Final presentation

Outils Formels Avancés 2024

AI fact checker

28 Mai 2024

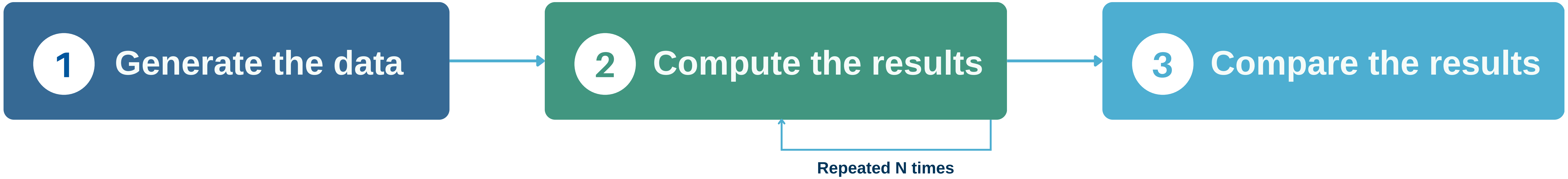
GOAL of the project

AI Fact Checker

The main goal is to use the programming logic to verify the truth of the LLM's answers.

- Find references
- Find non-factual answers from LLMs
- Define tests that measure the performance of LLMs

Pipeline



1 Generate the data

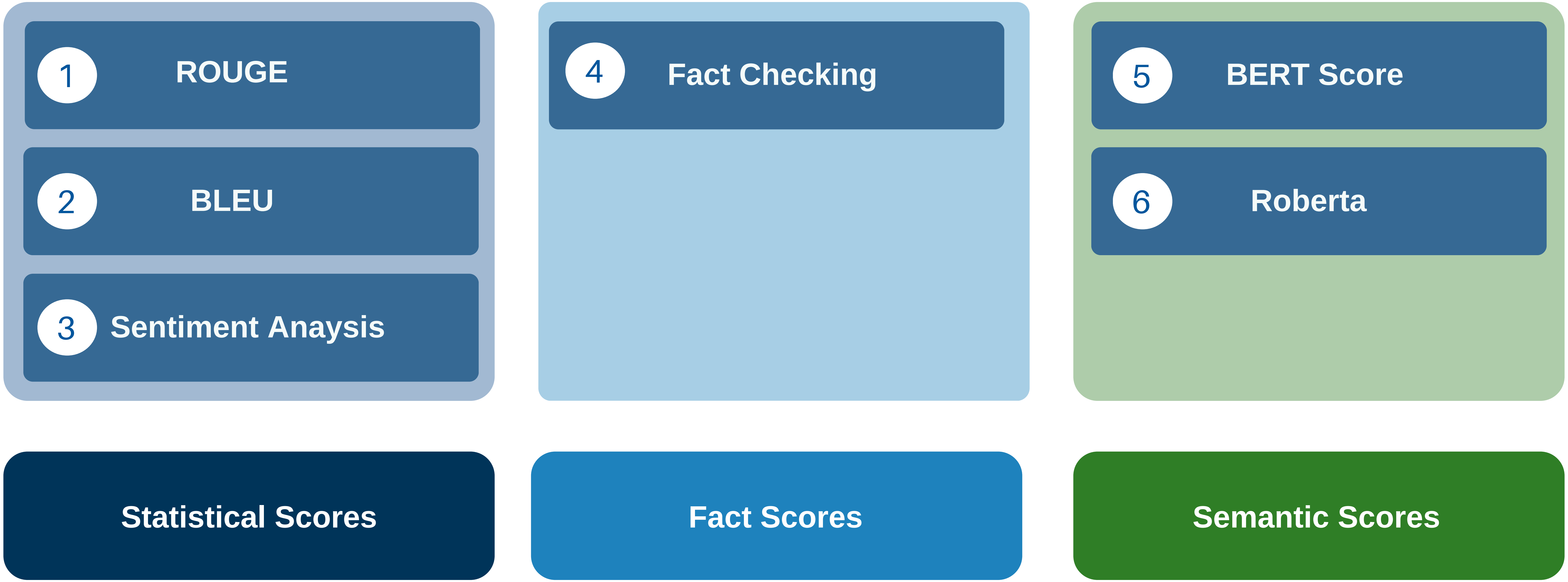
2 Compute the results

3 Compare the results

Repeated N times

N: *number of models / LLMs tested*

Evaluation Framework: Metrics



1 ROUGE

2 BLEU

3 Sentiment Anaysis

Statistical Scores

4 Fact Checking

Fact Scores

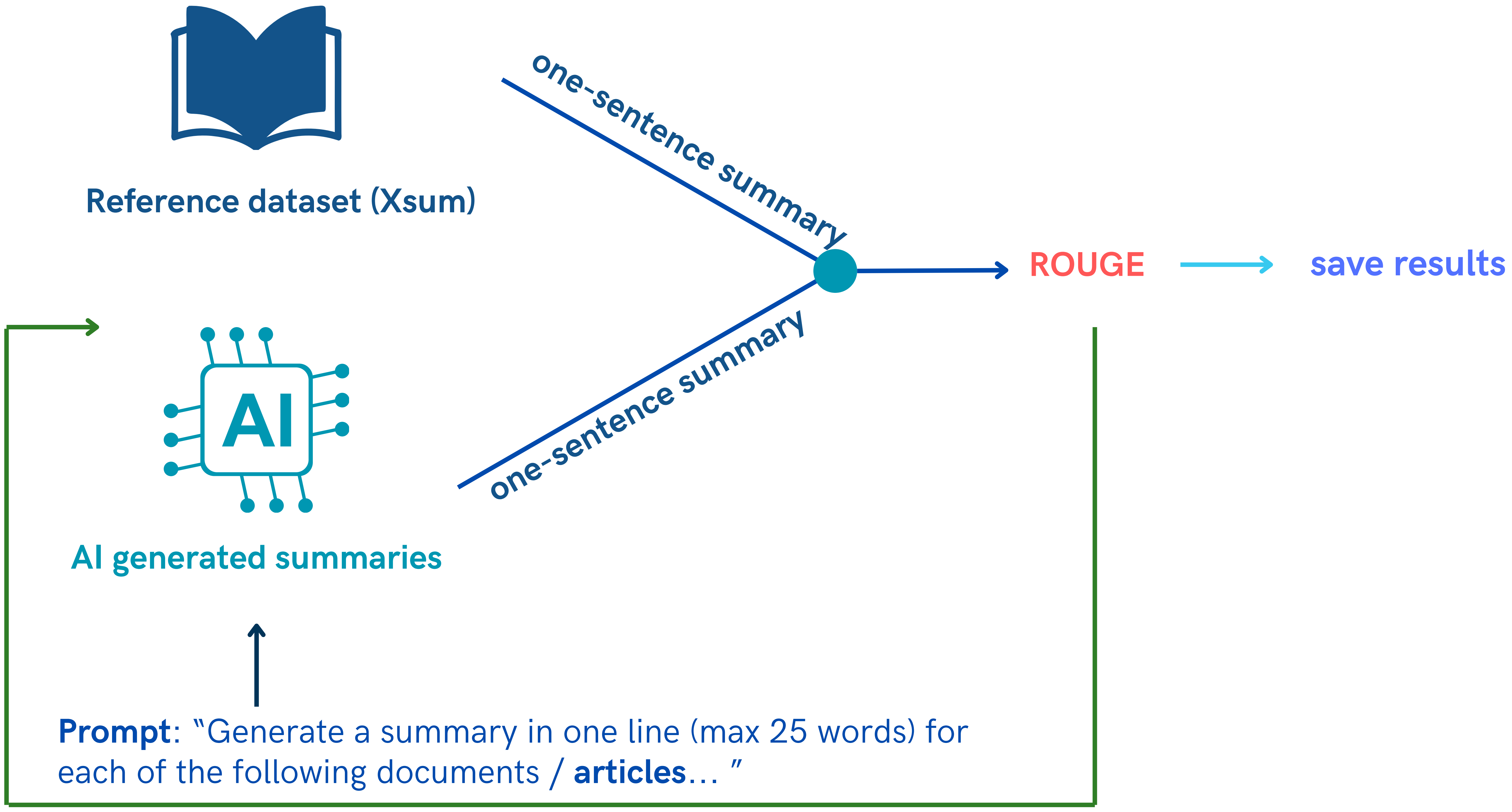
5 BERT Score

6 Roberta

Semantic Scores

1. The ROUGE metric: Text Summarization

Statistical Scores



Reference dataset (Xsum)

one-sentence summary

ROUGE

save results

AI generated summaries

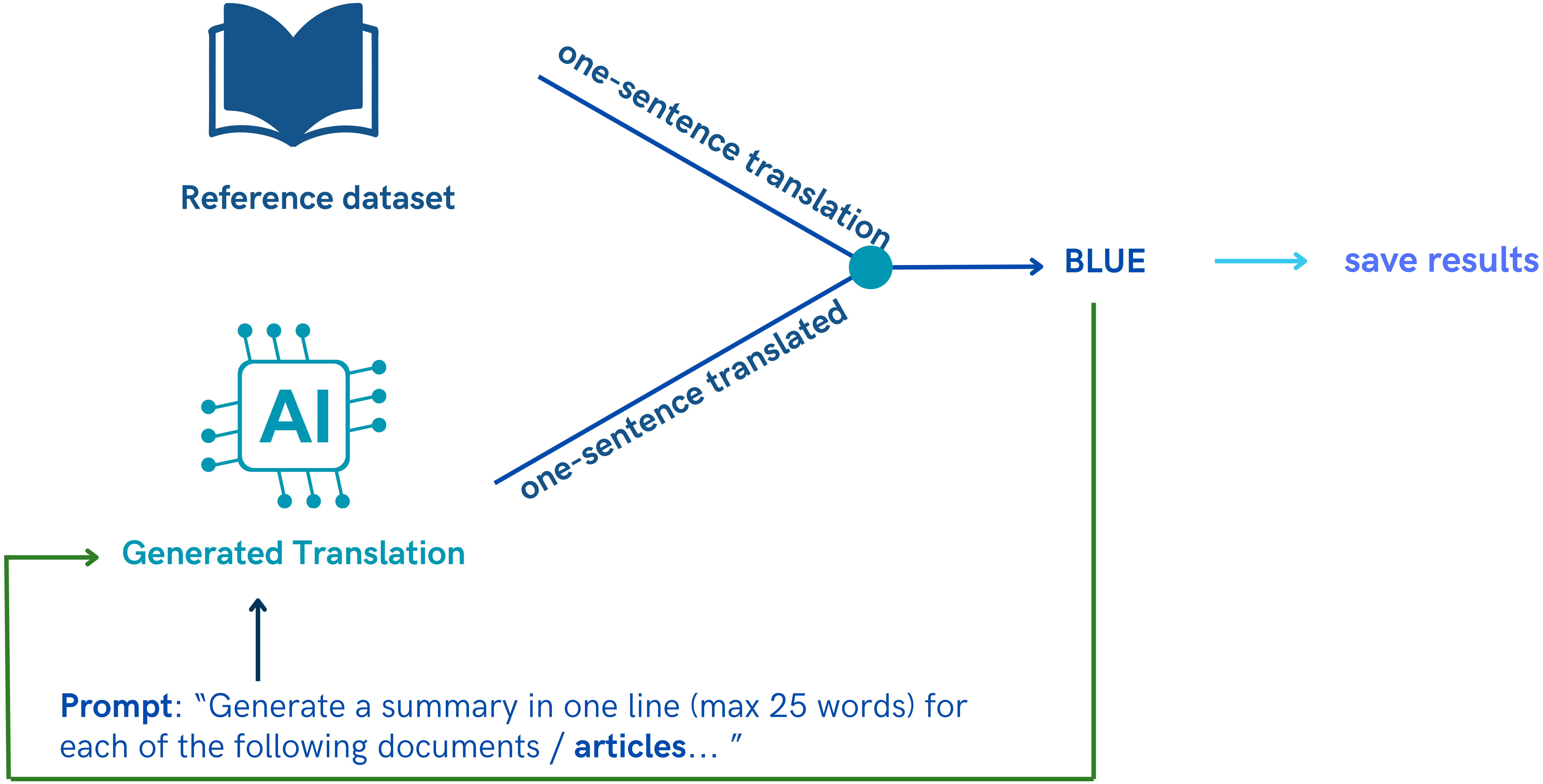
one-sentence summary

Prompt: "Generate a summary in one line (max 25 words) for each of the following documents / **articles**… "

<u>Evaluating 10 articles</u>

Unigram and LCS-gram

2. The BLEU metric: Text summarization / translation



Reference dataset

one-sentence translation

BLUE

save results

Generated Translation

one-sentence translated

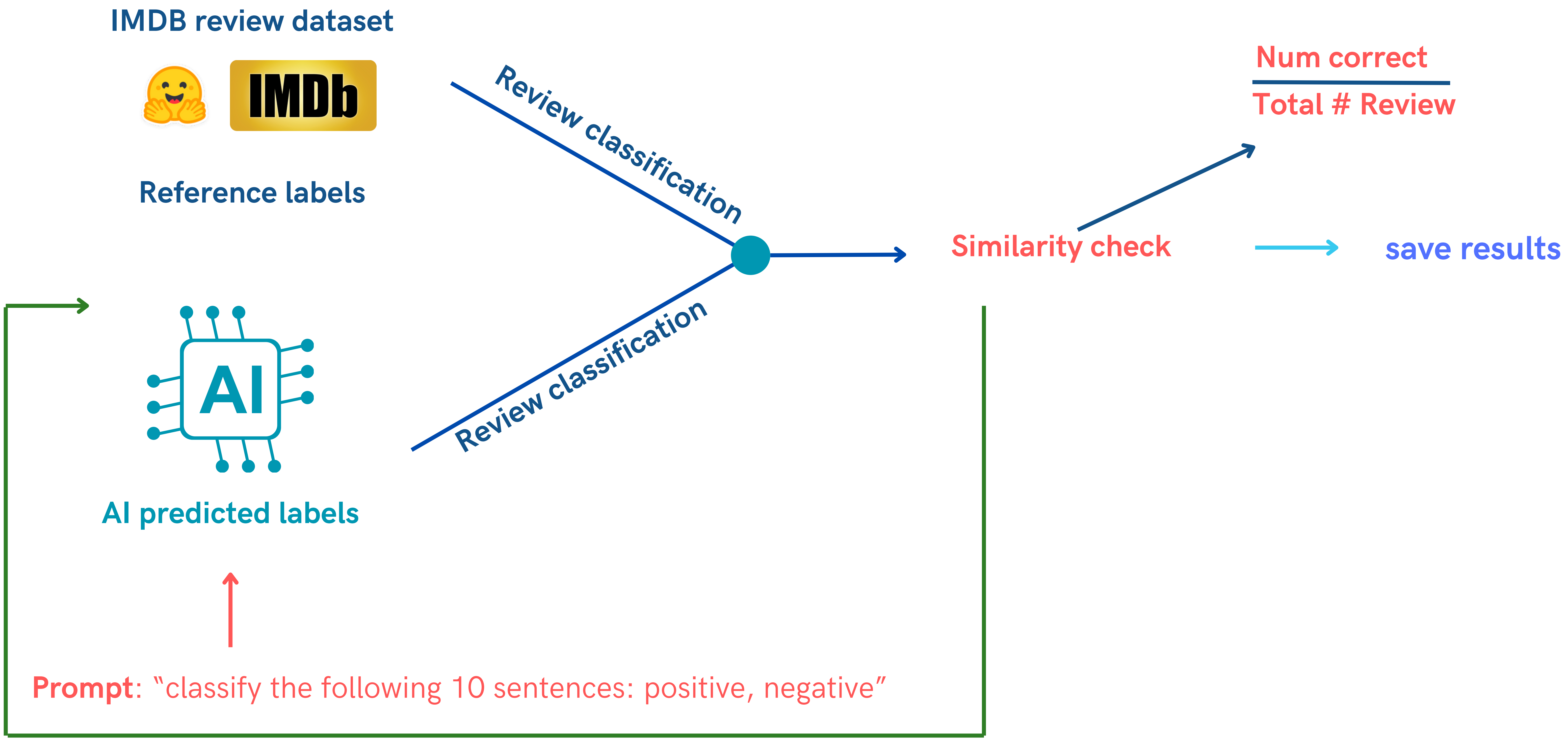
Prompt: "Generate a summary in one line (max 25 words) for each of the following documents / **articles**… "

Evaluating 10 articles

Average of precision of 1-gram, 2-gram,3-gram and 4-gram

3. Sentiment Analysis

Statistical Scores



IMDB review dataset

Reference labels

Review classification

Review classification

AI predicted labels

Similarity check

Num correct

Total # Review

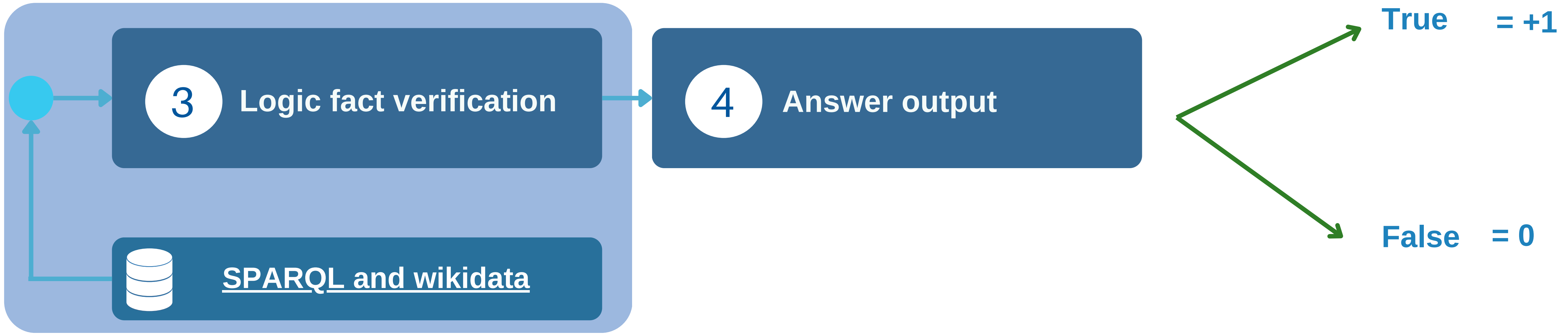
save results

Prompt: "classify the following 10 sentences: positive, negative"

Evaluating 10 random reviews

4. Fact Checker

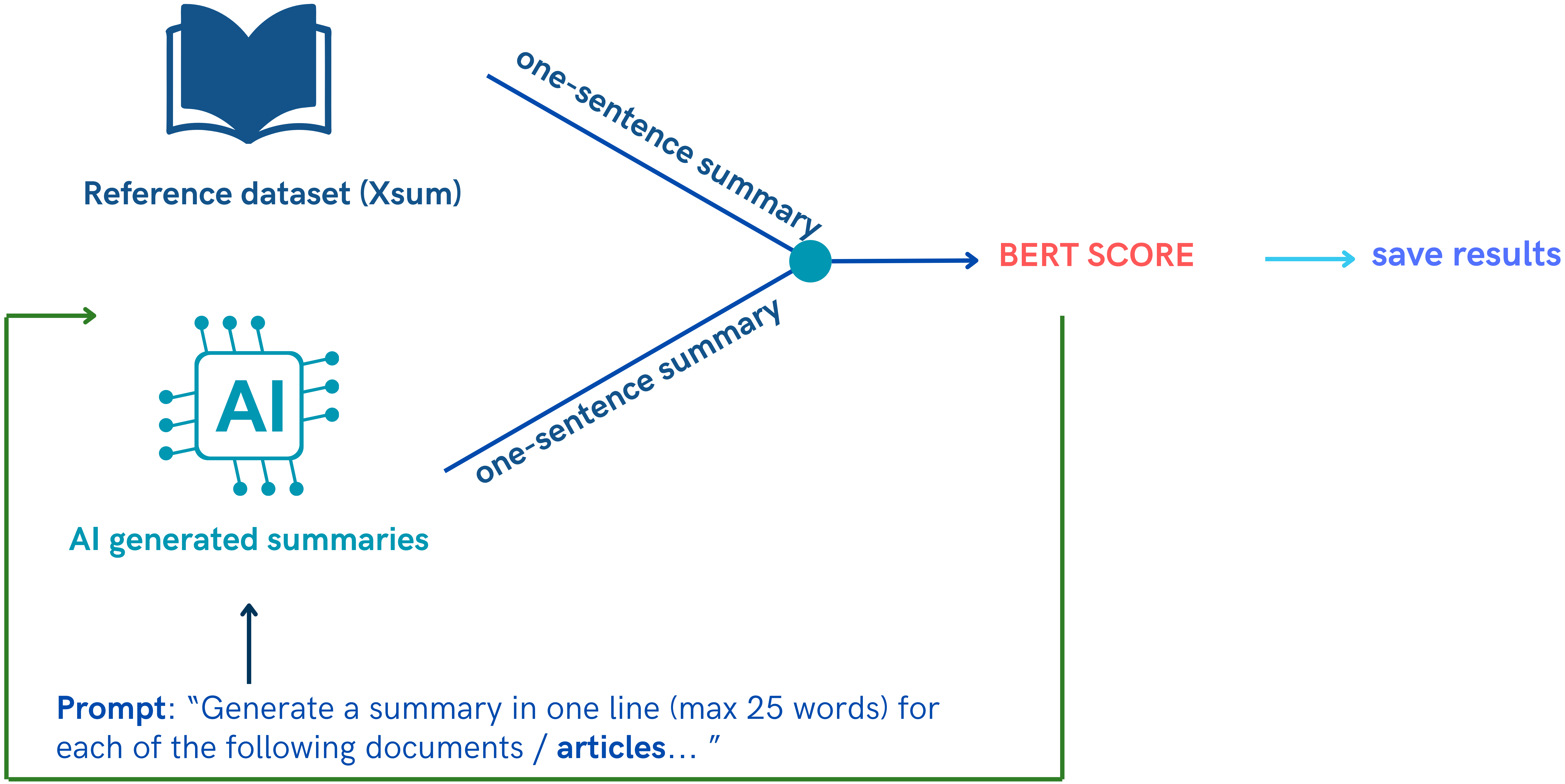
Fact Scores



Repeated with 10 different questions

5. BERT Score

Semantic Scores

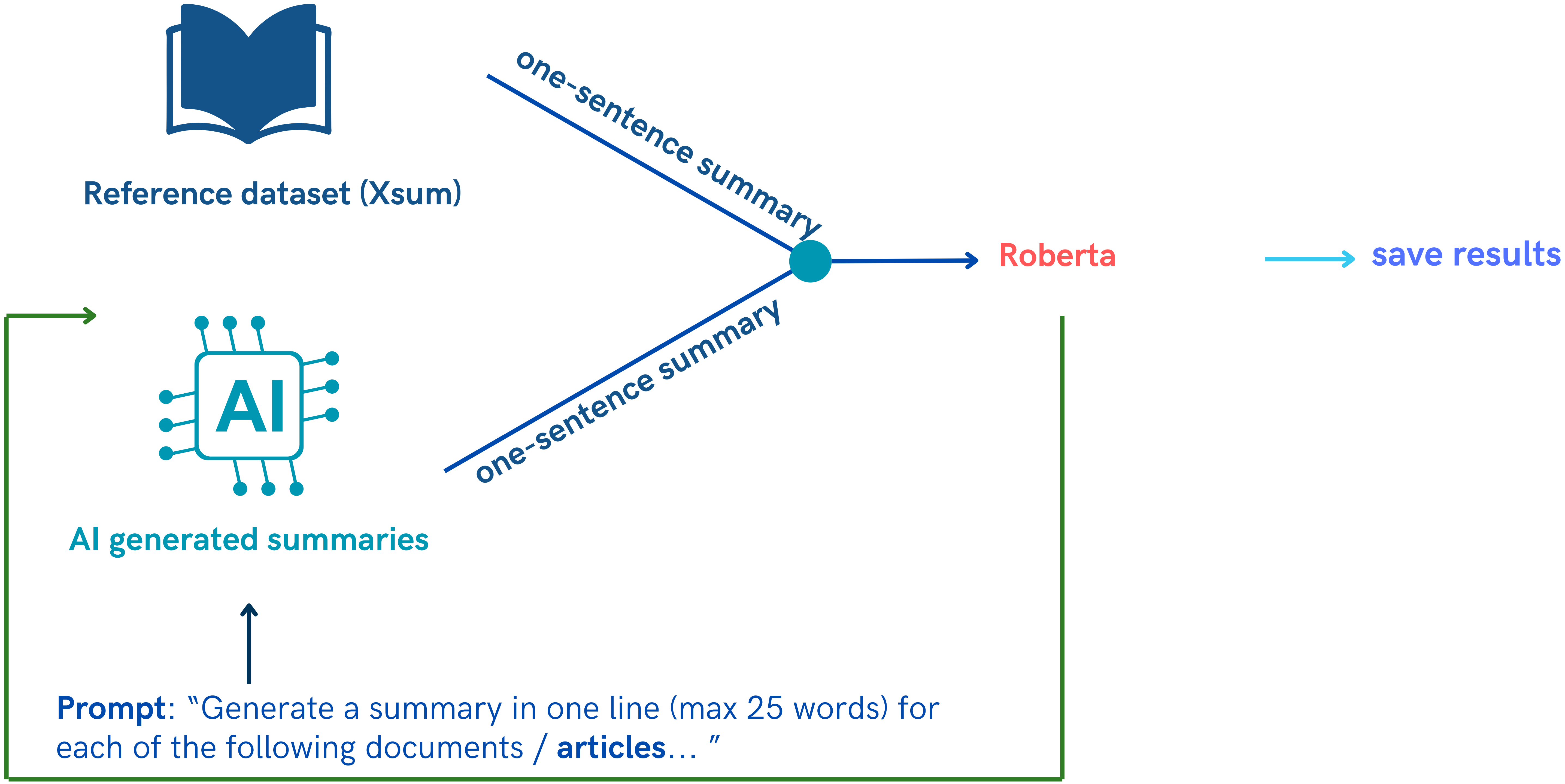


Prompt: "Generate a summary in one line (max 25 words) for each of the following documents / **articles**… "

Evaluating 10 articles

6. Roberta

Semantic Scores



Reference dataset (Xsum)

one-sentence summary

AI generated summaries

one-sentence summary

Roberta

save results

Prompt: "Generate a summary in one line (max 25 words) for each of the following documents / **articles**… "

Evaluating 10 articles

Computing results

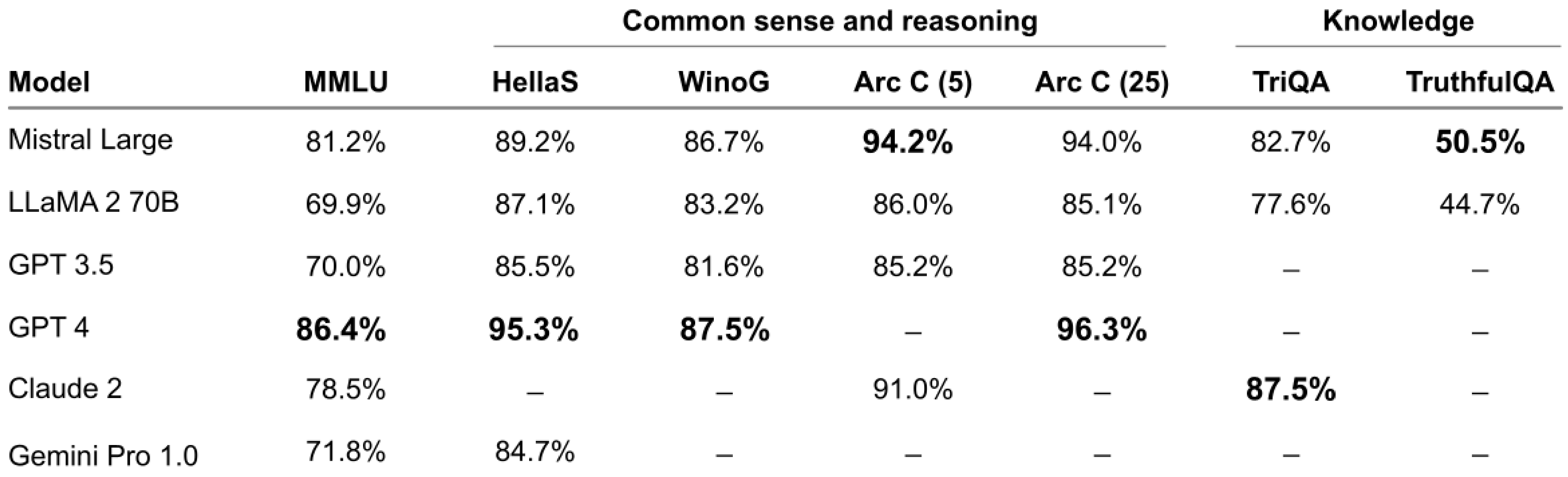
1. ROUGE
2. BLEU
3. Sentiment Anaysis
4. Fact Checking
5. BERT Score
6. Roberta

MISTRAL AI_

BLACKBOX AI

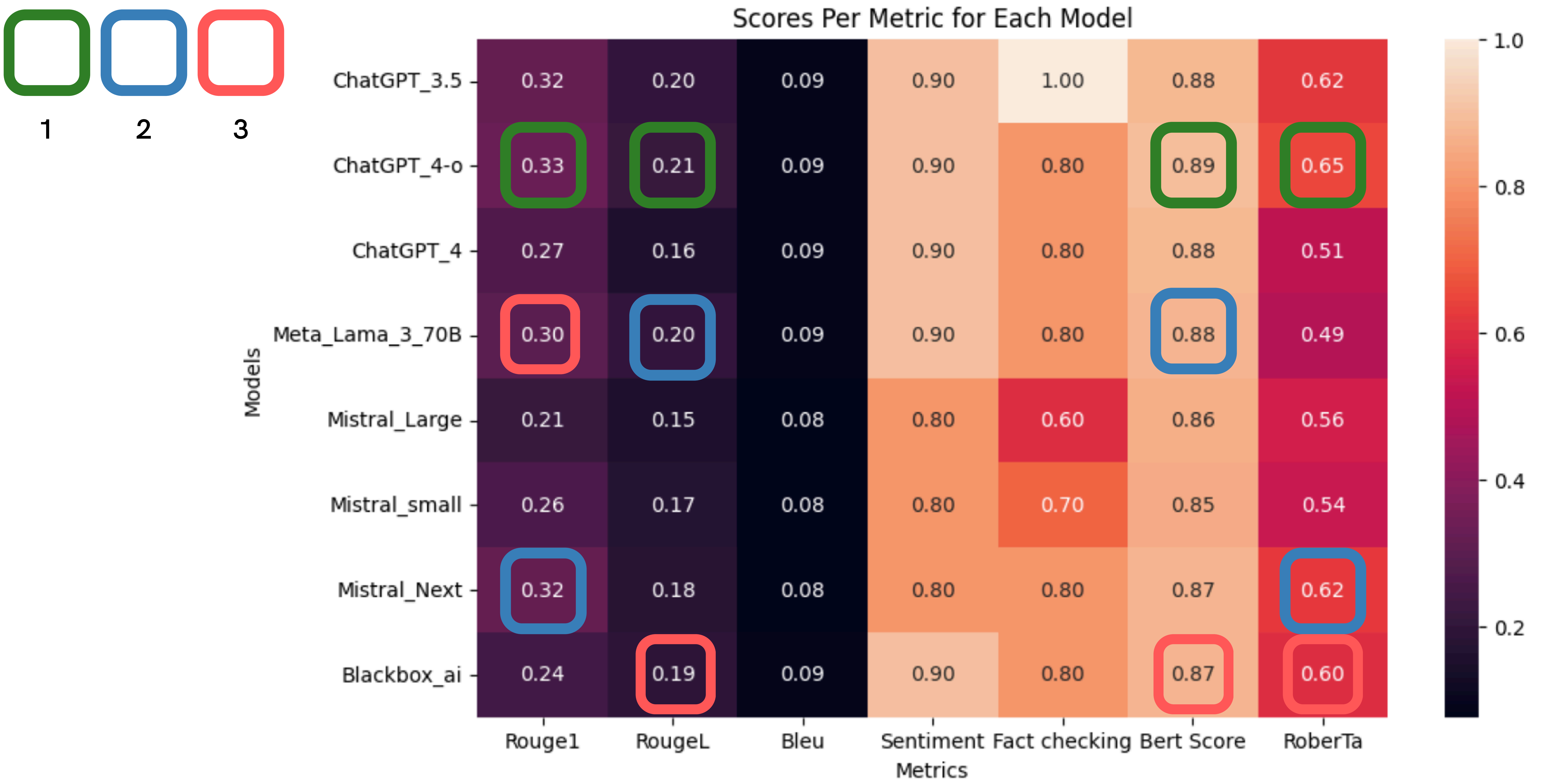
Comparing LLMS

GPT-4 Large Claude 2 Gemini Pro GPT-3.5 LLaMA 2 70B



		Common sense and reasoning				Knowledge	
Model	**MMLU**	**HellaS**	**WinoG**	**Arc C (5)**	**Arc C (25)**	**TriQA**	**TruthfulQA**
Mistral Large	81.2%	89.2%	86.7%	**94.2%**	94.0%	82.7%	**50.5%**
LLaMA 2 70B	69.9%	87.1%	83.2%	86.0%	85.1%	77.6%	44.7%
GPT 3.5	70.0%	85.5%	81.6%	85.2%	85.2%	–	–
GPT 4	**86.4%**	**95.3%**	**87.5%**	–	**96.3%**	–	–
Claude 2	78.5%	–	–	91.0%	–	**87.5%**	–
Gemini Pro 1.0	71.8%	84.7%	–	–	–	–	–

Comparing LLMS



Scores Per Metric for Each Model

Models	Rouge1	RougeL	Bleu	Sentiment	Fact checking	Bert Score	RoberTa
ChatGPT_3.5	0.32	0.20	0.09	0.90	1.00	0.88	0.62
ChatGPT_4-o	0.33	0.21	0.09	0.90	0.80	0.89	0.65
ChatGPT_4	0.27	0.16	0.09	0.90	0.80	0.88	0.51
Meta_Lama_3_70B	0.30	0.20	0.09	0.90	0.80	0.88	0.49
Mistral_Large	0.21	0.15	0.08	0.80	0.60	0.86	0.56
Mistral_small	0.26	0.17	0.08	0.80	0.70	0.85	0.54
Mistral_Next	0.32	0.18	0.08	0.80	0.80	0.87	0.62
Blackbox_ai	0.24	0.19	0.09	0.90	0.80	0.87	0.60

Findings and improvements

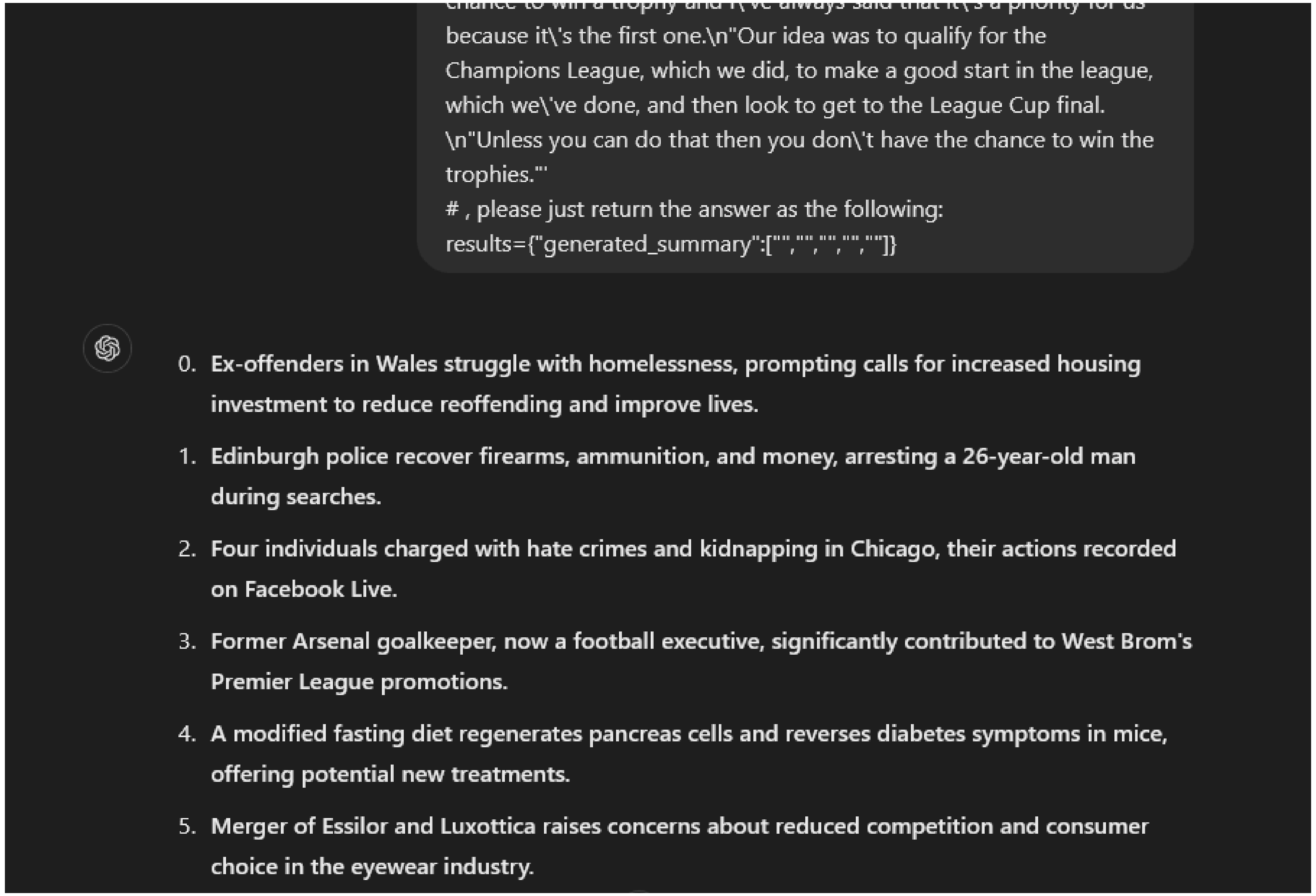
Findings

- Different metrics are interesting
- Weighting the metrics was not useful in the end
- Variability in Prompt Responses

Improvements for testing

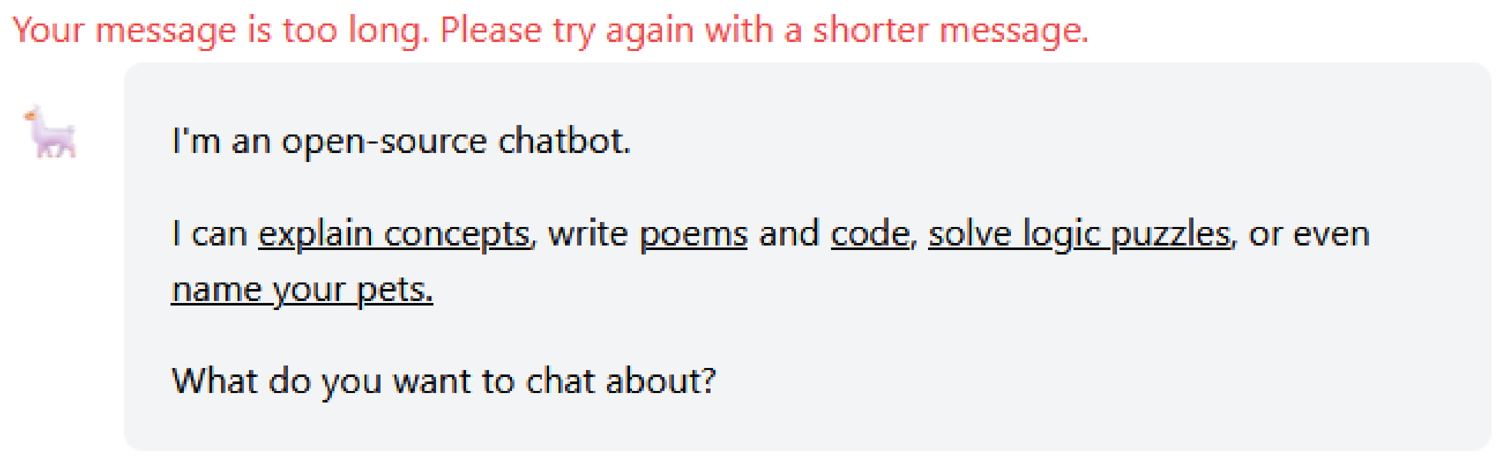
- Adaptive Prompt Engineering

Findings and improvements



chance to win a trophy and I\'ve always said that it\'s a priority for us because it\'s the first one.\n"Our idea was to qualify for the Champions League, which we did, to make a good start in the league, which we\'ve done, and then look to get to the League Cup final. \n"Unless you can do that then you don\'t have the chance to win the trophies."'
# , please just return the answer as the following:
results={"generated_summary":["","","","",""]}

0. Ex-offenders in Wales struggle with homelessness, prompting calls for increased housing investment to reduce reoffending and improve lives.
1. Edinburgh police recover firearms, ammunition, and money, arresting a 26-year-old man during searches.
2. Four individuals charged with hate crimes and kidnapping in Chicago, their actions recorded on Facebook Live.
3. Former Arsenal goalkeeper, now a football executive, significantly contributed to West Brom's Premier League promotions.
4. A modified fasting diet regenerates pancreas cells and reverses diabetes symptoms in mice, offering potential new treatments.
5. Merger of Essilor and Luxottica raises concerns about reduced competition and consumer choice in the eyewear industry.



Your message is too long. Please try again with a shorter message.

I'm an open-source chatbot.

I can explain concepts, write poems and code, solve logic puzzles, or even name your pets.

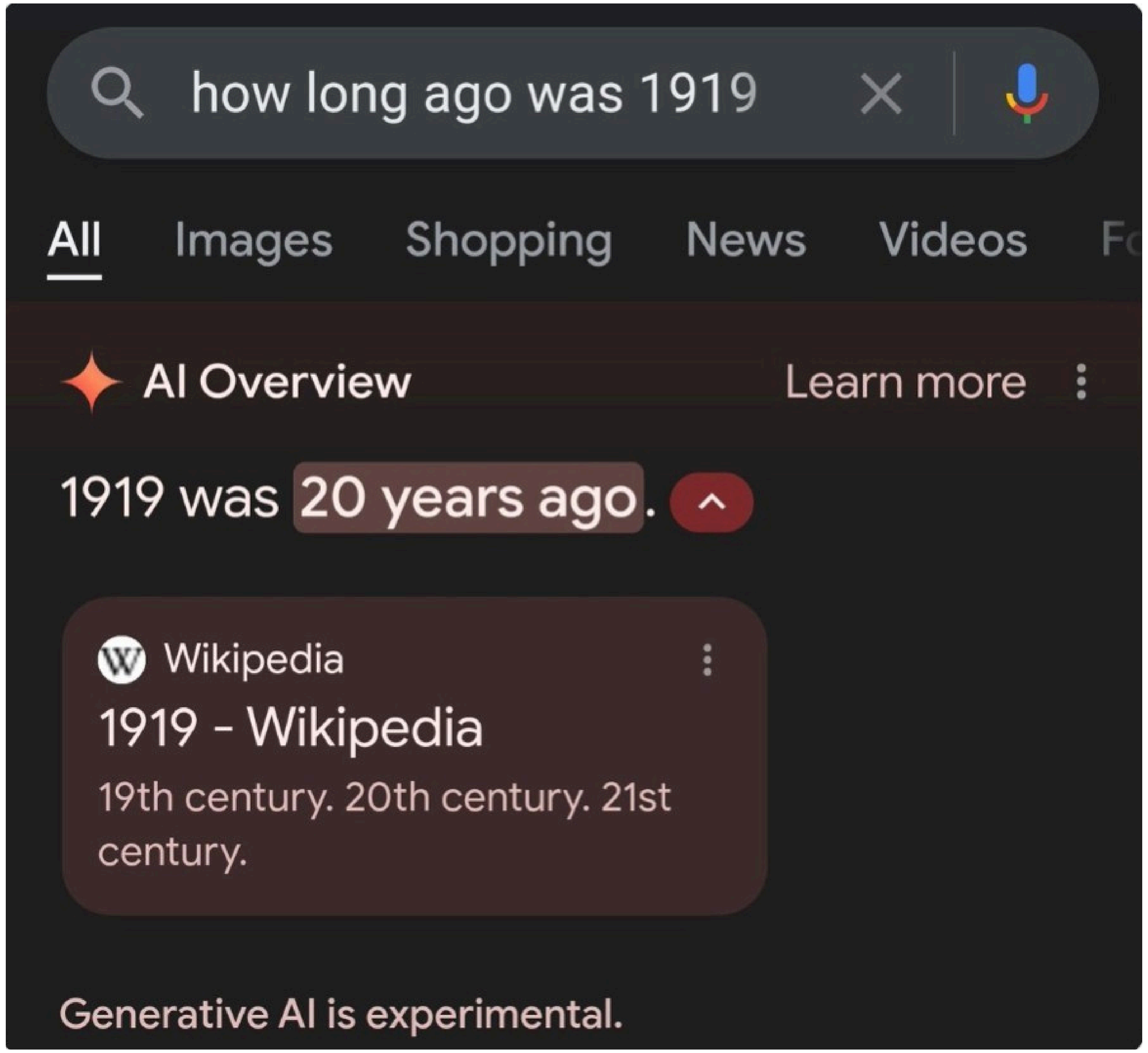
What do you want to chat about?

Improvement suggestion for LLMs

Improvements

- Post-process on answers
- Iterative answer generation
- Fine-tune on references
- Get them access to do calculation

State of the art :)



how long ago was 1919

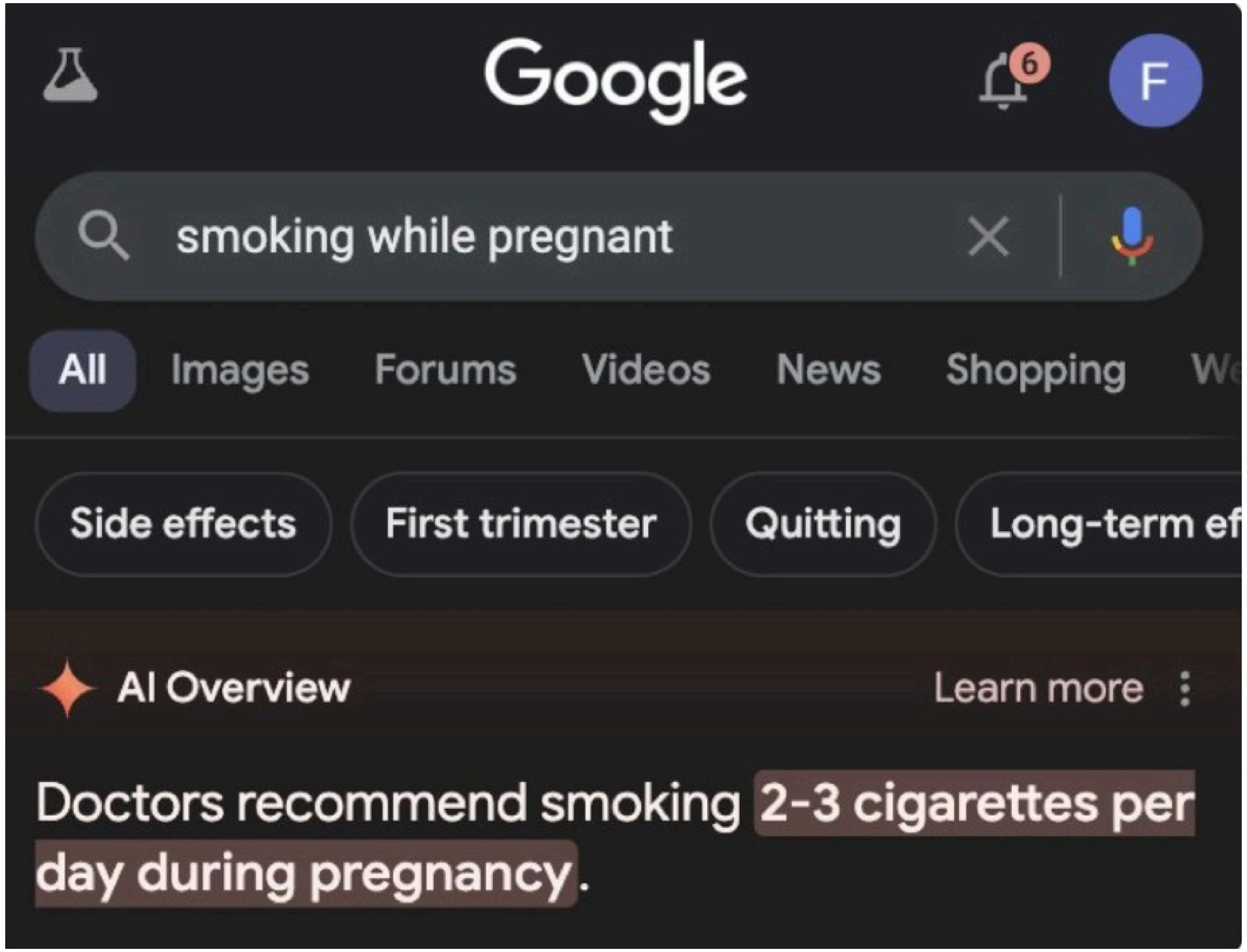
1919 was 20 years ago.

Wikipedia

1919 - Wikipedia

19th century. 20th century. 21st century.

Generative AI is experimental.



smoking while pregnant

Doctors recommend smoking 2-3 cigarettes per day during pregnancy.

Tentative Planning

1 12.03

- Brainstorming
- Determine Project Scope
- Presentation 1

2 26.03

- SparQL
- Defining metrics

3 16.04

- Re-iterating on metrics
- Refining queries

4 30.04

- Re-iterating on metrics
- Refining queries

5 14.05

- Finish documenting

6 28.05

- Final pres.

Final Presentation

End for semester!