auto-racionalização falha em novas classes. Assim, propusemos um novo método adaptativo de rótulo em duas etapas, que superou os métodos de ponta (como o GPT-4) em dois conjuntos de dados realistas de verificação de fatos. Esperamos que o trabalho desenvolvido nesta tese tenha um impacto positivo na implementação da verificação automatizada de fatos no mundo real.

Abstract

The Internet has become an integral part of our lives, connecting almost everyone globally. This vast connectivity allows false information to spread rapidly, potentially damaging the reputation of individuals, companies, and even nations. Given the significant impact of misinformation, automated fact-checking has become a critical area of research. However, automated fact-checking solutions have led to concerns from human fact-checkers, who generally do not trust machine-based decisions. Moreover, many algorithms have shown artificial intelligence problems: they are not explainable, can learn spurious features, and cannot generalize sometimes. Despite these challenges, human fact-checkers alone cannot keep pace with the ever-increasing volume of data generated each day, making machine assistance essential. Therefore, this thesis aims to bridge the gap between human fact-checkers and automated algorithms, so they can complement each other to achieve efficiency and explainability. In terms of efficiency, we developed methods for misinformation detection with text clustering and summarization, drastically reducing redundancy from raw social media data. We also explored efficient usage of data by selecting few annotated data or generating synthetic few-shot data for model fine-tuning. On the explainability front, we introduced question answering into the fact-checking pipeline, which can pinpoint the error location of a claim. In addition, to address the lack of annotated explanation data, we performed a large scale study on self-rationalization (the task of generating a label prediction and free-text explanation together) in out-of-distribution (OOD) scenarios. For the lack of reference explanations, we proposed a reference-free metric for the explanation evaluation. In the study, we showed that models can learn from a very small subset of data, and generalize comparably to models fine-tuned on the entire training data. We also showed that, for explanation generation, the quality of data is a key factor in having better OOD explanations. Further advancing explainability, we investigated improving self-rationalization for real-world fact-checking datasets. When encountering a dataset with more than three classes, self-rationalization fails to perform on new classes. Thus, we proposed a new two-step label-adaptive method, which outperformed state-of-the-art methods (such as GPT-4) on two realistic fact-checking datasets. We hope the work developed in this thesis will positively impact the deployment of automated fact-checking in the real world.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

With the prevalence of high-speed Internet, we live in the Digital Age where it has been easier than ever to gain and share information online. According to Digital Brazil 2024, around 87% of the Brazilian population uses the Internet, and the number of social media users takes up 66.3% of the population [1]. However, this surge in connectivity has led to a decline in trust: only 43% of Brazilians trust the news they consume online, even though 90% of the news is delivered through online platforms [2].

This huge gap in trust is largely due to the rampant spread of misinformation. In traditional media, journalists were the gatekeepers of news. Today, anyone can spread information with a single click, amplifying small ideas into significant movements. While social media has empowered voices and created financial opportunities, it has also been weaponized to spread falsehoods and create chaos.

Since the 2016 election of President Trump, "Fake News" has been a buzzword [3]. However, does "Fake News" literally mean a fake piece of news? In an article written by Claire Wardle, she says the term fake news has failed to cover our reality, in which most content people read is not fake or fabricated, and most of this cannot be described as "news". Instead, she and her group in First Draft[1] prefer to use the term "Information Disorder" [4], and based on two aspects — falseness and intention to harm — it can be further divided into:

- misinformation: False information shared without harmful intent.
- disinformation: False information deliberately shared with the intent to cause harm.
- malinformation: Information that is true but used maliciously to cause harm.

In this thesis, we focus on detecting false information (dis-/mis-information) through fact-checking, regardless of the intent[2].

In many cases, misinformation can cause a significant negative impact. For example, in Brazil, propaganda spread by official authorities towards ineffective scientific treatments (such as hydroxychloroquine [5]) put at risk patients' (already debilitated) health, drowning the public health system into chaos. Recently, misinformation about the Russian Invasion of Ukraine has also been on a rampage, intending to spread fear or hatred

[1] https://firstdraftnews.org

[2] We use the term misinformation throughout this thesis, regardless of the intent.

(many examples can be found in [6]). Despite efforts to combat false information, people are drawn to sensational headlines and eye-catching stories, accelerating the spread of misinformation and expanding its reach.

To address the problem, many automated solutions have been proposed to speed up the fight against misinformation. However, most algorithms treat it as a fake news classification problem in the research community. Among these algorithms, some utilize news articles' content [7], and others exploit additional information such as users' comments [8]. However, simply solving it by classification cannot surmount real-world issues, as these algorithms are likely over-relying on news writing styles and users' biases but fail to address the actual factual accuracy of the content.

Detecting misinformation requires more than classification; it requires factual verification by reasoning with external knowledge. This leads us to a more challenging but practical task: fact-checking. Figure 2.1 shows a typical automated fact-checking pipeline. It usually consists of four steps:

- Claim check-worthiness detection: Identify which claims are important enough to fact-check.
- Evidence retrieval: Gather relevant evidence from reliable sources, like Wikipedia.
- Evidence selection: Rank and select top evidence to ensure its relevance and accuracy.
- Veracity verification: Use machine learning to verify the claim based on the selected evidence.

Some works also propose adding other pipeline steps: 1) matching a claim with verified claims before checking to avoid repetitive work; 2) generating explanations after veracity verification to make the result more reliable.

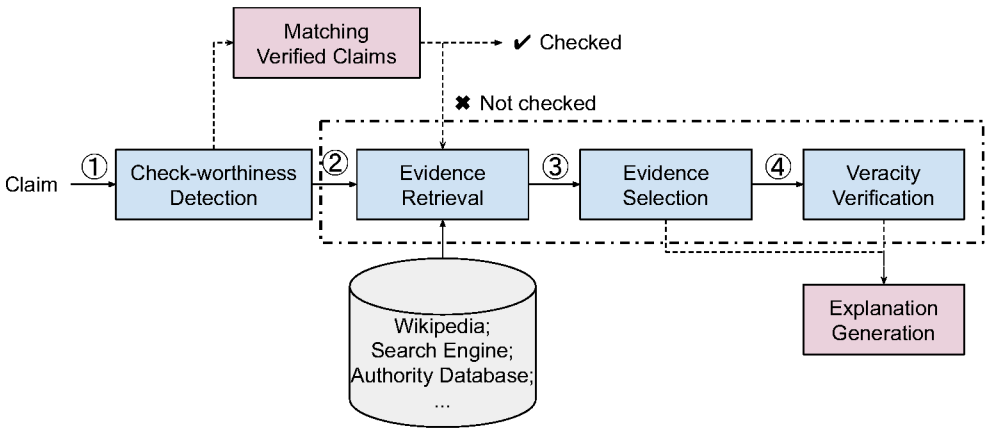


Figure 1.1: Automated Fact-checking Pipeline.

Fact-checking offers a more realistic solution for combating misinformation for several reasons: 1) this task has a more rigorous definition: check if a given piece of information is factually correct based on evidence retrieved from reliable sources; 2) fact-checking is

better integrated with journalists' jobs, which gives us more opportunities to collaborate with human fact-checkers directly; and 3) fact-checking usually deals with short claims, making it easier to identify factual inaccuracies compared to evaluating entire articles.

Many researchers have studied developing automated fact-checking methods. However, according to a recent survey [9], human fact-checkers do not trust the results from automated solutions. The reason is that automated methods are error-prone, and incorrect fact-checking could seriously damage organizations' reputations. However, fact-checkers cannot debunk every single post manually, as massive numbers of messages are posted on social media every second [10]. Guided by the literature and fact-checker collaborators, this thesis aims to bridge the gap between human fact-checkers and automated fact-checking solutions.

1.1 Research Objective

This thesis has two main goals: a) reducing human fact-checkers' effort in the fact-checking process; b) producing transparent or explainable fact-checking results to gain trust from fact-checkers.

1.2 Research Questions

To achieve the two main goals, this thesis is guided by the following research questions:

Fact-checking Efficiency

- RQ1: Given a large amount of raw text data, how can we speed up the fact-checking process to reduce fact-checkers' workload?
- RQ2: How do we use data efficiently to use less annotated data for model learning?

Fact-checking explainability

- RQ3: How can we modify the fact-checking process to make it more transparent to human fact-checkers?
- RQ4: In the absence of annotations, how can we leverage a different dataset to generate explanations for an unknown target dataset?
- RQ5: How do we evaluate generated explanations without any reference data?
- RQ6: How does our method work on real-world fact-checking datasets?

1.3 Main Contributions

Figure 1.2 summarizes the main contributions of this thesis, organized by our two main goals and four chapters.

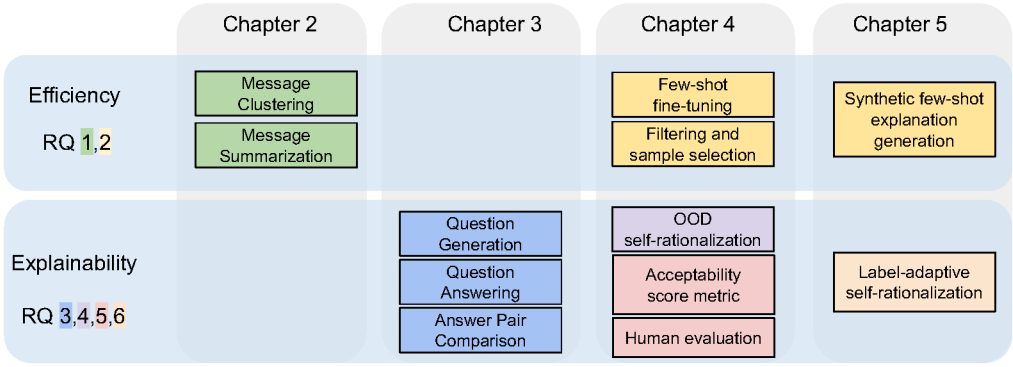


Figure 1.2: Main contributions of this thesis, organized by each chapter (columns) and the two research objectives (rows).

1.3.1 Reducing human effort in the fact-checking process

For the first goal (a), this thesis investigated two fronts: 1) developing methods to improve fact-checking efficiency and 2) learning from less annotated data.

To **improve fact-checking efficiency**, we proposed a method for reducing the amount of claims fact-checkers receive for verification (Chapter 2). A key observation that enables this task is that most posts from social media overlap extensively; many of them are slight modifications or paraphrases of other posts. This is due to social media's simple sharing design; a piece of news or a claim usually has multiple versions (copied, edited, re-shared) on social media, and they can spread rapidly from one platform to another. To approach this task, we proposed a novel semantic text clustering to group similar posts and then summarized them into their essential claims. This work reduced 28,818 original tweets to around 700 claims (more than 97%), as many tweets refer to the same claim.

For **efficient usage of data**, we explored few-shot learning with high quality data. Few-shot learning requires much less annotated data for training a machine learning model, which is important considering that human annotation is time-consuming and labor-intensive. Two main works developed in this thesis considered using very few numbers of annotated data for fact verification. In the OOD evaluation study (Chapter 4), we explored various numbers of shots for model fine-tuning. We showed that few-shot fine-tuned models with up to 128-shot have comparable performance with models fine-tuned on an entire dataset. In the label-adaptive self-rationalization approach (Chapter 5), we explored using few-shot synthetic explanations generated by Large Language Models (LLMs) for fine-tuning another smaller model, and the performance is also comparable to full-shot fine-tuning.

1.3.2 Improving automated fact-checking explainability

The second goal (b) focuses on having **explainable fact-checking results**, which is important for human fact-checkers to trust AI-based algorithms. Human fact-checked articles often contain detailed explanations for the decisions (examples can be found in

many fact-checking websites, such as Agência Lupa[3], Aos Fatos[4], among others). At the same time, most automated methods only produce a simple prediction. To help increase trust between human fact-checkers and automated solutions, this thesis provided two approaches for more explainable fact-checking.

Our first approach proposed to integrate **question answering (QA)** in the fact-checking process (Chapter 3). In detail, we proposed generating questions and answers from claims and answering the same questions from evidence. We also proposed an answer comparison model with an attention mechanism attached to each question. Despite promising results, there are some limitations of the question-answering approach: 1) Repetitive questions were generated from a claim, and some were of poor quality; more diverse questions are needed to improve them; 2) Current evidence is short, but real-world evidence is much longer; 3) To show that our generated questions and answers are accurate, we need to incorporate human-in-the-loop for QA quality evaluation.

To deal with these issues, Chapters 4 and 5 explored using **self-rationalization** methods to generate free-text explanations. Free-text explanations are useful as they are expressive and easy to understand. Despite their benefits, existing datasets often lack annotated explanation data, particularly in contexts such as fact-checking and hallucination detection, making it hard for training an explainable model.

In Chapter 4, we started by studying how to leverage existing explanation datasets to learn self-rationalization and evaluate models' out-of-distribution (OOD) performance (Chapter 4). We performed prompt-based fine-tuning with the T5-large and OLMo-7B and evaluated on 19 diverse OOD datasets across three tasks: natural language inference, fact-checking, and hallucination detection of abstractive summarization. We addressed the lack of reference explanations for evaluation by studying the effectiveness of the Acceptability score with a human evaluation, and comparing it against three LLM-based reference-free metrics.

Chapter 5 investigated how self-rationalization can be utilized on more realistic fact-checking datasets. In the OOD evaluation of self-rationalization, our fact-checking datasets (e.g., FEVER [11]) usually label claim veracity with three classes: SUPPORT, REFUTE, and NEI (not enough information), which is comparable to NLI labels (entailment, contradiction, and neutral). However, many real-world fact-checking datasets usually have different labeling schemes with the number of classes varying from 2-27 classes [12]. As the labeling scheme shifts from NLI tasks, directly applying self-rationalization with models pre-trained on NLI datasets performs poorly for fact-checking. Therefore, we proposed a label-adaptive learning approach to learn self-rationalization in two steps, allowing our model to adapt to a new domain more effectively than fine-tuning end-to-end self-rationalization directly.

[3] https://piaui.folha.uol.com.br/lupa/
[4] https://www.aosfatos.org

1.4 Thesis Organization

We organize the thesis as a compilation of articles published (or submitted for peer-review), with each chapter corresponding to one work. In total, we present four papers, three published in International Conferences and one submitted to a journal, under a second-around review. Chapter 2 introduces the clustering and summarization pipeline for reducing the amount of raw social media data for fact-checkers to process. Chapter 3 describes our approach for introducing question-answering into the fact-checking pipeline for more explainable fact-checking. Chapter 4 presents our large scale study on self-rationalization's Out-of-Distribution ability on 19 datasets. Chapter 5 describes our proposed method for adapting self-rationalization on realistic fact-checking datasets. Finally, Chapter 6 discusses the contributions by answering each research question and presents the limitations and potential future directions.

Chapter 2

Scalable fact-checking with human-in-the-loop

As misinformation becomes a growing concern to the public, news fact-checking organizations are also proliferating. However, the generation and spreading speed of the former is much faster than the latter. To fight misinformation, automated fact verification has received most attention in the literature. However, human fact-checkers often do not trust results from automated solutions [9]. The reason is that automated methods are error-prone, and incorrect fact-checking could seriously damage fact-checking organizations' reputations. Instead, what fact-checkers seek from automated methods is to scale-up manual fact-checking's speed. Indeed this is essential in fact-checking, as every day, billions of messages are posted on social media [1] and misinformation is ever increasing. Till now, most researchers have tried to handle this issue by checking if a post is worth-checking [13] to reduce the number of claims. However, this may not be enough as social media messages are noisy, and check-worthiness detection needs manual labeling which is bias prone.

To achieve our first goal of speeding up the fact-checking process, we developed an unsupervised method to reduce the number of claims fact-checkers receive. We notice that posts from social media overlap extensively; most of them are slight modifications or paraphrases of other posts. To exploit this observation, this work proposes to assist human fact-checkers by grouping semantically similar claims together and summarize them into single key claims. The grouping stage consists of separating posts into distinct claims. The summarization stage aims at reducing redundancy and formulating an informative and representative claim. This is the first work that addresses grouping and summarizing semantically similar messages together to scale up fact-checking, to the best of our knowledge. The contributions of this work are summarized as follows:

1. We propose a novel pipeline to filter redundancy in short messages and generate informative claims.

2. We propose a graph-based approach for the pipeline, combining community detection and graph-based extractive summarization that utilizes tweets metadata.

[1] https://tinyurl.com/nt4pxa6a

Experimental results show that the graph-based methods obtain the best performance.

3. We generate a graph of summaries to verify the clustering and summarization methods; the graph shows that the summaries are well-separated.

4. We brings humans back to the loop by assessing claims worthiness with a fact-checker specialist.

This work was published and presented at the International Workshop on Information Forensics and Security (WIFS 2021): "Jing Yang, Didier Vega-Oliveros, Taís Seibt and Anderson Rocha. Scalable Fact-checking with Human-in-the-Loop. IEEE International Workshop on Information Forensics and Security (WIFS), 2021."

2.1 Related Work

In this section, we review related work in increasing efficiency: claim check-worthiness detection and social media short message summarization.

2.1.1 Check-worthiness Detection

Check-worthiness detection is related to our work and serves a similar purpose — to reduce the number of claims to be checked. Currently, most check-worthiness detection work focus on checking claims related to political debates. The well-known ClaimBuster [14] extracts, ranks, and identifies essential factual claims from presidential debates sentences. CheckThat! Lab. [13] has hosted since 2018 an open detection task of check-worthy claims. The goal is to give check-worthiness scores to a list of sentences. The Prise de Fer's team [15] proposed a hybrid model with various sentence representations, including both syntactic and semantic features. The Copenhagen team [16] extracted sentence features by a Recurrent Neural Network (RNN) with Gated-Recurrent Units (GRU) memory units. They used contrastive sampling to select sentence pairs further trained for check-worthiness prediction. Our goal differs from check-worthiness detection; rather than predicting if a claim is check-worthy we provide human experts meaningful summarized claims to facilitate the arduous fact-checking task.

2.1.2 Social media message summarization

Social media post summarization brings another branch of related works. These studies are usually related to event/disaster discovery. For example, Rudra et al. [17] reported a framework to summarize messages from Twitter. Their method comprises two stages. The first selects essential tweets from the whole set. The second combines selected tweets and generates a new message by maximizing tweets' informativeness and avoiding redundancy. Another example is the systematic review of summarization on tweets for emergency

events reported by Dutta et al. [18]. The authors analyzed eight summarization methods and showed that different methods generated very different summaries.

Thus, although some tweets summarization methods have been proposed, there is still much room for improvement. To encourage research on this field, recently, Dusart et al. [19] proposed a large dataset for tweet summarization of events named ISSumSet, which contains 122 events. Each event has various related tweets labeled with different types and levels of importance. Tweets summarization is a challenging and vital task. Many messages are posted online daily, not to mention that they are noisy and multilingual. In fact-checking, we face similar challenges, and the summarization of posts to generate relevant claims may turn the task much more scalable.

2.2 Task Formulation

We define our task as follows: *given a set of social media posts, separate them into different groups based on their semantics; then, summarize all posts for each group to generate an informative claim that can represent the group to aid fact checkers later on.* Figure 2.1 shows the general pipeline for the task, consisting of two main steps: semantic clustering and content summarization/claim generation.

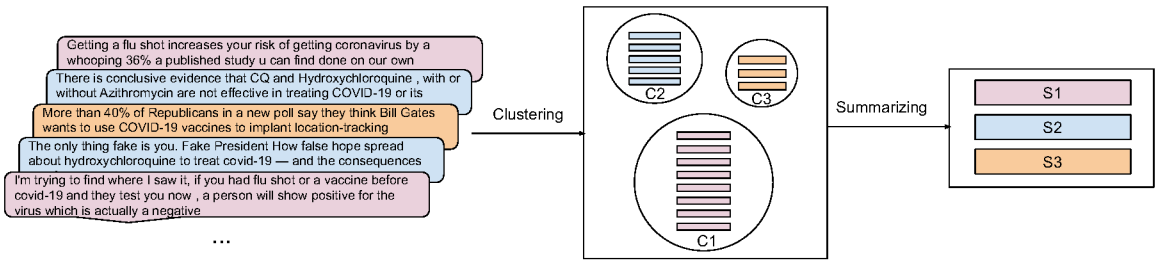


Figure 2.1: Unsupervised social media posts summarization pipeline. Given a set of posts, we perform two steps: semantic clustering and summarizing/claim generation. The first step groups posts into clusters and ranks them based on the number of posts in each cluster. The second step summarizes messages from each cluster to generate an informative summary.

To execute this pipeline, we propose an entirely graph-based method with the combination of community detection and extractive summarization. Additionally, we combine distance-based clustering and abstractive summarization to compare with the graph-based approach. Next, we detail the clustering and summarization methods.

2.2.1 Short Message Aggregation

Aggregation seeks to group short messages related to a single claim together. This group should include duplicated, near-duplicated, and paraphrased posts with same or opposite sentiments towards a claim. This is challenging because clustering is unsupervised; it is difficult to define the boundary for a short message to belong to one group or another.

State-of-the-art transformer-based language models are good at capturing semantic meanings of words. In our case, we need to capture semantic meaning of sentences,

therefore we leverage Sentence-Transformers[2] [20] for short messages embeddings as input for clustering.

There are different ways to perform aggregation or clustering. A standard clustering method is k-Means. However, it is not suitable in our case as each short message embedding has a dimension of at least 512, which makes k-Means relatively slow and, most importantly, it requires us to pre-define a specific number of clusters. Therefore, we adopt and compare two methods: Agglomerative clustering[3] and Leiden community detection[21] as they do not require the establishment of the number of clusters beforehand.

Agglomerative clustering Hierarchical clustering groups feature points based on their dissimilarity. The method starts with each point as a cluster and merging two clusters into one if their dissimilarity value is below a decision cutoff. This method is helpful because the number of clusters is unknown; we can control the decision cutoff to have smaller or larger clusters. First, we calculate a similarity matrix S of short message embeddings for the initial dissimilarity values, then provide $1 - S$ as the dissimilarity matrix. For the linkage criteria determining how the dissimilarity is calculated between two clusters, we choose the average dissimilarity between any two points in the two clusters.

Leiden community detection Leiden community detection is a graph-based clustering method that finds the best community partition in a graph [21]. It improves the convergence time of the Louvain algorithm with a smaller computational footprint, providing partitions focused on the micro-patterns of the communities that maximize the graph modularity. Formally, the graph $G(N, L)$ is formed by the set of nodes N representing each short message; and the set of links L, which represent the similarity weight between nodes. The construction process from the short message embedding to the graph calculates the similarity matrix among the vectors and then applies the ϵ-neighborhood method [22].

For both methods, we use the cosine similarity to compute the similarity matrix of all post representations. As both methods require a decision cutoff for similarities, we denote the threshold δ and $\epsilon = \delta$ as the ϵ-graph construction parameter.

2.2.2 Short message Summarization

For each cluster, our goal is to summarize its short messages to generate a claim. We leverage two types of summarization.

Extractive summarization For extractive summarization, we aim to select a representative short message from each cluster. In particular, we construct an ϵ-graph for each cluster and use centrality measures to rank the short messages in the cluster and select the most central one as the summary. The idea of using centrality measures is that central nodes are usually the more influential or representative in the graph [23]. We adopted two

[2] https://tinyurl.com/w4de7tvv
[3] https://tinyurl.com/vr3p25w3

methods: the Degree Centrality (DG) and the Multi-Centrality Index (MCI) [23]. The DG counts the total number of input/output connections of the nodes, and nodes with the highest DG centrality are known as hubs. The MCI considers multiple measures for finding the most relevant message in the cluster. In this work, we consider the Degree, PageRank, and Betweenness centrality, along with the number of reposts and likes of each message, for calculating the MCI.

Abstractive summarization For the abstractive summarization, we use two state-of-the-art transformer-based language models: BART[24] and T5[25] to generate a summary. The challenge of these two models is that the models' maximum input length is not long enough to fit all short messages in some clusters. Before feeding all the messages to the summary process, we remove duplicates and near-duplicates from each cluster to deal with this problem. We perform the agglomerative clustering with a higher similarity threshold within each cluster. Afterward, we have more sub-clusters in each cluster, and each sub-cluster contains only messages that are duplicates and near-duplicates. We randomly select one message from each sub-cluster as messages in the same sub-cluster can be treated as equivalent.

2.3 Evaluation and Analysis

In this section, we first describe our data collection and processing; then we present the results for clustering and summarization, respectively. After that, we show the human evaluation of the representativeness of the generated claims (summaries) by a journalist. Finally, we present examples of generated summaries and a visualization graph showing the similarity of the summaries.

2.3.1 Dataset

For the evaluation of our proposed pipeline, we adopt MM-COVID[26], a fake news detection dataset. Each news article in this dataset is accompanied by social media context: tweets, retweets, and replies. Here, we only use tweets content, as retweets are duplicates of tweets, and replies can be less related to the claim itself. We choose this dataset because we can use its labels for evaluation, as each news piece has a claim summarizing the news content. This news summary can be treated as the ground-truth for our short messages summary. We emphasize that although the dataset was proposed for supervised learning on text classification, we only use the labels for evaluation, i.e., our methods perform unsupervised learning all the time[4].

Through Twitter API[5], we collected 92,070 tweets associated with 2,227 news articles (around 12% tweets were removed from Twitter at the time of collection). Out of all tweets, 48,074 tweets (52.2%) associated with 1,092 news articles are in English. In this work, we consider only English tweets, but our pipeline can be easily adapted to other languages as long as trained language models are available. After collecting all the

[4]Our code is available at: https://github.com/jingyng/scalable-fact-checking

[5]https://developer.twitter.com/en/docs/twitter-api

tweets, we pre-processed them by removing duplicated tweets (those with the same id), user mentions, URLs, hashtags, and emojis.

One challenge of using this dataset for evaluation is that there are some mismatches between news claims and tweets content, i.e., a tweet associated with one news piece is not related to its news content. We show one example here:

> **News claim**: *Coronavirus is caused by 5G.*;
> **Tweet content**: *Recently, we have also had some misinterpret some CDC data related to deaths from COVID-19. Without a doubt, we know coronavirus has caused more than 400 deaths in Utah and over 177,000 in the United States.*

The example shows a tweet content not related to the news claim. This hinders us from using news claims as gold summaries for a tweet cluster. Therefore, we remove tweets less relevant to a news claim based on a relevance decision cutoff θ. This step is only necessary to evaluate summarization and does not need to be performed in real cases.

For calculating the relevance between tweets and their news summary, we rely upon BERTscore[27], as it has shown better performance than cosine similarity in [28]. We use the default model (*roberta-large*) and normalize the score[6]. After calculating the relevance between tweets and news summaries, we remove all tweets irrelevant to its news summary, with the threshold $\theta = 0.1$. We also filter out messages with less than 4 words, given that they do not contain meaningful information. After this process, we have 28,818 remaining tweets associated with 959 news original articles.

2.3.2 Clustering Evaluation

As we do not have ground-truth cluster labels, we rely upon the Silhouette coefficient metric[7] to evaluate the clustering results. This metric ranges from -1 to 1, with a higher value indicating a better-defined cluster with less overlap among clusters. To compare the clustering results, we consider two factors: embedding models and clustering methods.

Comparison of embedding models

A good embedding model is essential in clustering; it should map semantically similar messages closer in their feature representation space. To compare different embedding models, we set the decision cuttoff $\delta = 0.85$ and the clustering method to be Leiden community detection. Table 2.1 shows the results.

The clustering performance varies but all embedding models (except for *cardiffnlp/twitter-roberta-base*) lead to reasonable performances. Surprisingly, embedding model *cardiffnlp/twitter-roberta-base* only resulted in one big cluster, although it was pre-trained with tweets. All other models comprise about 700 clusters, less than the number of news, 959 indicating that some news claims are similar to each other.

[6] https://github.com/Tiiiger/bert_score
[7] https://tinyurl.com/3xfvsbck

Embedding model	# of clusters	Silh. Coef.
paraphrase-distilroberta-base-v2	701	0.76
paraphrase-mpnet-base-v2	677	0.73
paraphrase-MiniLM-L6-v2	707	0.75
nli-mpnet-base-v2	739	0.79
nli-roberta-base-v2	705	0.79
digitalepidemiologylab/covid-twitter-bert-v2	732	0.76
cardiffnlp/twitter-roberta-base	1	–

Table 2.1: Clustering results comparison between different embedding methods.

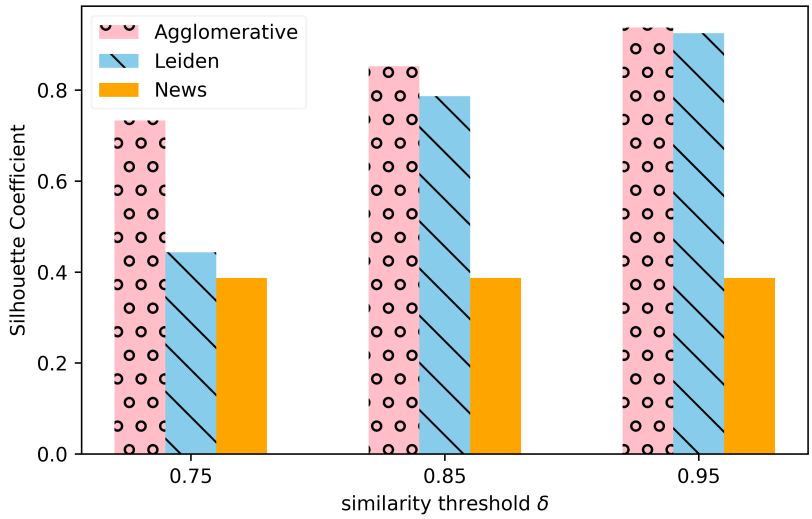


Figure 2.2: Clustering results varying the similarity threshold δ.

Comparison of clustering methods

As previously mentioned, we compare two clustering methods: Agglomerative clustering and Leiden community detection. As a comparison, we also consider the original posts clustering separated by news (i.e., each news corresponds to one cluster of posts). In Figure 2.2, we vary similarity threshold δ to compare clustering performance. We fixed the embedding model *nli-roberta-base-v2* as it performed best (see Table 2.1). The Silhouette coefficient increases when δ increases. When δ is close to 1, Agglomerative clustering and Leiden clustering methods yield about the same results because when δ is high they are only grouping near-duplicated tweets together. This indicates that the Silhouette coefficient can only partially evaluate clustering results, as we want to cluster posts that are semantically similar to each other, not just posts that contain similar words.

Distribution of news in clusters

To check if the method is indeed clustering posts related to one news claim, we analyze news in each cluster. The percentage of clusters with only one associated news claim for agglomerative and Leiden are 95.15% and 93.34%, respectively. We randomly examine

one cluster with more than one associated news piece (agglomerative method) to see if the news claims are similar. One example of news claims in a cluster is the following:

- *U.S. President Donald Trump or presidential candidate Joe Biden referred to the novel coronavirus pandemic as a time when "people are dying that have never died before."*
- *Donald Trump said about coronavirus, "People are dying who have never died before."*
- *Referring to the ongoing COVID-19 pandemic, U.S. President Donald Trump said, "People are dying today that have never died before."*

The three news claims are indeed related to one claim. This also explains why the number of clusters (agglomerative: 804, Leiden: 705) is less than the number of news articles (959).

2.3.3 Summarization Evaluation

After aggregating the posts, we perform the summarization. For the quantitative evaluation of summarization results, we use the F1-score of ROUGE-1, ROUGE-2, and ROUGE-L metrics[8]. ROUGE scores are common metrics for text summarization tasks. Given generated and reference summary pairs, ROUGE-1 and ROUGE-2 measure the overlap of unigram and bigram, respectively, and ROUGE-L measures the Longest Common Subsequence (LCS) between them. We use BERTscore to measure the semantic similarity between the generated and ground-truth summary pairs. For the informativeness of summaries, we use the average summary length.

Comparison of summarization methods

We combine two clustering methods (agglomerative clustering and Leiden community detection) and four summarization methods (BART, T5, DG, and MCI). We set the similarity threshold $\delta = 0.85$ for both clustering methods (Table 2.2). We consider news summaries to be ground-truth summaries as the tweets mention these news articles.

Table 2.2 shows that extractive summarization (DG and MCI) ouperforms abstractive summarization (BART and T5). This means most content are repetitions of news content, so the extractive summaries can be precisely the same as news summaries. However, models for the abstractive summaries try to generate fluent sentences by combining multiple different posts thus are longer and overlap less with news summaries. In terms of the abstractive approach, the average summary lengths for Leiden clusters are longer than the agglomerative ones. This is because in Leiden each cluster contain more posts, thus has in total fewer (around 100 less clusters) but bigger clusters. Therefore, we consider the Leiden method more suitable as it reduces redundancy without losing information, even though the scores of Agglomerative are slightly higher. Therefore, we conclude that our graph-based method performs the best among all combination of methods.

[8] https://tinyurl.com/3e5t8tpz

Summarization Method	ROUGE-1	ROUGE-2	ROUGE-L	BERT-score	Average Summary Length
Agglomerative+BART	0.53	0.41	0.49	0.91	22.03
Agglomerative+T5	0.51	0.39	0.47	0.90	23.99
Agglomerative+DG	**0.59**	**0.48**	**0.56**	**0.92**	**21.44**
Agglomerative+MCI	0.59	0.48	0.56	0.92	21.49
Leiden+BART	0.50	0.38	0.47	0.91	23.44
Leiden+T5	0.48	0.36	0.44	0.89	26.17
Leiden+DG	**0.59**	**0.48**	**0.55**	**0.92**	**21.48**
Leiden+MCI	0.58	0.47	0.55	0.92	21.53

Note: Average news summary length for Agglomerative and Leiden are 15.20 and 15.99 respectively. Number of clusters for Agglomerative and Leiden are 804 and 705, respectively.

Table 2.2: Summarization performance comparing different summarization methods

Table 2.3 shows an example of summaries for qualitatively illustration of the summarization process for Leiden clustering. All summaries essentially reduced the redundancy of posts, and capture the central claim of the tweets.

Analysis of the graph of summaries

To validate the robustness of our clustering and summarization methods, we construct a graph of summaries to see if there are similar summaries. We take the summaries generated from **Leiden+BART** and perform a Leiden community detection with similarity threshold of 0.75, then visualize the communities in the graph. We show the graph in Figure 2.3. In this graph, there are 609 communities. We can see that it is a sparse graph; most communities contain only a single node, indicating the clustering and summarization effectiveness.

To further check if the summaries in the same community are similar, we show two examples of all the summaries in a community for communities 1 and 2 (communities are sorted in descending order with the number of summaries) in Table 2.4. We can see that the summaries in one community are similar to each other and related to similar topics, but they are not related to one specific claim. Therefore we conclude that our clustering and summarization find reasonable and useful claims and a further reduction would risk losing information.

2.3.4 Human-in-the-Loop

We invite fact-checker journalists to evaluate the summarization methods, evaluating the four proposed methods. We also include the news summary (ground-truth summary) in the comparison. We asked the specialist to give a score ranging from 1-5 (one means the summary is not representative for the posts in the cluster, and five means the summary is very representative for the posts). We randomly select 50 clusters for each summarization

Original tweets from a cluster (11 out of 241 tweets after removing duplicates and near-duplicates)

1. Reupping this fact check –> How false hope spread about hydroxychloroquine to treat covid-19 — and the consequences that followed.
2. How Trump's false hope spread about hydroxychloroquine to treat covid-19 — and the consequences that followed - The Washington Post.
3. Coronavirus: How false hope spread about hydroxychloroquine to treat covid-19 - and the consequences that followed.
4. How false hope spread about hydroxychloroquine to treat covid-19 — and the consequences that followed. Four Pinocchios given by the WP.
5. Well done for "How false hope spread about hydroxychloroquine to treat covid-19 – and the consequences that followed" - highlighted in Journalism Matters survey on Excellence in Reporting Coronavirus.
6. How false hope spread about hydroxychloroquine to treat covid-19 — and the consequences that followed. Dr. Trump's medicine show: Why is he pushing an unproven drug? Follow the money.
7. For all you MAGA supporters who keep pushing the lie: "How false hope spread about hydroxychloroquine to treat covid-19 — and the consequences that followed."
8. Trump is making baseless, irresponsible medical recommendations based on rumor and social media idiocy. Analysis | How false hope spread about hydroxychloroquine to treat covid-19 — and the consequences that followed.
9. "How false hope spread about hydroxychloroquine to treat covid-19 — and the consequences that followed" - an excellent and important Fact Checker that explains how social media gave this dangerous info undeserved oxygen.
10. Trump and his enablers pushing dangerous and unproven medical advice as if they are doctors. Should be a law against this. How false hope spread about hydroxychloroquine to treat covid-19 — and the consequences that followed.
11. The only thing fake is you. Fake President How false hope spread about hydroxychloroquine to treat covid-19 — and the consequences that followed.

BART summarization
How false hope spread about hydroxychloroquine to treat covid-19 – and the consequences that followed. Four Pinocchios given by the WP.

T5 summarization
"how false hope spread about hydroxychloroquine to treat covid-19 – and the consequences that followed" "why is he pushing an unproven drug? follow the money" "trump is making baseless, irresponsible medical recommendations based on rumor"

DG summarization
How false hope spread about hydroxychloroquine to treat covid-19 — and the consequences that followed - msnNOW

MCI summarization
Is mystery How false hope spread about hydroxychloroquine to treat covid-19 — and the consequences that followed

News Summary (Gold)
President Trump has repeatedly touted the anti-malarial medications hydroxychloroquine and chloroquine as the much-needed solution to COVID-19

Table 2.3: An example of four summarization results on tweets in one cluster

All summaries in Community 1

1. What's a coronavirus superspreader?
2. How does the coronavirus work?
3. People with coronavirus may be most infectious in the first week of symptoms. SARS CoV2 COVID19.
4. COVID-19: Information on symptoms, transmission – Mayo Clinic News Network. Covid19 Corona Virus CoronavirusUSA: Terms to know.
5. GI Symptoms and Coronavirus (COVID-19) from. GI Symptoms of CO VID-19 from.
6. Coronavirus: How does the Covid-19 alert level system work?
7. What are the early symptoms of coronavirus (COVID-19)?
8. People with coronavirus may be most infectious in the first week of symptoms. That could lend more weight to the argument in favor of wearing a mask while in public.
9. Here are answers to key questions about the virus, including how to protect yourself and what to expect. What questions do you have about the new coronavirus?
10. Get an answer about the coronavirus, how does it kill, truth about masks, do they work, are pets safe, do HVAC systems spread the coronavirus, do quarantines work, what about cures, vaccines, treatment, how long will this coronavirus last, and more.

All summaries in Community 2

1. COVID-19 can be spread by people who do not have symptoms and do not know that they are infected. CDC recommends that you wear masks in public settings around people who don't live in your household and when you can't stay 6 feet away from others.
2. New Evidence Shows Wearing Face Mask Can Help Coronavirus Enter the Brain and Pose More Health Risk, Warn Expert. He stresses that only ill people should wear face masks.
3. The CDC recommends wearing a cloth face mask in public to help slow the spread of coronavirus. But the evidence for the efficacy of surgical or homemade masks is limited, and masks aren't the most important protection.
4. Dr. Russell Blaylock warns that not only do face masks fail to protect the healthy from getting sick, they also create serious health risks to the wearer.
5. The CDC does not recommend that asymptomatic, healthy people wear a facemask to protect themselves from respiratory diseases. Facemasks should be used by people who show symptoms of COVID-19 to help prevent the spread of the disease to others.
6. The CDC does not recommend that people who are healthy wear facemasks. It does recommend that those who are not healthy wear them.

Table 2.4: Examples of all BART summaries in a community of graph of summaries (community 1 and community 2)

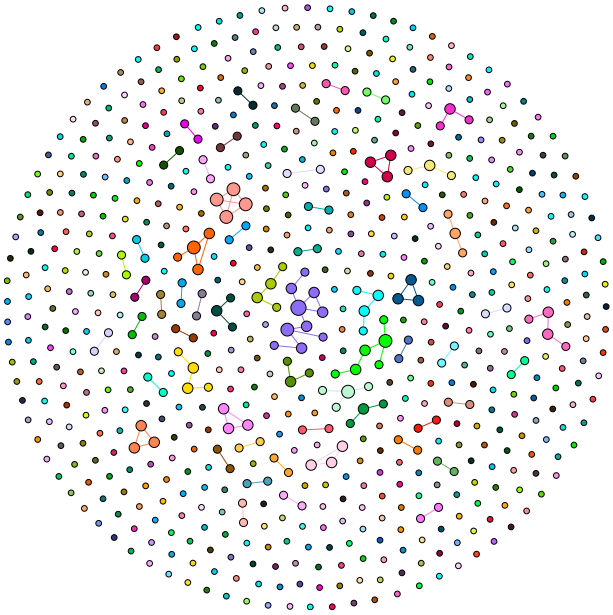


Figure 2.3: The communities from the graph of summaries generated by BART. Each connected component is a community.

method; the clustering method is Leiden. We average the scores along 50 clusters, and the average scores for DG, MCI, T5, BART, and news summaries are: 4.96, 4.96, 4.92, 4.90, and 4.68 respectively.

We can see that overall all summarization methods have an average score higher than 4, which means they are highly representative. Extractive methods' scores are slightly higher than abstractive ones as the latter sometimes bring additional comments, which are often wrong or are prejudiced. Surprisingly, the news summary, which we treat as ground-truth, has the lowest average score according to human evaluation. Specifically, some summaries received low scores because they do not offer sufficient information to obtain the claim in question or related to a similar but different claim.

2.4 Final Remarks

While automated fact-checking solutions are not near ready for deployment in real-world scenarios, it is key important to assist human checkers to improve speed and comprehensive inspection. Our approach fills the gap between manual and automated fact-checking through a two-step pipeline: grouping similar messages together and summarizing them into one claim, which a human will then check. We test our pipeline by combining two clustering and four summarization methods. The results show that the framework can largely reduce the number of original social media posts in more than 97% — from 28,818 tweets to 700 summary claims — and deliver more informative claims that enrich the knowledge about the clustered messages for the fact-checking process.

Chapter 3

Explainable Fact-checking Through Question Answering

In previous chapter, our work addressed fact-checking efficiency by reducing the amount of claims fact-checkers deal with. This step is before the actual fact-checking process. This chapter focuses on the second goal on building trust between machines and humans using such technology.

Explainable fact verification is key for modern automated fact-checking. Recent fact-checking datasets usually contain annotated explanations [29, 30] to address its importance. However, research on explainable fact-checking methods mainly focuses on text summarization [31, 29, 32] and, in such cases, explanations as summaries are not representative of real-world fact-checking explanations as they are not comparing the differences between claim and evidence to make conclusions.

Inspired by the QA works in checking factual consistency of documents and their summaries, we believe it is suitable for the fact-checking task, where we assess if claims are factually consistent with retrieved evidence. Therefore, we propose to leverage automated QA protocols and integrate them into the traditional fact-checking pipeline. As a result, we can provide explainable fact-checking results through question answering. The answer comparison model will predict a label and pinpoint the wrong part of a claim by showing which questions are more important for the decision. In this way, human fact-checkers can easily interpret the results and correct them if necessary. Our work differs from prior works [33, 34] because we not only generate question-answer pairs but also fully integrate QA protocols in the fact-checking pipeline to automatically compare answers and predict their labels. We compare the proposed method with several baselines, achieving state-of-the-art results but with the critical feature of adding explainability to the fact-checking process. We summarize our contributions as follows.

- We propose a novel pipeline for using question answering as a proxy for explainable fact-checking;
- We introduce an answer comparison model with an attention mechanism on questions to learn their importance on the claims.

from "Jing Yang, Didier Vega-Oliveros, Taís Seibt and Anderson Rocha. Explainable Fact-checking through Question Answering. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022."

3.1 Related Work

One popular line of work to increase trust for fact-checking is to generate textual explanations for the predicted results. Atanasova et al. [31] first proposed a pioneer work to generate explanations. They performed an optimization to learn together veracity prediction and explanation extraction from evidence. Subsequently, Kotonya et al. [29] proposed a joint extractive and abstractive text summarization method for explanation generation. The authors also published a survey specifically about generating fact-checking explanations [35].

Although generating explanations can provide more precise evidence to understand fact-checking decisions, existing systems lack a way to evaluate the explanations properly. Especially for explanations based on abstractive document summarization, researchers have shown that such models have problems of hallucination [36, 37], generating summaries factually inconsistent with their original document. To deal with this issue, several works have been proposed [38, 39, 40]. In particular, Pagnoni et al. [41] summarized different types of errors some models make and metrics used to evaluate them. Among these evaluation metrics, leveraging question answering (QA) as a proxy has been the focus of some work [39, 40]. The idea is to rely upon a question answering mechanism as an evaluation for the faithfulness of summaries. Wang et al. [40] extracted answers and questions from summaries and fine-tuned a QA model to generate answers from the documents; the answers from the document and its summary for the same questions are compared to determine the actual consistency of the summary. Recently, Nan et al. [42] proposed an improved method than [40], where instead of generating answers for both the summary and document, they model the likelihood of the summary and document conditioned on question-answer pairs generated from the summaries. Through this, the likelihood metric becomes suitable as a training objective to improve the factual consistency of summaries.

A few works have been proposed to leverage QA to help in fact-checking. For example, in PathQG [33], Wang et al. generated questions from facts. They accomplished this task in two steps: first, they identified facts from an input text to build a knowledge graph (KG) and then generated an ordered sequence as a query path; second, they utilized a seq2seq model to learn to generate questions based on the query path. The human evaluation showed that their model could generate informative questions. In another work, Fan et al. [34] generated question-answer pairs as a type of brief, along with passage brief and entity brief, and provided them to the human fact-checkers, aiming at improving their checking efficiency.

3.2 Proposed Methodology

We introduce question answering (QA) in the fact-checking process. Despite previous mentions of using QA for fact-checking, no previous work has explored integrating QA protocols in its pipeline. Our proposed solution is described as follows:

1) Given a claim C, generate multiple questions $Q_1,\cdots, Q_n$ and answers $A_1^C, \cdots, A_n^C$ from it;

2) Retrieve and re-rank evidence E based on the claim (and possibly questions);

3) For each question generated from 1), ask retrieved evidence for answers $A_1^E, \cdots, A_n^E$ respectively;

4) Compare the answer pairs (A_i^C, A_i^E) and transform the result into a label of SUPPORTS or REFUTES.

Our proposed pipeline leads to more explainability as we break down the fact-checking process into more steps, allowing a more fine-grained analysis of each part of the process (e.g., question generation, question answering, or answer comparison). In addition, through answer generation from claims and evidence, we vastly reduce the information (from claims and their evidence to only answer pairs) fed to the final classification model. Thus, the model learns from more direct and precise inputs.

To focus on how question answering empowers explainability, we use gold evidence instead of retrieved evidence. It means that for step 2), we take the gold evidence directly instead of retrieving them to focus on evaluating the other three stages of the problem. Future work will be dedicated to the retrieval by itself. Therefore, we focus on steps 1), 3), and 4) of the pipeline. Next, we detail the proposed methodology steps.

3.2.1 Question and Answer Generation

To generate questions from a text, answers for the text are usually provided first to generate more relevant questions [39, 40]. Answers are usually extracted based on named entities and noun phrases; then, questions are generated given the claim and answers. They can also be generated in parallel with questions [42]. We adopt the approach to generate questions and answers from claims simultaneously [42]. In particular, we follow the instruction of [42] to fine-tune the BART-large model to generate question-answer pairs (Q_1, A_1^C), $\cdots$, (Q_n, A_n^C) from a given claim C. Using beam search, 64 question-answer pairs are generated, then pairs are removed if the claim does not contain the answers. For answers of evidence E, we utilize a pre-trained extractive QA model to answer the questions generated previously from the claim. The model generates multiple answers, and we choose the one with the highest score (the most likely answer).

3.2.2 Answer Pair Comparison

For answer comparison, the token-level F1 score is usually used to measure similarity between answer pairs; however, it does not work when the two answers have non-overlapping

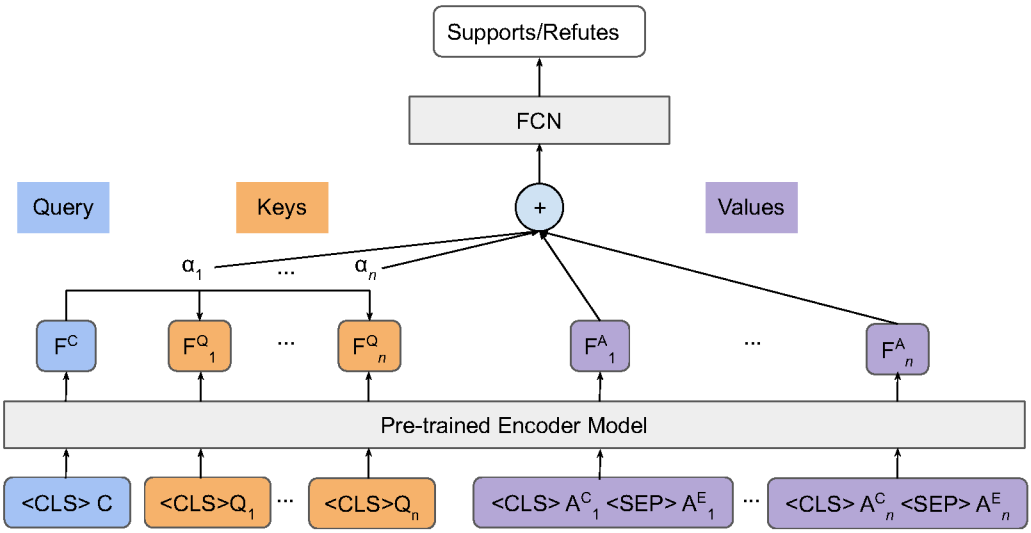


Figure 3.1: Answer comparison model with attention on questions. C represents a given claim, Q_i represents i_{th} questions, and $(A^C{}_i, A^E{}_i)$ represents i_{th} answer pairs for claim and evidence. n denotes the number of questions and answer pairs.

words but are semantically similar. We propose to fine-tune a transformer model to learn answer comparison. Considering that different questions have various purposes, they also vary in their importance. To account for this, we add attention to each question to learn the importance weights. The structure of the model is shown in Fig. 3.1.

Specifically, we rely on a pre-trained masked language model to encode the claim C, questions $Q_1, \cdots, Q_n$, and answer pairs $(A^C{}_1, A^E{}_1), \cdots, (A^C{}_n, A^E{}_n)$. For the answer pairs, we add a $\langle$SEP$\rangle$ token between two answers of the same question. We use one encoder model for encoding all the inputs, which means the weights are shared. After the encoding, we take the representation of the $\langle$CLS$\rangle$ token as each sentence embedding, thus transforming the claim, questions and answer pairs into features: F_C, F_1^Q, F_2^Q, $\cdots$, F_n^Q, and F_1^A, F_2^A, $\cdots$, F_n^A respectively, where n is the number of questions for each claim. Then we utilize additive attention proposed by Bahdanau et al. [43] to learn the importance of each question. We treat the claim as a query, questions as keys, and answers as values for each representation. The details are formulated as follows.

$$f_{\text{att}}\left(F^C, F^Q{}_j\right) = \mathbf{W}_3 \tanh\left(\mathbf{W}_1 F^C + \mathbf{W}_2 F^Q{}_j\right) \tag{3.1}$$

$$a_j = \text{softmax}(f_{\text{att}}\left(F^C, F^Q{}_j\right)) \tag{3.2}$$

$$\mathbf{F} = \sum_j a_j F_j^A \tag{3.3}$$

where f_{att} calculate the attention weight between F^C and F_j^Q $(j = 1, 2, \cdots, n)$, $\mathbf{W}_1$, $\mathbf{W}_2$ and $\mathbf{W}_3$ are learnable parameters.

In Eq. (3.1) and (3.2), attention weights are calculated and normalized by the softmax function. Then Eq. (3.3) uses a weighted sum to combine all answer features to the final feature F. Feature F is then fed to a fully connected layer to have the final prediction of SUPPORTS or REFUTES. Notice that all the information present in both claim and evidence are reduced into their respective answers. The claim and generated questions

take part in selecting the most relevant answers, assigning higher weights to them.

3.3 Experimental Setup

This section presents the dataset used in our experiment, the implementation details and the baselines for comparison.

3.3.1 Dataset

We adopt the Fool-Me-Twice (FM2) dataset [44], which comprises 12,968 claims and their associated evidence. FM2 is a recently published dataset collected through a multi-player game. In the game, one player generates a claim and tries to fool other players. The others have to decide if the claim is true or false based on evidence retrieved by the game before a timer runs out. The game's setting makes this dataset challenging as the players are motivated to generate claims hard to verify. FM2 is a more difficult and less biased dataset than the seminal dataset FEVER [11], in which a model can exploit specific words from the claim [45] to achieve reasonable accuracy (79.1% for two classes). In contrast, FM2 is shown not to have biases, a classification based only on claims resulted in low prediction accuracy (61.9%).

3.3.2 Implementation details

For question-answer pairs generation, we follow the code provided in [34] to fine-tune a BART-large model based on XSUM and CNNDM datasets[1]. For answer generation for evidence, we use the FARM framework from deepset[2] to generate answers from evidence, and the question-answering model is *deepset/electra-base-squad2*. For answer comparison, we use *microsoft/mpnet-base* model for encoding all input representation, as it has shown to perform well in question answering tasks [46]. As the question generation model does not output the same number of questions for every claim, we selected the first ten questions if the claim has more than 10; if the number of questions is less than 10, we repeat the first question until 10. We choose this quantity of questions because the average number of questions for each claim is 11.5. The hyperparameters for training the answer comparison models are: *number of epochs* = 5, *batch size* =32, *learning rate* = 2e-5, which is the standard for fine-tuning a masked language model, and *maximum token length* = 32. For statistical significance, we run each experiment 5 times and report the average and standard deviation. As the dataset is well-balanced, we use macro average accuracy as the evaluation metric.

3.3.3 Baselines

We set questions and answers for the baselines to be the same, only varying different answer comparison methods.

[1] https://bit.ly/3iBZyqR
[2] https://github.com/deepset-ai/FARM

- **Blackbox method:** we compare our results with the original proposed method in [44]. We refer to it as the *blackbox method* as they concatenate claim and evidence for the prediction without providing interpretability. We used the code provided by the authors[3] and ran it five times to have an average result.
- **QUALS score:** it is an automatic metric for checking factual consistency [42]. It does not generate answers for evidence. Instead, it calculates the likelihood of the evidence given the question-answer pair from the claim, compromising explainability.
- **Token level F1-score:** a standard metric for question-answer tasks. It counts words overlap between two answers.
- **BERTscore:** a common metric for measuring the similarity of two sentences. We use the default model *roberta-large*.
- **Cosine similarity:** a metric also used for measuring sentence similarity. We use sentence transformer *all-mpnet-base-v2* to embed the answers and calculate the cosine similarities between the embeddings.

Only the blackbox method requires training. The others are metrics to evaluate the answer pairs. These metrics calculate a score representing similarity for each answer pair, except for QUALS that outputs a score for all answers of the same claim. As each claim has several questions, we compute the average score for the claim and provide a threshold to convert the score to a binary label.

3.4 Results and Analysis

In this section, we first present our results in comparison with the baselines. Then, we analysis the effectiveness of our method with the attention visualization and an ablation study.

3.4.1 Comparison with baselines

We show the results with different baselines in Table 3.1. For the metric-based methods, we do a binary search to find the highest accuracy on the development set for the threshold selection.

The results show that training an answer comparison model specifically for the fact-checking task improves accuracy compared with the methods without training. Our attention-based method achieves slightly lower accuracy than the blackbox method. However, our method is more suitable for real-world applications than the blackbox one because: 1) our method essentially reduces the input needed for prediction while remaining almost the same accuracy, 2) we enable error analysis for fact-checking with several steps, and 3) our model additionally provides more explainability by learning the importance of each question.

[3] https://bit.ly/2ZO6CtR

Methods	Dev Acc	Test Acc
Blackbox (No X-AI)	**76.17±1.23**	**74.58±1.66**
QUALS (th=-1.2)	56.12	56.01
BERTscore (th=0.843)	58.68	62.32
cosine sim (th=0.305)	61.16	62.75
F1-score (th=0.06)	64.07	63.77
Attention C-Q-AA (ours, X-AI)	75.44±0.52	73.43±0.83

th: threshold

Table 3.1: Fact-checking label accuracy of different methods. 'X-AI' denotes Explainability capabilities.

3.4.2 Attention visualization

To illustrate how attention helps explainability, we show an example of our generated questions with attention weights and their answers from claim and evidence in Fig. 3.2. The question with the highest weight is bold, and the second-highest underlined. Although some answers are incorrect and there are non-matching answer pairs, the model can attend more on the questions and answers more relevant to the factuality of the claim, showing our approach's potential. We also see that because the claim is short, most questions are repetitive.

3.4.3 Ablation study

We carry out an ablation study to show if our attention mechanism improves performance compared with simple classification. Thus we remove the attention layer of our proposed attention model, the network structure is shown in Fig. 3.3. Specifically, to use all available questions, we concatenate all questions and all answers: so the model has two inputs $\langle\text{CLS}\rangle\ C\ \langle\text{SEP}\rangle\ Q_1Q_2\cdots Q_n$, and $\langle\text{CLS}\rangle\ A_1^CA_2^C\cdots A_n^C\ \langle\text{SEP}\rangle\ A_1^EA_2^E\cdots A_n^E$ (note here n can be different for different claims). As the inputs are concatenated, the maximum token length here is 128. Then through the embedding model, each input is transformed into a feature vector, and the two vectors are concatenated to be fed into the classification layer.

To study the effect of different components of our proposed model, we design the inputs as follows:

- C: only claims $\langle\text{CLS}\rangle\ C$;
- Q: only concatenated questions $\langle\text{CLS}\rangle\ Q_1Q_2\cdots Q_n$;
- AA: only answer pairs $\langle\text{CLS}\rangle\ A_1^CA_2^C\cdots A_n^C\langle\text{SEP}\rangle\ A_1^EA_2^E\cdots A_n^E$;
- Q-AA: concatenated questions $\langle\text{CLS}\rangle Q_1Q_2\cdots Q_n$ and answer pairs $\langle\text{CLS}\rangle\ A_1^CA_2^C\cdots A_n^C$ $\langle\text{SEP}\rangle A_1^EA_2^E\cdots A_n^E$;
- CQ-AA: our full model without attention (shown in Fig. 3.3).

Claim: *Lee MacPhail passed away at his home at the age of 92.*
Evidence: *MacPhail lived in Delray Beach, Florida, where he died November 8, 2012, at his home. He was 95. At time of his death he was the oldest living Hall of Famer.*

Question	Answer for claim	Answer for evidence	Attention weight (%)
How old was MacPhail when he died?	*92*	*95*	12.17
Where did MacPhail die?	*at his home*	*Delray Beach, Florida*	4.29
Who died at the age of 92?	*MacPhail*	*no_answer*	1.66
When did MacPhail die?	*at the age of 92*	*November 8, 2012*	3.23
What was MacPhail's age?	*92*	*95*	13.19
At what age did MacPhail die?	*92.*	*95*	18.87
Where did MacPhail die?	*his home*	*Delray Beach, Florida*	4.29
What age was MacPhail when he died?	***92.***	***95***	**20.67**
At what age did MacPhail die?	*92*	*95*	18.87
Who died at age 92?	*MacPhail*	*no_answer*	2.75

Predicted label: REFUTES **Gold label**: REFUTES

Figure 3.2: An example of our model generated questions, answer pairs, and attention weights. The question with the highest weight is in bold, and the second highest is underlined.

- Attention C-Q-AA: our full model with attention.

From the ablation study, we want to know how much each input affects the model's performance. In Table 3.2, we can see that with C or Q only, the model cannot perform well, indicating that the model can not rely solely on the claim information to achieve high accuracy. Our result agrees with the original paper, in which the model only with claims achieved an accuracy of 61.9%. Also, when adding Q and CQ information to AA, Q-AA and CQ-AA perform slightly better. This indicates that the model can learn most of the information from answer pairs only. It is reasonable because the answer pairs carry most of the critical information from both claims and evidence. Comparing CQ-AA with our attention-based C-Q-AA, we see that the attention mechanism can help increase performance because it uses claims and questions to weigh up essential answer pairs.

3.5 Limitations

- **Question Generation.** Generating diverse and relevant questions aiming at the factuality of a claim is challenging. Claims can be altered by changing the subject, object, time, place, actions, or even multiple editions together. In some cases, we observed that the questions have the problem of not recognizing complete phrases of the claim, and sometimes most questions of a claim are semantically similar because the claim is too short. For example in Fig. 3.1, we can see that most of the questions are paraphrases.

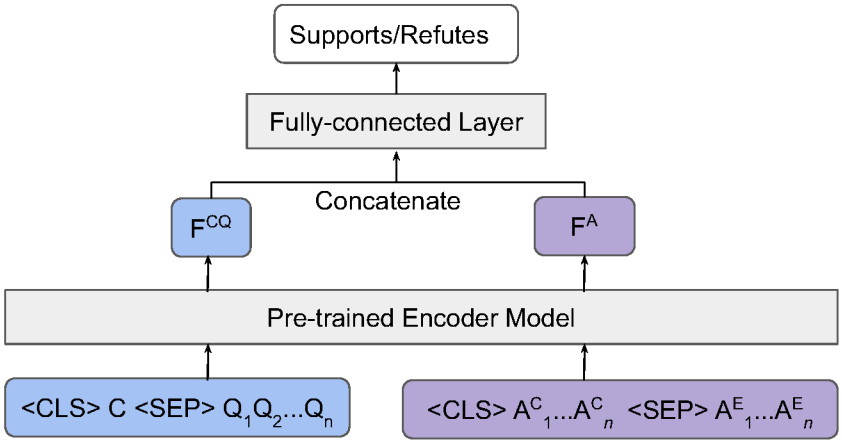


Figure 3.3: Answer comparison model without attention on questions. C represents a given claim, Q_i represents i_{th} questions, and $(A^C{}_i, A^E{}_i)$ represents i_{th} answer pairs for claim and evidence. n denotes the number of questions for claim C.

Inputs	Dev Acc	Test Acc
C	59.15±1.22	61.57±1.67
Q	56.22±1.37	56.90±0.90
AA	74.15±1.33	72.61±1.04
Q-AA	74.46±0.80	72.62±1.59
CQ-AA	74.88±0.81	72.89±1.14
Attention C-Q-AA	**75.44±0.52**	**73.43±0.83**

Table 3.2: Ablation study of the model without attention

Hence, better ways of generating questions and filtering less relevant and repetitive questions are needed to improve performance.

- **Question Answering.** Answering correctly giving the context is a non-trivial and crucial step in the pipeline. Unfortunately, state-of-the-art models can fail to answer correctly in some cases, as they require reasoning and logical thinking to calculate the correct answer from the context. We show a failing example here: **Evidence**: *Weber was born in Eutin, Bishopric of Lübeck, the eldest of the three children of Franz Anton von Weber and his second wife, Genovefa Weber, a Viennese singer.* **Question**: *How many siblings did Albert Weber have?* **Answer for evidence**: *three.*

  In the example, the model is not able to give the correct answer –*two*, because it is an extractive QA model, which is a limitation of this type of model. Nevertheless, the explainability provided by questions and answers gives us a better idea of which part is wrong in the claim and what could help us improve the model.

Reasoning over text is a very challenging task; other ways of transforming the claim into a format like tabular data [47] may also help simplify the reasoning and thus improve performance.

3.6 Final Remarks

In this chapter, we proposed a novel pipeline for using QA as a proxy for fact-checking. Based on this pipeline, we proposed an answer comparison model with an attached attention mechanism, which learns to attend critical questions with interpretability capabilities. Our ablation study showed that the model can achieve near state-of-the-art performance with only information from answer pairs. Thus, using QA, we can encourage the model to learn from more precise evidence; this can aid fact-checkers in better understanding models' decisions. Then, when necessary, they can compare the answers and make decisions for themselves.

Chapter 4

Self-Rationalization in the Wild: Large-scale Out-of-Distribution Evaluation of NLI related tasks

In the previous chapter, we proposed adding QA pairs to provide more explainable fact-checking results. Due to the lack of annotated QA pairs, we faced challenges in learning how to generate good questions and properly evaluate the generated questions. In addition, due to separate models for question generation, question answering, and answer comparison, an error from one model can aggregate to the next, causing inaccurate results. Thus we focus on a more straightforward approach – self-rationalization, generating free-text explanations along with the predictions.

Generating textual explanations has been a major focus in machine learning and NLP [48, 49, 50], as the explanations are expressive and do not require readers to have model-level knowledge to understand. One popular line of work is self-rationalization [51, 52], in which a model jointly generates the task label and a free-text explanation for the predicted label. Compared with highlighting words and phrases [53], free-text explanations can express unstated knowledge and common-sense in easily understandable forms. However, datasets containing annotated free-text explanations are rare due to expensive annotations.

A few datasets for free-text explanation generation [54, 55, 56, 57, 58] exist, with e-SNLI [54] being one of the seminal datasets in the NLI area. Based on SNLI [59], the dataset focuses on reasoning over fine-grained nuances of common-sense knowledge. However, datasets containing longer or more domain-specific text, such as fact-checking on real-world claims, lack annotated explanations [60, 61]. This poses severe challenges for (i) training and (ii) evaluating self-rationalizing models on these tasks. No large scale analysis exists to understand how well self-rationalization models can transfer from existing data to unknown datasets.

We fill the gap by learning self-rationalization from established sources with annotated explanations and evaluating its generalization performance on 19 out-of-distribution (OOD) datasets over three related tasks (see evaluation setup in Figure 4.1): NLI, fact-checking (FC) and hallucination detection of abstractive summarization (HDAS). NLI focuses on textual entailment within a controlled context, FC extends to reason over real-

world claims with retrieved evidence, and HDAS centers around machine-generated text. Our OOD datasets vary in *domains* (e.g., news, Wikipedia, social media, science), and *textual structures* (e.g., synthetic template-based, multiple premises, sentence compositions, long documents), presenting a diverse and challenging OOD setting (see details of each dataset in Table 4.1). Despite the popularity of LLMs, using them in a large exper-

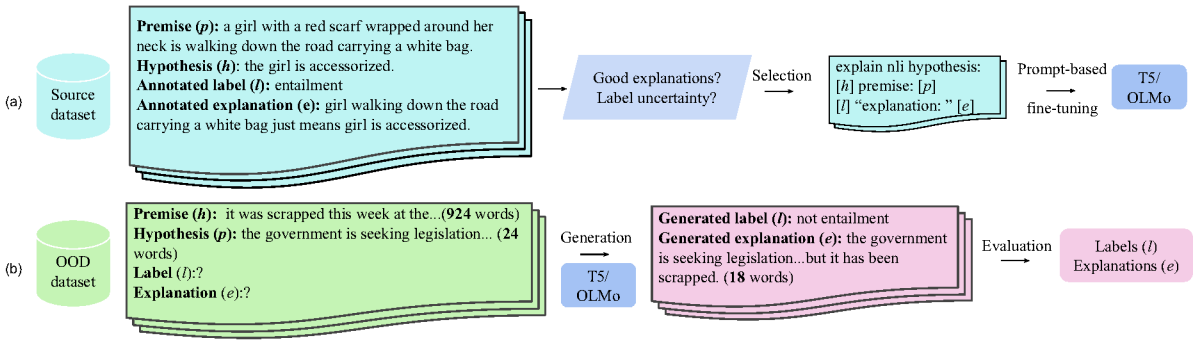


Figure 4.1: OOD evaluation pipeline of self-rationalization. The pipeline comprises two main parts. The first part (a) relates to **learning to self-rationalize** with a source dataset (Section 4.2); it involves sample selection and fine-tuning a generative model. The second part (b) relates to **OOD generation and evaluation** (Section 4.3); we evaluate the model on three categories of OOD tasks: NLI, fact-checking, and hallucination detection.

imental design is prohibitive, as they are computationally expensive to perform inference and evaluation, especially when the input text is long. Further, data contamination is a concern when performing evaluations on OOD datasets [62], as the training data of most LLMs are not transparent, such as Llama 2 [63] and GPT-4 [64]. To address this, we selected two open-source models—T5-Large [25] and OLMo-7B [65]—to study self-rationalization, both of which have fully transparent pretraining datasets. They also require fewer computational resources than many LLMs, allowing us to perform a large scale study.

We study the impact of data size and quality on OOD performance, focusing on these three factors: the source dataset for fine-tuning, the number of selected samples, and sample selection strategies for few-shot fine-tuning. To enhance the quality of generated explanations in OOD datasets, we introduce a new approach with an acceptability filtering model [66] to select better training samples. We address the lack of gold reference explanations by studying the effectiveness of the Acceptability score with a human evaluation and comparing it against three LLM-based reference-free metrics. Out of the automatic metrics, the Acceptability score correlates highest with humans in all three tasks. Our evaluation results show that: 1) OOD performances are comparable between models fine-tuned with few-shot selected samples and a full training set; 2) fine-tuning data source has a high impact on OOD performance, while sample selection has a lower impact; 3) higher Acceptability scores are associated with better label prediction performances, providing a new perspective on the task performance vs explainability trade-off.

The work described in this chapter was developed in collaboration with the Ubiquitous Knowledge Processing (UKP) Lab, under the supervision of Prof. Iryna Gurevych. It is

accepted in the Transactions of the Association for Computational Linguistics (TACL)[1], a renown open-source journal in the fields of Computational Linguistics and Natural Language Processing.

4.1 Related Work

This section presents the related work related to free-text explanation generation and its evaluation, and few-shot sample selection methods.

4.1.1 Free-text explanation generation and evaluation

Self-rationalization has been a popular approach for generating free-text explanations [51, 52, 67, 68, 69]. Wiegreffe et al. [51] shows that joint learning of label prediction and explanation generation results in explanations more aligned with predicted labels. Marasovic et al. [52] addressed the scarcity of annotated explanation data by using prompt-based fine-tuning on a few examples, though their evaluation was limited to in-distribution datasets. Few works have studied how such models can generalize to OOD. Zhou and Tan [70] studied how learning with few-shot instances with template-based explanations influences OOD generalization. Their OOD dataset (e-HANS) is limited with constructed templates based on the HANS dataset [71]. Ross et al. [67] studied the effect of self-rationalization on reducing models' reliance on spurious cues in out-of-domain datasets, and they showed that self-rationalization improves models robustness when fine-tuning data size is small. Yordanov et al. [72] studied the setup where the target dataset has few annotated free-text explanations but abundant labels. Their approach is limited to target datasets in which free-text explanations exist. In contrast to the above OOD evaluations, we focus on the OOD evaluation of self-rationalization for 19 diverse datasets, and our evaluation does not rely on reference explanations.

Reliable evaluation is crucial for explanation generation. Traditional metrics that measure text overlap with references have shown low correlation with human judgments [73], and reference explanations are not always available. Recent works, like TigerScore [74], Auto-J [75], and Themis [76], use LLMs as evaluators. These metrics rely on detailed instructions specifying evaluation aspects (e.g., relevance, accuracy, coherence) and formatted inputs for the task. The trained metric then generates a rating along with a textual analysis. To test their suitability for the explanation generated with self-rationalization, in this work, we study their correlations with human judgments.

4.1.2 Few-shot sample selection

Recent studies show that fine-tuning with smaller, high-quality datasets can outperform larger datasets [77, 78]. Li et al. [77] proposed to use a relatively small language model to evaluate and select a few instances for instruction-tuning on larger models. To select data to perform well in transfer learning, Xia et al. [78] proposed data selection for

[1]Copyright for TACL papers is held by the Association for Computational Linguistics, and articles are distributed under Creative Commons License CC-BY.

instruction-tuning on a target-specific domain. They show that training with 5% of the data outperforms training with the full dataset. The main constraint is that the validation set needs to be from the target domains. Chen and Mueller [79] proposed to improve data quality by estimating their model's confidence, and for the low-quality data, they either filter or correct them. Most methods for sample selection are designed to perform well on in-distribution or known target domains, and the goal is for better classification performance. In contrast, our work focuses on selecting data that should help OOD performance on both label prediction and explanation generation.

4.2 Learning to Self-rationalize

Figure 4.1 shows our out-of-distribution (OOD) evaluation pipeline. We first (a) fine-tune a language model on a source dataset to learn self-rationalization. Specifically, we require a fully annotated source dataset S, in which each instance contains input $x_s = (h_i, p_i)$ and output $y_s = (l_i, e_i)$, where h_i, p_i represent a hypothesis and premise pair, l_i and e_i represent the annotated label and explanation. We select m representative instances per class from S for fine-tuning by following a sample selection process. Our sample selection method deliberately restrains from using data from the OOD datasets, preserving them untouched. Finally, we fine-tune a language model to generate a label and explanation. In (b), we evaluate the fine-tuned model performance on OOD datasets (Section 4.3). Given an OOD dataset O, with instances $x_o = (h_j, p_j)$, where h_j, p_j represents a new hypothesis and premise pair, the fine-tuned model generates the label ($\hat{l_j}$) and explanation ($\hat{e_j}$).

4.2.1 Source dataset

To learn self-rationalization for NLI-related tasks, we select two large source datasets that contain explanations: (a) **e-SNLI** [54], derived from the NLI dataset SNLI [59] by adding human annotated explanations. (b) **e-FEVER** [80], originated from the fact-checking dataset FEVER [11] with GPT-3 generated synthetic explanations. To improve data quality, we heuristically filter out incorrect explanations from the dataset (around 14% of samples are removed from the training set) following the rules below:

- The explanation is: "The relevant information about the claim is lacking in the context." but the label is not NEI (NOT ENOUGH INFO).
- The explanation repeats the claim, and the label is not SUPPORTS.

We selected these two datasets as they are representative for our OOD datasets and have abundant explanations.

4.2.2 Acceptability-based sample selection

Inspired by Schiller et al. [81], we examine how varying the size and quality of fine-tuning data (source dataset) affects OOD performance. Since self-rationalization includes joint label prediction and explanation generation, we propose our method considering both the label and explanation quality:

Data filtering with acceptability score To improve explanation quality, we filter the fine-tuning data using the acceptability model from Wiegreffe et al. [66]. This model, trained on SNLI data, predicts whether a generated explanation is acceptable based on human judgment. We remove samples with acceptability scores (the predicted probability for the label "acceptable") below a 0.3 threshold.

Data selection For data quality estimation in label prediction, we adapt two methods from the literature: (1) **ambiguous**: Following Swayamdipta et al. [82], we select samples with high ambiguity, which has been shown to improve OOD generalization. Ambiguity is measured as the distance between an instance's predicted label probability and the mean of all predicted label probabilities using the pre-fine-tuning model (details in Appendix A.1.2). (2) **FastVote-*k*** [83]: A graph-based method to select diverse and representative samples. We use the recommended $k = 150$.

With the combined two steps (data filtering + selection), we denote the sample methods as **accept-ambiguous** and **accept-FastVote-*k***.

4.2.3 Fine-tuning on source datasets

For fine-tuning T5-Large, we use the standard NLI template from [52], which has been shown to give the best results for e-SNLI dataset with T5. The encoder and decoder prompts are (also shown in Figure 4.1) :

> **Input**: *explain nli hypothesis:* [hypothesis] *premise:* [premise]
> **Output**: [label] *"explanation: "* [explanation]

For fine-tuning OLMo-7B, as the model is relative large, we choose parameter-efficient tuning with LoRA [84] using the following instruction [85]. The response is in a JSON format to facilitate extraction of labels and explanations:

```
### Premise: [premise] Hypothesis: [hypothesis]
### Response: {"relationship": [label], "explanation": [explanation]}
```

For the number of shots, we compare 1, 2, 4, 8, 16, 32, 64, and 128 shots. To ensure robustness, we create five subsets from each source dataset, with 5,000 randomly selected samples per subset (with no overlap between subsets). We apply the sample selection methods from Section 4.2.2 to each subset and report the average results (see Appendix A.1.2 for additional fine-tuning details). In total, we fine-tuned 402 T5 models and 302 OLMo models[2].

Baselines We compare the few-shot fine-tuned models with two full-set fine-tuned models on e-SNLI and e-FEVER, respectively. In addition, we include the random sample

[2]For T5: 2 source datasets ×5 subsets ×8#shots ×5 sampling methods +2 full-shot models. For OLMo, we discard 1 and 2 shots as our primary results show that models fail to learn with too few examples.

selection baseline to compare few-shot sample selection methods.

4.3 OOD Self-rationalization

In this section, we introduce part (b) of the pipeline in Figure 4.1. For all fine-tuned models, we perform inference on all OOD datasets.

4.3.1 Out-of-Distribution datasets

For a comprehensive evaluation, we collect datasets that resemble the NLI task and divide them into three categories: **NLI**, Fact-checking (**FC**), and Hallucination Detection of Abstractive Summarization (**HDAS**). Table 4.1 lists the OOD datasets used (see Appendix A.1.1 for dataset details and pre-processing). To ensure no data contamination in our OOD evaluation, we specifically excluded datasets used for supervised fine-tuning of T5 [25]. OLMo model was pre-trained on Dolma [86] corpus, which contains data from diverse sources but is not fine-tuned with curated NLI datasets.

	OOD dataset	Size	#L.	Domain	#words (Hyp.)	#words (Pre.)	IAA
NLI	SICK [87]	4,906	3	news, image captions	10	10	0.84^{O}
	AddOneRTE [88]	387	2	news, image captions, forums, literature	13	12	0.77^{O}
	JOCI [89]	39,092	3	image captions, commonsense stories	6	14	0.54^{C}
	MPE [90]	1,000	3	image captions	4	48	0.70^{O}
	DNC [91]	60,036	2	events, named entities, puns, sentiments	5	19	-
	HANS [71]	30,000	2	template-based (synthetic)	6	9	-
	WNLI [92]	71	2	fiction books	7	21	-
	Glue Diagnostics [92]	1,104	3	news, Reddit, Wikipedia, academic papers	16	16	0.73^{F}
	ConjNLI [93]	623	3	Wikipedia	13	13	0.83^{C}
FC	Snopes Stance [60]	1651	3	Snopes (fact-checking platform)	16	126	0.70^{C}
	SciFact [94]	300	3	biomedicine, scientific articles	13	247	0.75^{C}
	Climate-FEVER [95]	1,381	3	climate change, Google searches	20	136	0.33^{K}
	VitaminC [96]	55,197	3	Wikipedia, COVID-19	13	28	0.71^{F}
	COVID-FACT [61]	4,086	2	Reddit, COVID-19	12	73	0.50^{C}
	FM2 [44]	1,380	2	Wikipedia	14	32	-
HDAS	FactCC [38]	503	2	news (CNN/DailyMail), rule-based	14	644	0.75^{C}
	QAGs CNNDM [40]	714	2	news (CNN/DailyMail), BART-based	16	318	0.51^{K}
	QAGs XSUM [40]	239	2	news (XSUM), BART-based	18	351	0.34^{K}
	XSUM Hallucination [37]	1,869	2	news (XSUM), 7 different models	19	361	0.92^{O}

Table 4.1: OOD datasets categories and details. NLI: yellow, FC: pink, and HDAS: blue. Hyp.: hypothesis, Pre.: premise, #words: number of words in average, IAA: inter-annotator agreement (numbers are from the original papers). L.: labels, C: Cohen's kappa, F: Fleiss's kappa, K: Krippendorff's alpha, O: other metrics, -: unspecified. The sizes are reported on test/dev split; if the split is not provided, we report and evaluate on the entire dataset.

NLI NLI datasets access models' ability to infer relationships between sentences, with challenges ranging from compositional meaning [87], adjective-noun composition [88], common-sense inference [89], to multiple premise entailment [90]. DNC [91] expands the challenge by incorporating diverse semantic phenomena into the NLI format. HANS [71]

and WNLI [92] are two adversarial datasets designed to reveal models' underlying heuristic biases. Glue Diagnostics [92] and ConjNLI [93] further diversify the NLI task, testing models against a wide array of linguistic challenges and over conjunctive sentences.

FC FC datasets aim to evaluate the veracity of claims against evidence from various sources, including fact-checking platforms [60], scientific articles [94], Wikipedia [96, 44], and information related to climate change and COVID-19 [95, 61]. The domain-specific nature of some datasets, such as SciFact's focus on biomedicine and Climate FEVER's on climate change, requires models to be domain-aware and handle evidence with varying granularity. FC datasets challenge models to evaluate the truthfulness of claims in real-world scenarios with applied NLI techniques. For all FC datasets, we use gold evidence, considering that retrieved evidence may change the gold label of the claim).

HDAS HDAS datasets encompass a variety of model-generated summaries, reflecting the evolving landscape of automatic text generation and its implications for information integrity. FactCC [38] challenges models to identify inaccuracies in summaries generated through five rule-based transformations. QAGS CNN and QAGS XSUM [40], derived from CNN/DailyMail and XSUM datasets, consist of summaries generated by the BART model [24]. XSUM Hallucination [37] contains factuality annotated summaries generated by seven models.

In comparison, the three tasks vary in objective, domain, and text length. NLI targets logical relationships between sentences, requiring models to handle linguistic subtleties and logic-based reasoning in a controlled textual context. FC focuses on real-world applicability, requiring external information and complex reasoning between sentences and documents. HDAS addresses the problems of automatic document summarization. Regarding text length, FC datasets typically have longer premises than NLI, with HDAS having the longest. Together, these datasets present a challenging NLI-related OOD scenario.

4.3.2 Inference on OOD datasets

During OOD inference, a fine-tuned model may not generate a label and explanation following the output template, due to poor generalizability. To address this, for T5 models, we take the first token to represent the predicted label. For label mapping, we focus on probabilities of tokens corresponding to our target labels: "entailment", "contradiction", "neutral", disregarding others[3]. The label is then determined based on the highest probability among these three tokens. For datasets that only include two classes ("entailment" and "non-entailment"), we merge the "contradiction" and "neutral" labels into the "non-entailment" label. Explanation extraction involves processing the entire token sequence. We search for the pattern "explanation: " to identify explanations. If absent, we treat all text after the first word as the explanation. For OLMo models, as we instruction-

[3]except for "entailment", as this word contains three word tokens: "en", "tail" and "ment", we take the token number of "en".

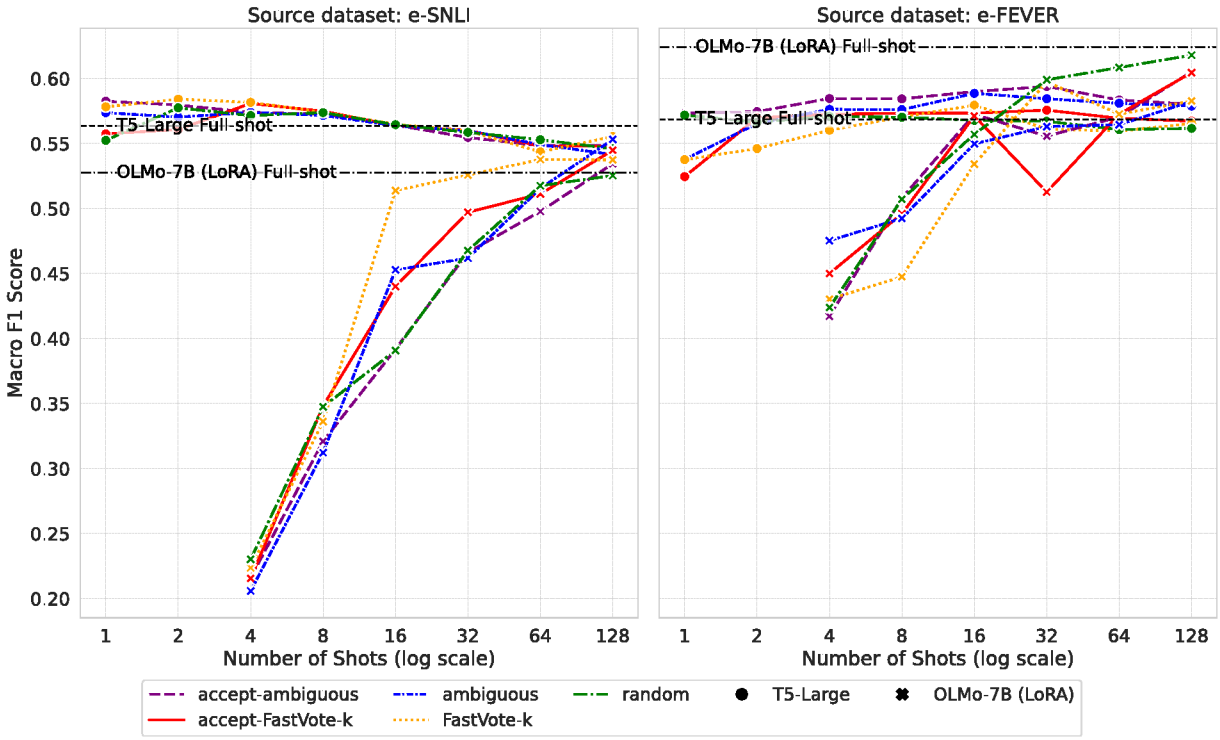


Figure 4.2: Average Macro F1 score across different number of shots and sample selection methods. Each point is the average of all 19 OOD datasets, and 5 models from the 5 subsets.

tuned the model to generate a json formatted output, we try to extract the labels and explanations by finding their keys, if not found, we set both to be none.

4.4 OOD Performance on Label Prediction

We compare the OOD label prediction performance of fine-tuned T5-Large and OLMo-7B models on two source datasets, considering various sample selection methods and number of shots, as shown in Figure 4.2. Label prediction performance is measured using the Macro F1 score.

T5 vs. OLMo: As shown in Figure 4.2, T5 and OLMo models exhibit distinct trends in label prediction performance as the number of shots increases. OLMo starts with low performance, improving almost monotonically with more shots. T5, however, shows less variation, starting with slightly higher performance and then reaching levels similar to full-shot models. This difference may be because of T5's pre-training on NLI datasets (MNLI, QNLI, RTE, CB), allowing it to handle NLI tasks effectively without much benefit from additional fine-tuning (see detailed discussion in Section 4.7.1). This is further indicted by the results: T5 full-shot fine-tuning with both source datasets have similar F1 scores, and neither yields better results than their best few-shot counterparts.

e-SNLI vs. e-FEVER: Overall, e-FEVER models achieve better average OOD F1 than e-SNLI, and the OLMo model fine-tuned on e-FEVER full-shot has the highest

Acronym	Source	Model	#Shots	Selection
$T^{Fev}_{64,AFk}$	e-FEVER	T5	64	accept-FastVote-k
$T^{Fev}_{128,R}$	e-FEVER	T5	128	random
$T^{Fev}_{128,Fk}$	e-FEVER	T5	128	FastVote-k
$T^{Fev}_{128,AFk}$	e-FEVER	T5	128	accept-FastVote-k
T^{Fev}_{Full}	e-FEVER	T5	Full	-
$T^{Sn}_{64,Fk}$	e-SNLI	T5	64	FastVote-k
$T^{Sn}_{64,AFk}$	e-SNLI	T5	64	accept-FastVote-k
T^{Sn}_{Full}	e-SNLI	T5	Full	-
$O^{Fev}_{16,AFk}$	e-FEVER	OLMo	16	accept-FastVote-k
$O^{Fev}_{128,AFk}$	e-FEVER	OLMo	128	accept-FastVote-k
O^{Fev}_{Full}	e-FEVER	OLMo	Full	-
$O^{Sn}_{128,AFk}$	e-SNLI	OLMo	128	accept-FastVote-k
O^{Sn}_{Full}	e-SNLI	OLMo	Full	-

Table 4.2: Selected models for human evaluation for the models **T**5 and **O**LMo. The left most column shows the acronym of the models, which will be used throughout the rest of the paper.

OOD F1 score. For e-SNLI, T5 and OLMo models reach similar performances at 128 shots, but the trends are the opposite. For e-FEVER, T5 models' performance tends to stabilize after just 2-shots, while OLMo models' performance continues to increase and eventually outperform T5 models.

Sample Selection As depicted in Figure 4.2, no sample selection method consistently outperforms others in label prediction. For T5, selection methods perform similarly, especially with e-SNLI, though "accept-ambiguous" is slightly better with e-FEVER. For OLMo, "FastVote-k" excels with e-SNLI, while "random" selection outperforms others with e-FEVER (after 32 shots), nearly matching full-shot performance. Surprisingly, "FastVote-k" and "ambiguous" do not surpass the random baseline, possibly due to outliers and training instability when using small numbers of samples [97, 83].

4.5 OOD Explanation Quality Evaluation

We evaluate the generated explanations using both human evaluation and reference-free automatic metrics, and analyze the correlation between them.

4.5.1 Human evaluation setup

Conducting a human study is challenging due to the extensive number of models and OOD datasets. Thus, we select three OOD datasets (SICK, VitaminC, XSUM Hallucination) representing NLI, FC, and HDAS, respectively. To study the impact of fine-tuning factors on OOD explanations, we select models that demonstrated high and comparable F1 scores

averaged across the three OOD datasets (see Figure A.2 in Appendix A.2 with the selected models highlighted). Table 4.2 lists the 13 selected mode details, with first column provides models' acronyms for across reference later (examples of generated explanations by the selected models can be found in Table A.1, A.2 and A.3 in Appendix A.1.6).

For instance selection, following Marasovic et al. [52], we shuffle each dataset and select the first 15 correctly predicted instances per class and model. This results in 1560 instances, including those with identical hypothesis-premise pairs but different model-generated explanations. Each instance is evaluated by three different workers, and each worker evaluate 10 instances, requiring in total 468 crowd-workers[4]. Evaluators are shown the hypothesis-premise pair, its relationship (gold label), and the generated explanation and then asked to answer two questions (see the evaluation page in Figure A.1 of Appendix A.1.4).

- Given the Hypothesis and Premise, does the Explanation justify the given Relationship (Single-selection)? Options: *Yes*, *Weakly Yes*, *Weakly No* and *No*.
- What are the shortcomings of the Explanation (Multi-selection)? Options: *Does not make sense*, *Insufficient justification*, *Irrelevant to the task*, *Too trivial (only repeating one of the sentences)*, *Contains hallucinated content (not present the premise)* and *None (only if the previous answer is Yes)*.

We calculate the average score of each instance from 3 evaluators by assigning the weight to the selected answers as follows [52, 72]: Yes: 1, Weakly Yes: 2/3, Weakly No: 1/3 and No: 0.

We use the Prolific platform for recruiting workers, and the open-source POTATO annotation tool [98] for the evaluation interface.

4.5.2 Evaluation with reference-free metrics

We propose to use the **Acceptability score**[5] [66] as a reference-free metric, considering it is designed for accessing NLI explanations. We choose the largest size of the model variance: T5-11B. The model assigns a score between 0 and 1. We compare this metric against the state-of-the-art NLG reference-free evaluation metrics:

- **Auto-J** [75]: trained with LLaMA-2-13B-chat model to evaluate LLM-generated responses. The metric generates an explanation for its judgment and a final integer rating from 1 to 10.
- **TigerScore** [74]: trained with LLaMA-2 on MetricInstruct dataset. We choose the larger size of the metric: TIGERScore-13B. It generates a breakdown error analysis and a final error score from 0 to -infinity (the smaller, the better).

[4]To select eligible participants, our screening requires participants to have at least an undergraduate degree, and primary language as English, with an approval rate above 99%. For high-quality evaluation, we inserted 2 attentions questions to filter out low-quality evaluations (an evaluation is rejected if the worker failed on both attention checks, or failed on one and contains invalid answers through our manual checking).

[5]In this paper, when mentioning the acceptability filter (T5-Large), we start with lowercase "a", and the Acceptability metric (T5-11B) capital "A".

Dataset	Auto-J	TigerScore	Themis	Accept.
SICK	-0.011	-0.220	0.400	**0.466**
VitaminC	0.163	-0.263	0.394	**0.469**
XSUM H.	0.223	-0.216	0.326	**0.475**
All	0.123	-0.219	0.387	**0.484**

Table 4.3: Spearman's correlation between human scores and automatic scores in different OOD datasets. All correlation coefficients are significant with $\rho < 0.001$, except for Auto-J on SICK.

- **Themis** [76]: trained with Llama-3-8B based on their constructed dataset NLG-Eval. It offers flexible aspect-based evaluations across different tasks. We tested three aspects—relevance, coherence, and consistency—and selected relevance due to its highest correlation with human judgments. The metric outputs an evaluation analysis and provides a scale rating from 1 to 5.

For all reference-free metrics, we calculate the scores for all samples in the datasets, given ground truth inputs (hypothesis, premise, and gold label). Appendix A.1.5 presents the instructions of the evaluation models.

4.5.3 Correlation between human evaluation and automatic evaluation metrics

Table 4.3 shows the Spearman's correlation[6] between human and reference-free metrics for the three OOD datasets. The Acceptability score (T5-11B) has the highest correlation with human evaluation for all datasets, followed by Themis, and Auto-J has the lowest. The highest correlations on all three datasets demonstrate the usability of the Acceptability score as a reference-free metric for the explanation evaluation of NLI-related tasks.

4.5.4 Evaluation results on selected models and instances

Dataset	Human	Themis	Accept.
SICK	**0.655**	**2.185**	**0.437**
VitaminC	0.621	2.183	0.363
XSUM H.	0.567	1.633	0.202
All	0.620	2.046	0.350

Table 4.4: Human scores and automatic scores in different OOD datasets.

The average scores of human evaluations in the three OOD datasets are shown in Table 4.4. The scores show that SICK has the highest explanation scores, with VitaminC

[6]We choose Spearman over Pearson correlation as Pearson correlation assumes variables to be continuous and from a normal distribution.

slightly lower than SICK's, and XSUM Hallucination the lowest, agreed by humans and two automatic metrics. This may be due to the extremely long premise/document in the XSUM dataset, making it difficult for the model to generate good explanations. For shortcomings of explanations, see the detailed results in Figure A.3 in Appendix A.2).

Table 4.5 shows the evaluation results on the 13 selected models. We include Acceptability and Themis scores as they have moderate correlations with humans. In addition, we show the average Acceptability score on all 19 datasets for overall results. We discuss the evaluation results regarding each factor in the following.

Model	Human	Themis	Accept. (3)	Accept. (19)
$T^{Fev}_{64,AFk}$	0.631	2.058	0.317	0.250
$T^{Fev}_{128,R}$	0.623	1.983	0.276	0.206
$T^{Fev}_{128,Fk}$	0.589	1.867	0.216	0.201
$T^{Fev}_{128,AFk}$	0.611	**2.092**	**0.328**	**0.256**
T^{Fev}_{Full}	**0.653**	1.958	0.309	0.191
$T^{Sn}_{64,Fk}$	0.621	2.133	0.369	0.259
$T^{Sn}_{64,AFk}$	**0.679**	**2.367**	0.418	0.281
T^{Sn}_{Full}	0.678	2.050	**0.519**	**0.343**
$O^{Fev}_{16,AFk}$	0.631	**2.417**	**0.423**	0.305
$O^{Fev}_{128,AFk}$	0.639	2.250	0.384	**0.307**
O^{Fev}_{Full}	**0.656**	1.917	0.311	0.219
$O^{Sn}_{128,AFk}$	**0.643**	**2.300**	**0.491**	**0.303**
O^{Sn}_{Full}	0.408	1.208	0.194	0.111

Table 4.5: Evaluation results on OOD datasets of the 13 selected models. 3 means on the three selected datasets, 19 means all datasets. Models are grouped by base models and source datasets.

T5 vs OLMo As shown in Table 4.5, the difference between the two base models is most pronounced with e-SNLI full-shot. T5 fine-tuned on full shot e-SNLI (T^{Sn}_{Full}) provides the best explanations (besides $T^{Sn}_{64,AFk}$), whereas OLMo on full-shot e-SNLI (O^{Sn}_{Full}) generates the worse explanations. This may be due to catastrophic forgetting in the OLMo model when fine-tuned on too many e-SNLI samples, as its few-shot version produces explanations comparable to those of the T5 model.

e-SNLI vs e-FEVER Most e-SNLI models outperform e-FEVER in explanation quality (under the same model type and number of shots), except for OLMO full-shot. This could be attributed to the higher quality of explanations in the e-SNLI source dataset, while e-FEVER explanations are generated by GPT-3 (see more detailed comparison in Section 4.7.2).

Few vs Full Overall, few-shot models achieved similar human scores to their full-shot counterparts, except for the OLMo full-shot e-SNLI model. Although full-shot models

showed slightly higher human scores, reference-free metrics favored the explanations generated by few-shot models, particularly for e-FEVER models.

Sample Selection As shown in Table 4.5, using the acceptability filter ("accept-FastVote-k") improves explanation quality compared with the same sample selection without the filter ("FastVote-k"); however, $\text{T}^{Fev}_{128,AFk}$ is not better than random selection ($\text{T}^{Fev}_{128,R}$) according to humans. Nevertheless, based on the scores from the two reference-free metrics, using the acceptability filter improves generated explanation quality (see more detailed discussion in Section 4.7.2).

4.6 Self-Rationalization in the Wild: Overall OOD Performance

A good self-rationalization model should perform well both on label prediction and explanation generation. Thus, we first evaluate the generated explanations from a large number of models using the Acceptability score (for all instances, we use the gold labels for calculating the Acceptability score). Due to computational constraints, we limit the number of shots to 4, 16, 64, 128, and full, with data selected from the first subset (the Acceptability scores across different number of shots and sample selections can be found in Figure A.4 of Appendix A.2). We then show models' overall performance considering both the F1 and Acceptability score. Finally, we select the best-performing models to demonstrate overall performance on the 19 OOD datasets.

4.6.1 Relationship between label prediction performance and explanation quality

Figure 4.3 shows the distribution of models under different fine-tuning factors, with the x-axis showing the Acceptability score and the y-axis the macro F1 score (scores are averaged over all datasets). We select the best models based on the Pareto fronts[7].

As depicted in Figure 4.3, higher Acceptability scores are usually associated with better F1 scores. Regarding each factor, we see that 1) OLMo models' OOD performances are less stable than T5 models' but achieve better results with higher numbers of shots; 2) Sample selection methods with the acceptability filter have higher Acceptability scores; 3) Comparing the source datasets, fine-tuning on e-SNLI in general achieve higher Acceptability scores while on e-FEVER yield better F1 scores (see more discussions on the impact of each factor in Section 4.7).

Regarding the best-performing models that consider both labels and explanations, two models are selected based on the Pareto front: $\text{O}^{Fev}_{128,AFk}$ (OLMo, 128 shots, accept-Fastvote-k, e-FEVER) and T^{Sn}_{Full} (T5, full-shot, e-SNLI). The first achieves the highest F1 score, while the second has the best Acceptability score, with both models performing competitively on the other metric.

[7]For each point if no other point is strictly higher in both scores, the point is part of the Pareto front. See definition in https://en.wikipedia.org/wiki/Pareto_front.

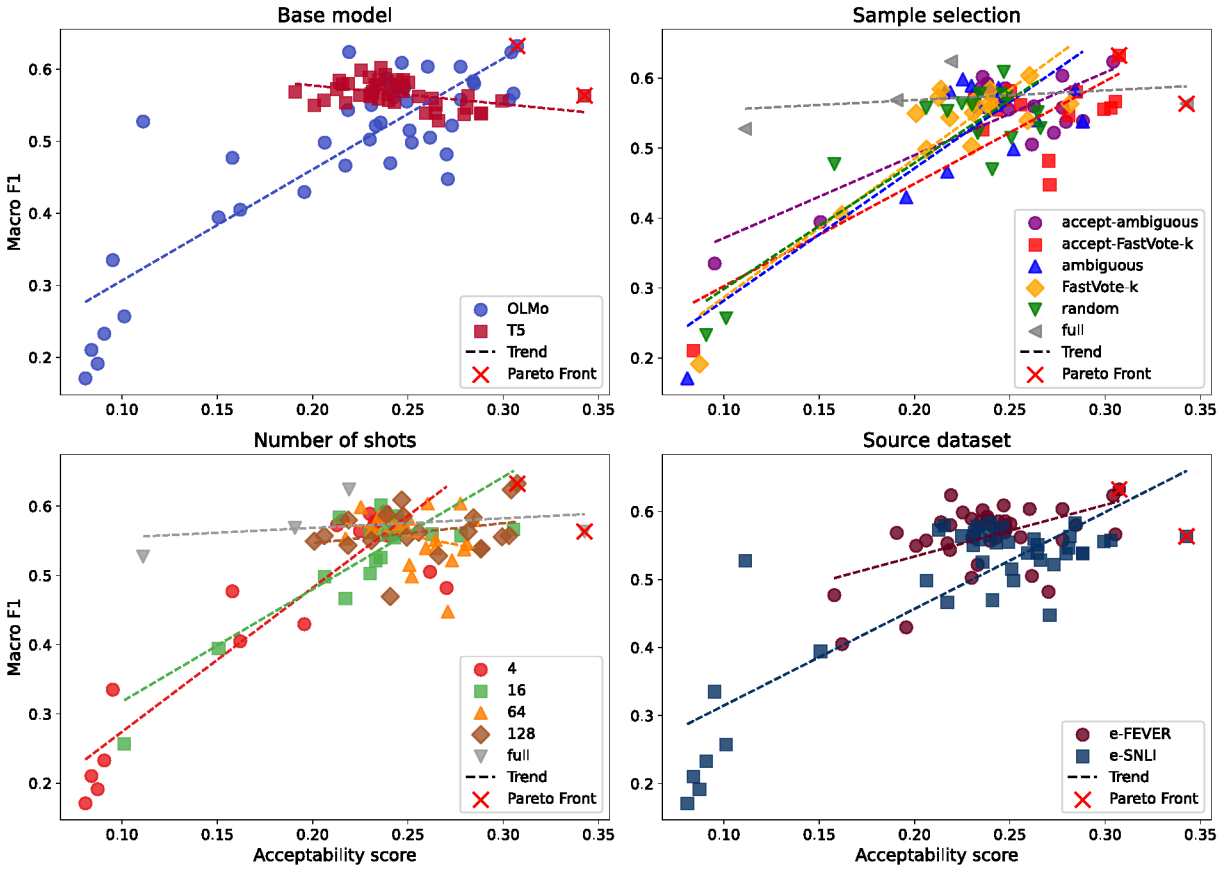


Figure 4.3: Distribution of models under different fine-tuning factors, with the x-axis showing the Acceptability score, and the y-axis the macro F1 score (scores are averaged over all datasets). The dashed lines are the estimated linear trends of the Acceptability score and macro F1 score.

4.6.2 Performance on the 19 OOD Datasets

Table 4.6 shows the F1 score and Acceptability score on the best models across each OOD dataset (state-of-the-art results on each dataset can be found in Table A.4 of Appendix A.2). As a comparison, we also include two other models with the same configurations as the best models but trained on a different source dataset: T^{Fev}_{Full} and $\text{O}^{Sn}_{128,AFk}$.

As shown in Table 4.6, the $\text{O}^{Fev}_{128,AFk}$ model achieves the highest F1 score on most OOD datasets, though its Acceptability score is slightly lower than that of the T^{Sn}_{Full} model. When comparing e-SNLI and e-FEVER fine-tuned models, e-FEVER models generally outperform in F1 scores on FC and HDAS datasets, with $\text{O}^{Fev}_{128,AFk}$ scoring about 10 percentile higher on average for FC (slightly less) and HDAS (slightly more). In terms of explanation generation, OLMo-based models exhibit better performance. Even on e-FEVER, OLMo achieves competitive scores across most OOD datasets, whereas the T5 model fine-tuned on e-FEVER (T^{Fev}_{Full}) produces the worst explanations, except for the HDAS task (this might also be due to the number of shots difference, as fine-tuned on more number of shots with e-FEVER do not always lead to better explanations). Finally, the Acceptability scores show a decreasing trend from NLI to HDAS tasks, consistent with previous human evaluation results (Table 4.4), where datasets with longer premises generally resulted in lower Acceptability scores.

	Macro F1 score				Acceptability score			
Dataset	$\mathbf{T}^{Sn}_{Full}$	T^{Fev}_{Full}	$\mathrm{O}^{Sn}_{128,AFk}$	$\mathbf{O}^{Fev}_{128,AFk}$	$\mathbf{T}^{Sn}_{Full}$	T^{Fev}_{Full}	$\mathrm{O}^{Sn}_{128,AFk}$	$\mathbf{O}^{Fev}_{128,AFk}$
SICK	58.5	**78.8**	55.4	65.1	**53.0**	18.5	47.5	40.2
AddOneRTE	72.3	**75.6**	65.0	72.0	44.5	9.3	**44.9**	39.4
JOCI	52.5	41.8	49.2	**53.7**	**51.9**	12.4	43.6	41.6
MPE	**68.7**	37.7	62.4	60.7	**49.8**	6.4	45.8	39.2
DNC	60.1	**66.9**	53.4	58.5	**35.1**	10.0	25.8	32.8
HANS	58.2	43.3	51.7	**65.9**	**38.6**	27.6	24.0	27.8
WNLI	35.0	32.4	42.1	**55.1**	29.9	22.7	**31.7**	28.0
Glue Diagnostics	57.9	59.3	57.7	**61.3**	**47.9**	29.0	42.7	41.9
Conj	62.6	**65.4**	58.1	56.9	**48.7**	30.4	41.4	38.7
Snopes Stance	36.8	44.1	45.7	**58.4**	**20.1**	9.9	18.1	20.1
SciFACT	60.7	62.5	56.2	**70.0**	25.7	17.6	22.5	**25.8**
Climate FEVER	46.9	47.5	42.4	**51.3**	**20.9**	12.8	18.4	20.8
VitaminC	55.8	**58.8**	55.3	56.5	**40.3**	29.8	39.2	37.2
COVID-Fact	63.3	65.9	55.3	**69.8**	**28.1**	12.2	19.8	23.5
FM2	70.2	71.7	76.0	**79.3**	38.4	24.1	**39.0**	38.1
FactCC	56.4	59.6	56.0	**65.2**	16.8	**27.6**	19.1	24.6
QAGS CNN	51.8	59.3	60.0	**72.5**	20.2	**26.4**	19.0	25.8
QAGS XSUM	55.0	59.3	61.4	**72.6**	**24.0**	15.9	19.0	23.0
XSUM H.	47.9	50.4	55.8	**56.9**	17.3	11.6	**17.6**	15.1
Avg NLI	58.4	55.7	55.0	**61.0**	**44.4**	18.5	38.6	36.6
Avg FC	55.6	58.4	55.2	**64.2**	**28.9**	17.7	26.2	27.6
Avg HDAS	52.8	57.1	58.3	**66.8**	19.6	**22.4**	17.9	22.1
Avg All	56.3	56.9	55.7	**63.2**	**34.3**	19.1	30.3	30.7

Table 4.6: Macro F1 and Acceptability Scores on each OOD Dataset on the best models ($\mathrm{O}^{Fev}_{128,AFk}$ and T^{Sn}_{Full}) and the different source dataset counterpart (T^{Fev}_{Full} and $\mathrm{O}^{Sn}_{128,AFk}$). The best score is bold, and second-best is underlined.

4.7 Discussions

This section explains the reasons for our earlier findings. First, we discuss how fine-tuning data and the model affect label prediction and explanation generation. Then, we analyze the relationship between label prediction performance and Acceptability score across the three OOD tasks.

4.7.1 Impact of fine-tuning dataset and base model on OOD label prediction

Source dataset Generally, OOD label prediction performance is better with models fine-tuned on the e-FEVER dataset. To explore the reasons, we show the F1 score per class for both ID and OOD test datasets (including cross-source and 9 OOD three-label datasets) in Table 4.7, based on $\mathrm{O}^{Sn}_{128,AFk}$ and $\mathrm{O}^{Fev}_{128,AFk}$ models. $\mathrm{O}^{Sn}_{128,AFk}$ (e-SNLI) model has a better ID performance (0.86) but generalizes poorly to OOD (0.54), whereas $\mathrm{O}^{Fev}_{128,AFk}$ (e-FEVER) model has a worse ID (0.69) but better OOD performance (0.59). For both source datasets, models perform better on e-SNLI test set than e-FEVER test set, indicating that e-FEVER is a harder dataset to learn. In addition, fine-tuning on e-FEVER helped improving performance on harder classes ("Neural (NEI)") and "Entailment (Sup-

ports)".

Source	Test Set	E.	N.	C.	A.
e-SNLI	ID (Sn)	**86.56**	**79.62**	**91.76**	**85.98**
	OOD (Fev)	78.17	38.65	68.82	61.88
	OOD (9)	59.26	49.56	51.97	53.60
e-FEVER	ID (Fev)	83.22	48.07	76.39	69.23
	OOD (Sn)	**89.04**	**78.18**	**86.63**	**84.61**
	OOD (9)	69.17	56.64	52.12	59.31

Table 4.7: F1 score performance on different test sets, contrasting the two source datasets. E.: entailment, N.: neutral, C.: contradiction, A.: average F1 score. Fev: e-FEVER, Sn: e-SNLI.

Base model We observed that T5 models' OOD label prediction performances are much more stable than OLMo. We believe it is due to two reasons: (1) T5 was fine-tuned for the supervised text-to-text language modeling objective [25] including NLI datasets, and FC and HDAS are relatively similar tasks. Since we formatted the claims/summaries and evidence/documents as hypothesis/premise pairs, T5 can perform relatively well with very few shots. On the downside, the model did not improve with more fine-tuning data (especially with e-SNLI). In contrast, although OLMo models started with low performance, they eventually outperformed T5 with increased number fine-tuning samples. (2) The prompt for fine-tuning T5 matches the one used during its original supervised fine-tuning on NLI datasets, so T5 models do not need to adapt to the format for predicting NLI labels. In contrast, OLMo models perform poorly with few samples due to output formatting issues (expected in JSON format with specific keys for labels and explanations).

4.7.2 Impact of fine-tuning data on OOD explanation quality

Source Dataset We observed that models fine-tuned on e-SNLI generally have higher OOD Acceptability scores (when having similar F1 scores). To understand the effect of fine-tuning data on OOD explanations, Table 4.8 compares the two source datasets based on input length (hypothesis, premise, and explanations), average Acceptability scores of the original data (128 shots), and Acceptability and F1 scores for ID and OOD test sets. The results, based on $O^{Sn}_{128,AFk}$ and $O^{Fev}_{128,AFk}$, show that the input length has a large impact on the ID Acceptability score, but the impact on OOD is minor (as it should depend on OOD input length). Despite lower OOD F1 scores, $O^{Sn}_{128,AFk}$ (e-SNLI) model has similar OOD Acceptability scores to $O^{Fev}_{128,AFk}$ (e-FEVER) model. This could be because part of the SNLI dataset was used to train the Acceptability model. Nevertheless, Acceptability score is more impacted by models' label prediction performance, as reflected by the F1 Scores.

Source	Input Length	Source Accept.	ID Accept.	OOD Accept.	ID F1	OOD F1
e-SNLI	38	**0.671**	**0.565**	0.262	**82.8**	54.3
e-FEVER	118	0.394	0.367	**0.263**	58.9	**59.9**

Table 4.8: Performance comparison across the two source datasets.

Data Filtering Our acceptability-based (T5-Large) filtering model had only slight impacts on label prediction but improved explanation quality, according to the Acceptability score. One hypothesis is that since the Acceptability score metric (T5-11b) is a larger version of the filter model (only differing in size), the metric may favor explanations generated from models fine-tuned on acceptability-filtered samples. To investigate this, we conducted an experiment using the Themis metric as the filter for selecting samples (called "Themis-FastVote-k"), filtering out samples with ratings below 3 (on a 1-5 scale). The experiment is based on the OLMo best model ($O^{Fev}_{128,AFk}$), and the results are shown in Table 4.9. The Acceptability score with "Themis-FastVote-k"(0.303) is similar to "accept-FastVote-k"(0.307), despite having a lower F1 score. This suggests that using the acceptability filter does not cause the Acceptability metric to overestimate explanations generated from the filtered data.

Selection	**Accept.**	**Themis**	**F1**
Themis-FastVote-k	0.303	3.027	58.24
accept-FastVote-k	0.307	2.774	63.24

Table 4.9: Evaluation results using Themis as a filter and as Acceptability a metric (T5-11B), compared to using acceptability as a filter (T5-Large) and Themis as a metric.

4.7.3 Relationship between label prediction performance and Acceptability score

In Figure 4.3, we observed a positive correlation between F1 and Acceptability scores across models. We analyze on the best e-SNLI and e-FEVER models to further explore the relationship between label prediction performance and the Acceptability score within a model. We calculated the average balanced accuracy (used instead of F1 to account for varying class counts across datasets) for each task within different Acceptability score ranges, shown in Figure 4.4. Among the three tasks, most HDAS samples have Acceptability scores below 0.3, while FC and NLI samples are distributed more evenly, indicating lower explanation quality in HDAS. When comparing source datasets, the e-SNLI model shows a steeper accuracy curve, suggesting that lower Acceptability scores often correspond to incorrect predictions of the model. In both models, the Acceptability score is positively linked to label prediction performance, especially in the lower score ranges (below 0.6).

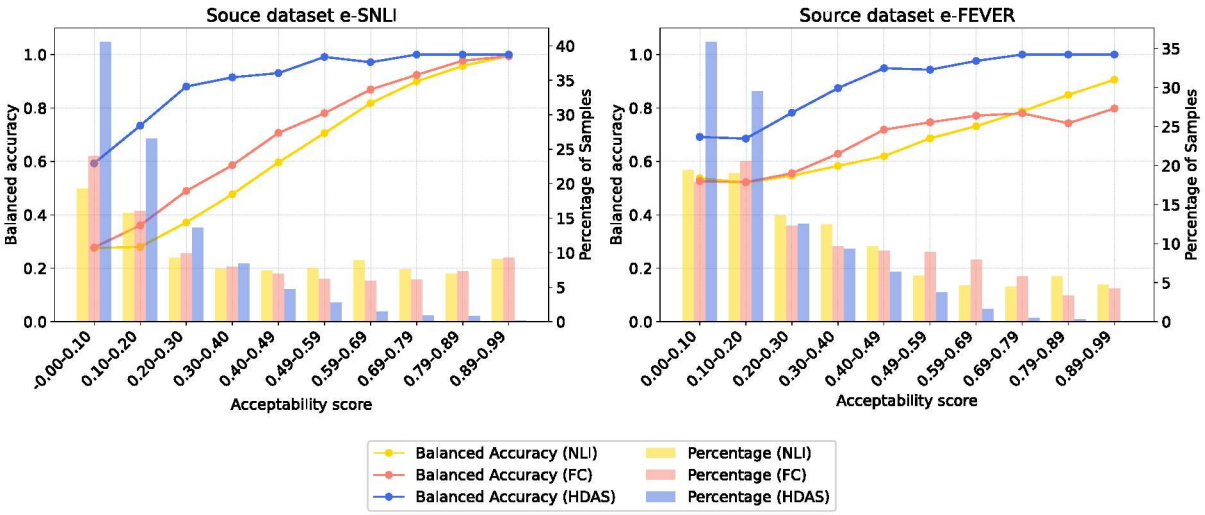


Figure 4.4: Distribution of label prediction accuracy (balanced) across different Acceptability score ranges. The left y-axis shows the balanced accuracy of samples from that Acceptability score range, and the right y-axis shows the percentage of samples in that range.

4.8 Final Remarks

This chapter investigated self-rationalization models' ability to generalize to NLI-related OOD tasks through the evaluation on 19 diverse datasets. We achieve this by fine-tuning T5-large and OLMo-7B under different configurations (varying fine-tuning dataset source, size, and instance selection strategies) to study the impact of data size and quality on OOD task performance and explanation quality. We also examined the Acceptability score as a reference-free metric for the generated explanation evaluation through a human evaluation. Through the study, we gained some important insights: i) fine-tuning a model on few-shot examples can perform surprisingly well in OOD datasets compared to fine-tuning on a large full-size dataset; ii) fine-tuning data source, compared to sample selection, has a larger impact on OOD performance; iii) Acceptability score is positively related to models label prediction performance.

Chapter 5

Label-Adaptive Self-Rationalization for Fact Verification and Explanation Generation

Self-rationalization, whereby models are trained to produce predictions and natural language explanations jointly, is a mainstream explainable approach for Natural Language Inference (NLI) tasks. In the previous chapter, we investigated how learning from existing annotated datasets generalizes to NLI-related OOD datasets, including six fact-checking ones. Our results showed that self-rationalization in its typical formulation is conditional to the target dataset labels being part of the language model pre/training [52, 99].

As an example, consider Figure 5.1. It depicts different methods performances on a recently released fact-checking dataset AVeriTec [30]. This dataset comprises four labels, besides the typical 3-class label (Support, Not Enougn Info (NEI), Refute), it includes a new one for "Conflict (Conflicting Evidence)". When performing zero-shot on the `T5-3B` (green bars), a model pre-trained with NLI datasets (fact verification is often considered similar to NLI) shows reasonable results on the "Support" and "Refute" classes but performs poorly on "NEI", and completely fails on the new "Conflict" class. Self-rationalization fine-tuned on `T5-3B`, depicted by the blue bars in Figure reffig:motivation, fails to learn the new class, resulting in low veracity prediction performance.

This problem is significant because most fact-checking datasets (e.g., FEVER [11]) usually label claim veracity with three classes: SUPPORT, REFUTE, and NEI (not enough information), which is comparable to NLI labels (entailment, contradiction, and neutral). However, many real-world fact-checking datasets usually have different labeling schemes with the number of classes varying from 2-27 classes [12] in some cases. As the labeling scheme shifts from NLI tasks, directly applying self-rationalization with models pre-trained on NLI datasets performs poorly for fact checking.

In this context, we propose a label-adaptive self-rationalization approach to tackle the challenge of the labeling shift for fact verification/checking. We first fine-tune a pre-trained model to learn the classification task with different labels; then, we fine-tune it again with labels and explanations to learn the self-rationalization task (explanations). Our results show that the 2-step formulation significantly outperforms direct self-rationalization learning by more than 20 percentage points (on the AVeriTec dataset)

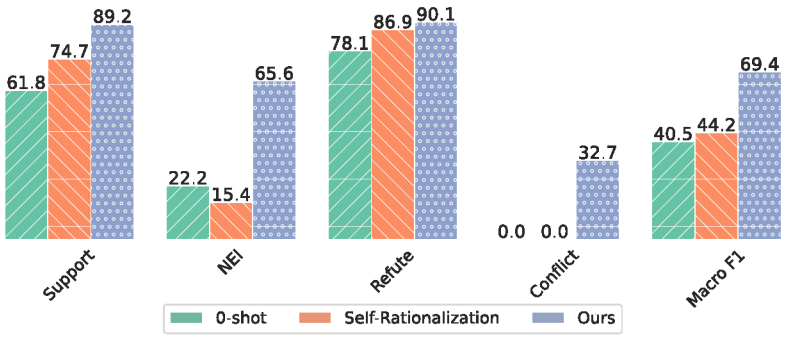


Figure 5.1: Models' performance on the AVeriTec dataset for each class (F1 score). 0-shot: zero-shot performance on `T5-3B`; Self-Rationalization: fine-tuned `T5-3B` model on joint labels and explanations. Ours: Label-adaptive Self-rationalization.

(Figure 5.1). This approach also achieves the best results compared to state-of-the-art methods.

In summary, our contributions herein are twofold:

- We propose a 2-step self-rationalization approach custom-tailored to the fact-checking domain;
- We propose to generate few-shot synthetic explanations by LLMs for step-2 self-rationalization, in case of lacking annotated explanations. In this case, the model's performance is comparable with the entire dataset.

This work is accepted and presented at the International Workshop on Information Forensics and Security (WIFS 2024) : © 2024 IEEE. Reprinted, with permission, from "Jing Yang and Anderson Rocha. Take It Easy: Label-Adaptive Self-Rationalization for Fact Verification and Explanation Generation. IEEE International Workshop on Information Forensics and Security (WIFS), 2024."

5.1 Related Work

This section presents available explainable fact-checking datasets in the literature and the most principled methods proposed to deal with this problem thus far.

5.1.1 Explainable Fact-checking Datasets

Explainability has been an important research front in fact-checking; however, only a few datasets are constructed for this task. LIAR-PLUS [100] was the first dataset by extending the LIAR [101] dataset with extracted justifications from PolitiFact fact-checking articles. Kotonya et al. [29] constructed a large dataset called PubHealth with claims about health topics collected from various fact-checking websites. e-FEVER [80] was a dataset based on FEVER [11], with synthetic explanations generated by `GPT-3`. A more recent dataset, AVeriTec, was released by Schlichtkrull et al. [30], in which the claims were also extracted from real-world fact-checking websites. Unlike previous explanation datasets, in which the

explanations are summarized versions of the evidence, AVeriTec justifications are human-written explanations that reason over the retrieved evidence in the form of questions and answers.

5.1.2 Explainable Fact-checking Methods

Summarization for explanation generation has been a popular approach, as most explanation datasets have their annotated explanations in the form of summarized evidence. Atanasova et al. [31] first proposed an extractive summarization approach based on the LIAR-PLUS datasets. They generate fact-checking explanations by selecting important sentences from the original fact-checking ruling comments. Kotonya et al. [29] used a joint extractive-abstractive summarization approach to generate human-understandable explanations based on their PubHealth dataset. Russo et al. [32] benchmarked extractive and abstractive approaches and showed that performing an extractive approach before abstractive yielded the best result. The problem with the summarization approach is that the summaries cannot build clear connections between the claim and evidence to draw a conclusion.

Another approach to generating explanations is through prompting of large language models (LLMs). Zhang et al. [102] proposed a prompting method (HiSS) to generate intermediate reasoning steps and a final prediction using `GPT-3.5`. The reasoning steps are composed of decomposed sub-claims followed by questions and answers related to each sub-claim. Zarharan et a. [85] tested zero-/few shot abilities of LLMs on the PubHealth dataset, and they showed that parameter-efficient fine-tuning on the `Mixtral-7B` outperformed `GPT-4` model. The main issue with using LLMs is that their pre-training data are not transparent, which can cause data contamination, i.e., the test dataset might have been seen during their pre-training, causing unreliable performances.

5.2 Methodology

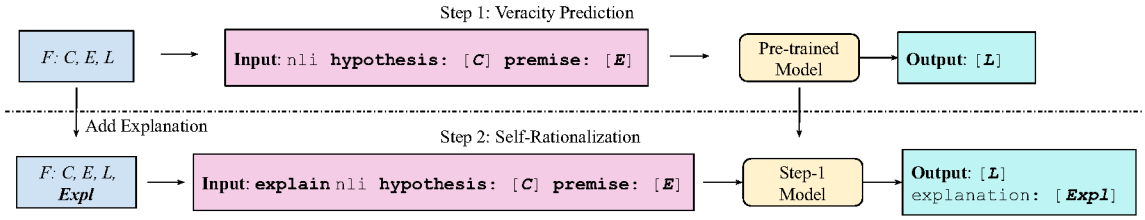


Figure 5.2: Label-adaptive self-rationalization 2-step pipeline. In step-1, the model learns veracity prediction with only provided labels; in Step-2, the model learns the self-rationalization task with both labels and explanations.

Provided with labels and explanations, directly fine-tuning for self-rationalization fails on newly added labels (as shown in Figure 5.1), thus we take a step-by-step approach to slowly adapt the model for the new domain and class. Our method is based on `T5-3B` model, as its size is comparable to many open large language models, and self-rationalization has been shown to perform well on T5 models [52, 103, 99].

5.2.1 Label-Adaptive Self-rationalization Learning

Our proposed approach is illustrated in Figure 5.2. It comprises two steps: in Step-1, the model learns to adapt to the new class with only provided labels; in Step-2, the model learns the self-rationalization task with labels and added explanations. We describe the details as follows:

Given a dataset $D = (C, E, L, Expl)$, with each sample $s_i = \{c_i, e_i, l_i, expl_i\}$, $c_i, e_i, l_i, expl_i$ represent a claim, evidence, label, and explanation, respectively, we perform two steps.

Step-1: Label Adaptation. We first adapt and fine-tune the `T5` model to generate the veracity label l_i. Given the input $x_i = \{c_i, e_i\}$, we follow the same standard prompt template that was used to pre-train `T5` for the NLI task ("claims" and "evidence" are mapped to "hypothesis" and "premise" as `T5` is more familiar with these words), as shown in first row of Figure 5.2.

Step-2: Self-Rationalization. After fine-tuning the model with the veracity prediction task, we now add gold explanation $expl_i$ to fine-tune the resulting `T5` model again after Step-1. Shown in second row of Figure 5.2, we change the encoder prompt to add the word "*explain*", and for the decoder prompt, a separation word "*explanation*", inspired by [99].

To simulate a realistic scenario with limited annotated explanations, we employ large language models (LLMs) to generate few-shot synthetic explanations. Specifically, we evaluate this task using `GPT-3.5-turbo-0125`, `GPT-4-turbo`, and `Llama-3-8B-Instruct`. We use the same prompt for generating the explanations with the three models, as shown below:

> ***System**: You are a fact-checking assistant. You should not simply repeat the claim or evidence, your answer should be concise and short.*
> ***User**: Given the evidence {evidence}, and claim {claim}. Please explain why the claim is {ground truth label}.*

5.2.2 Data Processing and Label Mapping

We perform experiments on two datasets with explanation annotations: AVeriTeC [30] and PubHealth [29]. We adopt these datasets as they better represent real-world fact-checking scenarios with 4-class annotations.

AVeriTeC: The dataset comprises claims from 50 fact-checking organizations. It is unique in the way that the evidence in AVeriTeC is composed of questions and answers extracted from retrieval of online websites. To facilitate training, we concatenate the questions and answers as follows. Given a piece of evidence $e = \{q_1(a_1, a_2, \cdots, a_i), \cdots, q_k(a_1, a_2, \cdots, a_j)\}$, we format it as "Question 1 : q_1 Answer 1 : $\{a_1\ a_2 \cdots a_i\} \cdots$ Question k : q_k Answer k : $\{a_1\ a_2 \cdots a_j\}$". The justifications are human-annotated to reason over a claim's given questions and answers.

PubHealth: The datasets contain claims from the health (biomedical) domain that are extracted from fact-checking and news review websites. The evidence consists of the full text from fact-checking articles or news reviews, with an average length exceeding 600 words, significantly longer than AVeriTeC's average of 120 words. Explanations for claim veracity are provided through fact-checking justifications or news summaries.

AVeriTeC	PubHealth	`GPT/Llama`	`T5`
Supported	TRUE	true	entailment
Not enough evidence	UNPROVEN	not enough information	neutral
Refuted	FALSE	false	contradiction
Conflicting evidence /cherry-picking	MIXTURE	partially true and false	mixture

Table 5.1: Label mapping scheme.

Dataset	AVeriTec (Train / Dev)	PubHealth (Train / Dev / Test)
entailment	848 / 122	5,078 / 629 / 599
neutral	282 / 35	291 / 41 / 45
contradiction	1,742 / 305	3,001 / 380 / 388
mixture	195 / 38	1,434 / 164 / 201
#words C	17 / 17	14 / 13 / 14
#words E	113 / 122	714 / 708 / 718

Table 5.2: Dataset details by each class.

We map the textual labels for different models as shown in Table 5.1. Specifically for `T5`, we align the labels with the NLI task naming scheme used during pre-training. For the "*Conflicting evidence*" label in AVeriTeC, we equate it to the "*MIXTURE*" class in PubHealth, which is "*partially true and false*" for `GPT/Llama` models.

The data statistics for each dataset are shown in Table 5.2; we removed instances that contain empty claims. Both datasets have very imbalanced classes, with less data with "*NEI (not enough evidence)*" and "*mixture*" classes.

5.3 Experimental setup

This section describes the implementation details, evaluation metrics, and baselines used in our experiments.

5.3.1 Implementation Details

In each fine-tuning experiment, we select the best model from the last epoch without using a validation set. For AVeriTec, we use a batch size of 4 and a max input length of 512. For PubHealth, due to the length of the evidence, we use a batch size of 2 and a max input length of 1024. All experiments are based on NVIDIA A100 GPUs. For `GPT-4` zero-shot baseline, we set the temperature to be 0.7, with a max output length of 200.

5.3.2 Evaluation Metrics

To evaluate the veracity prediction and explanation quality, we first extract the label and explanation from the generated text using the separator "`explanation: `". For

veracity prediction, we assess performance based on accuracy and macro F1 score. For explanations, we use both reference-based metrics (ROUGE scores and METEOR) and reference-free metrics. The latter is crucial in realistic scenarios where the test dataset lacks reference explanations for comparison. Specifically, we use the following reference-free metrics:

- Auto-J [104]: The metric is a model based on `LLaMA-2-13B-chat` by fine-tuning on judgments of LLM-generated responses with diverse user queries. It supports both single and pair-wise evaluations. We use it for single reference-free evaluations. The evaluation output comprises textual analysis and an overall quality rating between 1-10.

- TigerScore [105]: Another trained model-based metric that provides explainable evaluations for text generation tasks by following instructions. It outputs an overall error score ranging from 0 to -infinity, along with a textual analysis detailing the location and type of each detected error. We use the `TIGERScore-13B` model in our evaluation.

For the reference-free metrics, the input must be formatted using instruction-based prompts. Our instructions are similar to those used for generating synthetic explanations with LLMs. We evaluate the explanations based on ground truth labels.

5.3.3 Baselines

We compare our two-step approach (denoted as `2-R`, with `R` denoting Rationalization) with the following baselines:

1. `0-L`: zero-shot `T5-3B` baseline. As NLI datasets were used for `T5` pre-training, we formatted veracity prediction as an NLI task and prompted `T5-3B` to generate predictions. `L` denotes Label prediction.

2. `1-R`: Compared to `2-R`, this baseline model is directly fine-tuned with labels and explanations without first fine-tuning for the veracity prediction task.

3. `1-L`: veracity prediction model fine-tuned with labels only (Step-1 model). The model cannot generate explanations, thus is not included for explanation comparison.

4. Baseline approach by Schlichtkrull [30]. They have separate models for predicting veracity and generating explanations on the AVeriTec dataset, with the best results obtained with `BERT-Large` and `BART-Large`.

5. Baseline approach by Kotonya [29]. They also have separate models for the two tasks; on the PubHealth dataset, the best results are based on `SCIBERT` and `BERT` models.

6. Zarharan et al. [85]: They studied different LLMs' performance on the PubHealth dataset. All their models are based on summarized evidence to reduce the evidence length, using `GPT-3.5-turbo` for the summarization. The best results were achieved with parameter-efficient-fine-tuning (PEFT) on the `Mixtral-7B` model.

7. `GPT-4`. We conduct zero-shot prompting on `GPT-4-turbo` for the AVeriTec dataset. As reported in [85], `GPT-4`'s performance on the PubHealth dataset is directly reported in our work. We prompt the model to generate the output in JSON format to obtain the predicted veracity label and explanation, as illustrated below.

***System**: You are a helpful assistant designed to output JSON, formatted as "answer":, "reason:".*
***User**: Based on the evidence, determine if the claim is true, false, not enough information to confirm, or partially true and false.*
Evidence: [evidence]
Claim: [claim]
Options: - true - not enough information - false - partially true and false
Please provide your reason.

We directly refer to the numbers reported in the respective paper for baseline results. For explanation evaluation, Zarharan et al. [85] made their results publicly available, so we ran all evaluation metrics based on their released explanations.

5.4 Results and Discussions

We present results on veracity prediction and explanation generation in comparison with baselines; and the results of fine-tuning on few-shot synthetic LLM-explanations.

5.4.1 Veracity Prediction Performance

Table 5.3 shows the veracity prediction results on different baseline models and our `2-R` model. As expected, `0-L` (zero-shot on `T5-3B`) cannot predict the class "*mixture*" for either dataset. For AVeriTeC, our `2-R` model is comparable with `GPT-4`, with the best accuracy of 85.2%, while being a much smaller model. For PubHealth, the `1-L` model achieved the best performance, while `2-R` model slightly dropped (2%) on Macro F1 after learning to generate explanations. Both outperform the larger baseline models (`Mixtral-7B` and `GPT-4`). For both datasets, the `2-R` model improved performance (Macro F1) by more than 10 percentage points compared with the `1-R` model, showing that letting models learn the veracity task first greatly helps the model to adapt to the new domain with new classes. Specifically, the `1-R` model struggled with predicting classes "*neutral*" and "*mixture*", but with our label-adaptive approach (`2-R`), the model was able to improve predictions on these classes significantly.

	Model	S	N	R	M	F1	Acc.
AVeriTeC	BERT-Large[30]	48.0	59.0	74.0	15.0	49.0	49.0
	GPT-4	83.5	**65.9**	**91.5**	**45.5**	**71.6**	83.0
	0-L	64.8	22.2	78.1	0.0	40.5	62.0
	1-R	74.7	15.4	86.9	0.0	44.2	76.2
	1-L	87.5	59.0	89.3	29.5	66.3	83.4
	2-R	**89.2**	65.6	90.1	32.7	69.4	**85.2**
PubHealth	SCIBERT[29]	-	-	-	-	70.5	69.7
	Mistral-7B[85]	92.7	48.6	82.1	57.1	70.1	81.8
	GPT-4[85]	80.6	18.2	73.0	42.0	53.4	69.6
	0-L	65.0	2.8	42.9	0.0	27.7	48.7
	1-R	91.2	26.9	79.8	38.9	59.2	79.1
	1-L	**93.4**	**60.5**	**84.2**	**58.2**	**74.1**	**83.7**
	2-R	93.2	57.5	83.4	55.1	72.3	83.1

Table 5.3: Performance comparison on veracity prediction

5.4.2 Generated Explanation Quality

We show the evaluation of generated explanation quality in Table 5.4. For both datasets, GPT-4 generated explanations have the best scores on the reference-free metrics, indicating the reasoning abilities of GPT-4, although it has a tendency to be verbose (having the longest explanations on average). Our 2-R approach has the highest ROUGE scores, outperforming the baselines. For the AVeriTec dataset, the 2-R model generates better explanations than the 1-R model, as agreed by all metrics. For the PubHealth dataset, the scores for the two models are very similar, and both have the highest ROUGEs and METEOR scores. In general, the results show that fine-tuned models generate explanations that are better aligned with reference explanations, as the training data follow a similar pattern.

Overall, our 2-R approach achieves the highest veracity prediction performance and the best reference-based scores for explanations, outperforming LLMs and other state-of-the-art baselines.

5.4.3 Results from Synthetic Few-shot Explanations

To demonstrate the potential of our two-step approach in data-scarce scenarios, we test Step-2 with few-shot fine-tuning. We select 16 samples per class (64 samples total) to prompt an LLM to generate synthetic explanations. These samples and their generated explanations are then used to fine-tune the 1-L model. For robust results, we select few-shot samples with three different random seeds and report the results in average and standard deviation. The results for veracity prediction and explanation generation are shown in Table 5.5 and 5.6.

The veracity prediction results show that Step-2 with very few amount of data still achieve much better performance than end-to-end self-rationalization model (1-R), and perform comparably to the 2-R with full dataset fine-tuning. In terms of explana-

	Model	AJ↑	Tiger↓	ROUGEs	METEOR	#W
AVeriTeC	BART-Large[30]	-	-	-	**.28**	-
	GPT-4	**4.99**	**0.64**	25 / 9 / 19	.31	60
	1-R	3.45	2.06	27 / 10 /23	.24	18
	2-R (Ours)	3.61	1.87	**29 / 12 / 25**	.26	18
	Reference	3.54	1.48	-	-	22
PubHealth	BERT[29]	-	-	32 / 13 / 27	-	-
	Mistral-7B[85]	3.99	1.88	36 / 15 / 26	.29	73
	GPT-4[85]	**4.80**	**0.53**	26 / 8 / 17	.24	75
	1-R	3.63	2.34	43 / 24 / 34	**.37**	59
	2-R (Ours)	3.62	2.50	**43 / 24 / 35**	**.37**	59
	Reference	3.70	1.23	-	-	76

Table 5.4: Explanation evaluation with reference-free and reference-based metrics. #W means the average number of words in the explanations. Reference means gold explanation.

tion quality, the reference-free metrics indicate that the best explanations are from the 2-R(GPT-3.5), with a similar Auto-J score compared to the best, and the lowest TigerScore among few-shot models.

Surprisingly, the 2-R(GPT-4) model performs worse than both 2-R(GPT-3.5) and 2-R(Llama-3-8B), in contrast to Table 5.4, where GPT-4 model generated explanations are much better. We hypothesize that when generated text is long (2-R(GPT-4) model explanations are almost twice as long compared with the rest), it is more detailed but also more likely to contain errors.

We show an example of explanations generated by different models from the PubHealth dataset in Figure 5.7). We see that as the explanation becomes longer, models tend to hallucinate and makes more errors. In this sense, GPT-3.5 and Llama-3-8B generated explanations are better for having shorter explanations and thus less likely to make errors. This gap is particularly captured by TigerScore (Table 5.4), which measures the number of errors in the explanations.

5.5 Final Remarks

We proposed an effective two-step approach for joint fact-verification and explanation generation with self-rationalization. Our results show that having a label prediction step significantly helped the model to adapt to new classes and perform better. Our method with T5-3B outperformed larger models, including Mixtral-7B and GPT-4. We further utilized LLMs to generate few-shot synthetic explanations to fine-tune our T5-3B model, and it outperformed end-to-end self-rationalization models fine-tuned on the entire dataset. We also show that T5-3B models struggle with generating longer explanations when learning from GPT-4 explanations.

	Expl. Source	S	N	R	M	F1
AVeriTeC	2-R(GPT-4)	$83.1_{\pm 2.0}$	$52.9_{\pm 3.9}$	$86.4_{\pm 0.7}$	$29.8_{\pm 4.3}$	$63.1_{\pm 1.1}$
	2-R(GPT-3.5)	$86.2_{\pm 2.3}$	$\mathbf{61.0}_{\pm 1.6}$	$85.3_{\pm 2.2}$	$\mathbf{35.1}_{\pm 3.3}$	$\mathbf{66.9}_{\pm 1.0}$
	2-R(Llama-3-8B)	$83.0_{\pm 5.2}$	$58.1_{\pm 5.7}$	$85.9_{\pm 0.4}$	$35.0_{\pm 4.2}$	$65.5_{\pm 2.1}$
	2-R (orig.)	$\mathbf{86.5}_{\pm 2.8}$	$58.3_{\pm 4.1}$	$\mathbf{87.2}_{\pm 0.7}$	$30.6_{\pm 2.6}$	$65.6_{\pm 1.0}$
	1-R (orig., Full)	74.7	15.4	86.9	0.0	44.2
	2-R (orig., Full)	89.2	65.6	90.1	32.7	69.4
PubHealth	2-R(GPT-4)	$86.9_{\pm 1.0}$	$38.6_{\pm 2.0}$	$75.8_{\pm 1.6}$	$54.4_{\pm 1.1}$	$63.9_{\pm 1.2}$
	2-R(GPT-3.5)	$87.5_{\pm 1.5}$	$42.9_{\pm 2.3}$	$76.3_{\pm 1.6}$	$\mathbf{55.5}_{\pm 1.6}$	$65.5_{\pm 1.0}$
	2-R(Llama-3-8B)	$\mathbf{88.9}_{\pm 1.1}$	$46.1_{\pm 3.0}$	$\mathbf{78.6}_{\pm 2.6}$	$54.3_{\pm 2.2}$	$\mathbf{67.0}_{\pm 1.4}$
	2-R (orig.)	$86.1_{\pm 0.5}$	$\mathbf{46.7}_{\pm 2.4}$	$78.1_{\pm 2.3}$	$51.9_{\pm 0.3}$	$65.7_{\pm 1.1}$
	1-R (orig., Full)	91.2	26.9	79.8	38.9	59.2
	2-R (orig., Full)	93.2	57.5	83.4	55.1	72.3

Table 5.5: Veracity prediction results with few-shot Step-2 fine-tuning under different LLM-based synthetic explanations. All models are based on T5-3B. Orig. means original annotated explanations. Full means entire dataset fine-tuning, otherwise few-shot fine-tuning.

	Expl. Source	Auto-J↑	Tiger↓	ROGUE-1 / 2 / L	METEOR	#W.
AVeriTeC	2-R(GPT-4)	$\mathbf{4.51}_{\pm 0.04}$	$4.35_{\pm 0.45}$	$21_{\pm 0.7}$ / $8_{\pm 0.3}$ / $16_{\pm 0.4}$	$.29_{\pm 0.0}$	$84_{\pm 5}$
	2-R(GPT-3.5)	$4.42_{\pm 0.10}$	$\mathbf{2.49}_{\pm 0.25}$	$\mathbf{27}_{\pm 0.2}$ / $10_{\pm 0.0}$ / $\mathbf{20}_{\pm 0.1}$	$\mathbf{.30}_{\pm 0.0}$	$45_{\pm 1}$
	2-R(Llama-3-8B)	$4.31_{\pm 0.13}$	$2.90_{\pm 0.65}$	$27_{\pm 1.2}$ / $\mathbf{11}_{\pm 0.6}$ / $20_{\pm 0.9}$	$.29_{\pm 0.0}$	$45_{\pm 4}$
	2-R (orig.)	$3.38_{\pm 0.06}$	$2.70_{\pm 0.15}$	$25_{\pm 0.7}$ / $8_{\pm 0.4}$ / $20_{\pm 0.5}$	$.22_{\pm 0.0}$	$22_{\pm 3}$
	1-R (orig., Full)	3.45	2.06	27 / 10 /23	.24	18
	2-R (orig., Full)	3.61	1.87	29 / 12 / 25	.26	18
	Reference	3.54	1.48	-	-	22
PubHealth	2-R(GPT-4)	$\mathbf{4.47}_{\pm 0.03}$	$4.27_{\pm 0.30}$	$24_{\pm 0.3}$ / $7_{\pm 0.0}$ / $16_{\pm 0.1}$	$\mathbf{.25}_{\pm 0.0}$	$119_{\pm 3}$
	2-R(GPT-3.5)	$4.39_{\pm 0.01}$	$\mathbf{2.34}_{\pm 0.23}$	$27_{\pm 0.3}$ / $9_{\pm 0.2}$ / $18_{\pm 0.2}$	$.24_{\pm 0.0}$	$68_{\pm 5}$
	2-R(Llama-3-8B)	$4.29_{\pm 0.03}$	$2.89_{\pm 0.11}$	$27_{\pm 0.5}$ / $8_{\pm 0.2}$ / $18_{\pm 0.3}$	$.23_{\pm 0.0}$	$58_{\pm 2}$
	2-R (orig.)	$3.65_{\pm 0.03}$	$2.84_{\pm 0.26}$	$\mathbf{31}_{\pm 0.6}$ / $\mathbf{12}_{\pm 1.0}$ / $\mathbf{22}_{\pm 0.9}$	$\mathbf{.25}_{\pm 0.0}$	$56_{\pm 5}$
	1-R (orig., Full)	3.63	2.34	43 / 24 / 34	**.37**	59
	2-R (orig., Full)	3.62	2.50	43 / 24 / 35	.37	59
	Reference	3.70	1.23	-	-	76

Table 5.6: Explanation evaluation results with Step-2 few-shot fine-tuning under different LLM-based synthetic explanations. All models are based on T5-3B.

Claim (Mixture)
The Westmoreland News ran a KKK ad on the front page of their newspaper in April 2018.

Evidence (unimportant parts are omitted due to its length)
On 25 April 2018, the front page story of Virginia's Westmoreland News described the disturbing occurrence of KKK recruitment flyers in front home homes in Colonial Beach, VA [...] In a move that has generated significant online controversy, the Westmoreland News reproduced a full, un-redacted version of that ad in their story, leading some to brand it has free advertising for the KKK. As reported by Newsweek and local CBS affiliate WTVR, many readers of the paper were outraged: They posted the whole flyer as if they were giving them free advertising," Colonial Beach resident Betty Tate Thompson told WTVR. "It even had the number on flyer to contact the head of the KKK. Which I thought was totally ridiculous." [...] Plott said the editor called him back with an explanation. "She explained that she wanted to put it in to show people that those of us in Colonial Beach who may think there's not racism; there is and I think she didn't expect for this to blow up the way it did and with this outcome," explained Plott. While the ad was reproduced in full, The Westmoreland News did post a disclaimer on the story with this text: Westmoreland News in no way condones or supports the content or message of this flyer, nor does it condone or support any branch of the Loyal White Knights, or the KKK. Decisions on how to report the occurrence of KKK flyers in the media have not been made consistently in the past. [...] In each of those cases, it should be noted, the flyer was reproduced as an obvious photograph of a paper document. The front page image used by the Westmoreland Times, however, could be confused for a paid advertisement. Because that was not the case here, however, we rank the claim that the newspaper "ran an ad" for the KKK as a mixture.

`GPT-4`
The context refers to an occurrence in April 2018 where the Westmoreland News reproduced a KKK recruitment ad on its front page. The controversy discussed was about the decision to run this ad, which was perceived as free advertising for the KKK. Therefore, the claim is **supported** by the context.

`1-R` **(orig., Full)**
What's true: The Westmoreland News reproduced a full, un-redacted version of a KKK recruitment flyer in their front page story. What's false: The newspaper did not pay for the ad.

`2-R` **(orig., Full)**
What's true: The Westmoreland News reproduced a full, un-redacted version of a KKK recruitment flyer in their front page story. What's false: The newspaper did not run an ad for the KKK.

`2-R(GPT-4)`
The claim that the Westmoreland News ran a KKK ad on the front page of their newspaper in April 2018 is **mixture** because the evidence provided does not conclusively establish that the newspaper provided free advertising for the KKK. While the publication did reproduce a full, un-redacted version of the KKK ad in their story, the inclusion of the contact information and phone numbers on the flyer raised concerns among readers and sparked outrage on social media. Additionally, the editorial decision to run the ad in its entirety was made **after** significant discussion and debate amongst community members and law enforcement, who objected to the inclusion of such explicit recruitment materials in a news story. Therefore, the claim cannot be fully confirmed as it is made with certainty.

`2-R(GPT-3.5)`
The claim that The Westmoreland News ran a KKK ad on the front page of their newspaper in April 2018 is **partially true** because the newspaper included a full, un-redacted version of the ad in their report on the occurrence of KKK recruitment flyers in Colonial Beach, VA. The claim is **false** because the newspaper did not explicitly state that it was running an ad for the KKK, and the claim is also **false** because the editor provided a disclaimer on the article.

`2-R(Llama-3-8B)`
The claim is **partially true** because the Westmoreland News did reproduce a full, un-redacted version of the KKK ad in their story, which generated significant online controversy. While the newspaper did not explicitly endorse the KKK, it did state that it did not condone or support the group's message or message of racism. Additionally, the claim is **partially false** because the newspaper did not provide sufficient evidence to prove the claim.

Reference (Gold)
What's true: The Westmoreland News did publish a copy of a KKK flyer in full on their front page as part of their reporting on KKK flyers being distributed in the region. What's false: The document was not run as an ad in the paper, but was intended to be an illustration for their coverage of the KKK.

Table 5.7: An example of generated explanations from different models on PubHealth dataset. In the evidence and explanations, the accurate and relevant text is highlighted in color blue, while the inaccurate or hallucinated text is highlighted in red. Best viewed in color.

Chapter 6

Conclusions and Future Work

Misinformation has long been a persistent societal challenge, but the digital age has significantly amplified its scale and speed. The rise of Artificial Intelligence, particularly in language models, has further complicated the issue. These models, now capable of generating highly persuasive and realistic false content at extraordinary speed, make it increasingly difficult to distinguish truth from deception. However, the same advancements driving these challenges also present opportunities. Over the past few years, language models have grown remarkably in scale and capability, achieving impressive zero-shot performance. This progress has shifted our research efforts from designing smaller models to fine-tuning and prompting larger ones, unlocking their potential as tools to combat misinformation. By leveraging these models responsibly, we can develop innovative approaches to identify and mitigate the spread of false information while promoting truth and transparency.

Fact-checking is a crucial process in journalism for combating disinformation. While fully automated fact-checking solutions are not yet ready for real-world implementation, they can play a significant role in supporting human fact-checkers by enhancing the speed and thoroughness of the verification process. This collaborative approach is essential for improving the efficiency and effectiveness of combating misinformation. In this thesis, we focused on two goals: fact-checking efficiency and explainability.

For fact-checking efficiency, we proposed speeding up fact-checking process by grouping similar text messages together and summarizing them into one claim to reduce redundancy. Additionally, we explored few-shot learning techniques to minimize the amount of human-annotated training data required. Our research examined various numbers of fine-tuning samples, sampling methods, and the use of synthetic few-shot data for model fine-tuning.

For explainability, we explored using QA and self-rationalization methods for generating explainable fact-checking outputs. Our method was the first one proposed to address explainable fact-checking using QA. With self-rationalization, we performed a large-scale evaluation on 19 datasets containing diverse synthetic and real-world claims. Our thorough evaluation considered four different evaluation metrics and human evaluation with 468 crowd-workers.

In the remaining of the chapter, we revisit each research questions posed in Chapter 1 and present the insights we gained throughout our research journey. Finally, we discuss

the limitations of our research and potential future directions.

6.1 Revisiting the Research Questions

In advancing automated fact-checking, this thesis focused on two primary goals, guided by the research questions introduced in Chapter 1, Section 1.2. In light of the challenges, proposed solutions, results, and findings, we now revisit each question to demonstrate how they have been addressed.

6.1.1 RQ1: Given a large amount of raw text data, how can we speed up the fact-checking process to reduce fact-checkers' workload?

To address this research question, Chapter 2 focused on finding claims from raw data from social media (X.com) by grouping similar messages to summarize them into claims. Specifically, we first clean a set of social media posts (e.g., tweets) and build a graph of all posts based on their semantics. Then, we perform two clustering methods to group the messages for further claim summarization. Our results reduced 28,818 raw messages into 700 summary claims, effectively speeding up the claim discovery process of fact-checking.

6.1.2 RQ2: How do we use data efficiently to use less annotated data for model learning?

Annotating data specifically for writing tasks is challenging as it requires significantly more time and supervision than classification tasks. To address this issue, works usually need to use a small amount of annotated data for learning. In Chapters 4 and 5, we explored few-shot learning for generating fact-checking explanations, with various numbers of fine-tuning samples and methods for selecting high-quality samples. Our results (Chapter 4) showed that with up to 128 shots, models have comparable performances to a full training set (e.g., with half a million samples in the e-SNLI source dataset). We also explored generating synthetic explanations using LLMs for learning self-rationalization (Chapter 5), and showed that the model learned from the synthetic explanations can outperform the full-shot end-to-end self-rationalization model with human annotated explanations.

6.1.3 RQ3: How can we modify the fact-checking process to make it more transparent to human fact-checkers?

To increase fact-checking explainability, Chapter 3 addresses this by integrating question answering into the fact-checking pipeline. In particular, we proposed to generate questions and answers from claims and extract answers from the same questions from evidence. For answer pairs comparison, we proposed a model with attention mechanism attached to each question. With this, we break down the fact verification into several steps, aiding explainability as it allows more detailed analysis of each individual step.

6.1.4 RQ4: In the absence of annotations, how can we leverage a different dataset to generate explanations for the target dataset?

Our question answering approach for explainable outputs nevertheless has some limitations. When using multiple models for one solution, each model can aggregate the error, making the final results drastically wrong. Additionally, the attention model selected question and answer pairs may not always be correct and relevant. To address the issues with QA, we studied another approach: self-rationalization. In self-rationalization, a model jointly generates the task label and a free-text explanation for the predicted label. This allows one model to perform two tasks: label prediction and explanation generation. The main challenge of self-rationalization is the lack of annotated explanations. To handle this issue, Chapter 4 investigated how learning from existing explanation datasets generalizes to diverse datasets. Our results show that with few annotated explanations, models can effectively perform self-rationalization on some OOD datasets, compared to full dataset fine-tuned models.

6.1.5 RQ5: How do we evaluate generated explanations without any reference data?

Language generation tasks face evaluation challenges as metrics for evaluation usually require reference text for comparison. To make matters worse, traditional reference metrics such as BLEU and ROUGEs have been shown to correlate poorly with humans [73]. Thus, we investigated using the Acceptability score as our reference-free metric (Chapter 4). To further validate the effectiveness of the metric, we performed a human evaluation with 468 crowd-workers and showed that the Acceptability score had the highest correlation with humans, compared to three other state-of-the-art metrics.

6.1.6 RQ6: How does our method work on real-world fact-checking datasets?

In our self-rationalization OOD evaluation, the fact-checking datasets are commonly used three-class datasets. To work on more realistic fact-checking datasets, we extended self-rationalization to fact verification with four-class labels (Chapter 5). In detail, we proposed a two-step label adaptive approach: first, we fine-tuned a model to learn veracity prediction with annotated labels (step-1 model); then, we fine-tuned the step-1 model again to learn self-rationalization. This approach allows the model to adapt to a new domain more effectively than directly fine-tuning end-to-end self-rationalization. Our results show that our label-adaptive approach improves veracity prediction by more than ten percentage points (Macro F1) on both the PubHealth and AVeriTec datasets, outperforming the GPT-4 model. Our label-adaptive self-rationalization approach presents a promising direction for future research on real-world explainable fact-checking with different labeling schemes.

6.2 Limitations and Future Work

Lack of annotated data The scarcity of annotated data has posed significant challenges for both model training and result evaluation. In the initial work on grouping and summarization for claim generation (Chapter 2), there was an absence of reliable ground-truth labels for clustering and oracle summaries for summarization evaluation. The second study (Chapter 3), which focused on generating questions and answers for explainable fact-checking, similarly lacked ground-truth data to assess the effectiveness of the generation methods. To address the shortage of annotated explanations, Chapters 4 and 5 explored the use of a limited set of annotated explanations to learn self-rationalization for explanation generation. Chapter 5 also investigated reference-free evaluation metrics, a necessary step given the high cost and time demands of human evaluation. Future work could delve deeper into learning from synthetically generated data for model fine-tuning. Our findings indicated that many data points in a dataset are highly similar in terms of content and style, allowing for effective learning with less data. With up to 128 examples, we observed performance comparable to or even exceeding that of full-shot models; additional examples may further enhance results, a direction we suggest for future investigation. Moreover, future research could explore unsupervised evaluation metrics based on large language models (LLMs) as an alternative to human evaluation.

Instructions for model fine-tuning/prompting When using a generative model for classification, the naming of labels is an important factor that affects performance, as different models may have their own way of formatting the labels during pretraining. 2) We use the same instructions for different LLM models (Chapter 5), but there may be other instructions that help them generate more accurate explanations. Future work may focus on studying what models/instructions can generate better synthetic explanations for smaller models to learn from.

Reliability of evaluation metrics LLMs can generate fluent and plausible text, but also tend to generate long and creative content. When evaluating these models, current metrics mainly focus on the plausibility of the generated text. However, it is also important to consider other aspects such as consistency, coherence, and factuality. In our study, the highest correlation between reference-free metrics and humans is 0.484 (Chapter 4), which is moderate but not high enough to trust it completely. Future work may focus on designing better evaluation criteria to thoroughly investigate what kind of explanations users prefer or require.

Out-of-Distribution performance In our OOD evaluation study, we found that e-FEVER appeared to be a more challenging source dataset than e-SNLI (Chapter 4), as its model demonstrated worse ID but better OOD performance. Thus, future work may explore fine-tuning on harder tasks for better OOD generalization. Another promising direction is improving the explanations in the source datasets, with better ways of selecting good quality explanations or re-writing better few-shot explanations with LLMs.

Evidence retrieval In all our datasets, we used gold evidence instead of evidence retrieval to focus on fact verification. Future work can add the retrieval step to the pipeline instead of using gold evidence, as the retrieval is also a crucial part of fact-checking.

Datasets limited to English Since our selected datasets are sourced from English-only data, our methods are limited to English. Testing the approach to multilingual models and datasets is also a promising endeavor.

6.3 Research Outcomes

Finally, we list all the papers published, accepted or submitted during this Ph.D. research, in a reverse chronological order.

1. **Jing Yang**, Max Glockner, Anderson Rocha and Iryna Gurevych. Self-Rationalization in the Wild: A Large Scale Out-of-Distribution Evaluation on NLI-related tasks. **Accepted**in Transactions of the Association for Computational Linguistics (TACL), 2024.

2. **Jing Yang** and Anderson Rocha. Take It Easy: Label-Adaptive Self-Rationalization for Fact Verification and Explanation Generation. To appear in *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2024.

3. **Jing Yang**, José Nascimento, Gabriel Bertocco, Antonio Theophilo, Rafael Padilha, Aurea Soriano-Vargas, Fernanda A Andaló and Anderson Rocha. AI Knows What You Did Last Summer: Applications in Digital Forensics. In the Book Chapter of Computer Vision: Challenges, Trends, and Opportunities, pages 82–108, 2024.

4. João Phillipe Cardenuto, **Jing Yang**, Rafael Padilha, Renjie Wan, Daniel Moreira, Haoliang Li, Shiqi Wang, Fernanda Andaló, Sébastien Marcel and Anderson Rocha. The Age of Synthetic Realities: Challenges and Opportunities. *APSIPA Transactions on Signal and Information Processing*, pages 1–62, 2023.

5. José Nascimento*, João Phillipe Cardenuto*, **Jing Yang*** and Anderson Rocha. Few-shot Learning for Multi-modal Social Media Event Filtering. *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2022. * Equal contribution.

6. **Jing Yang**, Didier Vega-Oliveros, Taís Seibt and Anderson Rocha. Explainable Fact-checking through Question Answering. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

7. **Jing Yang**, Didier Vega-Oliveros, Taís Seibt and Anderson Rocha. Scalable Fact-checking with Human-in-the-Loop. *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2021.

8. Rafael Padilha, Antônio Theóphilo, Fernanda A. Andaló, Didier A. Vega-Oliveros, João P. Cardenuto, Gabriel Bertocco, José Nascimento, **Jing Yang** and Anderson Rocha. A Inteligência Artificial e os desafios da Ciência Forense Digital no século XXI. *Estudos Avançados 35*, pages 113-138, 2021.

Bibliography

[1] Simon Kemp. DIGITAL 2024: BRAZIL. https://datareportal.com/reports/digital-2024-brazil, 2024. [Online; accessed 17-Oct-2022].

[2] Rodrigo Carro. 2024 Digital News Report Brazil. https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2024/brazil, 2024. [Online; accessed 17-Oct-2024].

[3] Alexandre Bovet and Hernán A Makse. Influence of fake news in twitter during the 2016 us presidential election. *Nature communications*, 10(1):1–14, 2019.

[4] Claire Wardle. Understanding Information disorder. https://firstdraftnews.org/long-form-article/understanding-information-disorder/, 2019. [Online; accessed 22-May-2022].

[5] Paula Adamo Idoeta. A história de Bolsonaro com a hidroxicloroquina em 6 pontos: de tuítes de Trump à CPI da Covid. https://www.bbc.com/portuguese/brasil-57166743, 2021. [Online; accessed 09-September-2021].

[6] VoxCheck Team. VoxUkraine about war. https://voxukraine.org/en/category/voxukraine-informs/, 2022. [Online; accessed 23-May-2022].

[7] Xinyi Zhou, Jindi Wu, and Reza Zafarani. SAFE: Similarity-aware multi-modal fake news detection. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, pages 354–367. Springer, 2020.

[8] Stephane Schwarz, Antônio Theóphilo, and Anderson Rocha. EMET: Embeddings from multilingual-encoder transformer for fake news detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[9] Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. Automated fact-checking for assisting human fact-checkers. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2021.

[10] Infographic. Data Never Sleeps 11.0. https://www.domo.com/learn/infographic/data-never-sleeps-11, 2023. [Online; accessed 12-Nov-2024].

[11] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In Marilyn

Walker, Heng Ji, and Amanda Stent, editors, *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, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1074. URL https://aclanthology.org/N18-1074.

[12] Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. MultiFC: A Real-World Multi-Domain Dataset for Evidence-Based Fact Checking of Claims. In *Conf. on Empirical Methods in Natural Language Processing and the Intl. Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

[13] Pepa Atanasova, Lluís Màrquez, Alberto Barrón-Cedeño, Tamer Elsayed, Reem Suwaileh, Wajdi Zaghouani, Spas Kyuchukov, Giovanni Da San Martino, and Preslav Nakov. Overview of the CLEF-2018 CheckThat! lab on automatic identification and verification of political claims, task 1: Check-worthiness. In *Cross-Language Evaluation Forum (CLEF)*, 2018.

[14] Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 2017.

[15] Chaoyuan Zuo, Ayla Karakas, and Ritwik Banerjee. A hybrid recognition system for check-worthy claims using heuristics and supervised learning. In *Central Europe workshop (CEUR)*, 2018.

[16] Casper Hansen, Christian Hansen, Jakob Grue Simonsen, and Christina Lioma. Neural weakly supervised fact check-worthiness detection with contrastive sampling-based ranking loss. In *Cross-Language Evaluation Forum (CLEF)*, 2019.

[17] Koustav Rudra, Siddhartha Banerjee, Niloy Ganguly, Pawan Goyal, Muhammad Imran, and Prasenjit Mitra. Summarizing situational tweets in crisis scenario. In *ACM Conference on Hypertext and Social Media (ACMHT)*, 2016.

[18] Soumi Dutta, Vibhash Chandra, Kanav Mehra, Sujata Ghatak, Asit Kumar Das, and Saptarshi Ghosh. Summarizing microblogs during emergency events: A comparison of extractive summarization algorithms. In *Emerging Technologies in Data Mining and Information Security (IEMIS)*, Singapore, 2019.

[19] Alexis Dusart, Karen Pinel-Sauvagnat, and Gilles Hubert. ISSumSet: a tweet summarization dataset hidden in a trec track. In *Annual ACM Symposium on Applied Computing (SAC)*, 2021.

[20] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese bert-networks. In *Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3973–3983, Hong Kong, China, 2019.

[21] Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1):1–12, 2019.

[22] Ulrike Von Luxburg, Agnes Radl, and Matthias Hein. Hitting and commute times in large random neighborhood graphs. *The Journal of Machine Learning Research (JMLR)*, 15(1):1751–1798, 2014.

[23] Didier A Vega-Oliveros, Pedro Spoljaric Gomes, Evangelos E Milios, and Lilian Berton. A multi-centrality index for graph-based keyword extraction. *Information Processing & Management (IP&M)*, 56(6):102063, 2019.

[24] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

[25] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research (JMLR)*, 21:1–67, 2020.

[26] Yichuan Li, Bohan Jiang, Kai Shu, and Huan Liu. MM-COVID: A multilingual and multidimensional data repository for combating COVID-19 fake new. *arXiv preprint arXiv:2011.04088*, 2020. URL https://arxiv.org/abs/2011.04088v2.

[27] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations (ICLR)*, New Orleans, United States, 2019.

[28] Tamanna Hossain, Robert L Logan IV, Arjuna Ugarte, Yoshitomo Matsubara, Sean Young, and Sameer Singh. COVIDLIES: Detecting COVID-19 misinformation on social media. In *The 1st Workshop on NLP for COVID-19 (EMNLP)*, online, 2020.

[29] Neema Kotonya and Francesca Toni. Explainable automated fact-checking for public health claims. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[30] Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. Averitec: A dataset for real-world claim verification with evidence from the web. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[31] Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. Generating fact checking explanations. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

[32] Daniel Russo, Serra Sinem Tekiroğlu, and Marco Guerini. Benchmarking the Generation of Fact Checking Explanations. *Trans. of the Association for Computational Linguistics*, 11:1250–1264, 2023.

[33] Siyuan Wang, Zhongyu Wei, Zhihao Fan, Zengfeng Huang, Weijian Sun, Qi Zhang, and Xuan-Jing Huang. PathQG: Neural question generation from facts. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[34] Angela Fan, Aleksandra Piktus, Fabio Petroni, Guillaume Wenzek, Marzieh Saeidi, Andreas Vlachos, Antoine Bordes, and Sebastian Riedel. Generating Fact Checking Briefs. In *Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[35] Neema Kotonya and Francesca Toni. Explainable automated fact-checking: A survey. In *International Conference on Computational Linguistics (ACL)*, 2020.

[36] Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. Faithful to the original: Fact aware neural abstractive summarization. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[37] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

[38] Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. Evaluating the Factual Consistency of Abstractive Text Summarization. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.750. URL https://aclanthology.org/2020.emnlp-main.750.

[39] Esin Durmus, He He, and Mona Diab. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

[40] Alex Wang, Kyunghyun Cho, and Mike Lewis. Asking and answering questions to evaluate the factual consistency of summaries. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

[41] Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, 2021.

[42] Feng Nan, Cicero Nogueira dos Santos, Henghui Zhu, Patrick Ng, Kathleen McKeown, Ramesh Nallapati, Dejiao Zhang, Zhiguo Wang, Andrew O Arnold, and Bing Xiang. Improving factual consistency of abstractive summarization via question answering. In *Joint Conference of Annual Meeting of the Association for Computational Linguistics and International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, 2021.

[43] Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations (ICLR)*, 2015.

[44] Julian Eisenschlos, Bhuwan Dhingra, Jannis Bulian, Benjamin Börschinger, and Jordan Boyd-Graber. Fool Me Twice: Entailment from Wikipedia Gamification. In *Association for Computational Linguistics: Human Language Technologies (NAACL)*, 2021.

[45] Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. Towards debiasing fact verification models. In *Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

[46] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MPNet: Masked and permuted pre-training for language understanding. *arXiv preprint*, 2020.

[47] Vivek Gupta, Riyaz A Bhat, Atreya Ghosal, Manish Srivastava, Maneesh Singh, and Vivek Srikumar. Is my model using the right evidence? systematic probes for examining evidence-based tabular reasoning. *arXiv preprint*, 2021.

[48] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

[49] Jenny Kunz and Marco Kuhlmann. Properties and challenges of LLM-generated explanations. In Su Lin Blodgett, Amanda Cercas Curry, Sunipa Dev, Michael Madaio, Ani Nenkova, Diyi Yang, and Ziang Xiao, editors, *Proceedings of the Third Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 13–27, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.hcinlp-1.2. URL https://aclanthology.org/2024.hcinlp-1.2.

[50] Nitay Calderon and Roi Reichart. On Behalf of the Stakeholders: Trends in NLP Model Interpretability in the Era of LLMs. *arXiv preprint arXiv:2407.19200*, 2024. URL https://arxiv.org/abs/2407.19200.

[51] Sarah Wiegreffe, Ana Marasović, and Noah A. Smith. Measuring Association Between Labels and Free-Text Rationales. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10266–10284, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.804. URL https://aclanthology.org/2021.emnlp-main.804.

[52] Ana Marasovic, Iz Beltagy, Doug Downey, and Matthew Peters. Few-Shot Self-Rationalization with Natural Language Prompts. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 410–424, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.31. URL https://aclanthology.org/2022.findings-naacl.31.

[53] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.408. URL https://aclanthology.org/2020.acl-main.408.

[54] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. e-SNLI: Natural Language Inference with Natural Language Explanations. *Advances in Neural Information Processing Systems*, 31, 2018. URL https://proceedings.neurips.cc/paper/2018/file/4c7a167bb329bd92580a99ce422d6fa6-Paper.pdf.

[55] Cunxiang Wang, Shuailong Liang, Yue Zhang, Xiaonan Li, and Tian Gao. Does it Make Sense? And Why? A Pilot Study for Sense Making and Explanation. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4020–4026, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1393. URL https://aclanthology.org/P19-1393.

[56] Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. Social Bias Frames: Reasoning about Social and Power Implications of Language. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.486. URL https://aclanthology.org/2020.acl-main.486.

[57] Shourya Aggarwal, Divyanshu Mandowara, Vishwajeet Agrawal, Dinesh Khandelwal, Parag Singla, and Dinesh Garg. Explanations for CommonsenseQA: New Dataset and Models. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3050–3065, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.238. URL https://aclanthology.org/2021.acl-long.238.

[58] Wei-Lin Chen, An-Zi Yen, Hen-Hsen Huang, and Hsin-Hsi Chen. Learning to generate explanation from e-hospital services for medical suggestion. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2946–2951, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.260.

[59] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In Lluís Màrquez, Chris Callison-Burch, and Jian Su, editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1075. URL https://aclanthology.org/D15-1075.

[60] Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. A richly annotated corpus for different tasks in automated fact-checking. In *Conf. on Computational Natural Language Learning (CoNLL)*, 2019. doi: 10.18653/v1/K19-1046.

[61] Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. COVID-Fact: Fact Extraction and Verification of Real-World Claims on COVID-19 Pandemic. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2116–2129, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.165. URL https://aclanthology.org/2021.acl-long.165.

[62] Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.722. URL https://aclanthology.org/2023.findings-emnlp.722.

[63] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288*, 2023. URL https://arxiv.org/pdf/2307.09288.pdf.

[64] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal

Anadkat, et al. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*, 2023. URL https://arxiv.org/pdf/2303.08774.pdf.

[65] Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, William Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah Smith, and Hannaneh Hajishirzi. OLMo: Accelerating the science of language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15789–15809, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.841.

[66] Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. Reframing human-AI collaboration for generating free-text explanations. In *Association for Computational Linguistics: Human Language Technologies (NAACL)*, 2022. doi: 10.18653/v1/2022.naacl-main.47.

[67] Alexis Ross, Matthew Peters, and Ana Marasovic. Does self-rationalization improve robustness to spurious correlations? In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7403–7416, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.501. URL https://aclanthology.org/2022.emnlp-main.501.

[68] Aditya Srikanth Veerubhotla, Lahari Poddar, Jun Yin, György Szarvas, and Sharanya Eswaran. Few shot rationale generation using self-training with dual teachers. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4825–4838, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.297. URL https://aclanthology.org/2023.findings-acl.297.

[69] Sahana Ramnath, Brihi Joshi, Skyler Hallinan, Ximing Lu, Liunian Harold Li, Aaron Chan, Jack Hessel, Yejin Choi, and Xiang Ren. Tailoring Self-Rationalizers with Multi-Reward Distillation. In *International Conference on Learning Representations (ICLR*, 2024. URL https://openreview.net/forum?id=t8eO0CiZJV.

[70] Yangqiaoyu Zhou and Chenhao Tan. Investigating the Effect of Natural Language Explanations on Out-of-Distribution Generalization in Few-shot NLI. In João Sedoc, Anna Rogers, Anna Rumshisky, and Shabnam Tafreshi, editors, *Proceedings of the Second Workshop on Insights from Negative Results in NLP*, pages

117–124, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.insights-1.17. URL https://aclanthology.org/2021.insights-1.17.

[71] Tom McCoy, Ellie Pavlick, and Tal Linzen. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1334. URL https://aclanthology.org/P19-1334.

[72] Yordan Yordanov, Vid Kocijan, Thomas Lukasiewicz, and Oana-Maria Camburu. Few-Shot Out-of-Domain Transfer Learning of Natural Language Explanations in a Label-Abundant Setup. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3486–3501, 2022. doi: 10.18653/v1/2022.findings-emnlp.255. URL https://aclanthology.org/2022.findings-emnlp.255.

[73] Elior Sulem, Omri Abend, and Ari Rappoport. BLEU is not suitable for the evaluation of text simplification. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 738–744, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1081. URL https://aclanthology.org/D18-1081.

[74] Dongfu Jiang, Yishan Li, Ge Zhang, Wenhao Huang, Bill Yuchen Lin, and Wenhu Chen. TIGERScore: Towards Building Explainable Metric for All Text Generation Tasks. *arXiv preprint arXiv:2310.00752*, 2023. URL https://arxiv.org/abs/2310.00752.

[75] Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, hai zhao, and Pengfei Liu. Generative judge for evaluating alignment. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=gtkFw6sZGS.

[76] Xinyu Hu, Li Lin, Mingqi Gao, Xunjian Yin, and Xiaojun Wan. Themis: Towards Flexible and Interpretable NLG Evaluation. *arXiv preprint arXiv:2406.18365*, 2024. URL https://arxiv.org/abs/2406.18365.

[77] Ming Li, Yong Zhang, Shwai He, Zhitao Li, Hongyu Zhao, Jianzong Wang, Ning Cheng, and Tianyi Zhou. Superfiltering: Weak-to-Strong Data Filtering for Fast Instruction-Tuning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14255–14273, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.769.

[78] Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. LESS: Selecting influential data for targeted instruction tuning. In *International Conference on Machine Learning (ICML)*, 2024. URL https://openreview.net/forum?id=PG5fV50maR.

[79] Jiuhai Chen and Jonas Mueller. Automated Data Curation for Robust Language Model Fine-Tuning. *arXiv preprint arXiv:2403.12776*, 2024. URL https://arxiv.org/abs/2403.12776.

[80] Dominik Stammbach and Elliott Ash. e-FEVER: Explanations and summaries for automated fact checking. *Truth and Trust Online (TTO 2020)*, 2020.

[81] Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. Diversity Over Size: On the Effect of Sample and Topic Sizes for Argument Mining Datasets. *arXiv preprint arXiv:2205.11472*, 2022. URL https://arxiv.org/pdf/2205.11472.pdf.

[82] Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.746. URL https://aclanthology.org/2020.emnlp-main.746.

[83] Hongjin Su, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, et al. Selective Annotation Makes Language Models Better Few-Shot Learners. In *The Eleventh International Conference on Learning Representations*, 2022. URL https://openreview.net/pdf?id=qY1hlv7gwg.

[84] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

[85] Majid Zarharan, Pascal Wullschleger, Babak Behkam Kia, Mohammad Taher Pilehvar, and Jennifer Foster. Tell Me Why: Explainable Public Health Fact-Checking with Large Language Models. In *Workshop on Trustworthy Natural Language Processing (TrustNLP 2024)*, 2024.

[86] Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Evan Walsh, Luke Zettlemoyer, Noah Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse

Dodge, and Kyle Lo. Dolma: an Open Corpus of Three Trillion Tokens for Language Model Pretraining Research. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15725–15788, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.840.

[87] Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/363_Paper.pdf.

[88] Ellie Pavlick and Chris Callison-Burch. Most "babies" are "little" and most "problems" are "huge": Compositional Entailment in Adjective-Nouns. In Katrin Erk and Noah A. Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2164–2173, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1204. URL https://aclanthology.org/P16-1204.

[89] Sheng Zhang, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. Ordinal Common-sense Inference. *Transactions of the Association for Computational Linguistics*, 5:379–395, 2017. doi: 10.1162/tacl_a_00068. URL https://aclanthology.org/Q17-1027.

[90] Alice Lai, Yonatan Bisk, and Julia Hockenmaier. Natural Language Inference from Multiple Premises. In Greg Kondrak and Taro Watanabe, editors, *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 100–109, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing. URL https://aclanthology.org/I17-1011.

[91] Adam Poliak, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme. Collecting Diverse Natural Language Inference Problems for Sentence Representation Evaluation. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 67–81, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1007. URL https://aclanthology.org/D18-1007.

[92] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *7th International Conference on Learning Representations, ICLR 2019*, 2019. URL https://openreview.net/pdf?id=rJ4km2R5t7.

[93] Swarnadeep Saha, Yixin Nie, and Mohit Bansal. ConjNLI: Natural Language Inference Over Conjunctive Sentences. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8240–8252, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.661. URL https://aclanthology.org/2020.emnlp-main.661.

[94] David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or Fiction: Verifying Scientific Claims. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.609. URL https://aclanthology.org/2020.emnlp-main.609.

[95] Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. Climate-FEVER: A Dataset for Verification of Real-World Climate Claims. In *Tackling Climate Change with Machine Learning workshop at NeurIPS 2020*, 2020. URL https://www.climatechange.ai/papers/neurips2020/67.

[96] Tal Schuster, Adam Fisch, and Regina Barzilay. Get your vitamin C! robust fact verification with contrastive evidence. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.52. URL https://aclanthology.org/2021.naacl-main.52.

[97] Siddharth Karamcheti, Ranjay Krishna, Li Fei-Fei, and Christopher Manning. Mind your outliers! investigating the negative impact of outliers on active learning for visual question answering. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7265–7281, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.564. URL https://aclanthology.org/2021.acl-long.564.

[98] Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Apostolos Dedeloudis, Jackson Sargent, and David Jurgens. POTATO: The portable text annotation tool. In Wanxiang Che and Ekaterina Shutova, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 327–337, Abu Dhabi, UAE, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-demos.33. URL https://aclanthology.org/2022.emnlp-demos.33.

[99] Anonymous. Self-Rationalization in the Wild: A Large Scale Out-of-Distribution Evaluation on NLI-related tasks. https://openreview.net/forum?id=KYEdQdGvAR, 2024. Preprint available at OpenReview.

[100] Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. Where is your evidence: Improving fact-checking by justification modeling. In *First Workshop on Fact Extraction and Verification (FEVER)*, 2018.

[101] William Yang Wang. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In Regina Barzilay and Min-Yen Kan, editors, *Annual Meeting of the Association for Computational Linguistics*, 2017.

[102] Xuan Zhang and Wei Gao. Towards LLM-based Fact Verification on News Claims with a Hierarchical Step-by-Step Prompting Method. In *Intl. Joint Conf. on Natural Language Processing and the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 996–1011, 2023.

[103] Yordan Yordanov, Vid Kocijan, Thomas Lukasiewicz, and Oana-Maria Camburu. Few-Shot Out-of-Domain Transfer Learning of Natural Language Explanations in a Label-Abundant Setup. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 2022.

[104] Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Pengfei Liu, et al. Generative Judge for Evaluating Alignment. In *Intl. Conf. on Learning Representations (ICLR)*, 2023.

[105] Dongfu Jiang, Yishan Li, Ge Zhang, Wenhao Huang, Bill Yuchen Lin, and Wenhu Chen. TIGERScore: Towards Building Explainable Metric for All Text Generation Tasks. *Trans. on Machine Learning Research*, 2024. ISSN 2835-8856.

[106] Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. End-to-End Bias Mitigation by Modelling Biases in Corpora. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8706–8716, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.769. URL https://aclanthology.org/2020.acl-main.769.

[107] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/pdf?id=SkeHuCVFDr.

[108] Zeming Chen, Qiyue Gao, and Lawrence S. Moss. NeuralLog: Natural language inference with joint neural and logical reasoning. In Lun-Wei Ku, Vivi Nastase, and Ivan Vulić, editors, *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 78–88, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.starsem-1.7. URL https://aclanthology.org/2021.starsem-1.7.

[109] Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis only baselines in natural language inference. In Malvina Nissim, Jonathan Berant, and Alessandro Lenci, editors, *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/S18-2023. URL https://aclanthology.org/S18-2023.

[110] Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. End-to-end bias mitigation by modelling biases in corpora. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8706–8716, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.769. URL https://aclanthology.org/2020.acl-main.769.

[111] Najoung Kim, Roma Patel, Adam Poliak, Patrick Xia, Alex Wang, Tom McCoy, Ian Tenney, Alexis Ross, Tal Linzen, Benjamin Van Durme, Samuel R. Bowman, and Ellie Pavlick. Probing what different NLP tasks teach machines about function word comprehension. In Rada Mihalcea, Ekaterina Shutova, Lun-Wei Ku, Kilian Evang, and Soujanya Poria, editors, *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 235–249, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/S19-1026. URL https://aclanthology.org/S19-1026.

[112] Yuxiang Wu, Matt Gardner, Pontus Stenetorp, and Pradeep Dasigi. Generating data to mitigate spurious correlations in natural language inference datasets. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2660–2676, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.190. URL https://aclanthology.org/2022.acl-long.190.

[113] Payal Bajaj, Chenyan Xiong, Guolin Ke, Xiaodong Liu, Di He, Saurabh Tiwary, Tie-Yan Liu, Paul Bennett, Xia Song, and Jianfeng Gao. METRO: Efficient Denoising Pretraining of Large Scale Autoencoding Language Models with Model Generated Signals. *arXiv preprint arXiv:2204.06644*, 2022. URL https://arxiv.org/abs/2204.06644.

[114] Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. Evaluating the Logical Reasoning Ability of ChatGPT and GPT-4. *arXiv preprint arXiv:2304.03439*, 2023. URL https://arxiv.org/abs/2304.03439.

[115] Robert Wolfe, Isaac Slaughter, Bin Han, Bingbing Wen, Yiwei Yang, Lucas Rosenblatt, Bernease Herman, Eva Brown, Zening Qu, Nic Weber, et al. Laboratory-scale ai: Open-weight models are competitive with chatgpt even in low-resource settings. In *ACM Conference on Fairness, Accountability, and Transparency*, pages 1199–1210, 2024. URL https://dl.acm.org/doi/abs/10.1145/3630106.3658966.

[116] Yi Tay, Mostafa Dehghani, Vinh Q Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Dara Bahri, Tal Schuster, Steven Zheng, et al. UL2: Unifying Language Learning Paradigms. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/pdf?id=6ruVLB727MC.

[117] Jian Guan, Jesse Dodge, David Wadden, Minlie Huang, and Hao Peng. Language models hallucinate, but may excel at fact verification. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1090–1111, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.62. URL https://aclanthology.org/2024.naacl-long.62.

[118] Jiuding Yang, Hui Liu, Weidong Guo, Zhuwei Rao, Yu Xu, and Di Niu. Reassess summary factual inconsistency detection with large language model. In Sha Li, Manling Li, Michael JQ Zhang, Eunsol Choi, Mor Geva, Peter Hase, and Heng Ji, editors, *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 27–31, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.knowllm-1.3.

[119] Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. TRUE: Re-evaluating Factual Consistency Evaluation. In Song Feng, Hui Wan, Caixia Yuan, and Han Yu, editors, *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 161–175, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.dialdoc-1.19. URL https://aclanthology.org/2022.dialdoc-1.19.

Appendix A

Experimental Details for Self-Rationalization Fine-tuning

In the following, we describe more experimental details and supplementary results related to Chapter 4.

A.1 Category 1: Additional details

A.1.1 Data pre-processing

For the following datasets, we applied pre-processing as defined below:

AddOneRTE [88] We convert the mean human scores into two classes *entailed* (when the score is no less than 4) and *not_ entailment* (when the score is no greater than 3, anything between 3 and 4 are removed), following the literature convention [106].

Ordinal Common-sense Inference (JOCI) [89] We follow Karimi Mahabadi et al. [106] by mapping the labels *very likely* to *entailment*; *likely*, *plausible* and *technically possible* to *neutral*; and *impossible* to *contradiction*.

Multiple Premise Entailment (MPE) [90] We concatenate the premise sentences together to form one premise paragraph.

SciFact [94] The dataset does not have public available labels for test set, thus we use the dev set. We do not perform evidence retrieval and use the cited document abstracts as evidence.

Climate FEVER [95] We use the paragraph-level evidence labels.

FactCC [38] We map label *factual* as *entailment* and *non-factual* to *not_ entailment*.

QAGS CNN [40] We aggregate with majority voting from the provided human annotations.

QAGS XSUM [40] We aggregate with majority voting from the provided human annotations.

XSUM Hallucination [37] We aggregate with majority voting from the provided human annotations.

A.1.2 Ambiguous sample selection method

We input the (h_i, p_i) to the T5-large model, and take the probability of the first most likely output token, since the first token represent the classification label. We denote the probability as p_i. To select ambiguous samples, we calculate a mean probability score p_{mean} as follows:

$$p_{mean} = (p_{max} + p_{min})/2 \tag{A.1}$$

where p_{max} and p_{min} represents the highest and lowest probability score among all sample scores respectively. Then we re-calculate the score based on its absolute distance with p_{mean}:

$$p_i' = |(p_i - p_{mean})| \tag{A.2}$$

with the absolute distance, we re-rank the samples from low to high to select the most ambiguous ones. The lowest value represents the most ambiguous sample and the highest the least ambiguous.

A.1.3 Additional implementation details

For T5-Large model fine-tuning, we perform a hyper-parameter search over the learning rate for each number of shots for each source dataset separately, with random sample selection from the first subset. We select the learning rate based on the highest performance on the in-distribution validation set within 50 epochs. The performance is based on the summation of label accuracy and explanation BERTscore [107]. The same hyper-parameters are used for all sample selection methods, which share the same m and source dataset for fine-tuning. To calculate the labels' accuracy and explanations' BERTscore, we divide the output sequence into the label and explanation. With the template format, T5 learns to generate a text label, followed by a separation pattern, "explanation:", and then the explanation tokens. Thus, we take the token before the separation pattern as the text label and after as the explanation. During hyper-parameter search, we test these learning rates: 3e-7, 3e-6, 3e-5, and 3e-4. For the validation set in fine-tuning, we randomly select 300 samples in the original validation set as the in-distribution set, as the original one is too large; thus, validation takes much longer. We follow the same settings as FEB [52] for the validation instances; for the ones with more than one explanation annotated, we merge them into one sequence separated by [SEP] token.

For OLMo-7B fine-tuning with LoRA, we follow recommended hyperparameters studied in Zarharan et al. [85]: LoRA r and alpha values are both 16, the learning rate is 2e-4, and the optimizer is "paged_adamw_32bit". We fine-tune all few-shot models with 50 epochs and use the models from the last epoch. For full-shot fine-tuning, the number of epochs is ten instead of 50.

The sentence-transformer model used in embedding the input for the Fast-Vote-k method is *paraphrase-mpnet-base-v2*.

A.1.4 Human evaluation interface

The evaluation interface is shown in Figure A.1, including the task instruction, some examples, and the evaluation page.

A.1.5 Input template for explanation evaluation with the reference-free metrics

- **Acceptability score**

> *premise:* [premise] *hypothesis:* [hypothesis] *answer:* [gold label] *explanation:* [explanation]

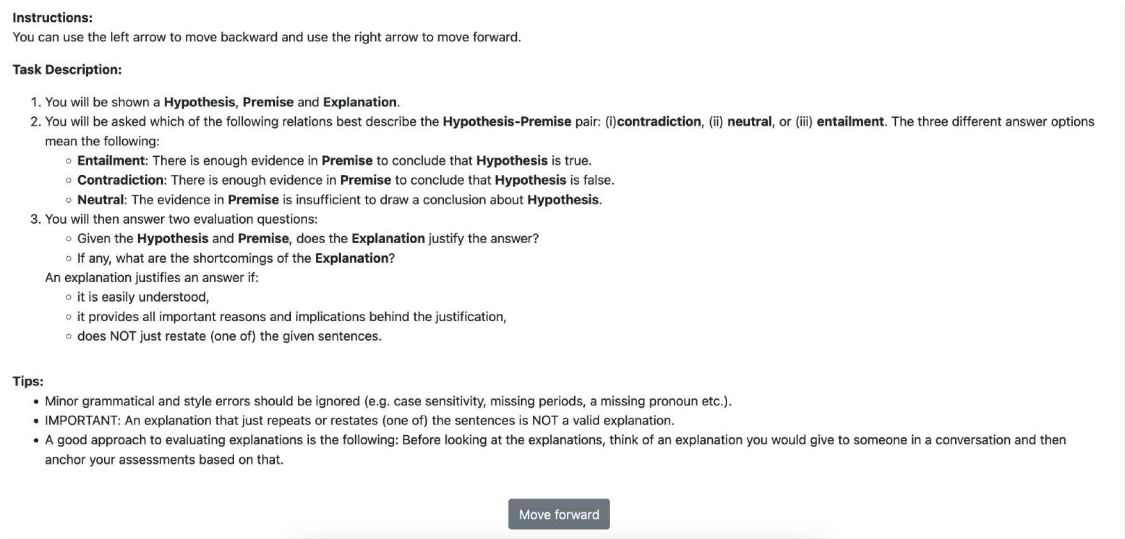
- **TigerScore** and **Auto-J**

> *Given a hypothesis and its premise, please explain why the hypothesis is entailment, neutral, or contradiction.*
> *Hypothesis:* [hypothesis], *Premise:* [premise].
> *Please explain why the hypothesis is* [gold label].

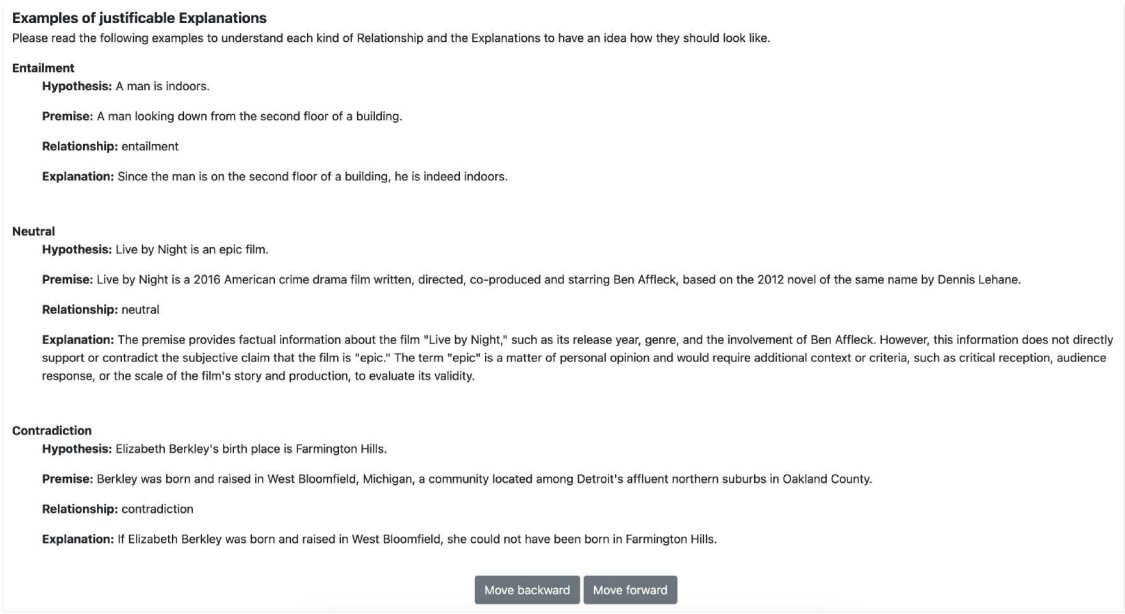
- **Themis (relevance aspect, input in JSON format)**

> {"task": "Controllable Generation", "aspect": "Coherence: Given the explanation for the relationship between the hypothesis and premise pair, how much does the generated explanation make sense?", "source_des": "Hypothesis and Premise Pair", "source": "Hypothesis: [hypothesis], Premise: [premise], please explain why the Hypothesis is [gold label].", "target_des": "Explanation", "target": [explanation]}

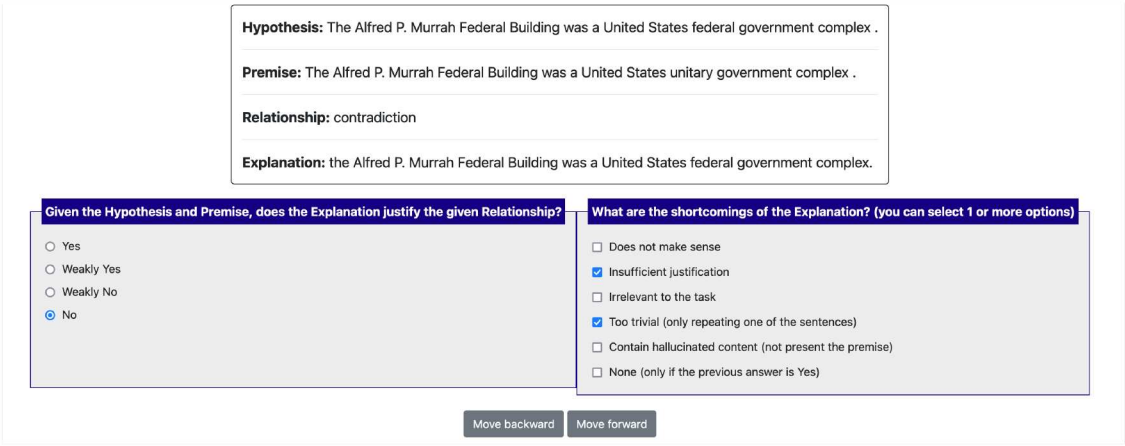
A.1.6 Generated explanations by different models and their evaluation scores



(a) Task instructions



(b) Examples



(c) The evaluation page

Figure A.1: Screenshots of human evaluation interface

Hypothesis (contradiction) A person in a blue jacket is jumping onto a tall cement wall
Premise The man is performing a large jump
$\mathbf{T}^{Fev}_{64,AFk}$ Human: 0.444, Accept.: 0.027 not all men wear a blue jacket.
$\mathbf{T}^{Fev}_{128,R}$ Human: 0.556, Accept.: 0.450 A man performing a large jump is not necessarily in a blue jacket and jumping onto a tall cement wall.
$\mathbf{T}^{Fev}_{128,Fk}$ Human: 0.889, Accept.: 0.036 "the relevant information about the claim is lacking in the context."
$\mathbf{T}^{Fev}_{128,AFk}$ Human: 1.0, Accept.: 0.494 The man is not necessarily wearing a blue jacket and the jump is not necessarily onto a tall wall.
$\mathbf{T}^{Fev}_{Full}$ Human: 0.222, Accept.: 0.036 "the relevant information about the claim is lacking in the context."
$\mathbf{T}^{Sn}_{64,Fk}$ Human: 0.556, Accept.: 0.520 the man is performing a large jump onto a tall cement wall.
$\mathbf{T}^{Sn}_{64,AFk}$ Human: 1.0, Accept.: 0.394 the man is performing a large jump onto a tall cement wall does not indicate that he is in a blue jacket.
$\mathbf{T}^{Sn}_{Full}$ Human: 1.0, Accept.: 0.842 a large jump does not imply that the man is jumping onto a tall cement wall.
$\mathbf{O}^{Fev}_{16,AFk}$ Human: 0.778, Accept.: 0.036 "the relevant information about the claim is lacking in the context."
$\mathbf{O}^{Fev}_{128,AFk}$ Human: 0.889, Accept.: 0.473 A person in a blue jacket is performing a large jump onto a tall cement wall
$\mathbf{O}^{Fev}_{Full}$ Human: 0.667, Accept.: 0.285 The man is performing a large jump, however it does not say what kind of jacket he is wearing.
$\mathbf{O}^{Sn}_{128,AFk}$ Human: 0.778, Accept.: 0.536 a man in a blue jacket is performing a large jump onto a tall cement wall.
$\mathbf{O}^{Sn}_{Full}$ Human: 0.667, Accept.: 0.805 The man is performing a large jump, but it does not say whether he is jumping onto a tall cement wall.

Table A.1: An example of generated explanations by the 13 selected models for human evaluation, with the instance randomly selected from the SICK dataset (all models have the correct prediction).

Hypothesis (entailment) Guns N ' Roses is an American hard rock band.
Premise Guns N ' Roses, often abbreviated as GnR, is an American hard rock band from Los Angeles formed in 1985.
$\mathbf{T}^{Fev}_{64,AFk}$ Human: 0.889, Accept.: 0.588 guns n roses is an american hard rock band.
$\mathbf{T}^{Fev}_{128,R}$ Human: 1.0, Accept.: 0.510 guns n'roses is an american hard rock band from los angeles formed in 1985.
$\mathbf{T}^{Fev}_{128,Fk}$ Human: 0.333, Accept.: 0.174 guns n'roses is an american hard rock band, however it does not say whether it is a rock band.
$\mathbf{T}^{Fev}_{128,AFk}$ Human: 0.778, Accept.: 0.527 guns n roses is an american hard rock band from los angeles formed in 1985.
$\mathbf{T}^{Fev}_{Full}$ Human: 1.0, Accept.: 0.556 guns n'roses is an american hard rock band from los angeles formed in 1985.
$\mathbf{T}^{Sn}_{64,Fk}$ Human: 0.778, Accept.: 0.498 the band was formed in 1985.
$\mathbf{T}^{Sn}_{64,AFk}$ Human: 0.889, Accept.: 0.573 the band is from Los Angeles and formed in 1985.
$\mathbf{T}^{Sn}_{Full}$ Human: 1.0, Accept.: 0.302 guns n'roses is an american hard rock band is a rephrasing of guns n'roses, often abbreviated as gnr, is an american hard rock band
$\mathbf{O}^{Fev}_{16,AFk}$ Human: 1.0, Accept.: 0.603 Guns N'Roses is an American hard rock band, formed in Los Angeles in 1985.
$\mathbf{O}^{Fev}_{128,AFk}$ Human: 0.889, Accept.: 0.556 Guns N'Roses is an American hard rock band from Los Angeles formed in 1985.
$\mathbf{O}^{Fev}_{Full}$ Human: 1.0, Accept.: 0.556 Guns N'Roses is an American hard rock band from Los Angeles formed in 1985.
$\mathbf{O}^{Sn}_{128,AFk}$ Human: 0.889, Accept.: 0.618 Guns N'Roses is a hard rock band.
$\mathbf{O}^{Sn}_{Full}$ Human: 0.111, Accept.: 0.088 Guns is hard to form a hard hard hard hard.

Table A.2: An example of generated explanations by the 13 selected models for human evaluation, with the instance randomly selected from the VitaminC dataset (all models have the correct prediction).

Hypothesis (entailment) a hospital trust is being investigated by the health watchdog over its finances.
Premise Monitor is looking into the financial sustainability of Southend University Hospital Foundation Trust “on behalf of patients”, the NHS regulator said. Finances became a concern when a planned £7.8m deficit grew, for which Monitor could see no recovery plan. The hospital blamed the larger-than-anticipated deficit on growth in demand and increased staff recruitment. Hospital chairman Alan Tobias OBE, said: “The overspend is a result of vital investment in services and more staff to ensure high standards of patient care as well as the rise in attendances. "We welcome this review to clearly demonstrate to Monitor - as well as patients and stakeholders - our future financial plans are both robust and sustainable.” The hospital has pledged to cut the deficit while maintaining “high quality care for local people” and said it was committed to returning to a surplus within three years. Monitor said its investigation will look into the state of the hospital trustś finances, assess the strength of its financial management and explore ways of improving its sustainability, on behalf of patients across the region. The investigation will also examine how the trust works with other local health and care organisations to respond to the financial challenges it faces. Katherine Cawley, regional director at Monitor, said: “People are relying on Southend to provide them with quality healthcare now and in the future, and we are concerned that this may be at risk because of the trustś poor financial performance. "We have launched this investigation to find out more about the financial situation at the trust and to establish what can be done to improve things.”
$\mathbf{T}^{Fev}_{64,AFk}$ Human: 0.667, Accept.: 0.267 monitor is looking into the financial sustainability of southend university hospital foundation trust “on behalf of patients”.
$\mathbf{T}^{Fev}_{128,R}$ Human: 0.889, Accept.: 0.276 Monitor is looking into the financial sustainability of Southend University Hospital Foundation Trust ‘on behalf of patients’.
$\mathbf{T}^{Fev}_{128,Fk}$ Human: 1.0, Accept.: 0.310 southend university hospital foundation trust is being investigated by monitor over its finances.
$\mathbf{T}^{Fev}_{128,AFk}$ Human: 0.889, Accept.: 0.319 a hospital trust is being investigated by monitor over its finances.
$\mathbf{T}^{Fev}_{Full}$ Human: 0.889, Accept.: 0.266 monitor is investigating the financial sustainability of southend university hospital foundation trust “on behalf of patients”.
$\mathbf{T}^{Sn}_{64,Fk}$ Human: 0.778, Accept.: 0.240 the health watchdog is looking into the financial sustainability of Southend University Hospital Foundation Trust “on behalf of patients”
$\mathbf{T}^{Sn}_{64,AFk}$ Human: 1.0, Accept.: 0.263 monitor is looking into the financial sustainability of southend university hospital foundation trust “on behalf of patients”
$\mathbf{T}^{Sn}_{Full}$ Human: 0.778, Accept.: 0.219 monitor is the health watchdog.
$\mathbf{O}^{Fev}_{16,AFk}$ Human: 1.0, Accept.: 0.262 a hospital trust is being investigated by the health watchdog over its finances.
$\mathbf{O}^{Fev}_{128,AFk}$ Human: 1.0, Accept.: 0.302 The hospital trust is being investigated by the health watchdog over its finances.
$\mathbf{O}^{Fev}_{Full}$ Human: 0.889, Accept.: 0.310 southend university hospital foundation trust is being investigated by monitor over its finances.
$\mathbf{O}^{Sn}_{128,AFk}$ Human: 1.0, Accept.: 0.358 Monitor is looking into the financial sustainability of Southend University Hospital Foundation Trust “on behalf of patients”,“explanation”:“The hospital trust’s poor financial performance is being investigated by the health watchdog over its finances.
$\mathbf{O}^{Sn}_{Full}$ Human: 0.444, Accept.: 0.151 The financial services watch the financial policy of the financial and financial management to the financial services to the financial services.

Table A.3: An example of generated explanations by the 13 selected models for human evaluation, with the instance randomly selected from the XSUM Hallucination dataset (all models have the correct prediction).

A.2 Category 2: Complementary results

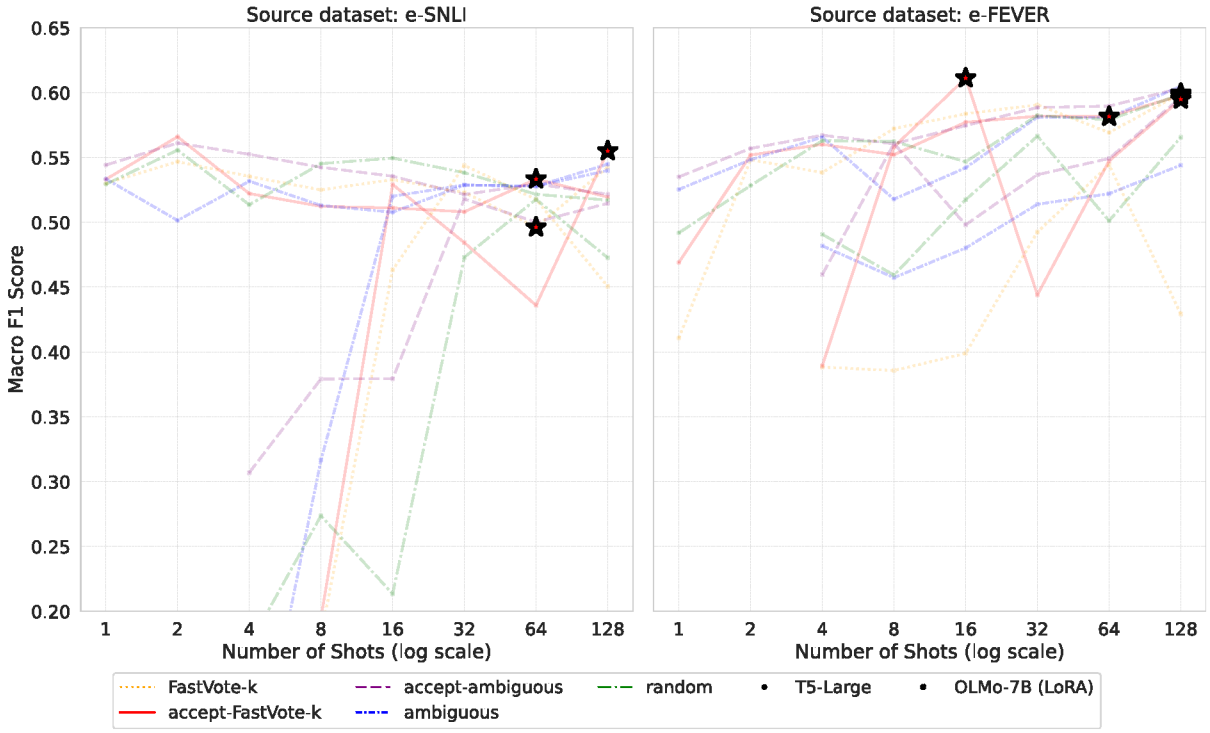


Figure A.2: F1 scores of the 3 selected OOD datasets (SICK, VitaminC, XSUM Hallucination) on models fine-tuned with data from the first subset. Models marked with the asterisks are the selected ones for human evaluation (besides the full-shot models which we all include). We did not consider 1- and 2-shots fine-tuned T5 models on e-SNLI, as we observed very low quality explanations in those models.

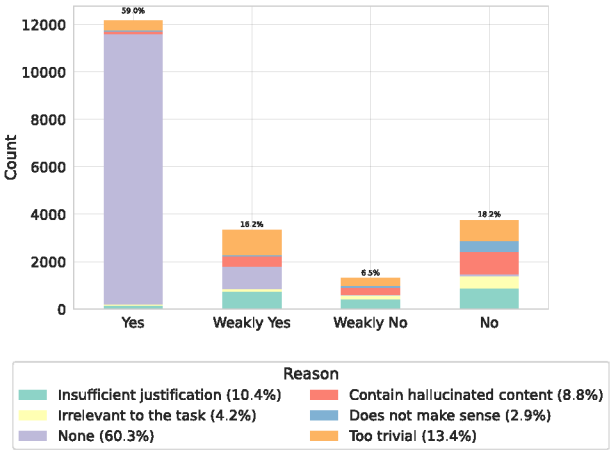


Figure A.3: Distribution of reasons of shortcomings from by four answers for the question "Does the explanation justify the answer?". The overall explanation quality is high according to the crowd workers, around 59% instances have "Yes" for the question "Does the explanation justify the answer?". The most common shortcoming across all answers is "Too trivial", followed by "Insufficient justification" and "Contain hallucinated content".

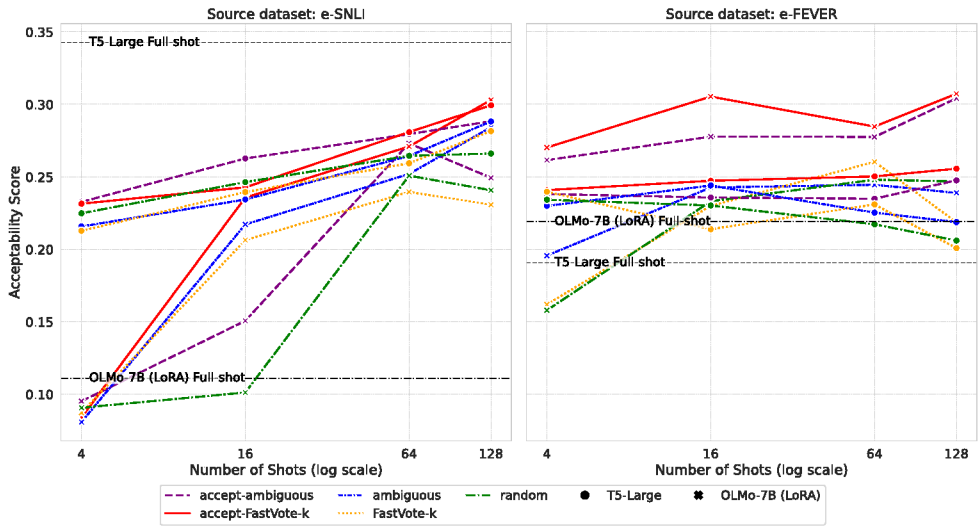


Figure A.4: Acceptability score across different number of shots and sample selection methods. Selection methods with "accept-" has highest Acceptability scores for all models on both source datasets.

Dataset	$\mathbf{T}^{Sn}_{Full}$	T^{Fev}_{Full}	$\mathrm{O}^{Sn}_{128,AFk}$	$\mathbf{O}^{Fev}_{128,AFk}$	MAJ	SOTA
SICK	57.1	82.4	53.7	64.2	56.9	90.3 [108]
AddOneRTE	88.6	88.4	81.9	85.5	85.3	92.2 [88]
JOCI	53.6	61.5	47.1	57.9	57.9	62.6 [109]
MPE	71.0	41.6	65.6	60.2	42.4	70.2 [110]
DNC	60.8	68.3	55.2	62.1	50.3	69.0 [111]
HANS	63.7	54.9	59.3	68.6	50.0	79.1 [112]
WNLI	45.1	43.7	49.3	56.3	56.3	85.6 [25]
Glue Diagnostics	60.1	61.9	58.2	62.7	41.7	57.0^{M} [113]
Conj	62.6	66.9	58.3	57.3	45.1	72.7 [114]
Snopes Stance	36.6	60.3	45.4	61.1	45.9	59.6^{F1} [60]
SciFACT	65.3	67.7	54.3	70.0	41.3	91.4^{F1} [94]
Climate FEVER	47.9	49.5	43.5	51.3	47.4	75.0 [115]
VitaminC	59.8	63.0	58.4	61.0	50.1	91.1 [116]
COVID-Fact	66.5	74.3	65.1	76.3	68.3	83.5 [61]
FM2	71.7	73.2	76.6	79.7	50.7	88.5 [117]
FactCC	88.3	89.3	68.6	79.1	87.7	91.3^{BA} [118]
QAGS CNN	75.6	78.2	62.9	76.8	74.4	81.3 [119]
QAGS XSUM	60.3	62.8	61.5	72.8	51.5	77.4 [119]
XSUM H.	58.9	62.4	82.9	80.0	90.1	66.4^{BA} [118]

Table A.4: Comparison of accuracy on the 19 OOD datasets with different models. MAJ: majority voting baseline, SOTA: state-of-the-art, M: Matthews coefficient, F1: F1 score, BA: balanced accuracy.

Appendix B

Copyright Notices

Scalable Fact-checking with Human-in-the-Loop

Conference Proceedings:
2021 IEEE International Workshop on Information Forensics and Security (WIFS)

Author: Jing Yang

Publisher: IEEE

Date: 07 December 2021

Thesis / Dissertation Reuse

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

BACK CLOSE WINDOW

Explainable Fact-Checking Through Question Answering

Conference Proceedings:
ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

Author: Jing Yang

Publisher: IEEE

Date: 23 May 2022

Thesis / Dissertation Reuse

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

BACK CLOSE WINDOW