



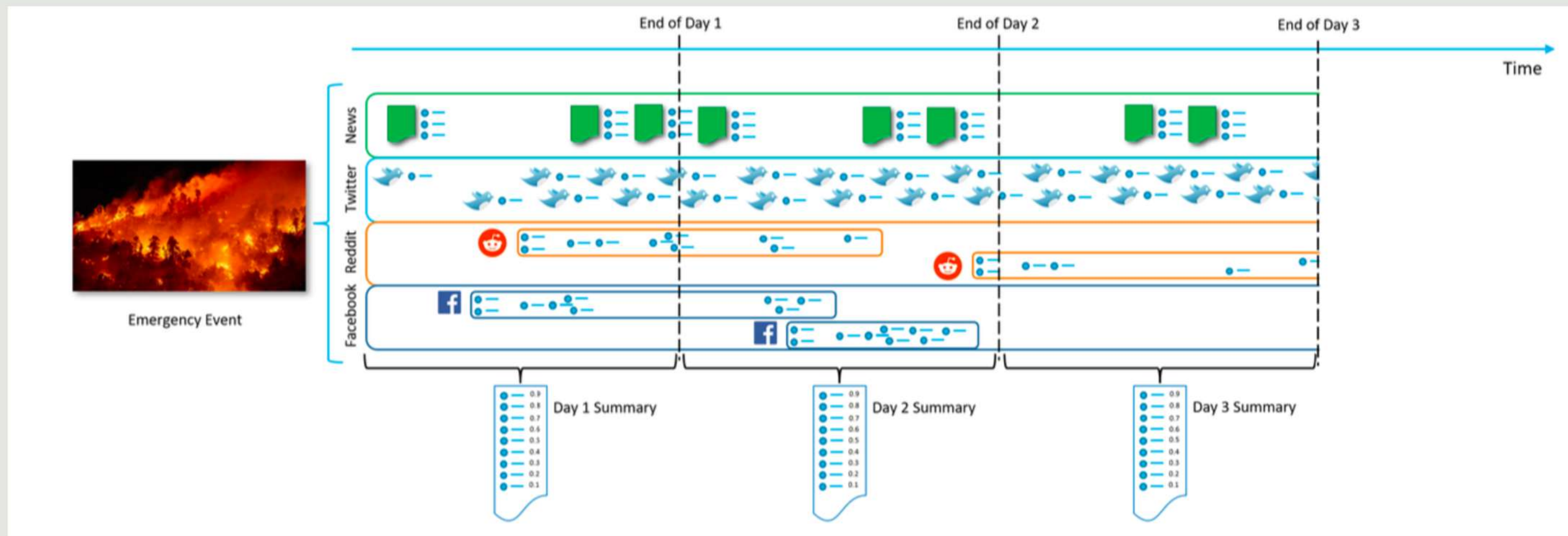
CRISISFACTS: DISASTER SUMMARIZATION

Presented by RescueRetrievers

INTRODUCTION

During crises, effective disaster response relies on timely and accurate information for dedicated personnel managing emergencies. Local crisis response teams, crucial in coordinating on-the-ground efforts, need transparent communication with stakeholders. To address this, the CrisisFACTS framework utilizes online content, creating real-time timelines to answer the key question: "What's happening on the ground?". The CrisisFACTS Track addresses these challenges, urging the research community to develop systems adept at multi-stream fact-finding and summarization.

CORE CRISIS SUMMARIZATION TASK



DATASET:

- **Event Definitions:** We provide content for multiple crises, including wildfires, hurricanes and flood events. The following is an example event definition:

```
{
  "eventID": "CrisisFACTS-001",
  "trecisId": "TRECIS-CTIT-H-092",
  "dataset": "2017_12_07_lilac_wildfire.2017",
  "title": "Lilac Wildfire 2017",
  "type": "Wildfire",
  "url": "https://en.wikipedia.org/wiki/Lilac_Fire",
  "description": "The Lilac Fire was a fire that burned
}
```

- **User Profiles:** The itemised list of general and event-type-specific queries representing a responder's information needs. These queries provide a method for filtering disaster-related content to only ICS-related facts. We also include TREC-IS category mappings for each query. Example queries include:

```
[{
  "queryID": "CrisisFACTS-General-q001",
  "indicativeTerms": "airport closed",
  "query": "Have airports closed",
  "trecisCategoryMapping": "Report-Factoid"
},
{
  "queryID": "CrisisFACTS-General-q002",
  "indicativeTerms": "rail closed",
  "query": "Have railways closed",
  "trecisCategoryMapping": "Report-Factoid"
},
{
  "queryID": "CrisisFACTS-General-q003",
  "indicativeTerms": "water supply",
  "query": "Have water supplies been contaminated",
  "trecisCategoryMapping": "Report-EmergingThreats"
},
...
{
  "queryID": "CrisisFACTS-Wildfire-q001",
  "indicativeTerms": "acres size",
  "query": "What area has the wildfire burned",
  "trecisCategoryMapping": "Report-Factoid"
},
{
  "queryID": "CrisisFACTS-Wildfire-q002",
  "indicativeTerms": "wind speed",
  "query": "Where are wind speeds expected to be high",
  "trecisCategoryMapping": "Report-Weather"
},
...
]
```

DATASET:

- **Summary Requests:** Each request lists the event ID, date to summarise, and the start and end timestamps bounding the requested summary. A multi-day disaster will have multiple such summary requests:

```
[{
  "eventID": "CrisisFACTS-001",
  "requestID": "CrisisFACTS-001-r3",
  "dateString": "2017-12-07",
  "startUnixTimestamp": 1512604800,
  "endUnixTimestamp": 1512691199
},
...,
{
  "eventID": "CrisisFACTS-001",
  "requestID": "CrisisFACTS-001-r4",
  "dateString": "2017-12-08",
  "startUnixTimestamp": 1512691200,
  "endUnixTimestamp": 1512777599
}]
```

Fig 4. Example Summary Requests

- **Content Streams:** A set of content snippets from online streams. Each snippet is approximately a sentence in length and may not be relevant to the event—i.e., **content streams are noisy**. Each snippet has an associated event ID, stream identifier (*CrisisFACTS-Event ID-Stream Name-Stream ID-Snippet*), source, timestamp and a piece of text. Your summary should be built from these items.

```
[{
  "event": "CrisisFACTS-001",
  "streamID": "CrisisFACTS-001-Twitter-14023-0",
  "unixTimestamp": 1512604876,
  "text": "Big increase in the wind plus drop in humidity",
  "sourceType": "Twitter"
},
{
  "event": "CrisisFACTS-001",
  "streamID": "CrisisFACTS-001-Twitter-27052-0",
  "unixTimestamp": 1512604977,
  "text": "Prayers go out to you all! From surviving 2 ma",
  "sourceType": "Twitter"
},
{
  "event": "CrisisFACTS-001",
  "streamID": "CrisisFACTS-001-Twitter-43328-0",
  "unixTimestamp": 1512691164,
  "text": "If you're in the San Diego area (or north of i",
  "sourceType": "Twitter"
}]
```

Fig 5. Three Event Snippets for Event CrisisFACTS-001

DATASET:

Example Abstractive Output

Examples of system output are as follows:

```
"requestID": "CrisisFACTS-001-r3",
"factText": "Increased threat of wind damage in the San D
"unixTimestamp": 1512604876,
"importance": 0.71,
"sources": [
  "CrisisFACTS-001-Twitter-14023-0"
],
"streamID": null,
"informationNeeds": ["CrisisFACTS-General-q015"]
```

Fig 6. Example System Output with Abstractive Facts. The `streamID` field is empty as this fact may not appear in the dataset verbatim. It is, however, supported by one Twitter message.

Example Extractive Output

```
{
  "requestID": "CrisisFACTS-001-r3",
  "factText": "Big increase in the wind plus drop in hu
  "unixTimestamp": 1512604876,
  "importance": 0.71,
  "sources": [
    "CrisisFACTS-001-Twitter-14023-0"
  ],
  "streamID": "CrisisFACTS-001-Twitter-14023-0",
  "informationNeeds": ["CrisisFACTS-General-q015"]
}
```

Fig 7. Example System Output with Extractive Facts. The `streamID` field is populated with the Twitter document from which this text was taken.

OUR APPROACH

- **RETRIEVAL AND RANKING:** WE USED THE WEIGHT MODEL AS BM25 AND THEN GAVE THE IMPORTANCE SCORE TO THE DOCUMENT/MULTI-STREAM DATA COLLECTION.
- **RERANKING:** WE COUNT THE IMPORTANCE SCORE AS A RERANKING PROCESS.
- **SUMMARIZATION:**WE CONDUCTED TWO TYPES OF SUMMARIZATION—EXTRACTIVE AND ABSTRACTIVE—EMPLOYING MODELS SUCH AS BART, BERT.

RETRIEVAL

- **INPUT DATA PROCESSING:**

- SOURCES: TWITTER, FACEBOOK, REDDIT, AND NEWS ARTICLES.
- STEPS: TEXT NORMALIZATION (REMOVE URLS, HASHTAGS), TOKENIZATION, AND METADATA EXTRACTION (TIMESTAMPS, STREAM IDS, EVENT IDS).

- **QUERY-BASED RETRIEVAL:**

- MATCH RESPONDER QUERIES (E.G., "HAVE AIRPORTS CLOSED?") WITH MESSAGES.
- BM25 MODEL: RANK MESSAGES BASED ON RELEVANCE USING TERM FREQUENCY, INVERSE DOCUMENT FREQUENCY, AND DOCUMENT LENGTH.

- **IMPORTANCE SCORING: (BASELINE)**

- QUERY MATCH COUNT: COUNT QUERIES EACH MESSAGE MATCHES.
- RELEVANCE SCORE SUM: AGGREGATE BM25 SCORES ACROSS QUERIES TO COMPUTE OVERALL IMPORTANCE

RE-RANKING & SUMMARIZATION

- **RE-RANKING:**

- REFINE RANKINGS USING CUMULATIVE RELEVANCE SCORES FROM THE IMPORTANCE SCORING STAGE.
- OUTPUT: A REFINED LIST OF RELEVANT MESSAGES.

- **EXTRACTIVE SUMMARIZATION:**

- TOP-K SELECTION: CHOOSE THE TOP-K MESSAGES (E.G., TOP 50) BASED ON SCORES.
- DIRECT INCLUSION: INCLUDE SELECTED MESSAGES VERBATIM IN THE SUMMARY.

SUMMARIZATION

- **QUERY-BASED RETRIEVAL**
 - **INITIAL FILTERING WITH BM25:**
 - USES BM25 TO RETRIEVE A SMALLER, RANKED SUBSET OF MESSAGES RELEVANT TO THE QUERY (E.G., TOP 200).
 - ENSURES EFFICIENCY BY NARROWING THE DATASET FOR DEEPER ANALYSIS.
- **IMPORTANCE SCORING WITH BERT**
 - **QUERY-FACT PAIRING:**
 - EACH MESSAGE IS PAIRED WITH RELEVANT QUERIES (E.G., QUERY: "WHAT AREAS ARE BEING EVACUATED?").
 - **BERT RELEVANCE SCORING:**
 - - COMPUTES SEMANTIC SIMILARITY BETWEEN QUERY AND MESSAGE USING CONTEXTUAL EMBEDDINGS.
- **CUMULATIVE IMPORTANCE:**
 - AGGREGATE SCORES FOR MESSAGES MATCHING MULTIPLE QUERIES TO CALCULATE TOTAL IMPORTANCE.

SUMMARIZATION

- **RE-RANKING WITH BERT**
 - **RE-RANKING:**
 - FACTS BASED ON CUMULATIVE BERT SCORES (E.G., HIGHER SEMANTIC RELEVANCE = HIGHER RANK).
- **EXTRACTIVE SUMMARIZATION**
 - **TOP-K SELECTION:**
 - SELECT THE TOP-K MESSAGES (E.G., TOP 50) FOR THE SUMMARY BASED ON RE-RANKED SCORES.

SUMMARIZATION

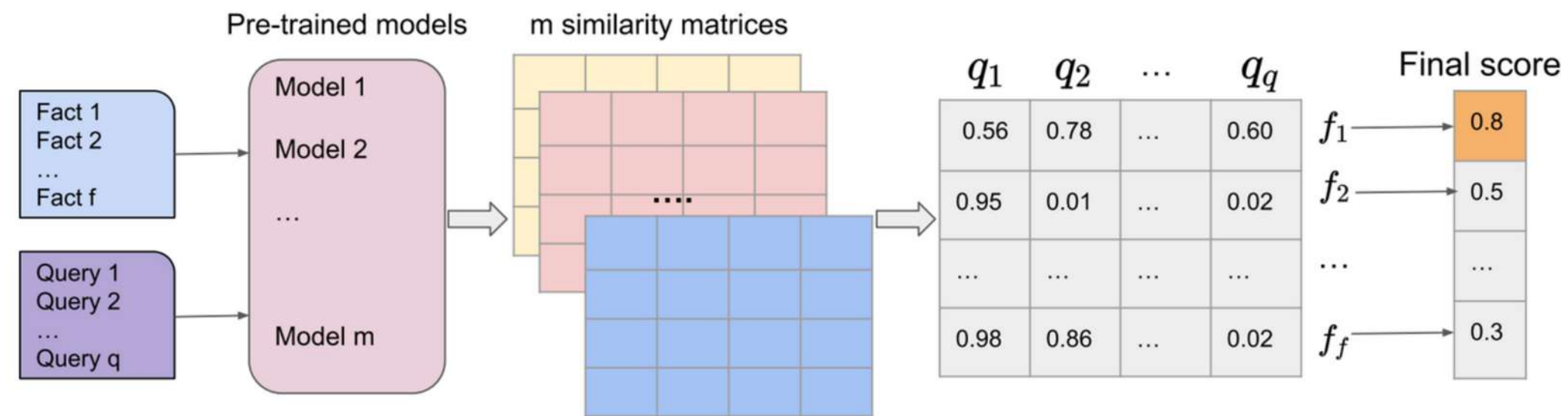


Fig. 1: Overview of steps to compute importance scores of facts.

SUMMARIZATION

- **ABSTRACTIVE SUMMARIZATION:**

- **TOP-K SELECTION:**

- WE HAVE USED THE BASELINE IMPORTANCE SCORE AND RE-RANKING APPROACH HERE.
 - BASED ON THE IMPORTANCE SCORE WE ARE GOING TO FIND THE TOP-K MOST RELEVANT FACT

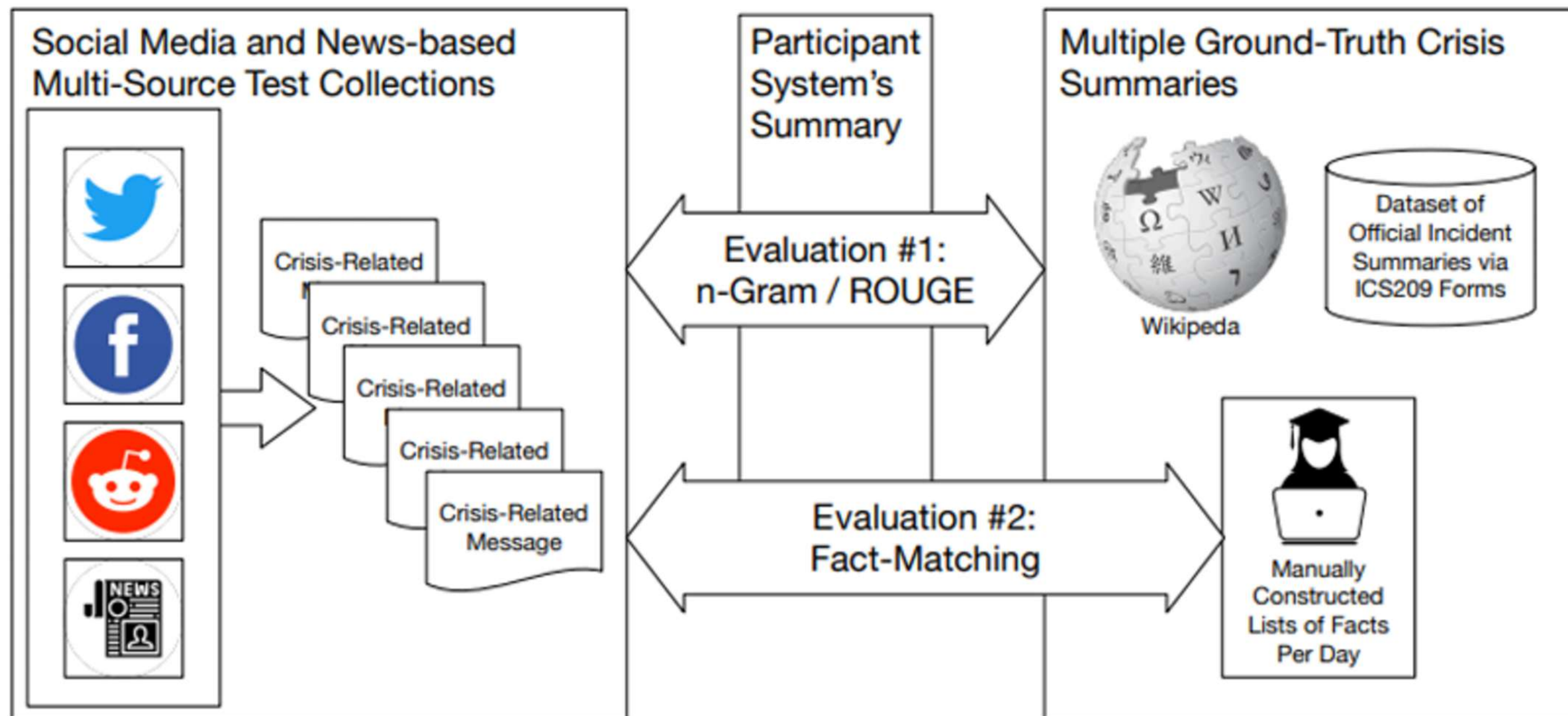
- **ABSTRACTIVE SUMMARIZATION:**

- **BART:** A SEQUENCE-TO-SEQUENCE MODEL FINE-TUNED FOR SUMMARIZATION TASKS. EFFECTIVE FOR GENERATING COHERENT AND FLUENT SUMMARIES.

- **MODEL INPUT:**

- CONCATENATE THE SELECTED MESSAGES INTO A SINGLE INPUT TEXT.
 - FEED THE INPUT TEXT TO THE ABSTRACTIVE SUMMARIZATION MODEL.

TREC'S EVALUATION



AUTOMATIC EVALUATION BY TREC

- **TREC'S AUTOMATIC EVALUATION PROCESS SELECTS THE TOP-K DOCUMENTS ESSENTIAL FOR SUMMARIZING EACH DAY OF AN EVENT, WITH K BEING DYNAMIC FOR EACH DAY.**
- **THESE DAILY TOP-K DOCUMENTS ARE THEN COMBINED TO CREATE A COMPREHENSIVE SUMMARY FOR THE ENTIRE EVENT.**
- **TREC EVALUATES THE GENERATED SUMMARIES BY COMPARING THEM TO GOLD-STANDARD SUMMARIES, INCLUDING ICS, NIST, AND WIKIPEDIA SUMMARIES.**
- **THE EVALUATION SYSTEM PROVIDES BERT-SCORE AND ROUGE-SCORE FOR EACH SUMMARY RUN, OFFERING INSIGHTS INTO THE PERFORMANCE OF THE SUMMARIZATION METHODS.**

ROUGE-SCORE

ROUGEScore: RECALL-ORIENTED UNDERSTUDY FOR GISTING EVALUATION, WIDELY USED FOR TEXT SUMMARIZATION EVALUATION.

- **ROUGEMETRICS:**

- **ROUGE-1 (UNIGRAM OVERLAP)**: MEASURES THE PRECISION AND RECALL OF UNIGRAMS (INDIVIDUAL WORDS) IN THE GENERATED SUMMARY COMPARED TO THE REFERENCE SUMMARY.

- **ROUGE-2 (BIGRAM OVERLAP)**: SIMILAR TO ROUGE-1 BUT EVALUATES BIGRAMS, CAPTURING THE PRECISION AND RECALL OF CONSECUTIVE WORD PAIRS.

- **ROUGE-L (LONGEST COMMON SUBSEQUENCE)**: MEASURES THE OVERLAP OF THE LONGEST COMMON SUBSEQUENCE (LCS) BETWEEN THE GENERATED AND REFERENCE SUMMARIES, PROVIDING A MEASURE OF CONTENT OVERLAP.

BERT-SCORE

- **THE BERT-SCORE**, PIVOTAL IN SUMMARIZATION EVALUATION, IS CALCULATED BY COMPARING TOKENIZED WORDS BETWEEN SYSTEM-GENERATED SUMMARIES AND GOLDSTANDARDS.
- THIS METRIC DELVES IN TO **PRECISION, RECALL, AND F1 SCORE**, OFFERING ANUANCED ASSESSMENT OF SUMMARIZATION QUALITY. PRECISION GAUGES THE ACCURACY OF IDENTIFIED TOKENS, RECALL MEASURES RELEVANCE CAPTURE, AND F1 SCORE BALANCES BOTH ASPECTS.
- BERT-SCORE'S IMPORTANCE LIES IN ITS ABILITY TO PROVIDE COMPREHENSIVE INSIGHTS BEYOND BASIC OVERLAP METRICS, MAKING IT A VALUABLE TOOL FOR REFINING AND OPTIMIZING SUMMARIZATION MODELS, ULTIMATELY ENSURING THE GENERATION OF HIGH-QUALITY AND CONTEXTUALLY RELEVANT SUMMARIES

COMPARISON OF OUR BEST AND TREC'S BEST RESULT

- IN OUR ICS SUMMARIZATION, OUR BEST BERT-SCORE STANDS AT 0.4431, CLOSELY TRAILING TREC'S BEST AT 0.4591, WHILE OUR BEST ROUGE SCORE REACHES 0.0515 COMPARED TO TREC'S BEST AT 0.0581. BOTH SCORES EXHIBIT PROMISING AND COMMENDABLE PERFORMANCE.
- MOVING TO NIST SUMMARIES ,OUR IMPLEMENTED CODE YIELDS NOT ABLE RESULTS WITH A BEST BERT-SCORE OF 0.5566 AND A ROUGE SCORE OF 0.1307. INCOMPARISON, TREC'S BEST SCORES SLIGHTLY EDGE OURS WITH 0.5642 FOR BERT AND 0.1471 FOR ROUGE. NEVER THE LESS,OUR IMPLEMENTATION COMPETES ADMIRABLY, SHOW CASING ROBUST PERFORMANCE.
- FOR WIKIPEDIA SUMMARIES,OUR BEST BERT-SCORE AND ROUGE SCORE HIT 0.5273 AND 0.0267, RESPECTIVELY. WHILE TREC'S BEST SCORES ARE HIGHER AT 0.5646 FOR BERT AND 0.0362 FOR ROUGE,OUR RESULTS HOLD THEIR GROUND, PERFORMING WELL IN COMPARISON TO SOME OF THE TOP SUBMISSIONS.

RESULTS

Baseline	ICS	NIST	WIKI
Bert_score/Rouge	0.4403/0.0398	0.5422/0.1229	0.5106/0.0248
TREC Best	0.4591/0.0581	0.5642/0.1471	0.5646/0.0362

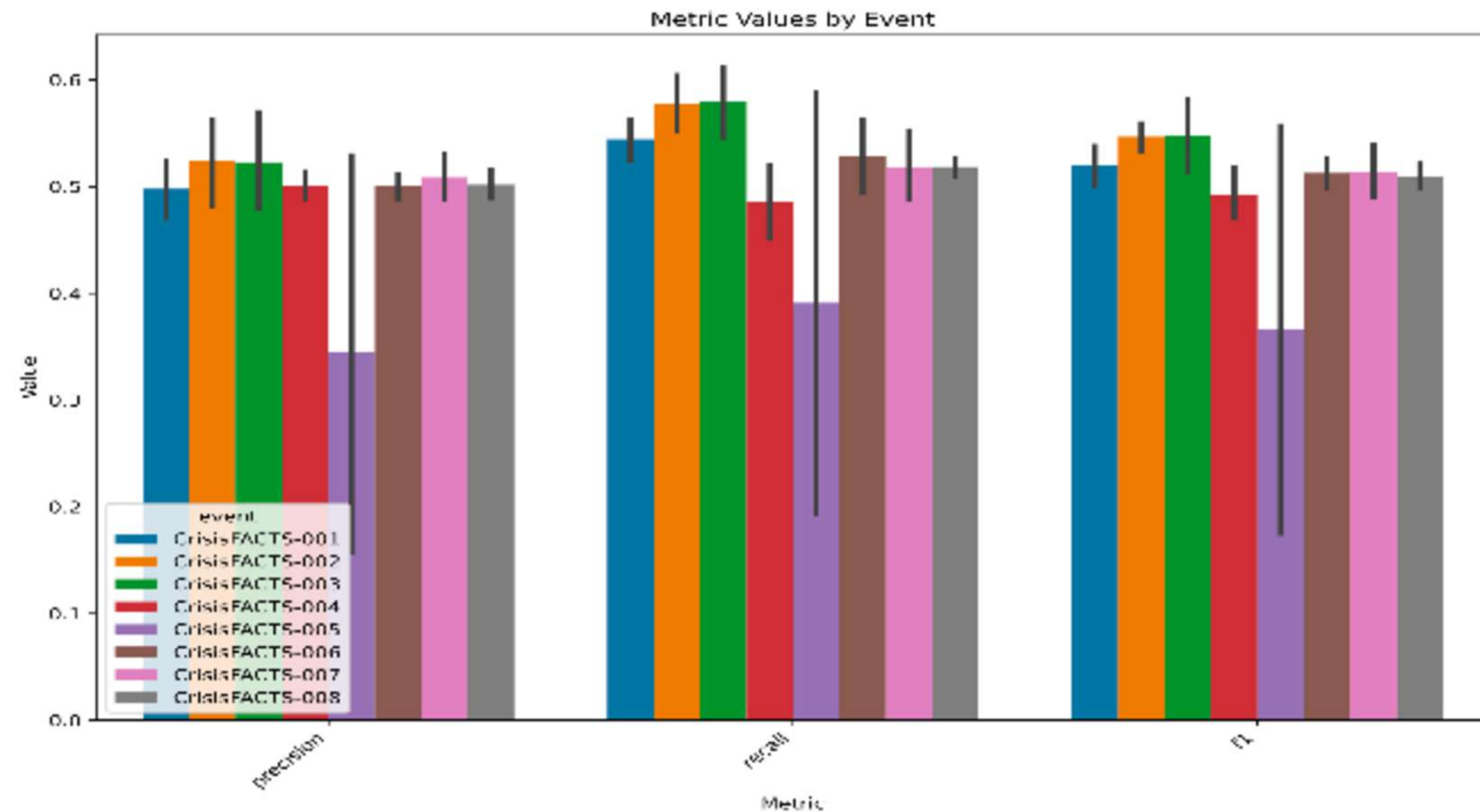
Bert	ICS	NIST	WIKI
Bert_score/Rouge	0.4431/0.0418	0.5566/0.1307	0.5273/0.0267
TREC Best	0.4591/0.0581	0.5642/0.1471	0.5646/0.0362

BART	ICS	NIST	WIKI
Bert_score/Rouge	0.4428/0.0515	0.5271/0.1146	0.5046/0.0213
TREC Best	0.4591/0.0581	0.5642/0.1471	0.5646/0.0362

OBSERVATIONS

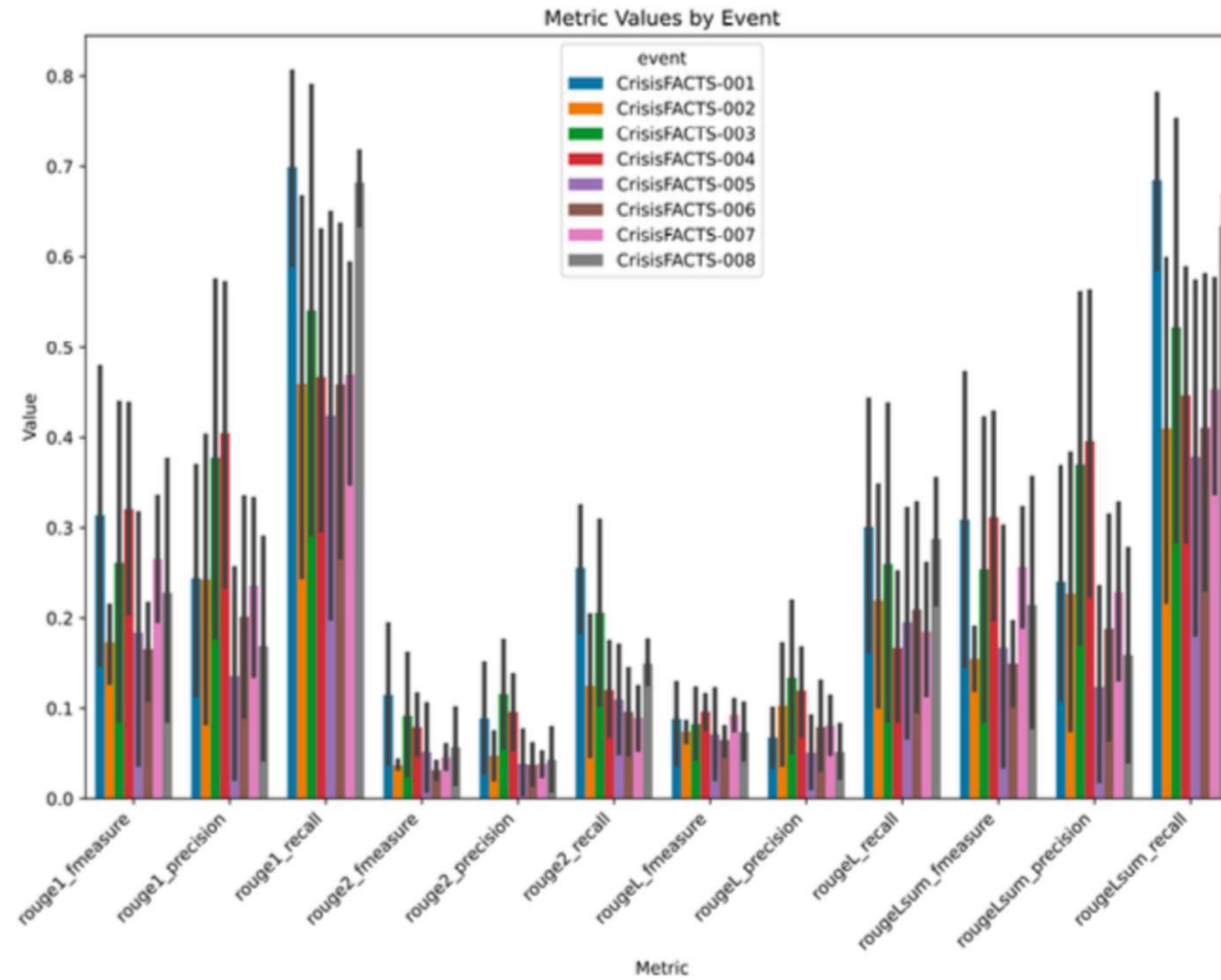
- THE EXTRACTIVE SUMMARIZED RUNS DEMONSTRATE SUPERIOR RESULTS COMPARED TO THE ABSTRACTIVE SUMMARIZED RUNS.
- THIS DISCREPANCY CAN BE ATTRIBUTED TO THE FACT THAT, IN ABSTRACTIVE SUMMARIZATION, MULTIPLE MODELS ARE EMPLOYED TO ABSTRACT SENTENCES.
- ABSTRACTIVE SUMMARIZATION USES MULTIPLE MODELS TO CREATE SUMMARIES, BUT DURING EVALUATION, THE GOLD SUMMARIES (THE BEST POSSIBLE SUMMARIES) ARE VERY DETAILED AND SPECIFIC. THIS MAKES IT CHALLENGING FOR THE ABSTRACTIVE MODELS TO MATCH THE LEVEL OF DETAIL AND PRECISION IN THE GOLD STANDARDS.

TREC'S EVALUATION

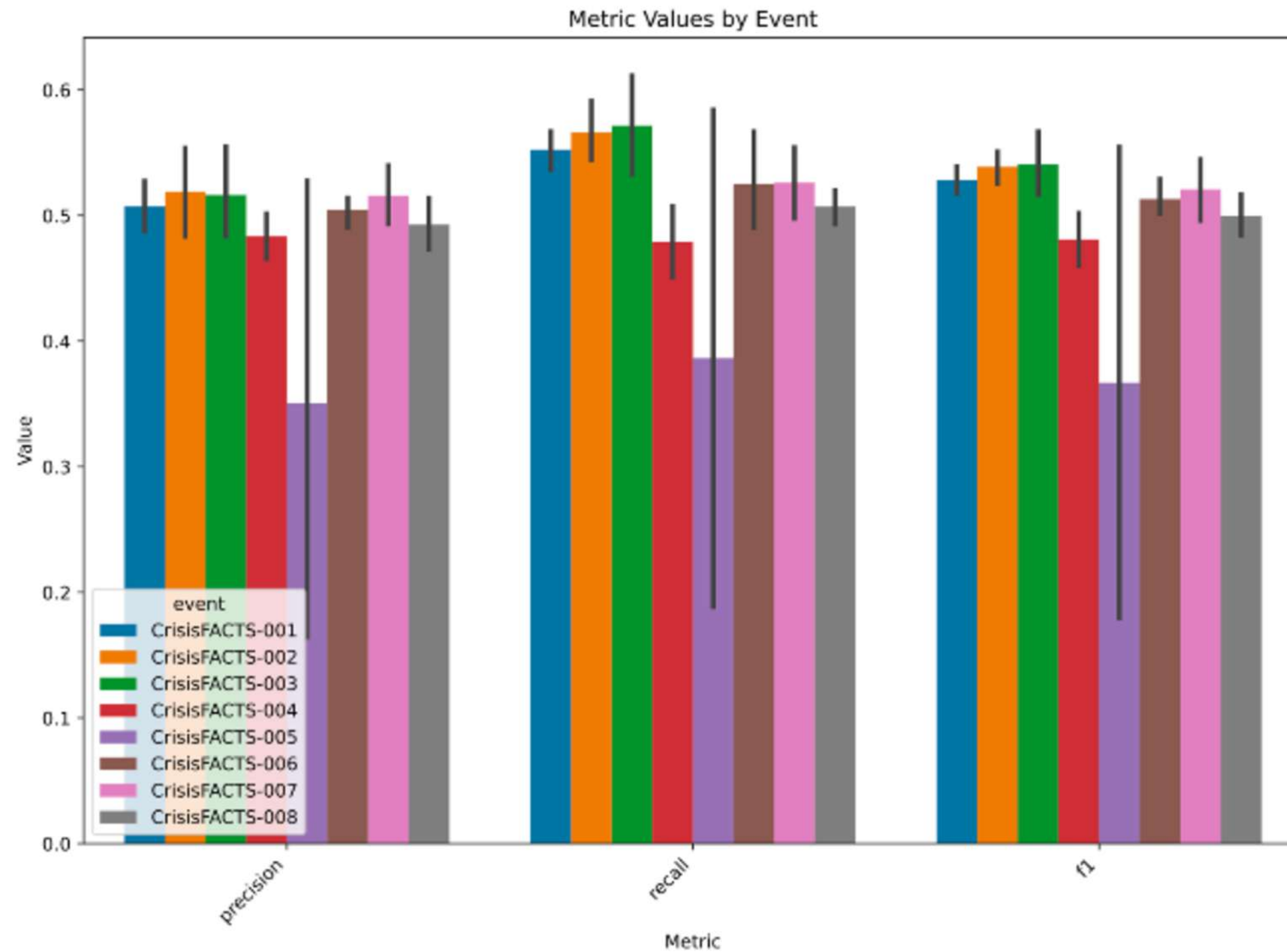


BERTScore Extractive Summarization Baseline

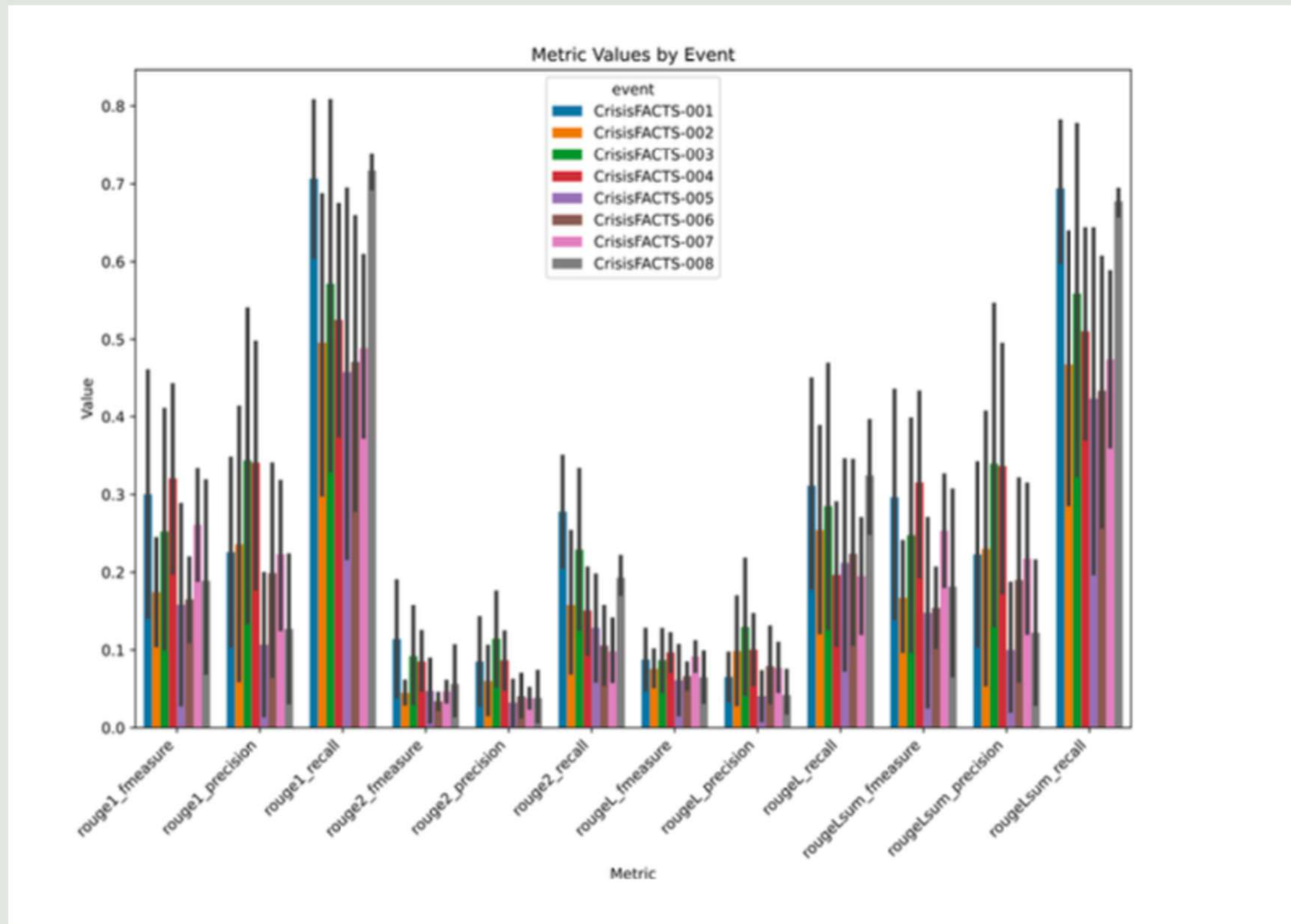
Rouge-Score (Extractive Summarization)



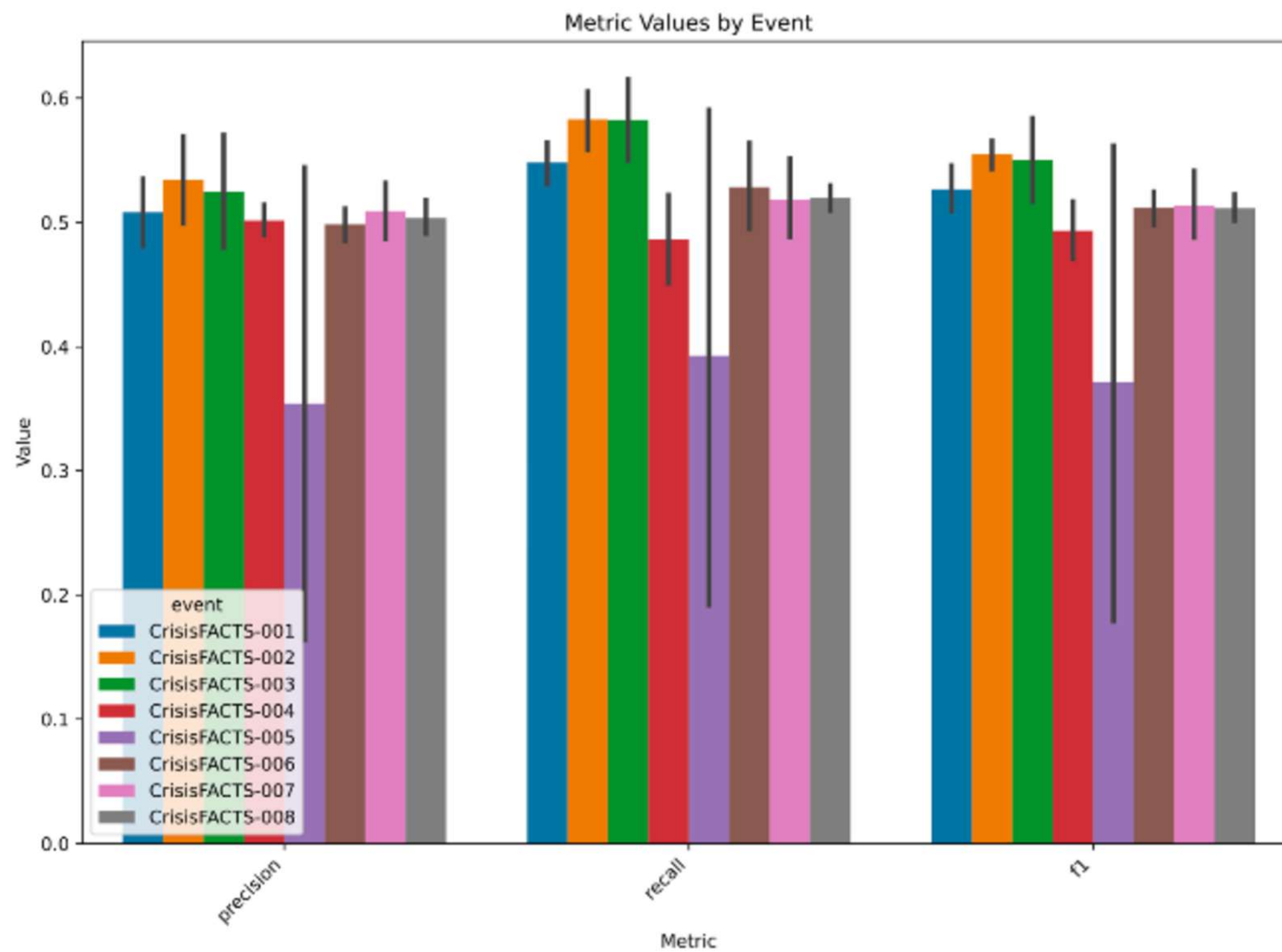
BERTScore(BART Summarization)



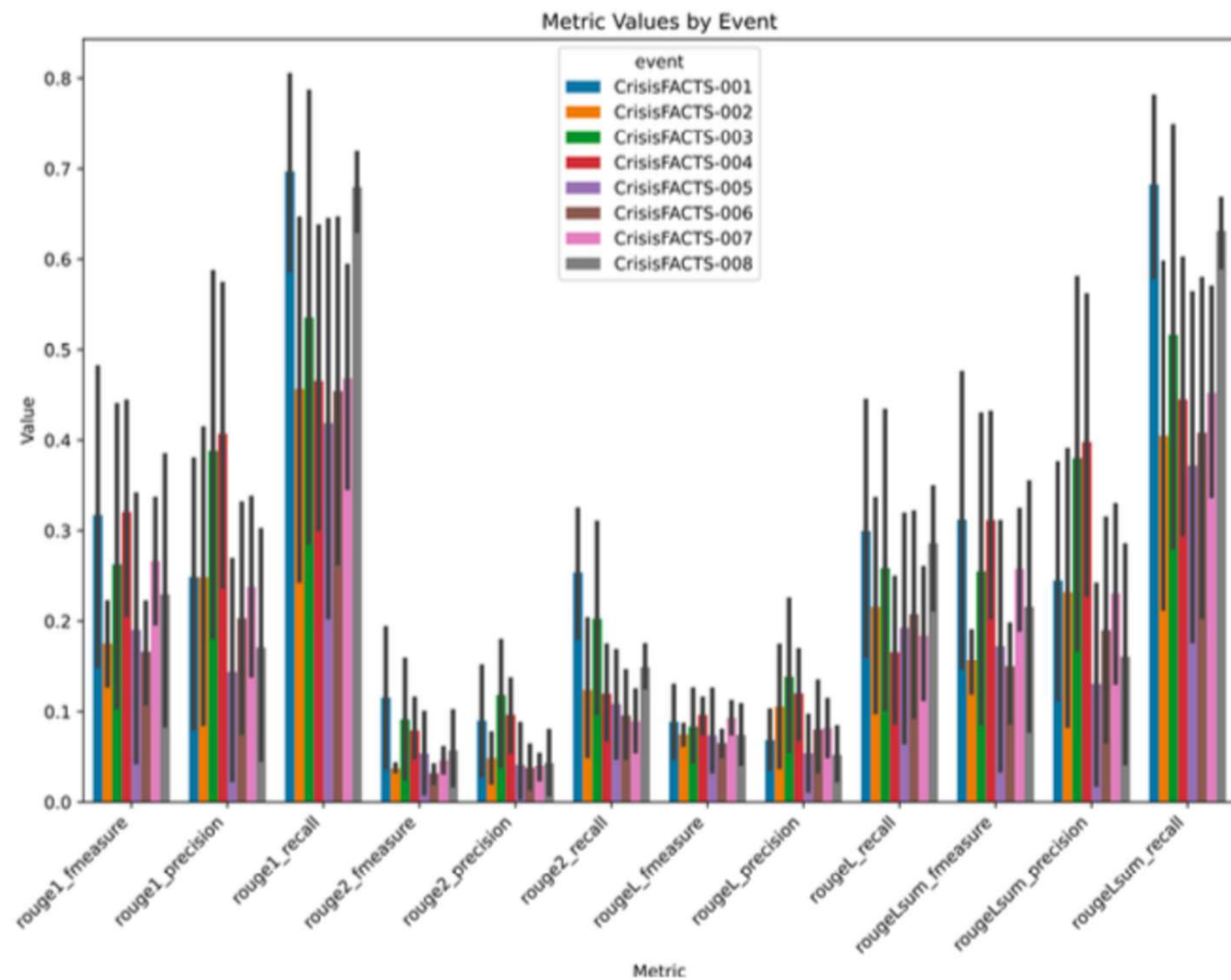
Rouge-Score (BART Summarization)



BERTScore(BERT Summarization)



Rouge-Score(BERT Summarization)




CONCLUSION

- The project focuses on CrisisFACTS, enhancing document relevance through queries and evaluating extractive and abstractive summaries with TREC scripts.
- Extractive summarization outperforms abstractive due to richer content and better alignment with gold summaries, though models like BART show promise.
- A static top-50 approach limits comparisons with TREC's dynamic method, making it hard to assess BART models.



Thank You



Aadesh Minz - 202103002
Asish Joel - 202103015
Raj Kariya - 202103048