RU22Fact: Optimizing Evidence for Multilingual Explainable Fact-Checking on Russia-Ukraine Conflict

Yirong Zeng[1], **Xiao Ding**[1]*, **Yi Zhao**[1], **Xiangyu Li**[2],
Jie Zhang[2], **Chao Yao**[2], **Ting Liu**[1], **Bing Qin**[1]
[1] Research Center for Social Computing and Information Retrieval, Harbin Institute of Technology
[2] Academy of Cyber, China
{yrzeng,xding,yzhao,tliu,qinb}@ir.hit.edu.cn
{lixiangyu1101,zhangjie9108,yaochao}@outlook.com

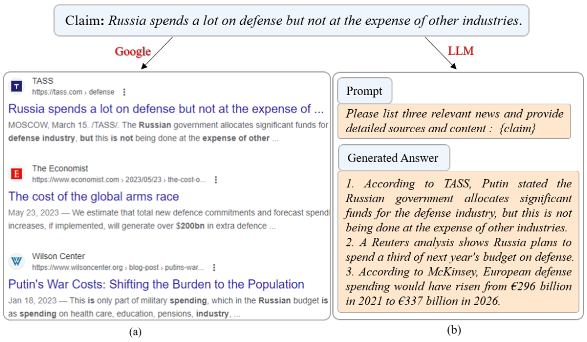
Abstract

Fact-checking is the task of verifying the factuality of a given claim by examining the available evidence. High-quality evidence plays a vital role in enhancing fact-checking systems and facilitating the generation of explanations that are understandable to humans. However, the provision of both sufficient and relevant evidence for explainable fact-checking systems poses a challenge. To tackle this challenge, we propose a method based on a Large Language Model to automatically retrieve and summarize evidence from the Web. Furthermore, we construct RU22Fact, a novel multilingual explainable fact-checking dataset on the Russia-Ukraine conflict in 2022 of 16K samples, each containing real-world claims, optimized evidence, and referenced explanation. To establish a baseline for our dataset, we also develop an end-to-end explainable fact-checking system to verify claims and generate explanations. Experimental results demonstrate the prospect of optimized evidence in increasing fact-checking performance and also indicate the possibility of further progress in the end-to-end claim verification and explanation generation tasks.

Keywords: fact-checking, evidence, explainability, large language models

1. Introduction

As information quickly spreads through social media, fake news become an urgent social issue and even a means of warfare. For example, conspiracy theories about Ukrainian and US bioweapons research during the Russian-Ukrainian conflict emerged (Bacio Terracino and Matasick, 2022). To combat fake news, automated fact-checking becomes an essential task, which aims to verify the factuality of a given claim based on the collected evidence. Figure 1 (a) illustrates a real-world claim[1] that has been verified using a search engine as a basis.


Claim: *Russia spends a lot on defense but not at the expense of other industries.*

TASS
https://tass.com › defense
Russia spends a lot on defense but not at the expense of ...
MOSCOW, March 15. /TASS/. The **Russian** government allocates significant funds for **defense industry**, **but** this **is not** being done at the **expense of other** ...

The Economist
https://www.economist.com › 2023/05/23 › the-cost-o...
The cost of the global arms race
May 23, 2023 — We estimate that total new defence commitments and forecast spendi increases, if implemented, will generate over $**200bn** in extra defence ...

Wilson Center
https://www.wilsoncenter.org › blog-post › putins-war...
Putin's War Costs: Shifting the Burden to the Population
Jan 18, 2023 — This **is** only part of military **spending**, which in the **Russian** budget **is** as **spending** on health care, education, pensions, **industry**, ...

Prompt
Please list three relevant news and provide detailed sources and content : {claim}

Generated Answer
1. According to TASS, Putin stated the Russian government allocates significant funds for the defense industry, but this is not being done at the expense of other industries.
2. A Reuters analysis shows Russia plans to spend a third of next year's budget on defense.
3. According to McKinsey, European defense spending would have risen from €296 billion in 2021 to €337 billion in 2026.

Figure 1: An example of fact verification, based on two different pieces of evidence. (a) a return from a search engine (e.g., Google), and (b) a reply generated by LLMs (e.g., New Bing).

Traditional fact-checking systems follow a pipeline approach that involves an evidence document retrieval module and a claim verification module (Kotonya and Toni, 2020a; Vlachos and Riedel, 2014). Although most researchers assume that evidence has been properly identified and focus on subsequent steps (Krishna et al., 2022; Liu et al., 2020; Nie et al., 2019), it is crucial to recognize the significant role of evidence in fact-checking (Schuster et al., 2020).

In fact-checking, it is natural to verify the claim in all the collected documents (Xiong et al.; Khattab et al., 2021), resulting in a substantial memory footprint due to storage requirements. To tackle this concern, Thorne et al. (2018) proposes the extraction of evidence documents from Wikipedia that are relevant to a claim, followed by the selection of the most pertinent sentences from these documents to produce the evidence. However, the crowd-sourced claims from this study introduce lexical biases, such as the excessive presence of explicit negation and unrealistic misinformation. Recent research retrieves real-world claims from fact-checking websites and considers search snippets (Gupta and Srikumar, 2021; Augenstein et al., 2019) or retrieved documents (Hu et al., 2023) provided by search engines as evidence to mitigate this issue. Nonetheless, as depicted in Figure 1 (a), search snippets often fail to give sufficient information to verify the claim (Hu et al., 2023), and the retrieved documents frequently contain a substan-

*Corresponding authors
[1] https://tass.com/defense/1589173

LREC-COLING 2024, pages 14215–14226
20-25 May, 2024.

tial amount of irrelevant information. The substandard content retrieved by search engines forces the need for an evidence extractor before the fact verification stage, which would result in error-cascading concerns. However, providing sufficient and relevant evidence for the fact-checking system is an unresolved challenge.

In response to the aforementioned challenge, we consider introducing Large Language Models (LLMs) given their excellent performance in natural language understanding (Bubeck et al., 2023; OpenAI, 2023). As illustrated in Figure 1 (b), LLMs have more potential to produce more relevant and sufficient information than search snippets. We propose an LLMs-driven method to automatically retrieve and summarize documents from the Web to produce precise evidence with less noise, and refer to the evidence obtained through this method as optimized evidence. On this basis, we construct RU22Fact, a novel multilingual explainable fact-checking dataset. It contains 16,033 examples, each containing real-world claims, optimized evidence, and referenced explanations about the Russia-Ukraine conflict in 2022. We build an explainable fact-checking system to establish the baseline performance and experimental results show that there is room for future improvements in this end-to-end fact-checking and explanation generation task. Experimental results demonstrate the prospect of optimized evidence in increasing fact-checking performance, while there is a challenge to solve the problem of generating end-to-end claim verification and explanations[2]. Our main contributions are summarized as follows:

- We propose an LLMs-driven method to automatically acquire sufficient and relevant evidence from the web. To our knowledge, we are the first to explore optimized evidence in the fact-checking system.
- We construct RU22Fact, a novel multilingual explainable fact-checking dataset. This dataset includes optimized evidence to support end-to-end claim verification and human-understandable explanation generation.

2. Related Work

2.1. Fact-Checking System

Normally, when verifying a claim, systems often operate as a pipeline consisting of an evidence document retrieval module, and a claim verification module (Kotonya and Toni, 2020a; Vlachos and Riedel, 2014). Most existing methods follow this framework and mainly focus on the last stages (Liu et al., 2020; Krishna et al., 2022; Nie et al., 2019). However, we argue that an optimized evidence document is also critical to building a fact-checking system. At present, there are two main ways to carry out evidence document retrieval. The first is to extract evidence documents related to a claim by entity link (Thorne et al., 2018), or by TF-IDF (Hanselowski et al., 2018) from the knowledge base (e.g., Wikipedia) or fact-checking websites, and then select the most relevant sentences from the documents to produce evidence (Wan et al., 2021; Aly and Vlachos, 2022). Nevertheless, the source of evidence limits its broad application. The second is to regard search snippets returned by search engines as evidence (Gupta and Srikumar, 2021; Hu et al., 2023). Although it can verify claims from various sources under real-world scenarios, the low-quality evidence from the search snippet limits the performance of the fact-checking system. Different from these methods, we proposed an automated LLMs-based evidence document retrieval method to produce optimized evidence for building a better fact-checking system.

2.2. Fact-Checking Dataset

We group existing fact-checking datasets into two categories: synthetic and real-world. Synthetic datasets (e.g., Fever (Thorne et al., 2018), Feverous (Aly et al., 2021), Hover (Jiang et al., 2020)), consider Wikipedia as the source of evidence and annotate the sentences of articles as evidence. Although these datasets have made a significant contribution to fact-checking, crowd-sourced claims from this line of work are written with minimal edits to reference sentences, leading to strong lexical biases. Thus, real-world efforts (Hanselowski et al., 2019; Kotonya and Toni, 2020b) extract summaries accompanying fact-checking articles about claims as evidence. Nevertheless, using fact-checking articles restricts evidence to a single source, and they are not available during inference, which is not ideal for developing automated fact-checking systems. To address this issue, some researchers regard search snippets (Gupta and Srikumar, 2021; Augenstein et al., 2019) or retrieved documents (Hu et al., 2023) returned by search engines as evidence. However, the low-quality content returned by search engines limits the performance of the system. For explainability in the dataset, most existing methods are dedicated to producing extractive explanations (e.g., explanations for veracity predictions about inputs to the system (Lu and Li, 2020; Wu et al., 2020)), which is unfriendly to humans. Recent researchers have formulated the explanation generation task as an abstract summarization problem for human understanding (Liu and Lapata, 2019; Kotonya and Toni, 2020b; Yao et al., 2022).

As shown in Table 1, in this paper, we construct a

[2] Data are available at https://github.com/zeng-yirong/ru22fact.

Dataset	Evi.	Exp.
Synthetic		
Fever (Thorne et al., 2018)	Wiki.	Ex.
Feverous (Aly et al., 2021)	Wiki.	Ex.
Hover (Jiang et al., 2020)	Wiki.	Ex.
Real-world		
MultiFC (Augenstein et al., 2019)	FCA	Ex.
PubH (Kotonya and Toni, 2020b)	FCA	Ab.
EFact (Hu et al., 2023)	SE	Ex.
XFact (Gupta and Srikumar, 2021)	SE	Ex.
RU22Fact	LLMs	Ab.

Table 1: Comparison of Fact-Checking Datasets. ”Evi.” and ”Exp.” are abbrs for Evidence and Explanation. ”Ex.”, ”Ab.”, and ”Wiki.” are abbrs for ”Extractive explanation”, ”Abstract summarization” and ”Wikipedia”. ”FCA” and ”SE” are abbrs for fact-checking articles and search engines.

fact-checking dataset, containing real-world claims, high-quality evidence, and referenced explanations, generating explanations as an abstract summarization task.

3. Evidence Analysis

To explore the critical role of evidence in fact-checking, we conducted both manual analysis and experimental analysis.

In experimental analysis, we conducted an exploratory experiment. XFact (Gupta and Srikumar, 2021) is a multilingual fact-checking dataset and contains evidence consisting of documents retrieved by a search engine, which sometimes fails to provide sufficient information for fact-checking. To produce optimized evidence for claims in XFact, we retrieve and summarize documents from the Web for each claim by LLMs. Then we extend XFact with optimized evidence and verify claims based on the evidence from search engines or optimized evidence. We implement the following experiment according to Gupta and Srikumar (2021).

1. **Attention-based Evidence Aggregator (Attn-EA)**: Aggregation of evidence using an attention-based model that operates on evidence documents retrieved by a search engine. For comparison, we utilize optimized evidence for attention-based evidence aggregator **(+OE)** .

2. **Augmenting metadata (+Meta)**: Concatenate additional key-value metadata with the claim text by representing it as a sequence. We also implement an optimized evidence-based evidence aggregator enhanced by metadata **(+ Meta + OE)**.

The results are shown in Table 2, from which we can find that the model with optimized evidence achieves better performance compared to the model with retrieved documents by a search engine. This indicates that optimized evidence can provide more sufficient and relevant evidence for fact-checking to improve its performance, and demonstrates the prospect of optimized evidence to solve the fact-checking problem.

Model	α_1	α_2	α_3
Attn-EA	38.9	15.7	16.5
Attn-EA+Meta	41.9	15.4	16.0
Attn-EA+OE	40.37	**17.29**	18.90
Attn-EA+Meta+OE	**42.71**	17.14	**19.59**

Table 2: Average F1 scores of the model. α_1 ,α_2 and α_3 is the different test sets in XFact. α_1 is distributionally similar to the training set, α_2 is out-of-domain test set and α_3 is the zero-shot test set.(%)

Evidence	Sufficiency	Relevance
Original Evidence	2.35(0.53)	3.02(0.57)
Optimized Evidence	3.51(0.50)	4.28(0.63)

Table 3: The manual evaluation of original evidence and optimized evidence. **Sufficiency** and **Relevance** represent the average scores of 100 samples. Kappa values are represented in brackets.

In manual analysis, we conducted a manual evaluation of the following aspects of evidence: 1) sufficiency: there is sufficient information in the evidence to verify the claim; 2) relevance: each sentence in the evidence relates to the claim. Each aspect is given a score of 1 to 5. We compared the original evidence and the optimized evidence in extended XFact and randomly sampled 100 samples for manual evaluation. We utilize Fleiss’ Kappa (Fleiss, 1971) to assess the inter-annotator agreement. The result is shown in Table 3, it shows that the optimized evidence obviously outperforms the original evidence in these two aspects, which indicates that the optimized evidence is better at fact-checking.

4. Dataset Construction

In this section, we introduce the whole procedure of dataset construction. We construct a multilingual explainable fact-checking dataset, named RU22Fact, of 16k real-world claims related to the 2022 Russia-Ukraine conflict, including conflict coverage, energy crisis, and related stories (e.g., humanitarianism, conspiracy theory, politics). To combat fake news about the Russia-Ukraine conflict in different countries and languages, the proposed dataset contains four languages: English, Chinese, Russian, and Ukrainian. An example dataset entry is shown in Table 4.

Claim	*1,000,000 Ukraine soldiers wiped out.*
Evidence	*I found a claim on social media that 1,000,000 Ukraine soldiers were wiped out. However, according to a fact-check by PolitiFact, this claim has no official backing. United States and European officials estimate that as many as 120,000 Ukrainian soldiers have died or been injured in the war.*
Explanation	*United States and European officials estimate as many as 120,000 Ukrainian soldiers have died or been injured in the war. We find no basis for a ”1 million” estimate.*
Label	Refuted
Date	May 25, 2023
Claimant	Facebook posts
Language	English

Table 4: A example from RU22Fact. Labels and explanations are provided during training but need to be inferred during evaluation.

4.1. Data Collection

To obtain sufficient claims related to the Russia-Ukraine conflict, we collect claims from two sources: fact-checking websites (e.g., Politifact[3], Chinafactcheck[4], Lenta[5]) and credible news release websites (e.g., CNN[6], People's Daily Online[7], TASS[8]). We consider websites that included Russian-Ukrainian conflict-related claims and eventually choose ten fact-checking websites and six news release websites, shown in Table 5

As a starting point, we first query the Russia-Ukraine conflict topic for each website. For websites without such a topic, we search for relevant content using keywords related to the Russia-Ukraine conflict. We scrape fact-checked claims from fact-checking websites and headline claims from news release websites, then take the fact-checking justification from fact-checking websites as referenced explanations for the veracity label of the claim. For headline claims, we summarize the news article by LLMs and check them manually to be referenced explanations. All claims were published between February 2022 and June 2023. In addition to the claim and referenced explanations, we crawl metadata related to each claim such as claimant and date of the claim. Initially, we scraped 39K claims, amounting to 9,037 fact-checked claims from fact-checking websites, and 30,412 news headline claims from news release websites.

4.2. Data Processing

Dataset Filtering. There are two major challenges in using the crawled data directly: 1) standardizing the labels and 2) cleaning the claims and explanations in the dataset. The initial data contains 46 labels. Referring to Hanselowski et al. (2019), we review the rating system of the fact-checking websites along with some examples and manually mapped these labels to three categories, including *Supported*, *Refuted*, and *NEI (Not Enough Information)*. For headline claims from news release websites, we assume that they are verified and labeled these *Supported* due to reputable sources, and each claim is assigned one of the three label categories. In data cleaning, we filter out claims longer than 50 characters to avoid multiple statements in a claim, and we also filter out shorter than 5 words in English, Russian, Ukrainian and 5 characters in Chinese to provide complete semantics in a claim. We remove explanations that are less than the length of the claim because it is difficult to provide qualified explanations. To alleviate label leakage in some claims, we remove the claims that contain unique keywords associated with the label.

Optimizing Evidence. To provide both sufficient and relevant evidence that differs from prior works, we propose an LLMs-driven method to automatically retrieve and summarize documents from the Web to produce optimized evidence. The detailed description is shown in section 5.1.

4.3. Task Definition

As we have automatically retrieved and summarized documents to produce optimized evidence, We introduce an end-to-end fact-checking approach to verify the claim, instead of a pipeline, taking into consideration the potential issue of error cascading in the pipeline. Specifically, we explore two subtasks in the proposed dataset, end-to-end claim verification, and explanation generation.

- **Claim Verification**: The Claim Verification task is to predict the label (Supported, Refuted, or NEI) of the claim based on the provided evidence.

- **Explanation Generation**: Given an input claim and optimized evidence, as well as the label, the goal of Explanation Generation is to generate a short paragraph to explain the ruling process and justify the label.

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

[4] https://chinafactcheck.com/

[5] https://lenta.ru/

[6] https://edition.cnn.com/

[7] http://www.people.com.cn/

[8] https://tass.com/

Source Type	Website	Quantity	Language	Total
Fact-checking website	politifact.com	2,782	English	6,035
	snopes.com	806	English	
	factcheck.afp.com	175	English	
	stopfake.org/en	300	English	
	stopfake.org/uk	1,082	Ukrainian	
	lenta.ru	259	Russian	
	factcheck.kz	58	Russian	
	factpaper.cn	167	Chinese	
	chinafactcheck.com	389	Chinese	
	vp.fact.qq.com	17	Chinese	
News release website	edition.cnn.com	3,890	English	9,998
	bbc.com	739	English	
	tass.ru	2,000	Russian	
	pravda.com.ua	2,430	Ukrainian	
	people.com.cn	702	Chinese	
	xinhuanet.com	237	Chinese	

Table 5: The distribution of data sources for the RU22Fact.

Language	Train	Dev	Test	Total
English	6,082	867	1,741	8,690
Chinese	1,055	152	305	1,512
Russia	1,621	231	465	3,399
Ukrainian	2,458	350	704	2,430
Total	11,217	1,600	3,216	16,033
#Support Labels		10,081		
#Refuted Labels		4,651		
#NEI Labels		1,301		

Table 6: Statistics of RU22Fact.

As a result, we collected 16,033 samples covering four languages: English, Chinese, Russian, and Ukrainian. We split the whole dataset into training, development, and test sets. Detailed statistics of the dataset are illustrated in Table 6. Each entry consists of a real-world claim, optimized evidence, referenced explanation, and meta-data (e.g., date, claimant).

5. Fact-Checking System

In this section, we describe the explainable fact-checking system we built. The framework is illustrated in Figure 2, which consists of three components: Evidence Optimization, Claim Verification, and Explanation Generation. Next, we will describe the details of each component.

5.1. Evidence Optimization

We propose an LLMs-driven method to automatically retrieve and summarize retrieved documents from the Web to produce optimized evidence consisting of some sentences. Specifically, it first queries the search engine by a claim and then scratches the retrieved documents. Aiming to make full use of worthy information and remove irrelevant information, subsequently, designing a prompt carefully to summarize the retrieved documents by a single LLM. In practice, we utilize an LLM that can connect to the Internet, such as New Bing[9] or Spark[10], which can retrieve and summarize documents from the Web, and finish all processes in a single step, taking into account possible error cascades. We query LLM with a carefully designed prompt, such as *"Please list five relevant news and provide detailed sources and content:*$\{$*claim*$\}$*"*. We provide up to five pieces of evidence for each claim in RU22Fact.

5.2. Claim Verification

Based on the optimized evidence, we further design a claim verification module to predict the truthfulness of each input claim. Given an input claim C and its text evidence $E = \{s_1, s_2, ..., s_l\}$, where s_k denotes the k-th sentence in evidence, we utilize a text encoder to encode the claim and sentences in the evidence. We feed claim C and sentences E independently to the text encoder and utilize the representations of CLS tokens as their contextual representations: $X_C \in \mathbb{R}^D$ and $X_E = \{x_{s_1}, x_{s_2}, ..., x_{s_l}\} \in \mathbb{R}^{D \times L}$, where D denotes the embedding size and L is the sentence number in the evidence. We then pair each sentence with the input claim and detect the stance of the sentence towards the claim. As Figure 2 shows, we first compute an attention distribution between the claim and the sentence by using X_C as query, s_k as key and value, to compute cross attention and obtain the stance representation X_{s_k2C}.

$$X_{s_k2C} = Attention(s_k, X_C).$$

[9] https://www.bing.com/new

[10] https://xinghuo.xfyun.cn/spark

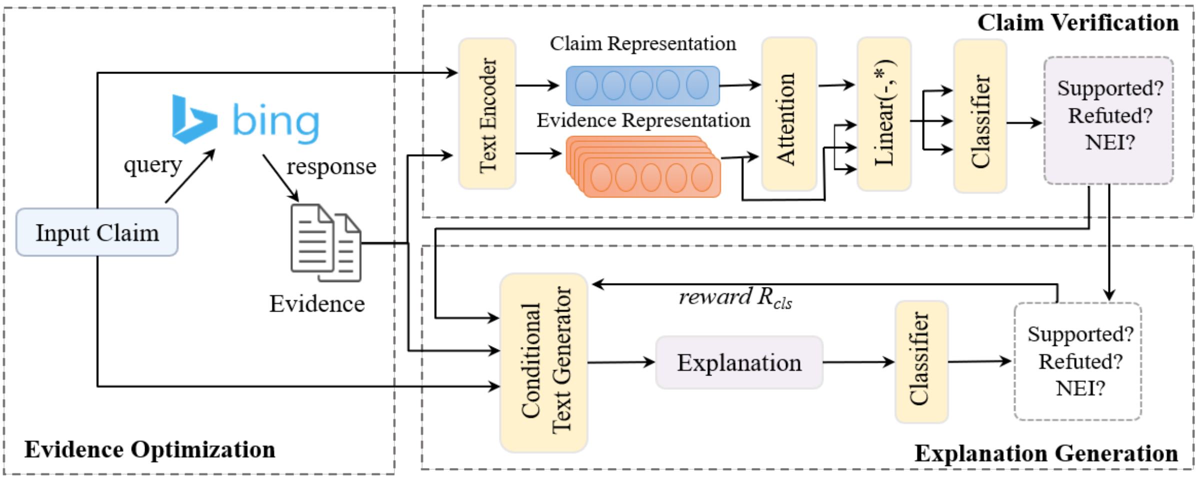


Figure 2: Overview of system framework. It consists of an auto evidence optimization module, a claim verification module, and an explanation generation module.

We further obtain the stance representation H_{s_k2C} of sentence s_k towards claim X_C by concatenating X_{s_k2C} and s_k, feeding them to a linear layer:

$$H_{s_k2C} = Linear(X_{s_k2C} : s_k),$$

where [:] denotes concatenation operation. In the end, we average the overall stance representation and then feed the result to a linear classifier to predict the label with a cross-entropy objective.

5.3. Explanation Generation

To generate a human-understandable explanation for fact-checking prediction, we generate explanations as abstractive summarization and utilize a conditional text generator to generate an explanation by considering the input claim, the predicted label, and the evidence. Further, we incorporate a truthfulness reward based on a classification layer and then optimize the generation model with reinforcement learning to ensure the generated explanation is consistent with the label (Yao et al., 2022). As depicted in Figure 2. Specifically, given an input claim C, label y, and evidence $E = \{s_1, s_2, ..., s_l\}$, we concatenate them into an sequence X. Then we feed X as input to conditional text generator and optimize generator for generating explanation $S = \{s_1, s_2, ...s_q\}$ close to the referenced explanation $\widetilde{S} = \{\widetilde{s}_1, \widetilde{s}_2, ..., \widetilde{s}_q\}$. We take the gold label as input during training and the predicted label during evaluation. The training objective is to minimize the following negative log-likelihood:

$$\mathcal{L}_g = -\sum_i \log(p\left(\tilde{s}_i \mid \tilde{s}_{1:i-1}, X; \phi\right)).$$

To ensure the generated explanation is consistent with the label of the claim, we introduce a truthfulness reward. Specifically, we pre-train a truthfulness classification model, which takes the generated explanation as input and outputs a confidence score for each candidate's label. In practice, we take BERT (Devlin et al., 2019) as a classifier.

$$\boldsymbol{p}(\tilde{y} \mid S) = \operatorname{Softmax}_i\left(\operatorname{classifier}_\theta(S)\right).$$

We take the difference between the confidence score of the correct answer and the wrong answer as reward R_{cls} and apply it to policy learning.

$$R_{cls} = \boldsymbol{p}\left(\tilde{y}_C \mid S\right) - \sum_{\tilde{y}_j \neq \tilde{y}_C} \boldsymbol{p}\left(\tilde{y}_j \mid S\right),$$

where $\tilde{y}_C$ is the gold label of C, $\tilde{y}$ and $\tilde{y}_j$ is the predicted label.

6. Experiment

We conduct experiments to evaluate the performance of two tasks: Claim Verification and Explanation Generation in the proposed dataset RU22Fact.

6.1. Claim Verification

We adopt three different text encoders: 1) Multilingual BERT (FCS_{mBert}), a multilingual variant of BERT (Devlin et al., 2019). 2) XLM-RoBERTa (FCS_{XLM-R}), a multilingual version of RoBERTa pretrained on CommonCrawl data containing 100 languages (Conneau et al., 2019). 3) DistilBERT (FCS_{dBert}), a distilled version of the BERT based multilingual model (Sanh et al., 2019).

To analyze the proposed dataset, we adopt the following different settings to conduct claim verification experiments in the fact-checking system, FCS: 1) only consider the claim without evidence,

FCS_{claim}; 2) only consider evidence without claim, $FCS_{evidence}$; 3) consider the claim with random evidence, FCS_{re}. Random evidence denotes random sampling evidence for each claim in RU22Fact. We utilize FCS_{mBert} as the text encoder for these settings.

The experimental result is assessed against precision (Pr), recall (Rc), and macro F1 metrics. The result is shown in Table 7. The performance of FCS_{claim} performs worse than FCS_{mBert}, indicating that the claim lacks sufficient information for claim verification. FCS_{re} performs worse than FCS_{mBert} and similarly to FCS_{claim}, indicating there is no obvious bias in the evidence. When using random evidence, the model tends to focus on the claim rather than the evidence. $FCS_{evidence}$ performs better than FCS_{claim} and similarly to FCS_{mBert}, indicating there is more useful information in the optimized evidence than in the claim.

According to the results, we find that FCS_{mBert} achieves similar performance compared to FCS_{dBert} and FCS_{XLM-R}, with a macro F1 score of approximately 60%. However, there is still room for further improvement in the proposed dataset RU22Fact. The challenges include low resource language processing, and the label distribution is uneven.

Settings	Pr	Rc	F1
FCS_{claim}	65.71	59.01	57.93
$FCS_{evidence}$	68.07	62.33	59.68
FCS_{re}	55.33	59.84	57.35
FCS_{XLM-R}	57.56	**63.57**	59.91
FCS_{dBert}	**74.30**	63.49	60.40
FCS_{mBert}	58.31	62.91	**60.56**

Table 7: Performance of Claim Verification. The claim and evidence are concatenated and input into the text encoder in FCS, FCS_{XLM-R}, FCS_{dBert}, FCS_{mBert} represent three different text encoders used in the fact-checking system (FCS). (%)

6.2. Explanation Generation

We adopt two different conditional text generators in this section: 1) Bart-large-cnn (FCS_{bart}) (Lewis et al., 2019), a transformer encoder-decoder model with a bidirectional encoder and an autoregressive decoder, fine-tuned on CNN Daily Mail (See et al., 2017). 2) T5-base (FCS_{t5}), a Text-to-Text Transfer Transformer model, which is a versatile and efficient pre-trained model for various natural language processing tasks (Raffel et al., 2020). We also add GPT-3.5-turbo-0613 ($GPT3.5$, an AI chat mode based on the GPT-3.5-series model that generates responses based on user input (OpenAI, 2022), as the interpreter generator for comparison. We fine-tune FCS_{bart} and FCS_{t5} to generate the explanation in the fact-checking system, and prompt $GPT3.5$ to generate the explanation. We use two methods for evaluating the quality of explanations generated: automated evaluation and qualitative evaluation.

Settings	Rouge1	Rouge2	RougeL	BLEU
FCS_{bart}	34.17	16.28	32.08	9.56
FCS_{t5}	32.24	15.17	30.47	8.92
$GPT3.5$	36.56	18.90	34.10	17.05

Table 8: ROUGE and BLEU scores for generated explanation via our explainable fact-checking system. (%)

6.2.1. Automated Evaluation

We evaluate the generated explanation by ROUGE (Lin, 2004) and BLEU (Papineni et al., 2002), and use the F1 values for ROUGE-1, ROUGE-2, and ROUGE-L.

The results are shown in Table 8. From the results, we find that fine-tuned FCS_{bart} performs better than FCS_{t5} in the fact-checking system, and $GPT3.5$ achieves the best performance because part of the referenced explanation in the dataset comes from itself.

6.2.2. Qualitative Evaluation

Evaluation using ROUGE and BLEU does not present a complete picture of the quality of these explanations, therefore, we introduce three desirable coherence properties for machine learning explanations and evaluate the quality of the generated explanations against them (Kotonya and Toni, 2020b). More about these three coherence properties is shown in Appendix A.

- **Strong Global Coherence**. It holds for a generated fact-checking explanation, every sentence in the explanatory text must entail the claim.
- **Weak Global Coherence**. It holds for a generated fact-checking explanation, no sentence in the explanatory text should contradict the claim (by entailing its negation).
- **Local Coherence**. The generated explanation satisfies local consistency if each sentence in the explanatory text does not contradict each other.

We employ human evaluation to assess the quality of the referenced explanation and the generated explanations for these properties. We randomly sampled 100 samples from the test set of RU22Fact, and five annotators to evaluate them according to these three properties.

Method	SGC	WGC	LC
		Human	
$Referenced\ Explanation$	**58.23**	**96.63**	**93.38**
FCS_{bart}	43.97	92.08	90.14
FCS_{t5}	40.29	91.62	90.47
$GPT3.5$	53.29	95.73	90.52
BERT;MNLI			
$Referenced\ Explanation$	37.72	**60.17**	**60.36**
FCS_{bart}	45.52	59.94	53.02
FCS_{t5}	**46.64**	58.77	49.79
$GPT3.5$	37.18	59.74	54.84
RoBERTa;MNLI			
$Referenced\ Explanation$	**24.13**	**94.96**	**90.03**
FCS_{bart}	20.26	93.31	87.51
FCS_{t5}	19.46	94.26	85.41
$GPT3.5$	21.78	94.65	88.56

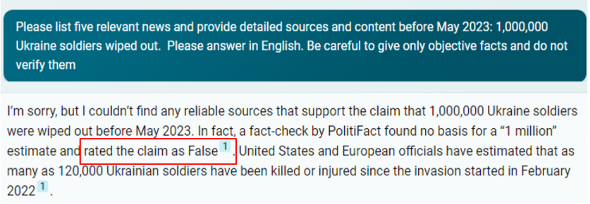
Table 9: The results of the qualitative evaluation in three properties, strong global coherence (*SGC*), weak global coherence (*WGC*), and local coherence (*LC*) properties. (%)

Also, we conduct a computational evaluation of the three properties using NLI (natural language inference (Dagan et al., 2022)). In practice, we use two pretrained NLI models: BERT trained in MNLI (the multi-genre natural language inference corpus (Williams et al., 2018)) and RoBERTa trained in MNLI.

The results of the qualitative evaluation are shown in Table 9. The referenced explanation achieves almost the best results on three properties and three NLI models, which implies that the referenced explanation is of higher quality than others. RoBERTa trained in MNLI performs better than BERT trained in MNLI, which means RoBERTa is a better approximation of these three properties. NLI models are reliable approximations of weak global coherence and local coherence, and they seem to be a poor approximation for strong global coherence.

7. Conclusion

In this paper, we first analyze the challenge of providing sufficient and relevant evidence for fact-checking, and then propose an LLMs-driven method to automatically retrieve and summarize documents from the Web to produce optimized evidence. An analytical experiment indicates that optimized evidence can provide more sufficient and relevant information for building a better fact-checking system. Furthermore, we construct a novel multilingual explainable fact-checking dataset named RU22Fact, including real-world claims, optimized evidence, and referenced explanation. To establish the baseline performance, we build an explainable fact-checking system based on RU22Fact. Experimental results demonstrate the prospect of optimized evidence to increase fact-checking performance including claim verification and explanation generation.



Please list five relevant news and provide detailed sources and content before May 2023: 1,000,000 Ukraine soldiers wiped out. Please answer in English. Be careful to give only objective facts and do not verify them

I'm sorry, but I couldn't find any reliable sources that support the claim that 1,000,000 Ukraine soldiers were wiped out before May 2023. In fact, a fact-check by PolitiFact found no basis for a "1 million" estimate and rated the claim as False [1]. United States and European officials have estimated that as many as 120,000 Ukrainian soldiers have been killed or injured since the invasion started in February 2022 [1].

Figure 3: An example of information leakage, and the red box indicates label leakage.

8. Limitations

Several limitations should be considered in this paper, though this paper provides a step forward in fact-checking.

- **Information Leakage**: There is possible information leakage when retrieving documents from the web. To alleviate this problem, we add some restrictions to a prompt, such as "*Be careful to give only objective facts and do not verify them*". Nevertheless, It sometimes fetches snippets of fact-checking articles if the claim comes from a fact-checking website, which can also lead to information leakage. An example is shown in Figure 3, there is a label leakage in the red box.

- **Low-resource Languages**: The dataset proposed in this work covers claims related to the Russia-Ukraine conflict of 2022, a worldwide topic that is not limited to high-resource languages. However, our work covers only four languages and has less data in non-English languages, which limits fact-checking in low-resource languages.

- **Domain Generalization**: In this article, our data set is a topic related to the Russia-Ukraine conflict. It might not work well in the same way for other topics, and it requires further research.

9. Acknowledgements

We thank the anonymous reviewers for their constructive comments, and gratefully acknowledge the support of the National Natural Science Foundation of China under Grants U22B2059 and 62176079, Natural Science Foundation of Heilongjiang Province under Grant Y02022F005.

10. Bibliographical References

Rami Aly, Zhijiang Guo, Michael Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. Feverous: Fact extraction and verification over unstructured and structured information. In *35th Conference on Neural Information Processing Systems, NeurIPS 2021*. Neural Information Processing Systems foundation.

Rami Aly and Andreas Vlachos. 2022. Natural logic-guided autoregressive multi-hop document retrieval for fact verification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6123–6135, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. Multifc: A real-world multi-domain dataset for evidence-based fact checking of claims. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4685–4697.

Julio Bacio Terracino and Craig Matasick. 2022. Disinformation and russia's war of aggression against ukraine: Threats and governance responses. *Organization for Economic Co-Operation and Development. Available online: https://www. oecd. org/ukraine-hub/policy-responses/disinformation-and-russia-s-war-of-aggression-against-ukraine-37186bde/(accessed on 14 January 2023)*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Ido Dagan, Dan Roth, Fabio Zanzotto, and Mark Sammons. 2022. *Recognizing textual entailment: Models and applications*. Springer Nature.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Ashim Gupta and Vivek Srikumar. 2021. X-fact: A new benchmark dataset for multilingual fact checking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 675–682.

Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. 2019. A richly annotated corpus for different tasks in automated fact-checking. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 493–503.

Andreas Hanselowski, Hao Zhang, Zile Li, Daniil Sorokin, Benjamin Schiller, Claudia Schulz, and Iryna Gurevych. 2018. Ukp-athene: Multi-sentence textual entailment for claim verification. *EMNLP 2018*, page 103.

Xuming Hu, Zhijiang Guo, Guanyu Wu, Lijie Wen, and Philip S Yu. 2023. Give me more details: Improving fact-checking with latent retrieval. *arXiv preprint arXiv:2305.16128*.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. Hover: A dataset for many-hop fact extraction and claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3441–3460.

Omar Khattab, Christopher Potts, and Matei Zaharia. 2021. Baleen: Robust multi-hop reasoning at scale via condensed retrieval. *Advances in Neural Information Processing Systems*, 34:27670–27682.

Neema Kotonya and Francesca Toni. 2020a. Explainable automated fact-checking: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5430–5443.

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

Amrith Krishna, Sebastian Riedel, and Andreas Vlachos. 2022. Proofver: Natural logic theorem proving for fact verification. *Transactions of the Association for Computational Linguistics*, 10:1013–1030.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

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

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740.

Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. 2020. Fine-grained fact verification with kernel graph attention network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7342–7351.

Yi-Ju Lu and Cheng-Te Li. 2020. Gcan: Graph-aware co-attention networks for explainable fake news detection on social media. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 505–514.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6859–6866.

OpenAI. 2022. About openai.

OpenAI. 2023. Gpt-4 technical report.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140).

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Tal Schuster, Roei Schuster, Darsh J Shah, and Regina Barzilay. 2020. The limitations of stylometry for detecting machine-generated fake news. *Computational Linguistics*, 46(2):499–510.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: A large-scale dataset for fact extraction and verification. In *2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2018*, pages 809–819. Association for Computational Linguistics (ACL).

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 workshop on language technologies and computational social science*, pages 18–22.

Hai Wan, Haicheng Chen, Jianfeng Du, Weilin Luo, and Rongzhen Ye. 2021. A dqn-based approach to finding precise evidences for fact verification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1030–1039.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Lianwei Wu, Yuan Rao, Yongqiang Zhao, Hao Liang, and Ambreen Nazir. 2020. Dtca: Decision tree-based co-attention networks for explainable claim verification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1024–1035.

Wenhan Xiong, Xiang Li, Srini Iyer, Jingfei Du, Patrick Lewis, William Yang Wang, Yashar Mehdad, Scott Yih, Sebastian Riedel, Douwe Kiela, et al. Answering complex open-domain questions with multi-hop dense retrieval. In *International Conference on Learning Representations*.

Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. 2022. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. *arXiv preprint arXiv:2205.12487*.

Appendix

Appendix A. Coherence Property

Figure 4 demonstrates the three coherence properties schematically in graphical form. Figure 5 demonstrates examples of the three coherence properties.

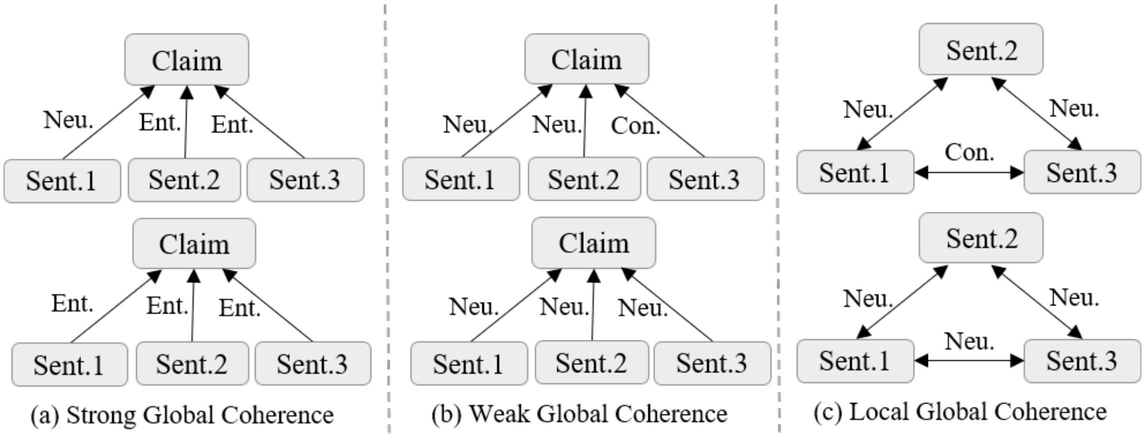
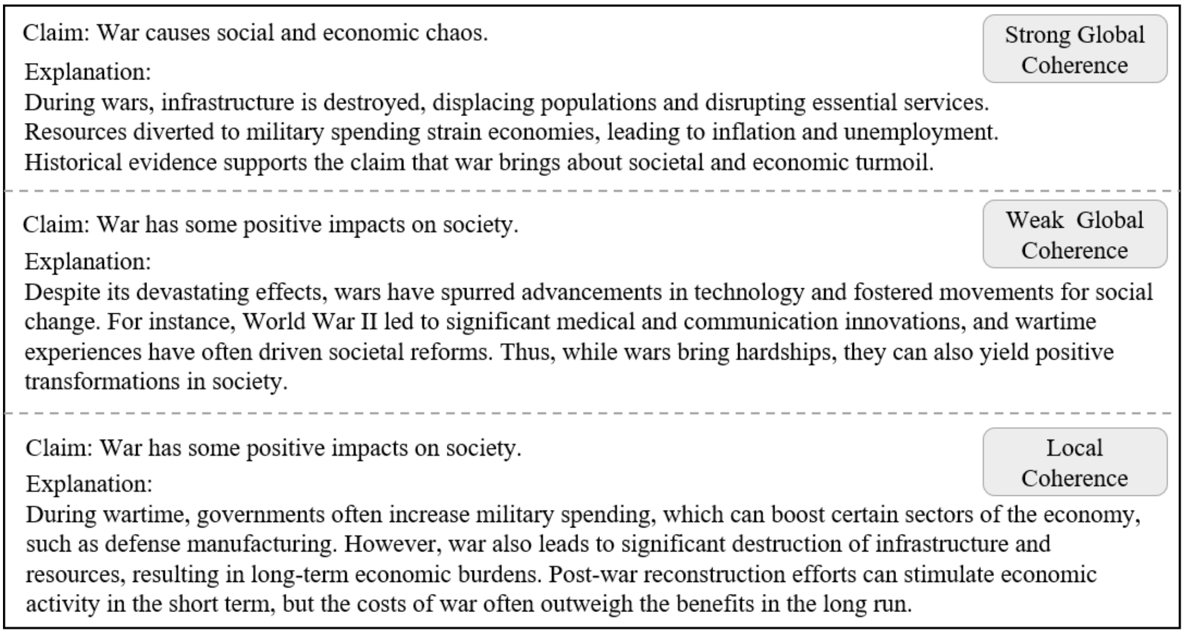


Figure 4: Schematic representations of strong global coherence, weak global coherence and local coherence. ”Neu.”, ”Ent.” and ”Con.” are abbrs for neutral, entails and contradicts. In each column, the upper part means coherence cannot be satisfied, and the lower part means coherence is satisfied.



Claim: War causes social and economic chaos.

Strong Global Coherence

Explanation:
During wars, infrastructure is destroyed, displacing populations and disrupting essential services.
Resources diverted to military spending strain economies, leading to inflation and unemployment.
Historical evidence supports the claim that war brings about societal and economic turmoil.

Claim: War has some positive impacts on society.

Weak Global Coherence

Explanation:
Despite its devastating effects, wars have spurred advancements in technology and fostered movements for social change. For instance, World War II led to significant medical and communication innovations, and wartime experiences have often driven societal reforms. Thus, while wars bring hardships, they can also yield positive transformations in society.

Claim: War has some positive impacts on society.

Local Coherence

Explanation:
During wartime, governments often increase military spending, which can boost certain sectors of the economy, such as defense manufacturing. However, war also leads to significant destruction of infrastructure and resources, resulting in long-term economic burdens. Post-war reconstruction efforts can stimulate economic activity in the short term, but the costs of war often outweigh the benefits in the long run.

Figure 5: Examples of strong global coherence, weak global coherence and local coherence.