



中国科学院自动化研究所  
模式识别实验室  
New Laboratory of Pattern Recognition



ACL 2024  
*Bangkok, Thailand*

# EX-FEVER: A Dataset for Multi-hop Explainable Fact Verification

Huanhuan Ma, Weizhi Xu, Yifan Wei, Liuji Chen, Liang Wang,  
Qiang Liu, Shu Wu, Liang Wang

# Fact Verification

---

- In the era of mobile internet, the proliferation of misleading or fake information has raised great concern in human society.
- Fact verification, also known as fact-checking, is the task of determining the veracity of claims by finding supporting evidence.
- It plays a crucial role in combating misinformation and maintaining the integrity of public discourse.
- Current research on automatic fact verification, using deep learning methods, focuses only on accuracy improvement while neglecting explainability, a crucial ability of an automatic fact verification system.

# EX-FEVER Dataset Overview

Datasets	Hops	Explainable	Class
HOVER	2-3-4	✗	2
FEVER	1-2	✗	3
e-FEVER	1-2	✓	3
EX-FEVER	2-3	✓	3

Table 1: Related Datasets Comparison

## Claim

John Mayer is an American singer-songwriter whose debut EP was later re-released by an American record label owned by Sony Music Entertainment.

## Golden Explanation

John Mayer is an American singer-songwriter who released his first extended play, *Inside Wants Out*. *Inside Wants Out* is the debut EP by John Mayer that was later re-released by Columbia Records. Columbia Records is an American record label owned by Sony Music Entertainment.

## Golden Document

John Mayer, *Inside Wants Out*, Columbia Records

**Label SUPPORT**

Key features: 60,000+ complex multi-hop claims

Dataset composition:

- Verification labels (SUPPORTS, REFUTES, NOT ENOUGH INFO)
- Explanatory annotations

# Dataset construction

- We use Top 50,000 popular Wikipedia pages and Create multi-hop reasoning paths using hyperlinks.
- We hired annotators, trained them with detailed guidelines
- We reviewed and refined their work through quality inspections. In total, we collected 60,000 annotated claims, each with a verdict and explanation

The screenshot displays a web interface for dataset construction. At the top, three document panels are shown side-by-side:

- The\_Woman\_in\_Red\_(1984\_film)**: A paragraph about the 1984 film, mentioning Gene Wilder, Charles Grodin, Gilda Radner, Joseph Bologna, Judith Ivey, and Kelly LeBrock. The name 'Kelly LeBrock' is highlighted in yellow.
- Kelly\_LeBrock**: A paragraph about the actress, mentioning her debut in 'The Woman in Red' and other films like 'Weird Science' and 'Hard to Kill'.
- Hard\_to\_Kill**: A paragraph about the 1990 film, mentioning Bruce Malmuth, Steven Seagal, and the plot.

Below the documents, there are six text boxes for claim verification, arranged in a 2x3 grid:

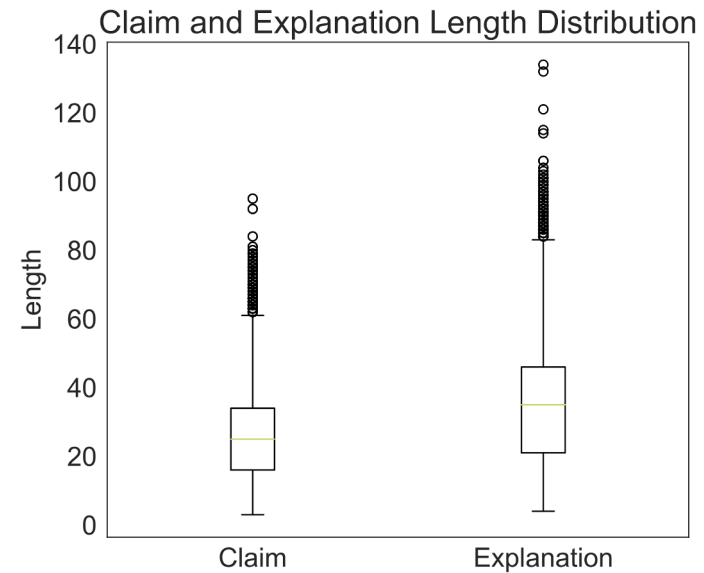
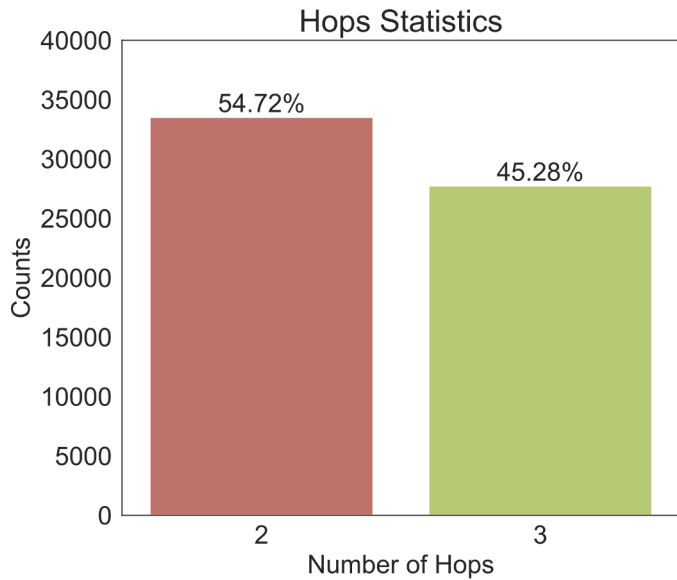
- Claim\_SUP**: The Woman in Red is a romantic comedy film including the American actress and model who also starred in a 1990 American action thriller film directed by Bruce Malmuth.
- Explanation\_SUP**: The Woman in Red is a romantic comedy film including Kelly LeBrock. Kelly LeBrock is an American actress and model who also starred in the film Hard to kill. Hard to Kill is a 1990 American action thriller film directed by Bruce Malmuth.
- Claim\_REF**: The Woman in Red is a 1984 American romantic comedy film including the American actress and model who also starred in a 1990 American horror-thriller film directed by Bruce Malmuth.
- Explanation\_REF**: The Woman in Red is a 1984 American romantic comedy film including Kelly LeBrock. Kelly LeBrock is an American actress and model who also starred in the film Hard to kill. Hard to Kill is a 1990 American action, not horror thriller film directed by Bruce Malmuth.
- Claim\_NEI**: NO.2document is discarded. The Woman in Red is a romantic comedy film including the American actress and model who also starred in a 1990 American action thriller film directed by Bruce Malmuth.
- Explanation\_NEI**: The Woman in Red is a romantic comedy film including Kelly LeBrock. Hard to Kill is a 1990 American action thriller film directed by Bruce Malmuth. No information shows that Kelly LeBrock is an American actress and model who also starred in the film Hard to kill.

At the bottom right, there are two buttons: 'Valid' (highlighted in blue) and 'Invalid'. Below this is a horizontal scrollbar. At the very bottom, there is a '质检反馈' (Quality Inspection Feedback) section with a timestamp '质检1 06.13 14:29'.

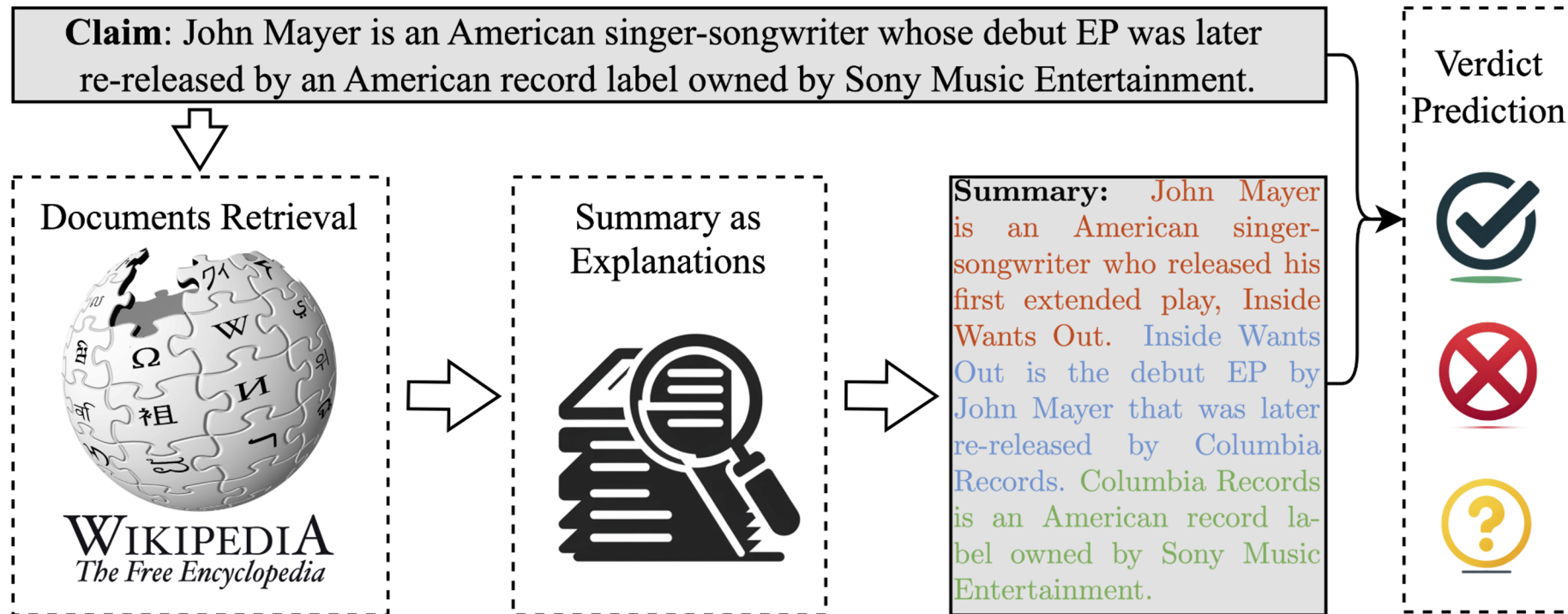
# Dataset Characteristics

Table1: Data Statistics with different number of hops and different label classes. The average claim length and explanation length in word level are reported.

Hops	SUP	REF	NEI	Claim	EXP
2 Hops	11053	11059	11412	21.63	28.39
3 Hops	9337	9463	8941	30.69	43.45
Total	20390	20522	20353	25.73	35.21



# Benchmark System Architecture



The baseline system comprises three stages: document retrieval, summary generation as explanations, and verdict prediction. The system produces two main outputs: a veracity label and a summary that serves as an explanation for the prediction.

# Experimental Setup

---

- **Model selection**

## Document Retrieval

- Rule-based: TF-IDF
- Neural-based:
  - BERT-based model
  - MDR model

## Explanatory Stage

- BART Fine-tuned on dataset's training split

## Verdict Prediction

- BERT
- GEAR - graph-based text reasoning model

- **Evaluation Metrics**

## Document Retrieval

- Exact match score (EM)
- Recall@k

## Explanation Generation

- ROUGE score

## Verification

- Accuracy
- F1 score

# Results & Analysis

Table2: Retrieve Model Performance Comparison

Model	EM	Hit@6	Hit@12	Hit@30
MDR	43.3	55.00	60.90	68.60
BERT-based	32.4	66.12	70.28	73.98

Table3: Generated Summary Metrics Comparison

Model	Length	rouge1	rouge2	rougeL	rougeLsum
MDR	54.79	54.88	41.34	49.42	53.02
BERT-based	46.05	46.88	32.80	35.52	44.41
Explanation from ChatGPT					
GPT-0example	58.05	52.28	33.74	48.13	49.89
GPT-3example	48.56	59.98	42.85	57.66	55.61

Table4: Verify Model Comparison. The accuracy (%) of each model is reported

Model	Val	Test	Test On Golden	Train With Golden
Gear@BERT-based	54.96	54.71	53.08	61.05
Gear@MDR	59.68	58.89	53.98	-
BERT@BERT-based	68.07	67.65	76.69	99.29
BERT@MDR	73.86	73.34	76.89	-
HOVER@MDR	46.58	45.41	33.79	-

## Overall Conclusions

1. Retrieval model quality significantly impacts system performance
2. EM score is crucial due to text generation model constraints
3. Current graph-based methods may lack true reasoning capabilities
4. High-quality explanations are vital for accurate verdict prediction



# Large Language Model Exploration

---

**LLMs as actors:** direct fact-checking

**LLMs as planners:** decomposing complex claims

Table4: Use LLM as an actor or a planner. The accuracy (\%) of each model is reported.

Type	Model	Close	Open	Gold
Actor	ClaimOnly	45.78	-	-
	w/o exp	-	-	47.91
	w/ exp	-	-	47.92
	1 shot	-	-	47.91
	3 shots	-	-	58.69
Planner	ProgramFc	47.30	51.70	64.90

## Findings:

Despite extensive training data, LLMs require additional knowledge to perform well on this task. Incorporating few-shot examples proves effective. Large models excel in generating guides to assist other models in making judgments, rather than making predictions directly.

# Conclusion

---

- **Dataset Introduction:** We present a publicly accessible fact-checking dataset, EX-FEVER, with over 60,000 multi-hop claims and detailed annotations for understanding veracity assessments.
- **System Design:** Our comprehensive system includes retrieval, summarization for explanation, and verification stages, highlighting the dataset's significance.
- **LLM Investigations:** Preliminary studies with the GPT-3.5-turbo model show that using LLMs as planners yields better performance than as actors, particularly in generating explanations.
- **Improvement Potential:** Despite the capabilities of LLMs, there is substantial room for enhancement in the fact-checking process.
- **Benchmarking Value:** EX-FEVER serves as a crucial benchmark for advancing explainable multi-hop fact-checking, aiding in reliability and informed decision-making across various fields.

Thanks