LIS at CheckThat! 2025: Multi-Stage Open-Source Large Language Models for Fact-Checking Numerical Claims

Quy Thanh Le

PhD student at Aix-Marseille University, France

Maamar El Amine Hamri

Assoc. Professor at Aix-Marseille University, France

Ismail Badache

Assoc. Professor at Aix-Marseille University, France

Aznam Yacoub

Asst. Professor at University of Windsor,Canada

Table of contents

1 Task description

2 Overview of our proposed models

3 Experimental Results

4 Conclusion

Task description

- Automated fact-checking is crucial for reducing misinformation spread by social media. In particular, **verifying numerical claims** is especially important, as they often appear more trustworthy—a phenomenon known as the **numeric truth effect** [1].

- Existing fact-checking datasets rarely focus on numerical claims. Task 3 uses the extended QuanTemp dataset [2], the first real-world multilingual dataset for this purpose. The main objective of this task is to assess the veracity of claims containing numerical information are **True**, **False**, or **Conflicting**, based on the retrieved evidence.

[1] N. Sagara, E. Peters, Consumer understanding and use of numeric information in product claims, in: D. R. Deeter-Schmelz (Ed.), Proceedings of the 2010 Academy of Marketing Science (AMS) Annual Conference, Springer International Publishing, Cham, 2015, pp. 245–245

[2] V. V, A. Anand, A. Anand, V. Setty, Quantemp: A real-world open-domain benchmark for factchecking numerical claims, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’24, Association for Computing Machinery, New York, NY, USA, 2024, p. 650–660

Overview of our proposed models

1. Goal of our models

- To address this task, we propose a two-stage fact-checking framework to classify the claims based on the retrieved evidences. The first stage consists of an evidence retrieval module. In the second stage, we implement a veracity prediction module.

- Our study builds on a key hypothesis shaped by open-source LLM democratization: Are LLMs capable of performing accurate fact-checking on the dataset provided in this task.

Overview of our proposed models

2. Open-source LLMs Used in Experiments

- Open-source LLMs offer notable benefits in terms of cost-efficiency, transparency, and community collaboration. Therefore, we employ two open-source LLMs, alongside an embedding model built on LLM foundations.

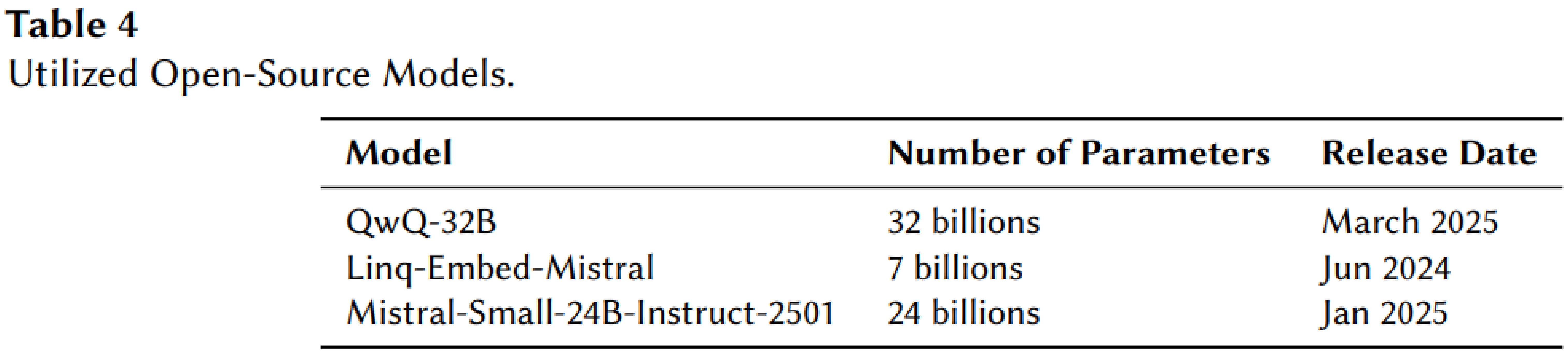


Table 4
Utilized Open-Source Models.

Model	Number of Parameters	Release Date
QwQ-32B	32 billions	March 2025
Linq-Embed-Mistral	7 billions	Jun 2024
Mistral-Small-24B-Instruct-2501	24 billions	Jan 2025

Overview of our proposed models

3. Datasets

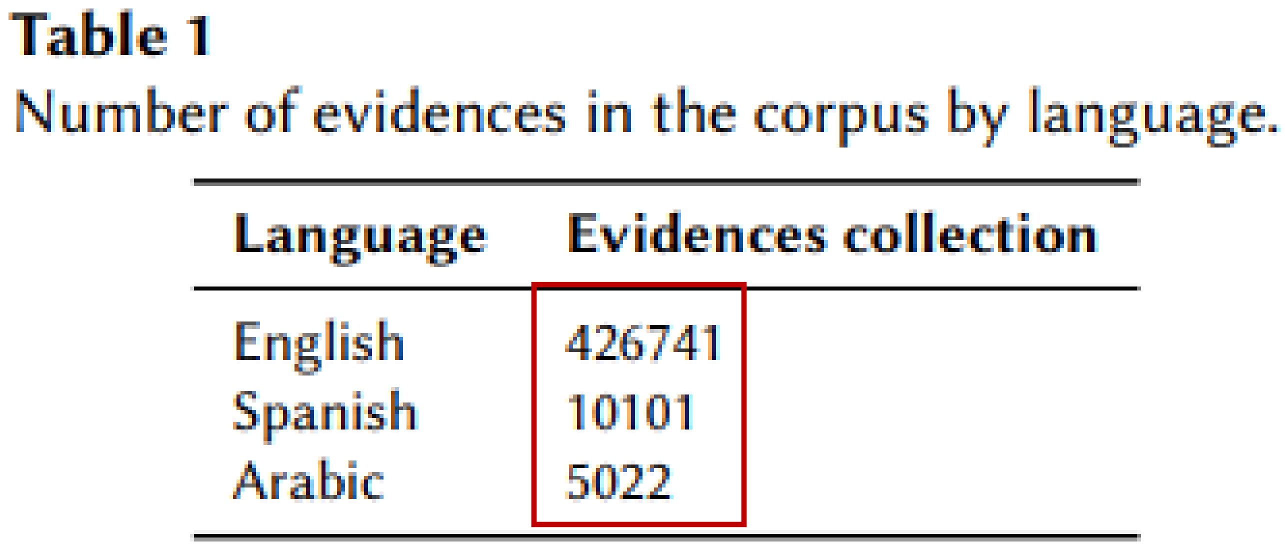


Table 1
Number of evidences in the corpus by language.

Language	Evidences collection
English	426741
Spanish	10101
Arabic	5022

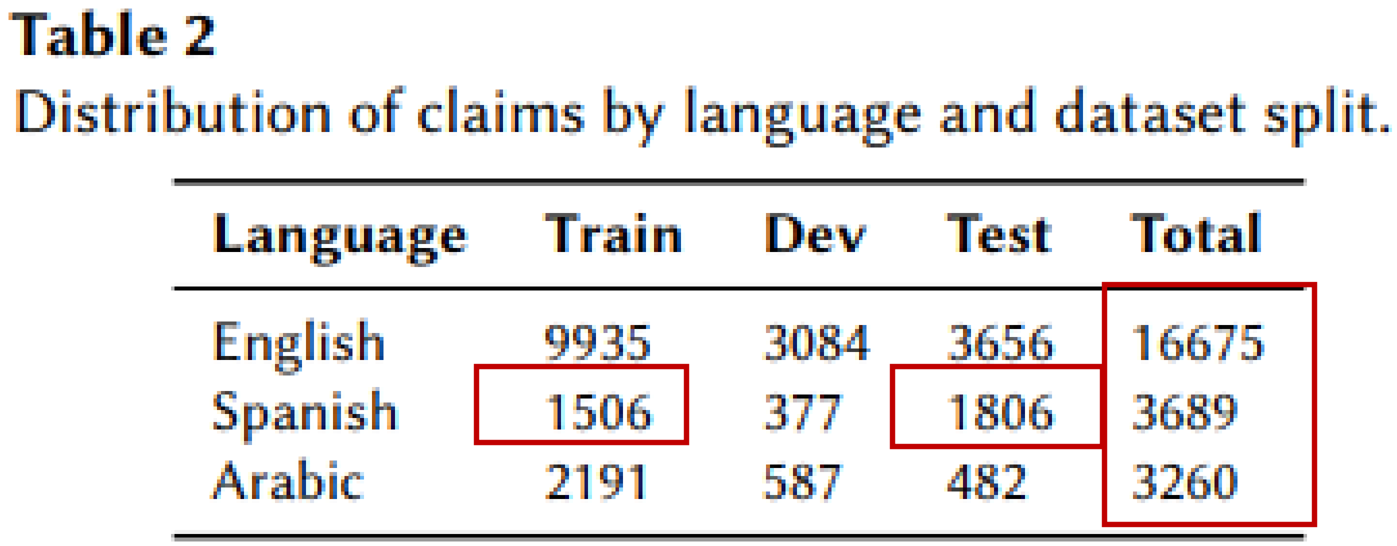


Table 2
Distribution of claims by language and dataset split.

Language	Train	Dev	Test	Total
English	9935	3084	3656	16675
Spanish	1506	377	1806	3689
Arabic	2191	587	482	3260

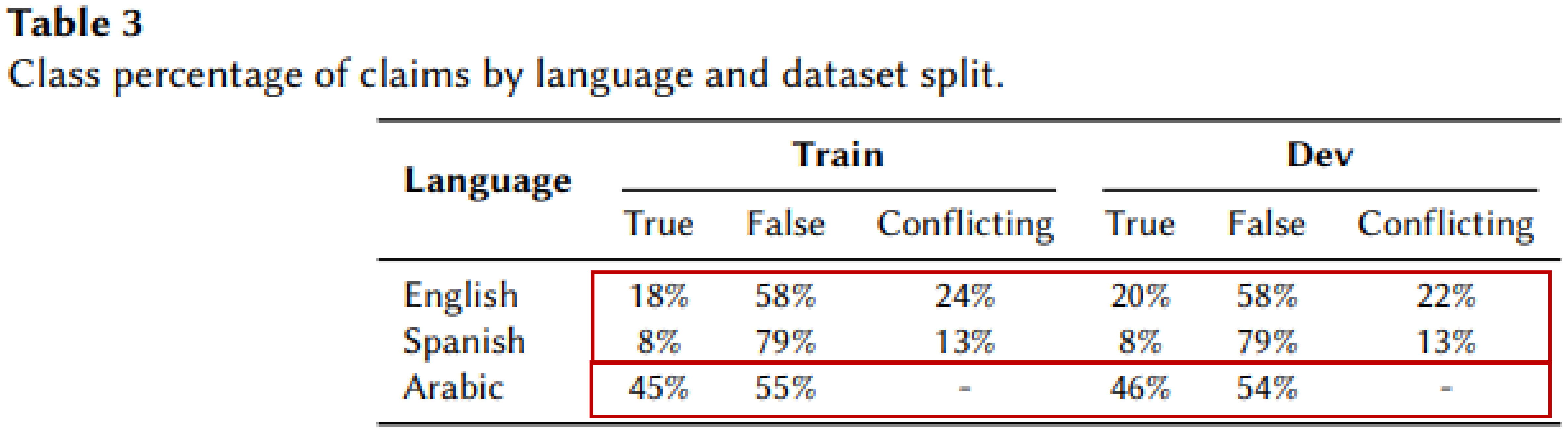


Table 3
Class percentage of claims by language and dataset split.

Language	Train			Dev		
	True	False	Conflicting	True	False	Conflicting
English	18%	58%	24%	20%	58%	22%
Spanish	8%	79%	13%	8%	79%	13%
Arabic	45%	55%	-	46%	54%	-

Overview of our proposed models

4. Methodology

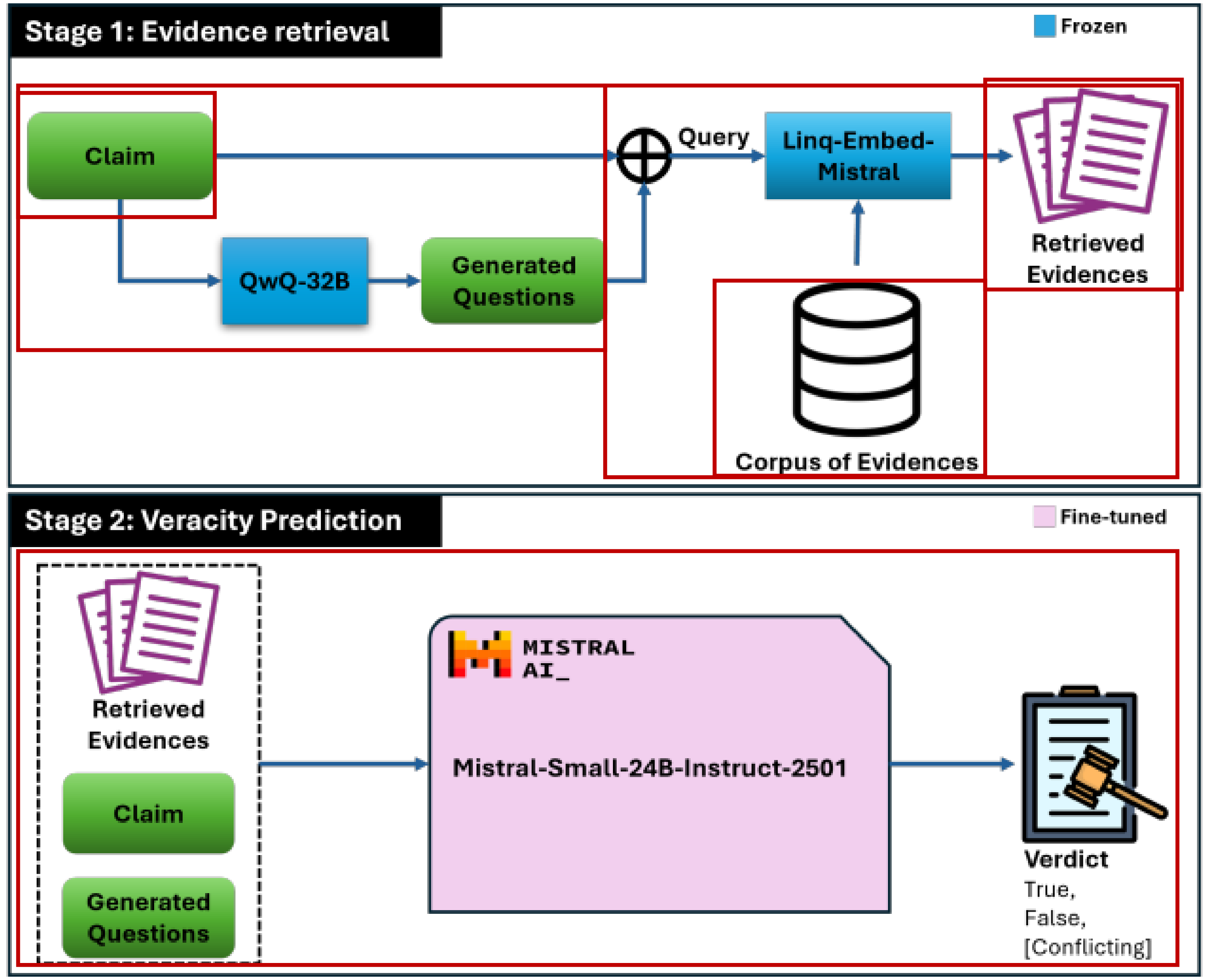


Figure 1: Two-stage inference pipeline for our claim verification system.

Overview of our proposed models

4. Methodology

- Like the baseline, it consists of two stages, but differs in the queries and the model used

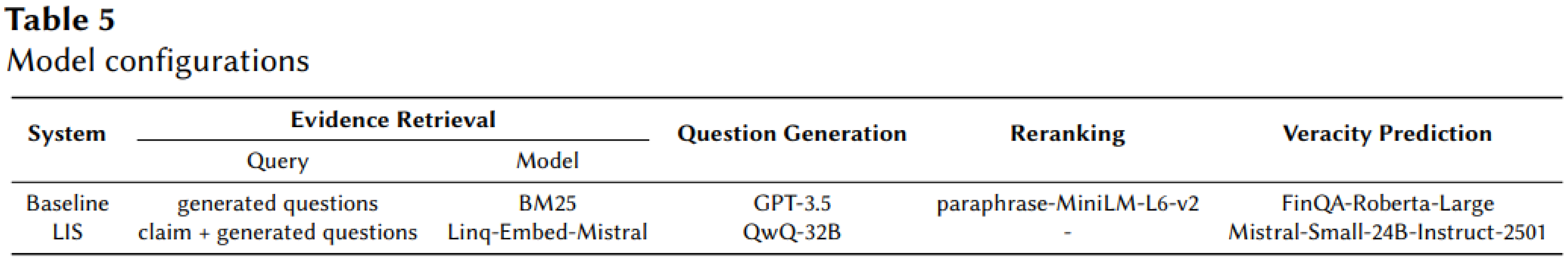


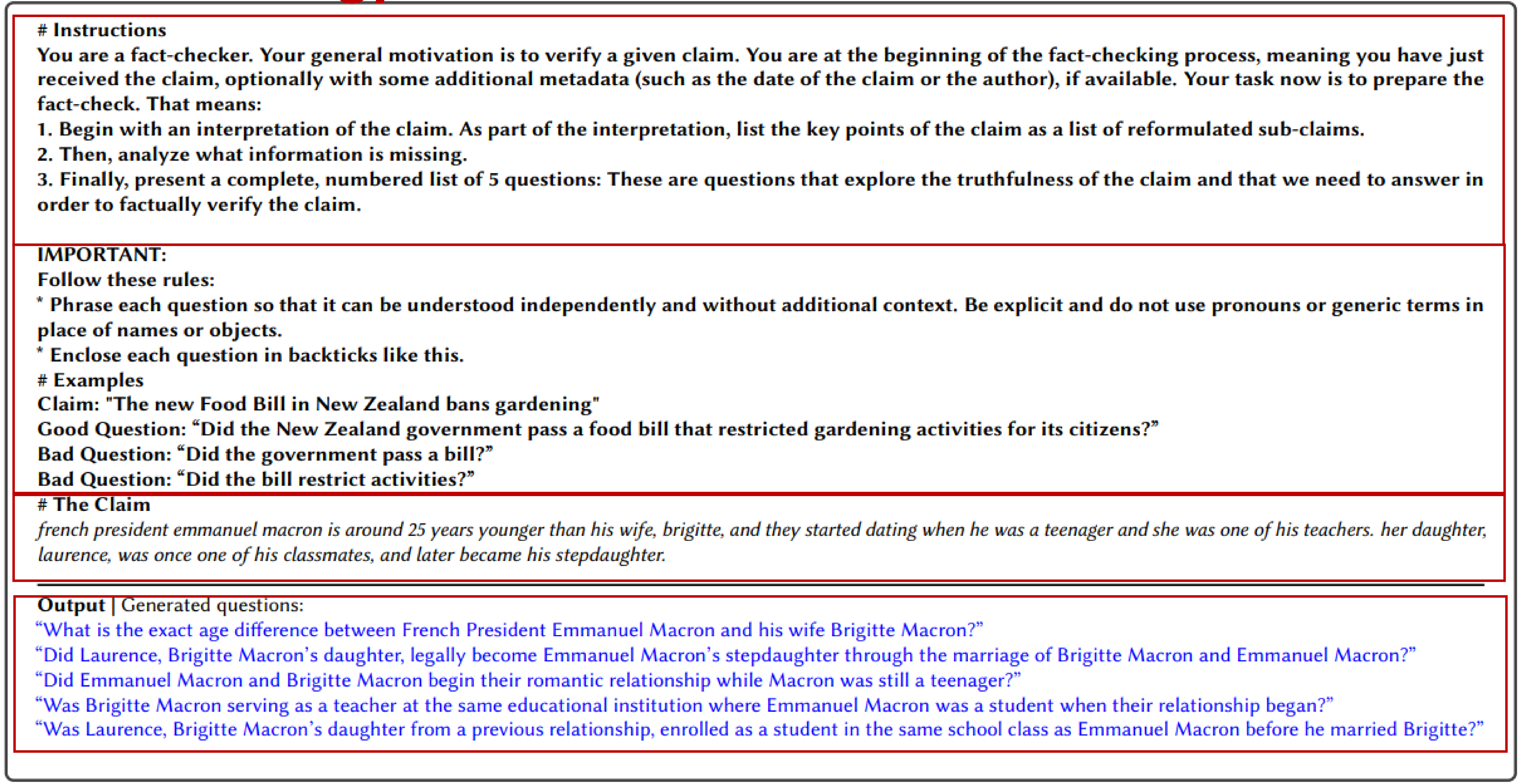
Table 5
Model configurations

System	Evidence Retrieval		Question Generation	Reranking	Veracity Prediction
	Query	Model			
Baseline	generated questions	BM25	GPT-3.5	paraphrase-MiniLM-L6-v2	FinQA-Roberta-Large
LIS	claim + generated questions	Linq-Embed-Mistral	QwQ-32B	-	Mistral-Small-24B-Instruct-2501

- To ensure consistency, the same pipeline described above is used to retrieve evidences for the training, development, and test sets in all languages

Overview of our proposed models

4. Methodology



# Instructions

You are a fact-checker. Your general motivation is to verify a given claim. You are at the beginning of the fact-checking process, meaning you have just received the claim, optionally with some additional metadata (such as the date of the claim or the author), if available. Your task now is to prepare the fact-check. That means:

1. Begin with an interpretation of the claim. As part of the interpretation, list the key points of the claim as a list of reformulated sub-claims.
2. Then, analyze what information is missing.
3. Finally, present a complete, numbered list of 5 questions: These are questions that explore the truthfulness of the claim and that we need to answer in order to factually verify the claim.

IMPORTANT:

Follow these rules:

* Phrase each question so that it can be understood independently and without additional context. Be explicit and do not use pronouns or generic terms in place of names or objects.

* Enclose each question in backticks like this.

# Examples

Claim: "The new Food Bill in New Zealand bans gardening"

Good Question: “Did the New Zealand government pass a food bill that restricted gardening activities for its citizens?”

Bad Question: “Did the government pass a bill?”

Bad Question: “Did the bill restrict activities?”

# The Claim

french president emmanuel macron is around 25 years younger than his wife, brigitte, and they started dating when he was a teenager and she was one of his teachers. her daughter, laurence, was once one of his classmates, and later became his stepdaughter.

Output | Generated questions:

“What is the exact age difference between French President Emmanuel Macron and his wife Brigitte Macron?”

“Did Laurence, Brigitte Macron’s daughter, legally become Emmanuel Macron’s stepdaughter through the marriage of Brigitte Macron and Emmanuel Macron?”

“Did Emmanuel Macron and Brigitte Macron begin their romantic relationship while Macron was still a teenager?”

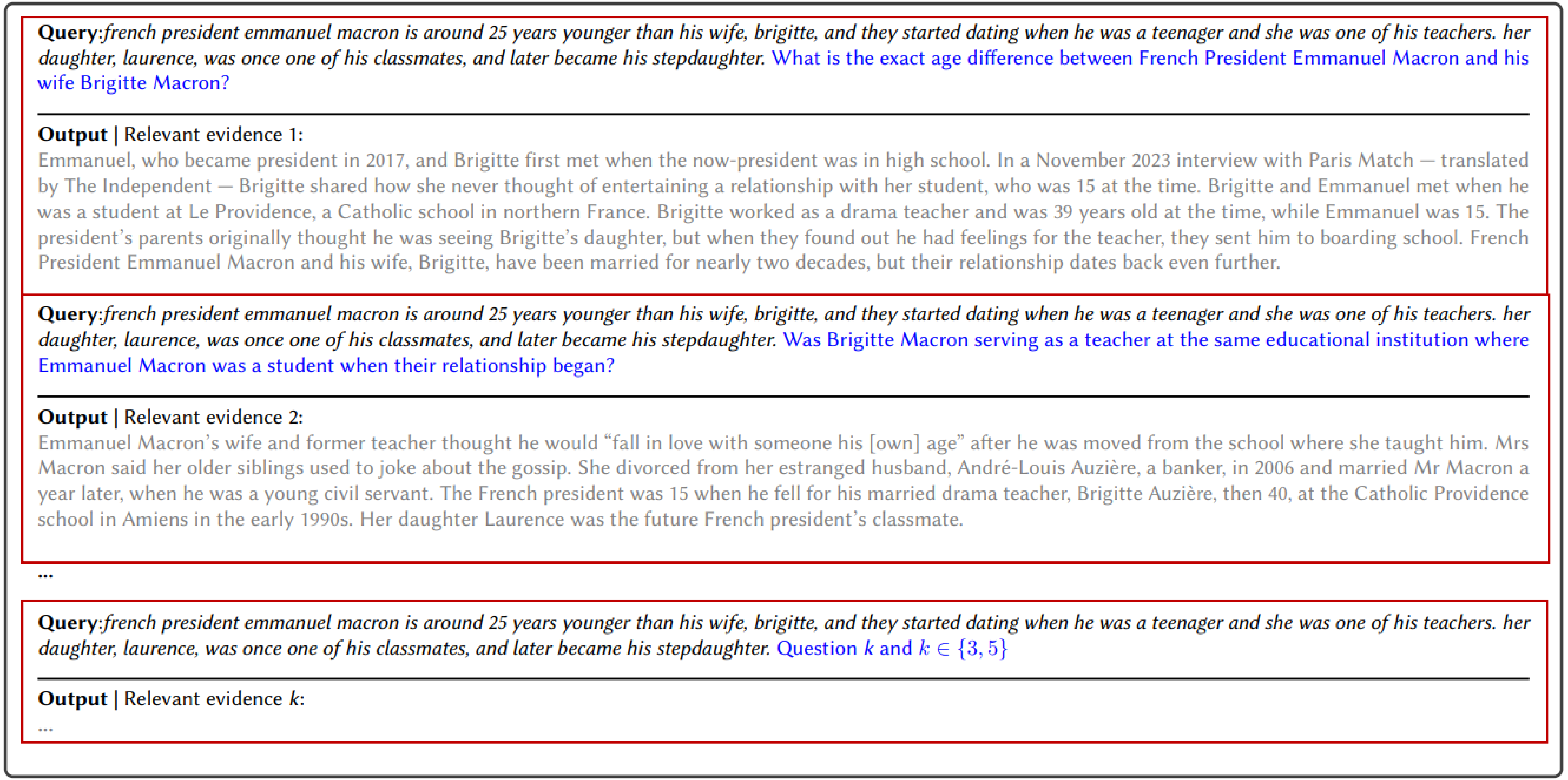
“Was Brigitte Macron serving as a teacher at the same educational institution where Emmanuel Macron was a student when their relationship began?”

“Was Laurence, Brigitte Macron’s daughter from a previous relationship, enrolled as a student in the same school class as Emmanuel Macron before he married Brigitte?”

Figure 2: An example of an instruction prompt and its output for question generation.

Overview of our proposed models

4. Methodology



Query:*french president emmanuel macron is around 25 years younger than his wife, brigitte, and they started dating when he was a teenager and she was one of his teachers. her daughter, laurence, was once one of his classmates, and later became his stepdaughter.* What is the exact age difference between French President Emmanuel Macron and his wife Brigitte Macron?

Output | Relevant evidence 1:

Emmanuel, who became president in 2017, and Brigitte first met when the now-president was in high school. In a November 2023 interview with Paris Match — translated by The Independent — Brigitte shared how she never thought of entertaining a relationship with her student, who was 15 at the time. Brigitte and Emmanuel met when he was a student at Le Providence, a Catholic school in northern France. Brigitte worked as a drama teacher and was 39 years old at the time, while Emmanuel was 15. The president's parents originally thought he was seeing Brigitte's daughter, but when they found out he had feelings for the teacher, they sent him to boarding school. French President Emmanuel Macron and his wife, Brigitte, have been married for nearly two decades, but their relationship dates back even further.

Query:*french president emmanuel macron is around 25 years younger than his wife, brigitte, and they started dating when he was a teenager and she was one of his teachers. her daughter, laurence, was once one of his classmates, and later became his stepdaughter.* Was Brigitte Macron serving as a teacher at the same educational institution where Emmanuel Macron was a student when their relationship began?

Output | Relevant evidence 2:

Emmanuel Macron's wife and former teacher thought he would "fall in love with someone his [own] age" after he was moved from the school where she taught him. Mrs Macron said her older siblings used to joke about the gossip. She divorced from her estranged husband, André-Louis Auzière, a banker, in 2006 and married Mr Macron a year later, when he was a young civil servant. The French president was 15 when he fell for his married drama teacher, Brigitte Auzière, then 40, at the Catholic Providence school in Amiens in the early 1990s. Her daughter Laurence was the future French president's classmate.

...

Query:*french president emmanuel macron is around 25 years younger than his wife, brigitte, and they started dating when he was a teenager and she was one of his teachers. her daughter, laurence, was once one of his classmates, and later became his stepdaughter.* Question k and $k \in \{3, 5\}$

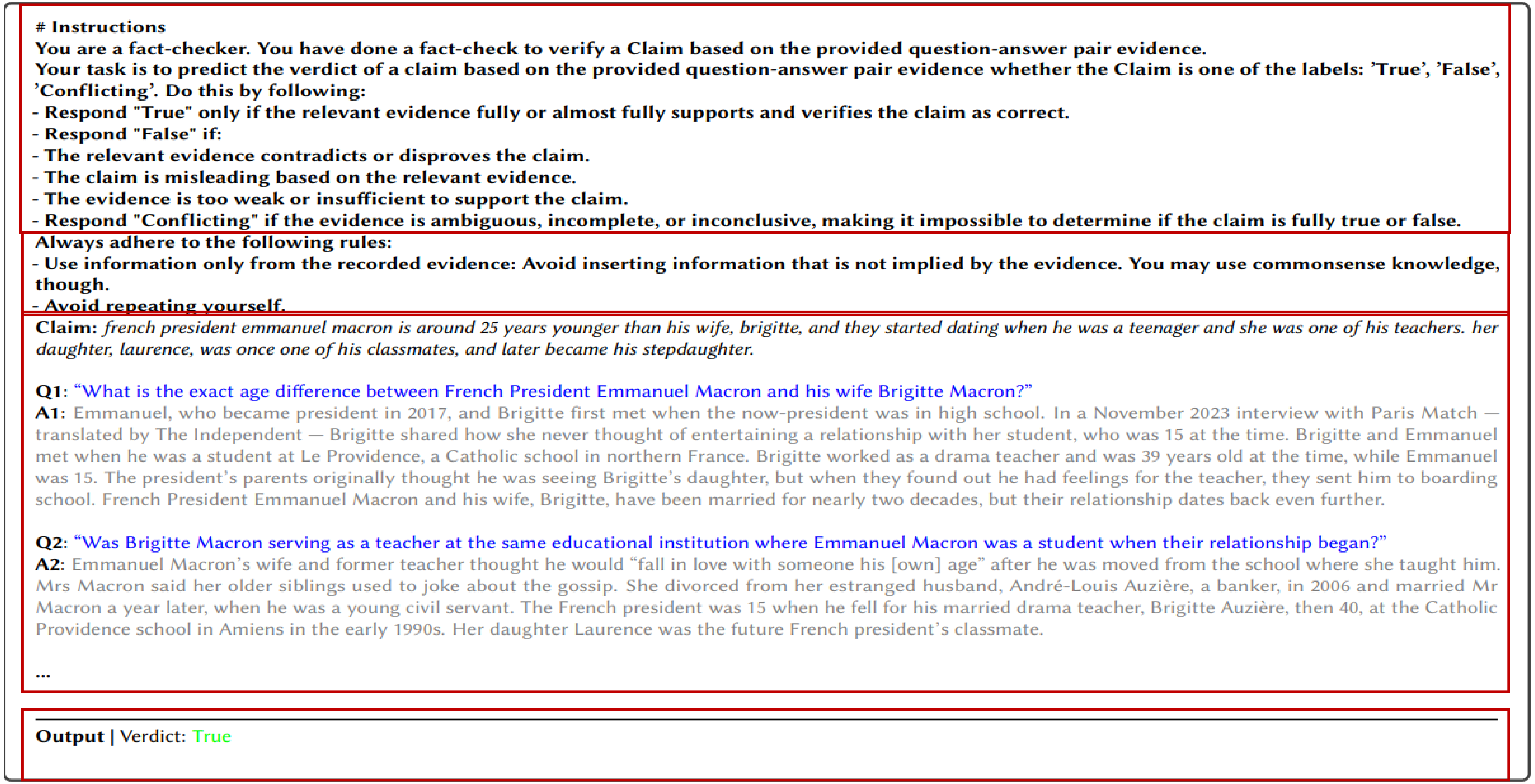
Output | Relevant evidence k:

...

Figure 3: Example of query formation by combining a claim with a generated question to search for relevant evidences.

Overview of our proposed models

4. Methodology



# Instructions
You are a fact-checker. You have done a fact-check to verify a Claim based on the provided question-answer pair evidence.
Your task is to predict the verdict of a claim based on the provided question-answer pair evidence whether the Claim is one of the labels: ’True’, ’False’, ’Conflicting’. Do this by following:
- Respond "True" only if the relevant evidence fully or almost fully supports and verifies the claim as correct.
- Respond "False" if:
- The relevant evidence contradicts or disproves the claim.
- The claim is misleading based on the relevant evidence.
- The evidence is too weak or insufficient to support the claim.
- Respond "Conflicting" if the evidence is ambiguous, incomplete, or inconclusive, making it impossible to determine if the claim is fully true or false.

Always adhere to the following rules:
- Use information only from the recorded evidence: Avoid inserting information that is not implied by the evidence. You may use commonsense knowledge, though.
- Avoid repeating yourself.

Claim: *french president emmanuel macron is around 25 years younger than his wife, brigitte, and they started dating when he was a teenager and she was one of his teachers. her daughter, laurence, was once one of his classmates, and later became his stepdaughter.*

Q1: “What is the exact age difference between French President Emmanuel Macron and his wife Brigitte Macron?”
A1: Emmanuel, who became president in 2017, and Brigitte first met when the now-president was in high school. In a November 2023 interview with Paris Match — translated by The Independent — Brigitte shared how she never thought of entertaining a relationship with her student, who was 15 at the time. Brigitte and Emmanuel met when he was a student at Le Providence, a Catholic school in northern France. Brigitte worked as a drama teacher and was 39 years old at the time, while Emmanuel was 15. The president’s parents originally thought he was seeing Brigitte’s daughter, but when they found out he had feelings for the teacher, they sent him to boarding school. French President Emmanuel Macron and his wife, Brigitte, have been married for nearly two decades, but their relationship dates back even further.

Q2: “Was Brigitte Macron serving as a teacher at the same educational institution where Emmanuel Macron was a student when their relationship began?”
A2: Emmanuel Macron’s wife and former teacher thought he would “fall in love with someone his [own] age” after he was moved from the school where she taught him. Mrs Macron said her older siblings used to joke about the gossip. She divorced from her estranged husband, André-Louis Auzière, a banker, in 2006 and married Mr Macron a year later, when he was a young civil servant. The French president was 15 when he fell for his married drama teacher, Brigitte Auzière, then 40, at the Catholic Providence school in Amiens in the early 1990s. Her daughter Laurence was the future French president’s classmate.

...

Output | Verdict: True

Figure 4: Example of instruction prompt and output for veracity prediction.

Experimental Results

1. Implementation

- All experiments ran on the LIS cluster with NVIDIA A100 GPUs (80GB). The evidence retrieval phase took approximately 8 hours, while the question generation phase required around 2 days.

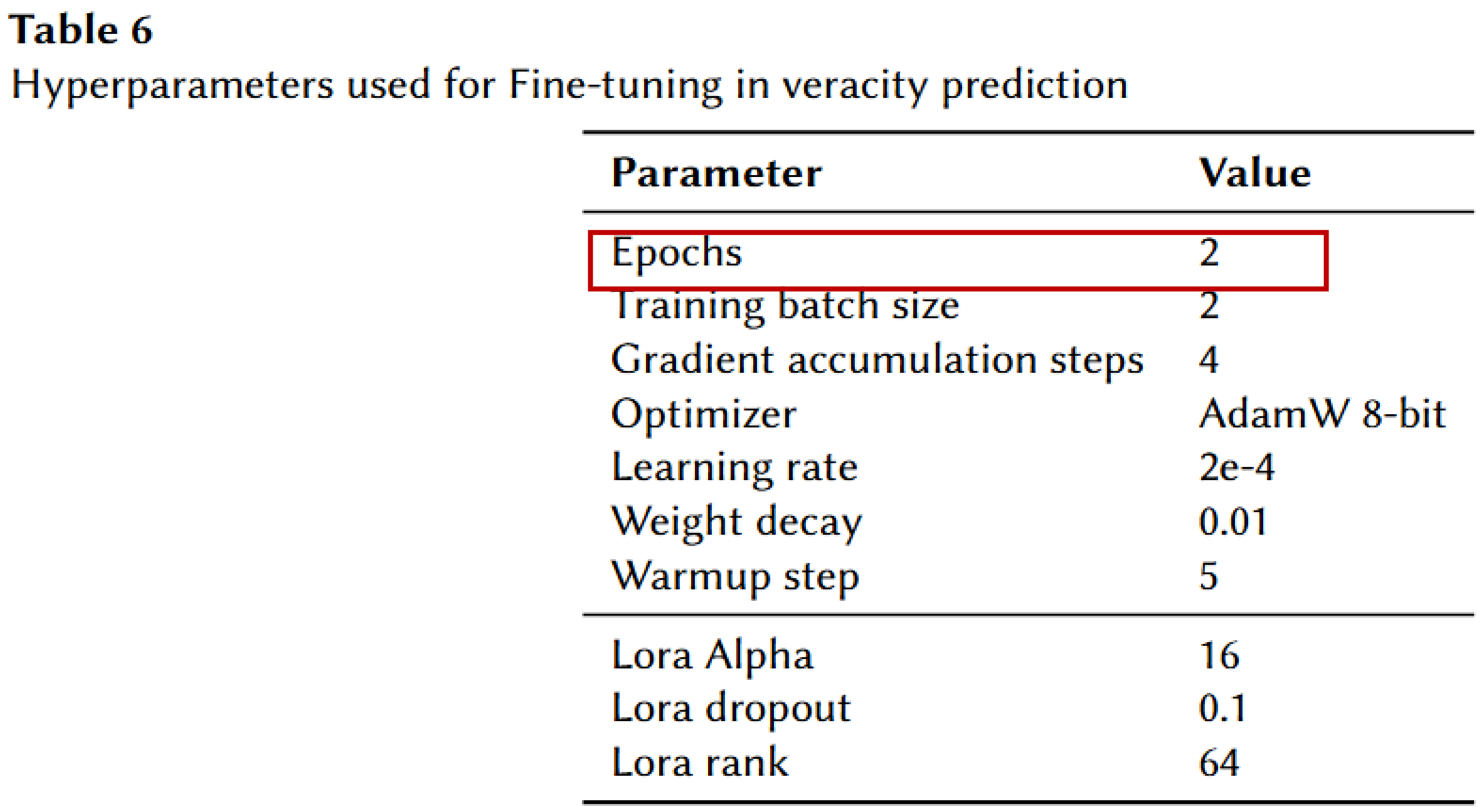


Table 6
Hyperparameters used for Fine-tuning in veracity prediction

Parameter	Value
Epochs	2
Training batch size	2
Gradient accumulation steps	4
Optimizer	AdamW 8-bit
Learning rate	2e-4
Weight decay	0.01
Warmup step	5
Lora Alpha	16
Lora dropout	0.1
Lora rank	64

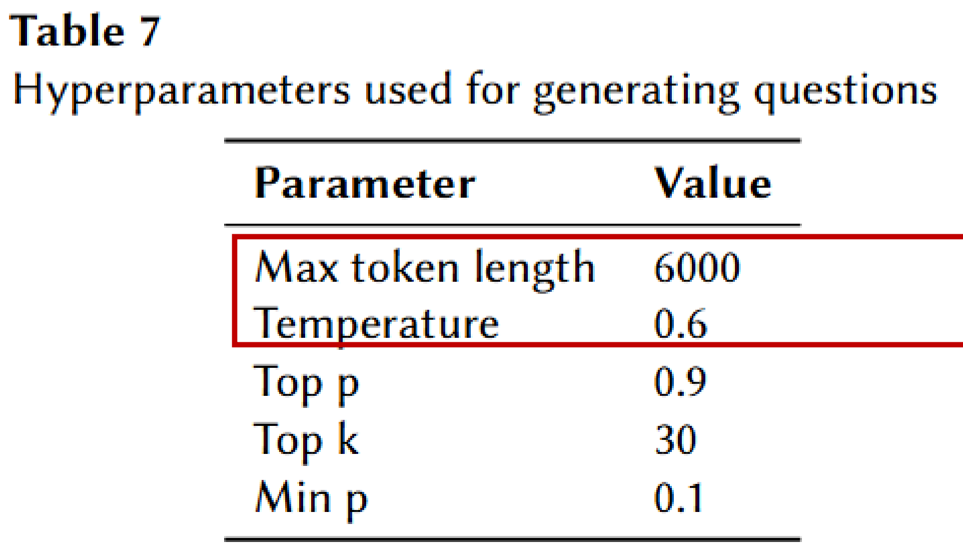


Table 7
Hyperparameters used for generating questions

Parameter	Value
Max token length	6000
Temperature	0.6
Top p	0.9
Top k	30
Min p	0.1

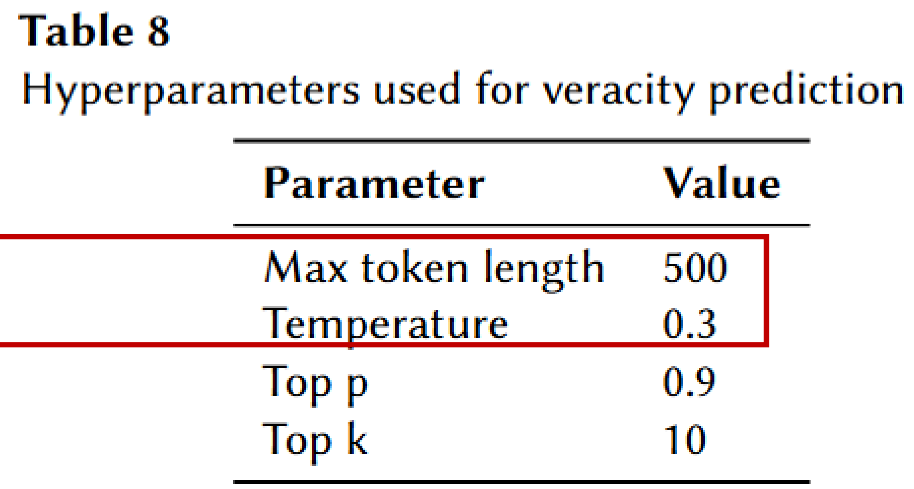


Table 8
Hyperparameters used for veracity prediction

Parameter	Value
Max token length	500
Temperature	0.3
Top p	0.9
Top k	10

Experimental Results

2. Results and Discussions

1. English

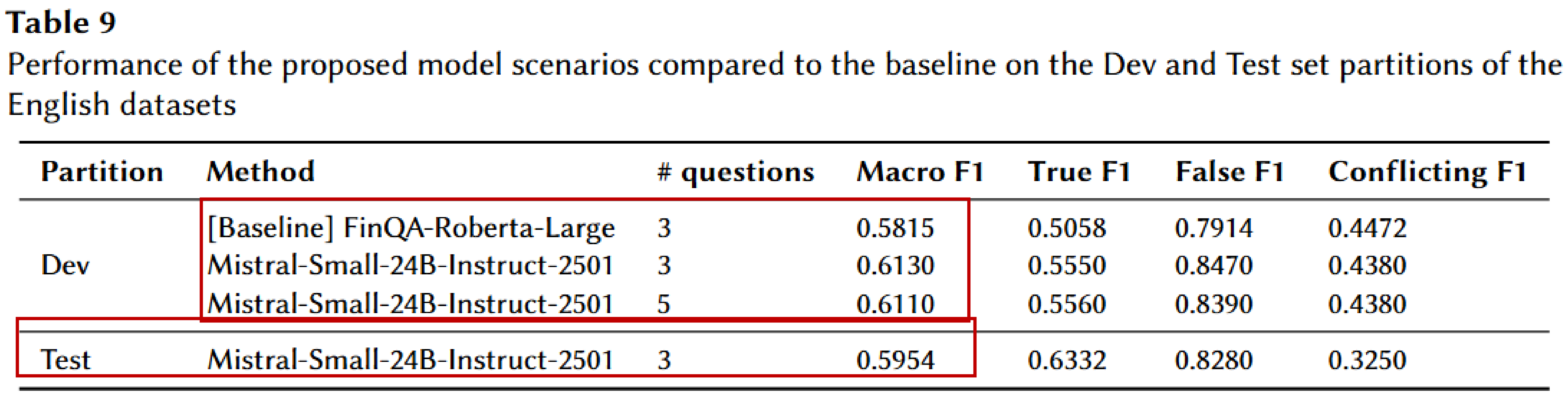


Table 9
Performance of the proposed model scenarios compared to the baseline on the Dev and Test set partitions of the English datasets

Partition	Method	# questions	Macro F1	True F1	False F1	Conflicting F1
Dev	[Baseline] FinQA-Roberta-Large	3	0.5815	0.5058	0.7914	0.4472
	Mistral-Small-24B-Instruct-2501	3	0.6130	0.5550	0.8470	0.4380
	Mistral-Small-24B-Instruct-2501	5	0.6110	0.5560	0.8390	0.4380
Test	Mistral-Small-24B-Instruct-2501	3	0.5954	0.6332	0.8280	0.3250

- **Submitted model:** 3-question model, macro-F1 = 59.54%, ~1.5% lower than dev set, ranked **1st**, outperforming runner-up by ~3% and lowest-ranked by ~24%.
- **Baseline comparison:** Not possible; official Task 3 baseline results **not yet released**.

Experimental Results

2. Results and Discussions

2. Spanish & Arabic

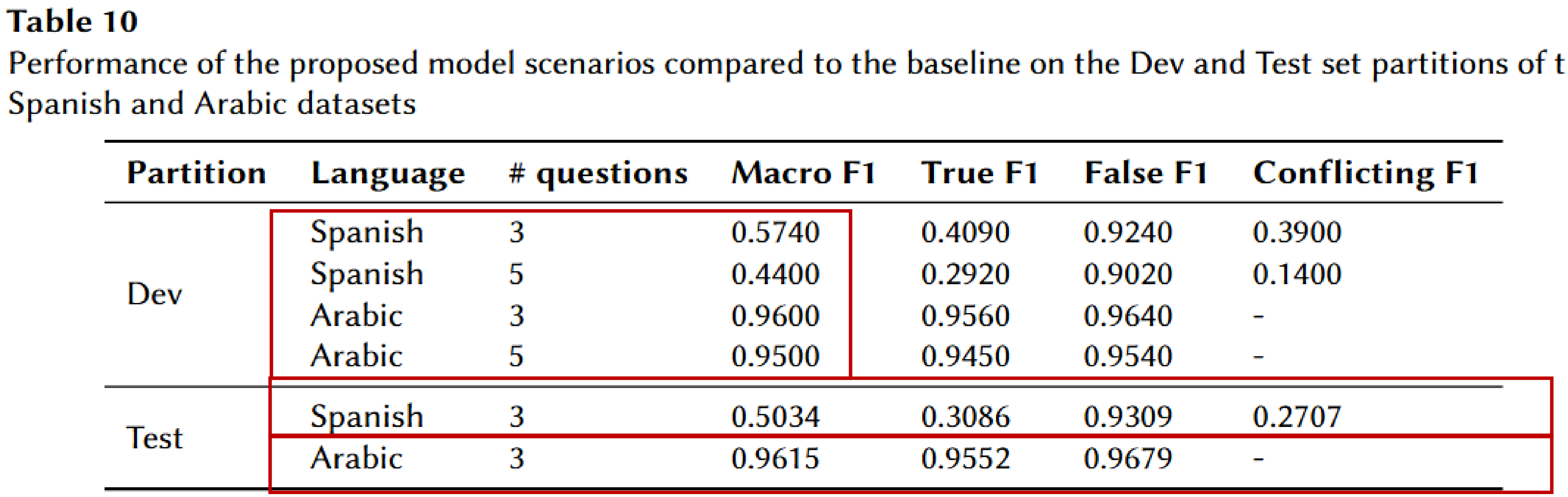


Table 10
Performance of the proposed model scenarios compared to the baseline on the Dev and Test set partitions of the Spanish and Arabic datasets

Partition	Language	# questions	Macro F1	True F1	False F1	Conflicting F1
Dev	Spanish	3	0.5740	0.4090	0.9240	0.3900
	Spanish	5	0.4400	0.2920	0.9020	0.1400
	Arabic	3	0.9600	0.9560	0.9640	-
	Arabic	5	0.9500	0.9450	0.9540	-
Test	Spanish	3	0.5034	0.3086	0.9309	0.2707
	Arabic	3	0.9615	0.9552	0.9679	-

- **Spanish**: 3-question model → macro-F1 = 50.34% (~7% drop vs dev set), **1st place**, leading over runner-up by ~13% and lowest-ranked by ~25%.

- **Arabic:** 3-question model → consistently produced 2 labels; macro-F1 = 96.15% (+0.15% vs dev set), **1st place,** ahead of runner-up by ~33% and lowest-ranked by ~60%.

Conclusion

- **Task & Languages**: Verifying numerical claims in English, Spanish, Arabic
- **Pipeline**: Evidence retrieval + veracity prediction using instruction-following LLMs and effective fine-tuning strategi as LORA
- **Consistency**: Same methodology applied across all three languages; only team fully participating in all
- **Performance**: Ranked 1st on leaderboard, outperforming 19 teams; superior to baseline NLI model (FinQA-Roberta
- **Limitations**:
  - Only one question configuration tested (3 questions)
  - Fine-tuning limited to monolingual models
  - Arabic dataset uses 3-verdict prompt, though only 2 labels exist → may cause non-applicable
- **Future Work**:
  - Test more open-source LLMs.
  - Explore diverse question configurations.
  - Fine-tune multilingual models for improved performance.

Q&A