Step-by-Step Claim Verification Using LLMs and Knowledge Graphs

Ivana Hacajová

24.6.2024, Master's Thesis Final Presentation

Chair of Software Engineering for Business Information Systems (sebis)
Department of Computer Science
School of Computation, Information and Technology (CIT)
Technical University of Munich (TUM)
wwwmatthes.in.tum.de

Outline

Motivation

Research Questions

Methods

- Base question-guided step-by-step pipeline
- Predicate pipeline
- Knowledge graph pipeline

Results

Conclusion

Motivation

Need to verify claims

Internet is flooded by often misleading or untrue claims about health, climate change, politics, public affairs, ...

LLMs

They can generate highly readable textual outputs, can be used as a source of information, can answer questions and perform tasks when shown how to

Knowledge graphs

Can serve as a reliable knowledge source, as opposed to LLMs which often hallucinate

Domain specificity

The recent claim verifiers using LLMs were not tested on domain specific datasets

Research Questions

RQ1

How can the use of LLMs help claim verification?

RQ2

Does leveraging knowledge from knowledge graphs and structured reasoning improve performance?

RQ3

How do different domains compare in this task?

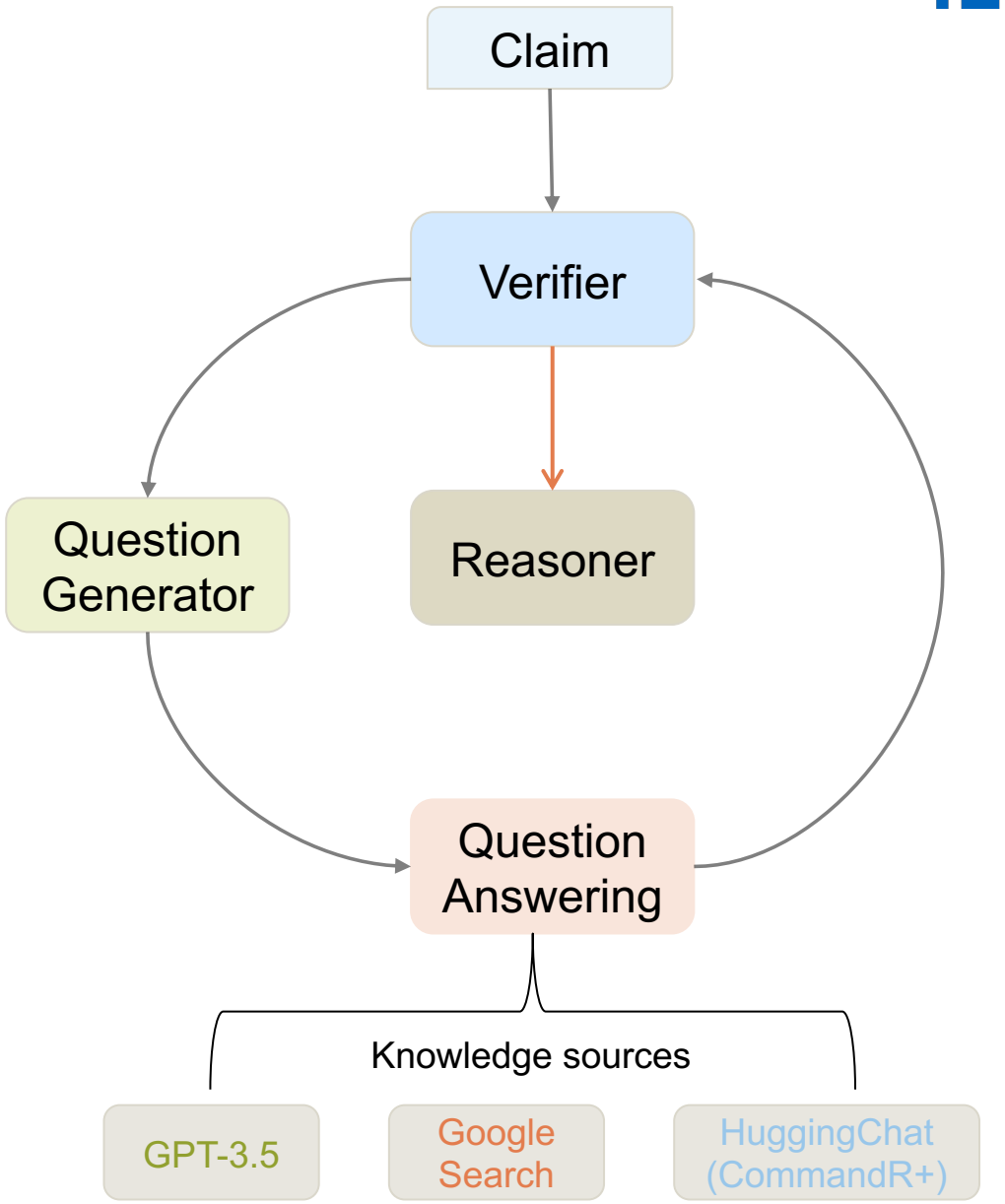
Methods

Base Pipeline

- Based on QACheck
- Generates a simple question from a claim, finds an answer, repeats until enough evidence is collected

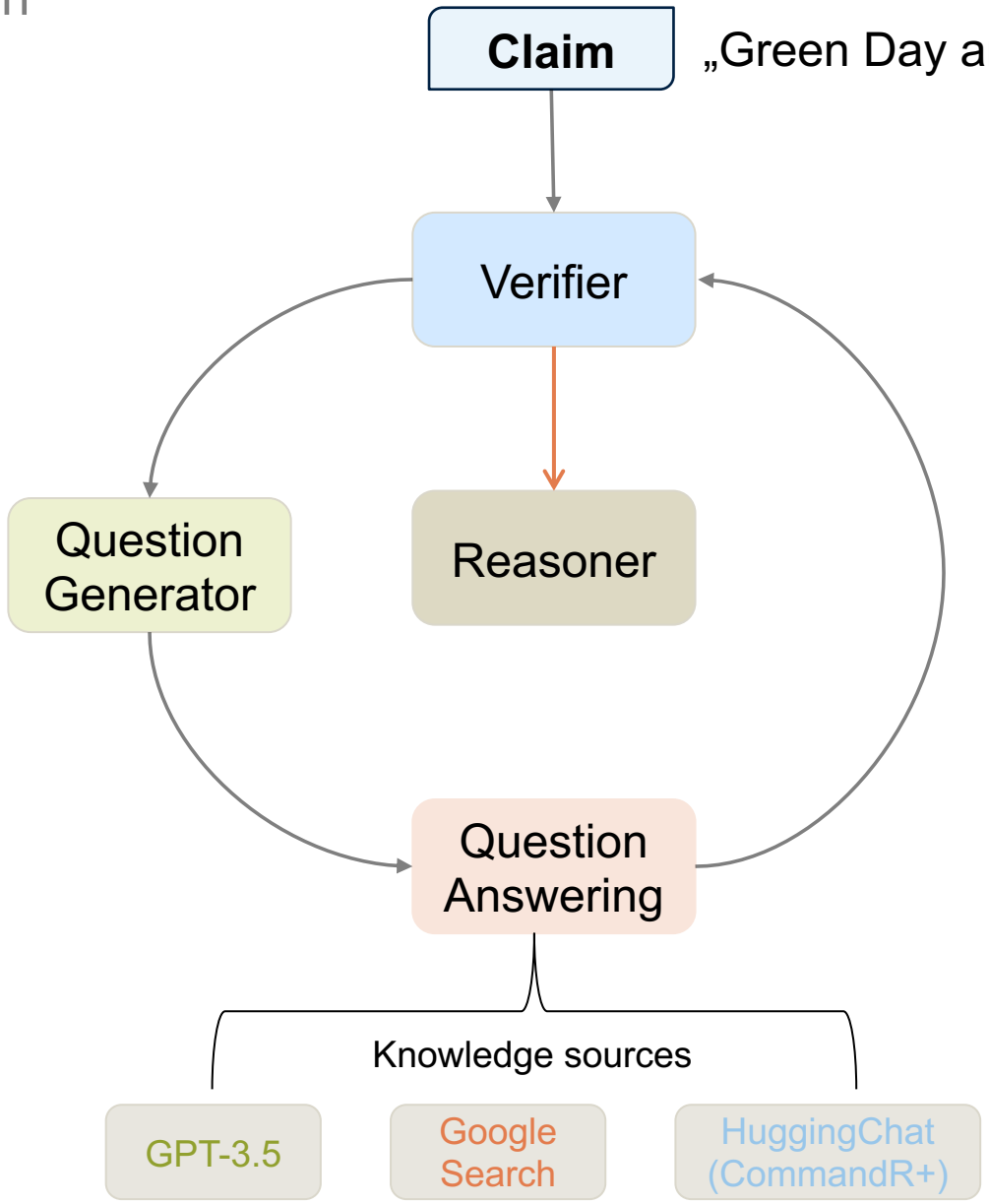
Main features:

- LLM guides each step
- Does not need training or fine-tuning
- Does not require labeled data and evidence to train
- The collection of questions and answers serves the purpose of explanaiton of the veracity label
- In-context learning used to perform intermediate tasks



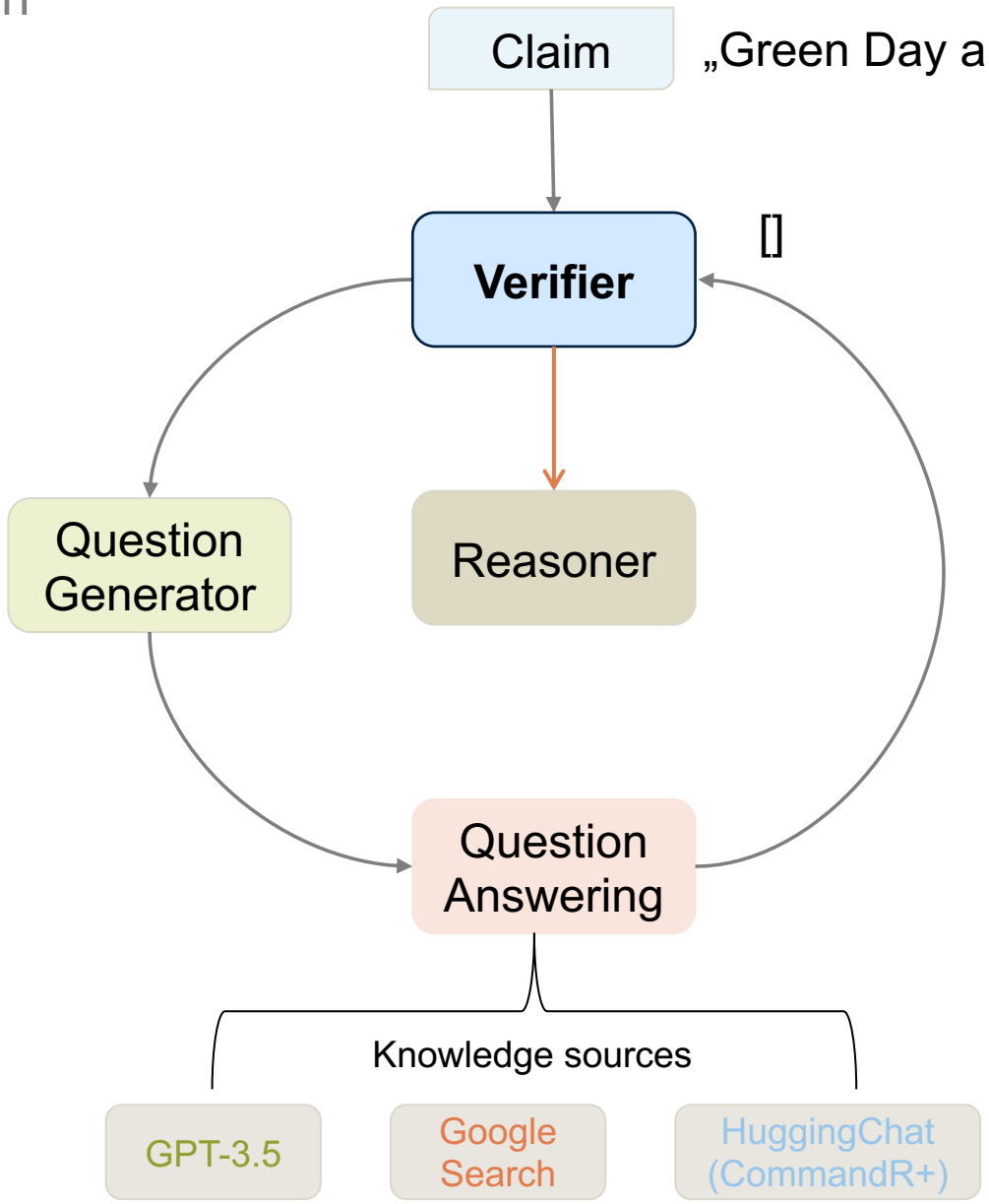
Methods

Base Pipeline - demonstration



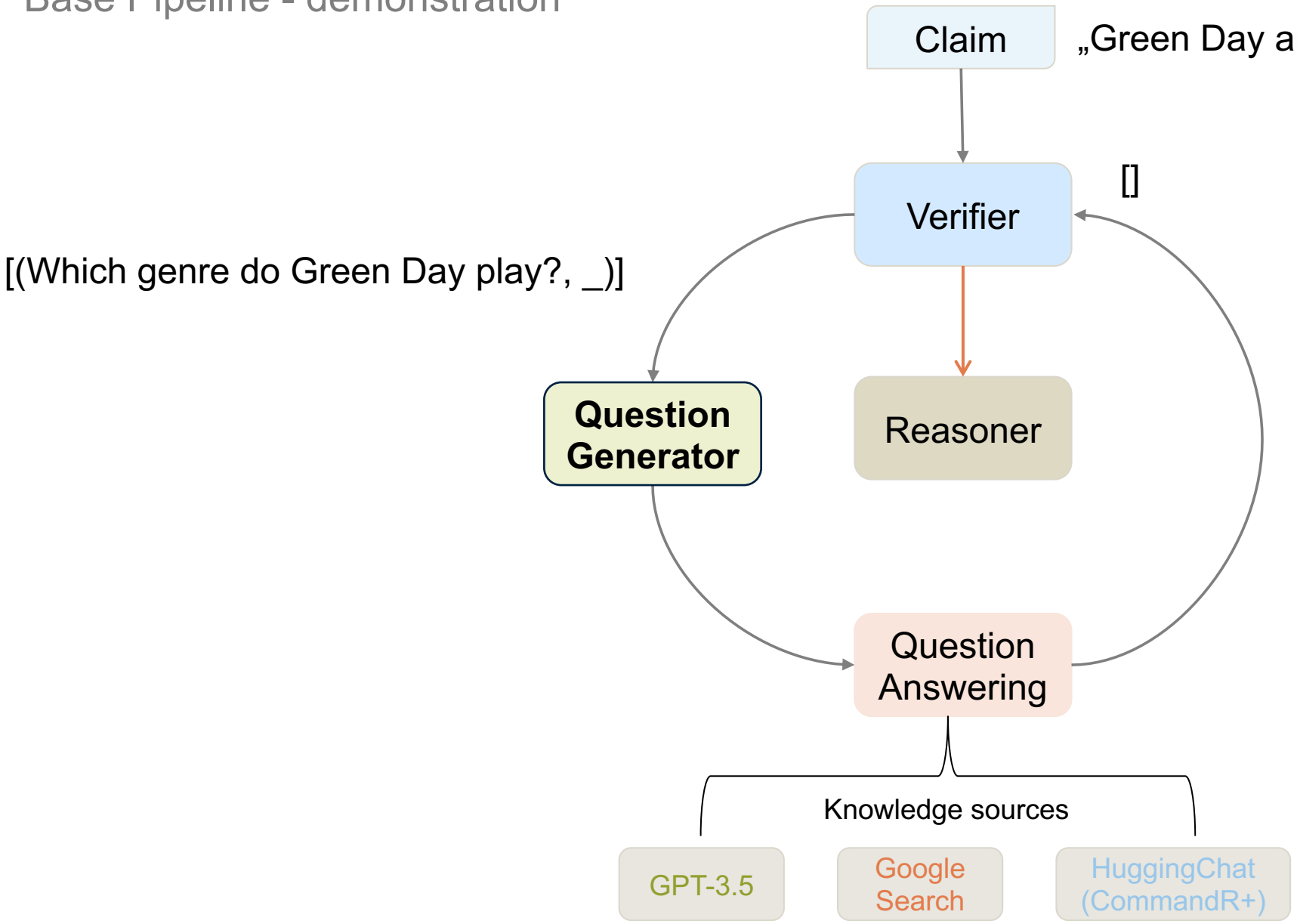
Methods

Base Pipeline - demonstration



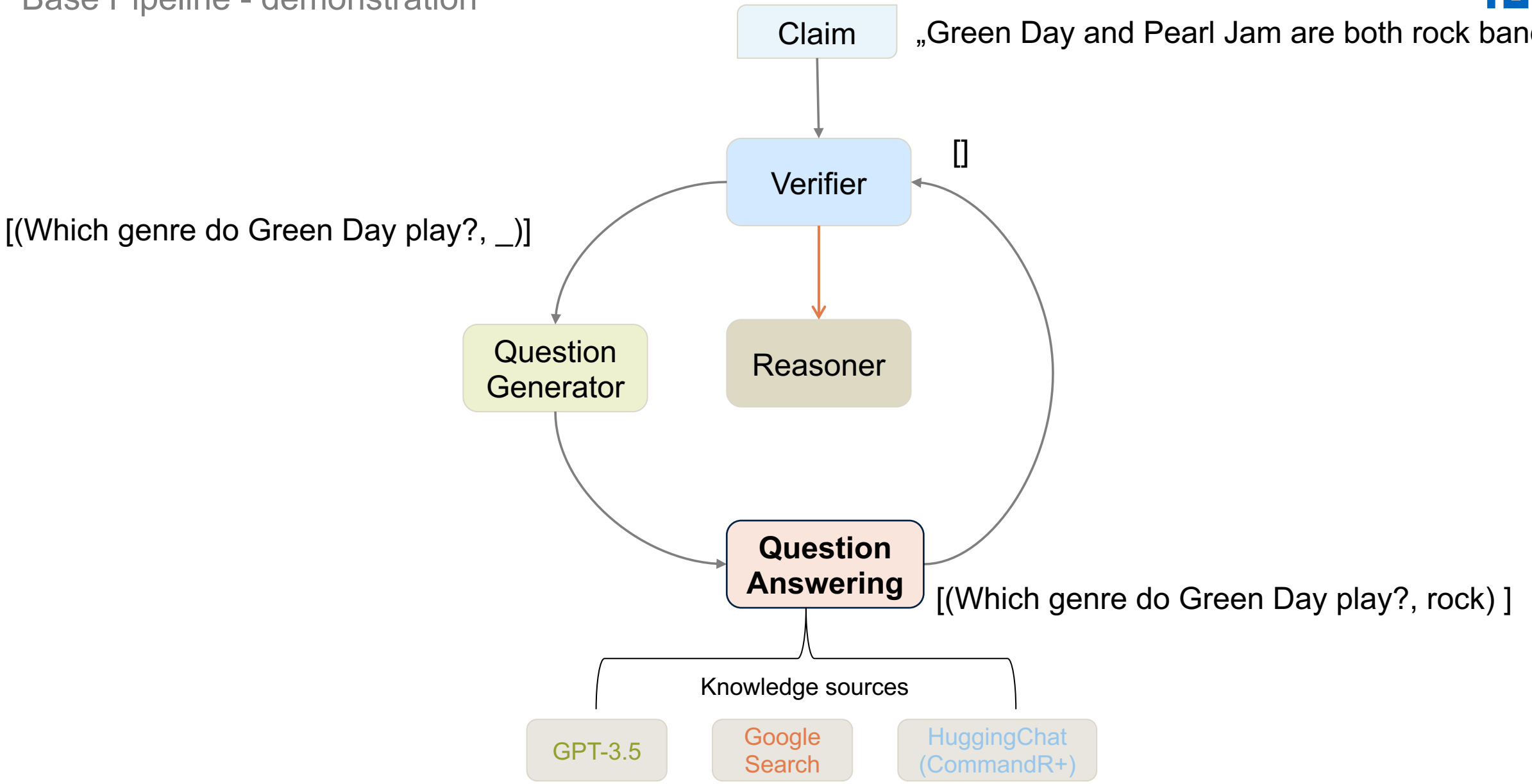
Methods

Base Pipeline - demonstration



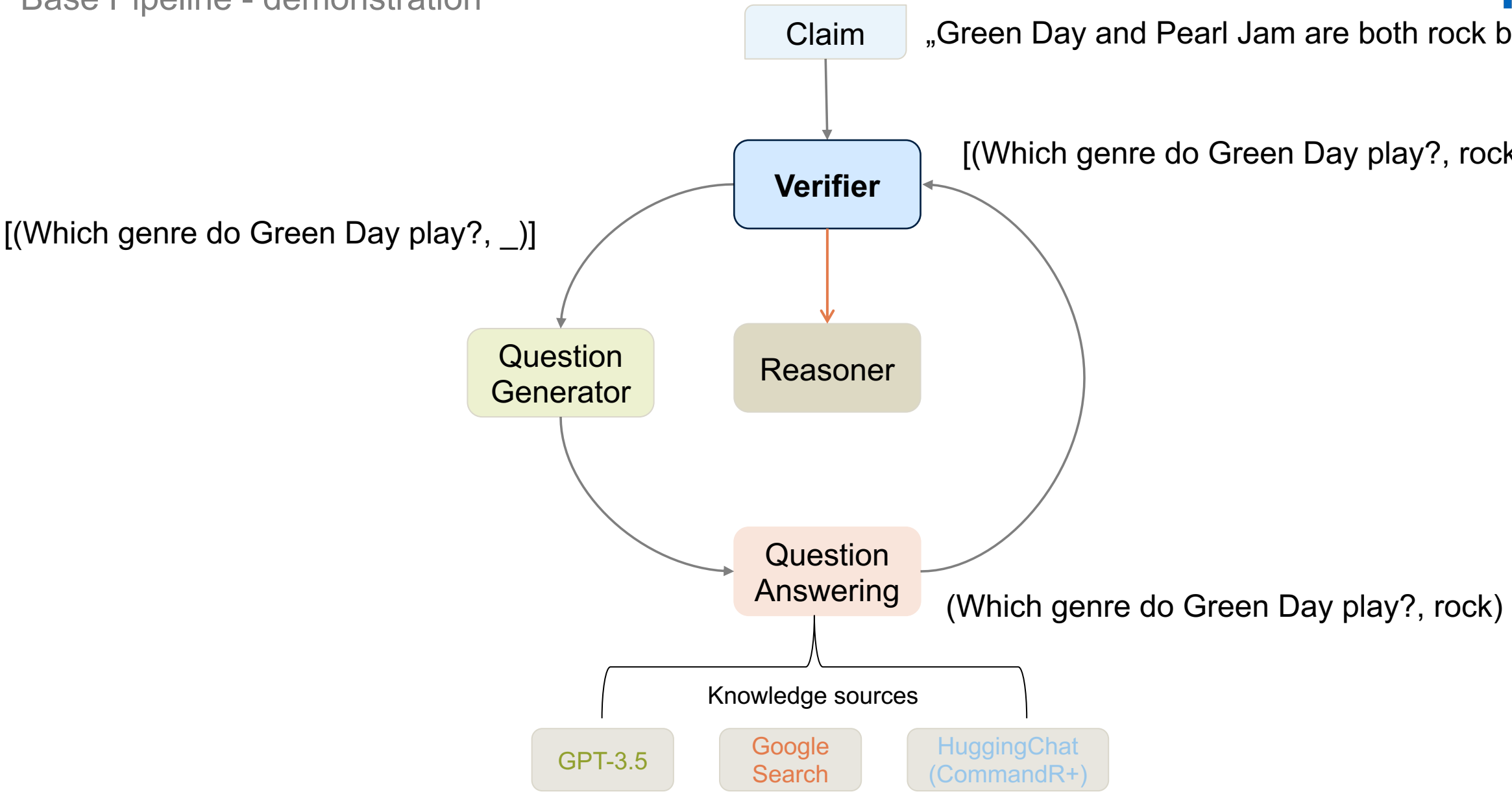
Methods

Base Pipeline - demonstration



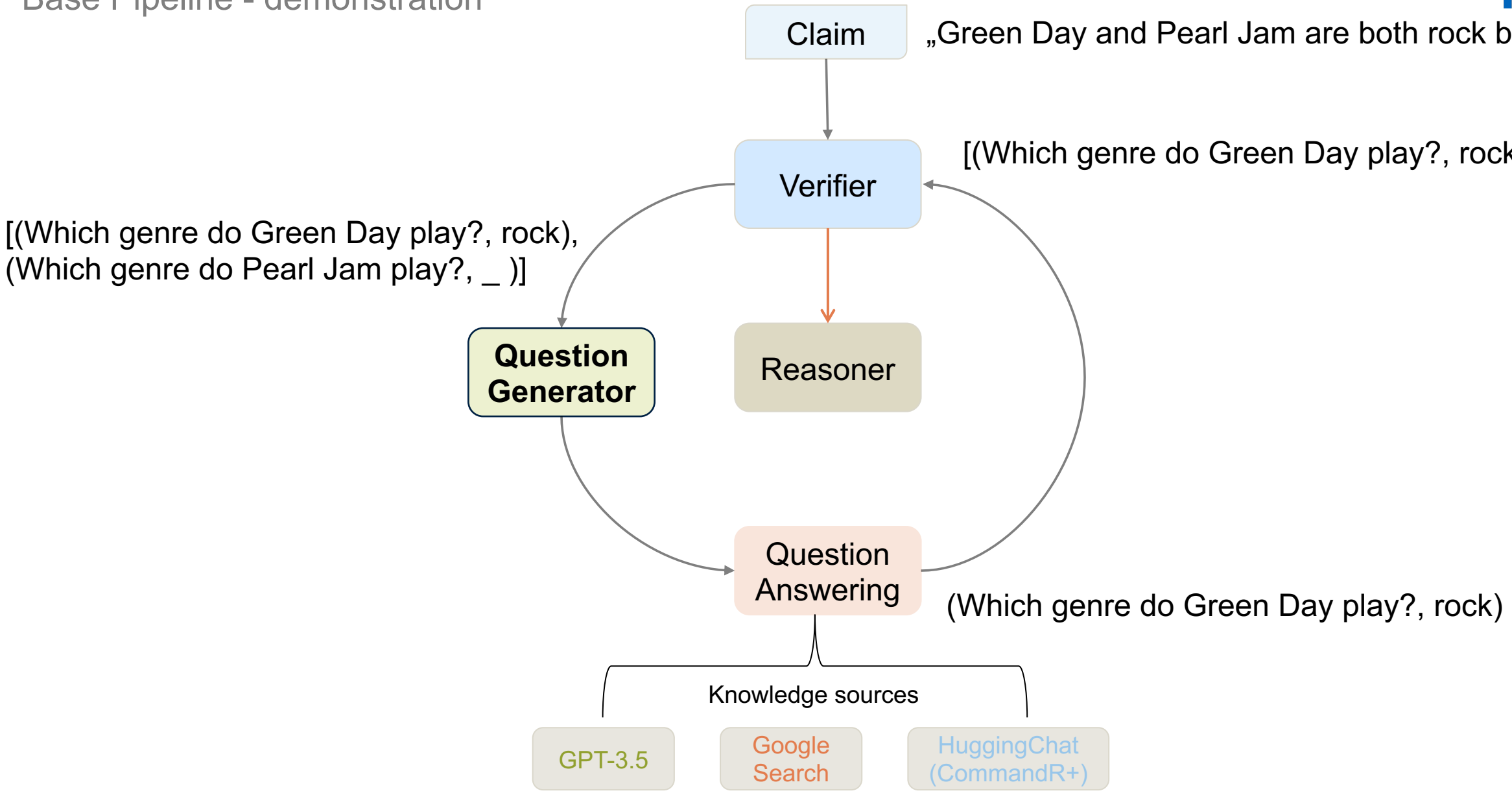
Methods

Base Pipeline - demonstration



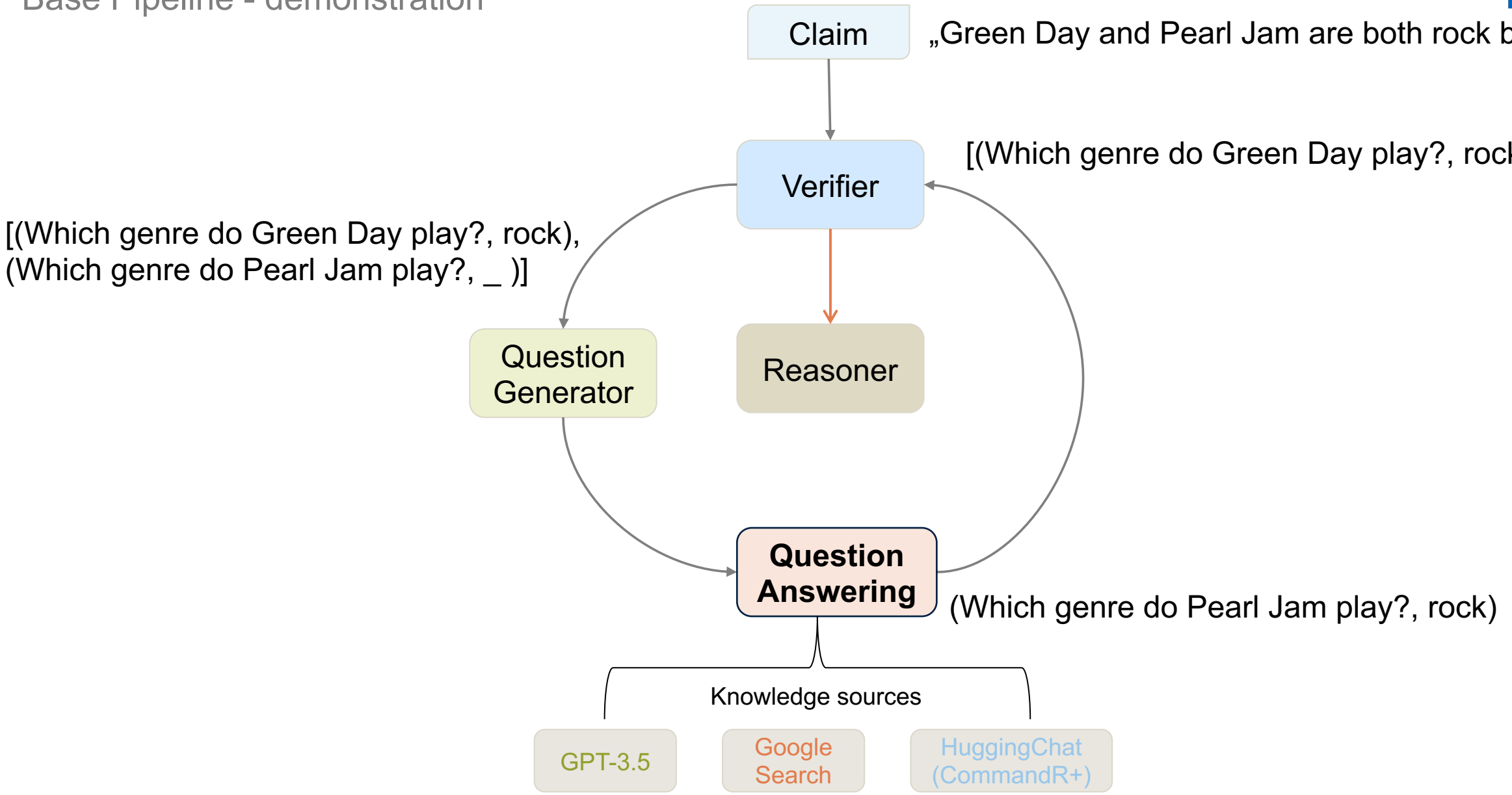
Methods

Base Pipeline - demonstration



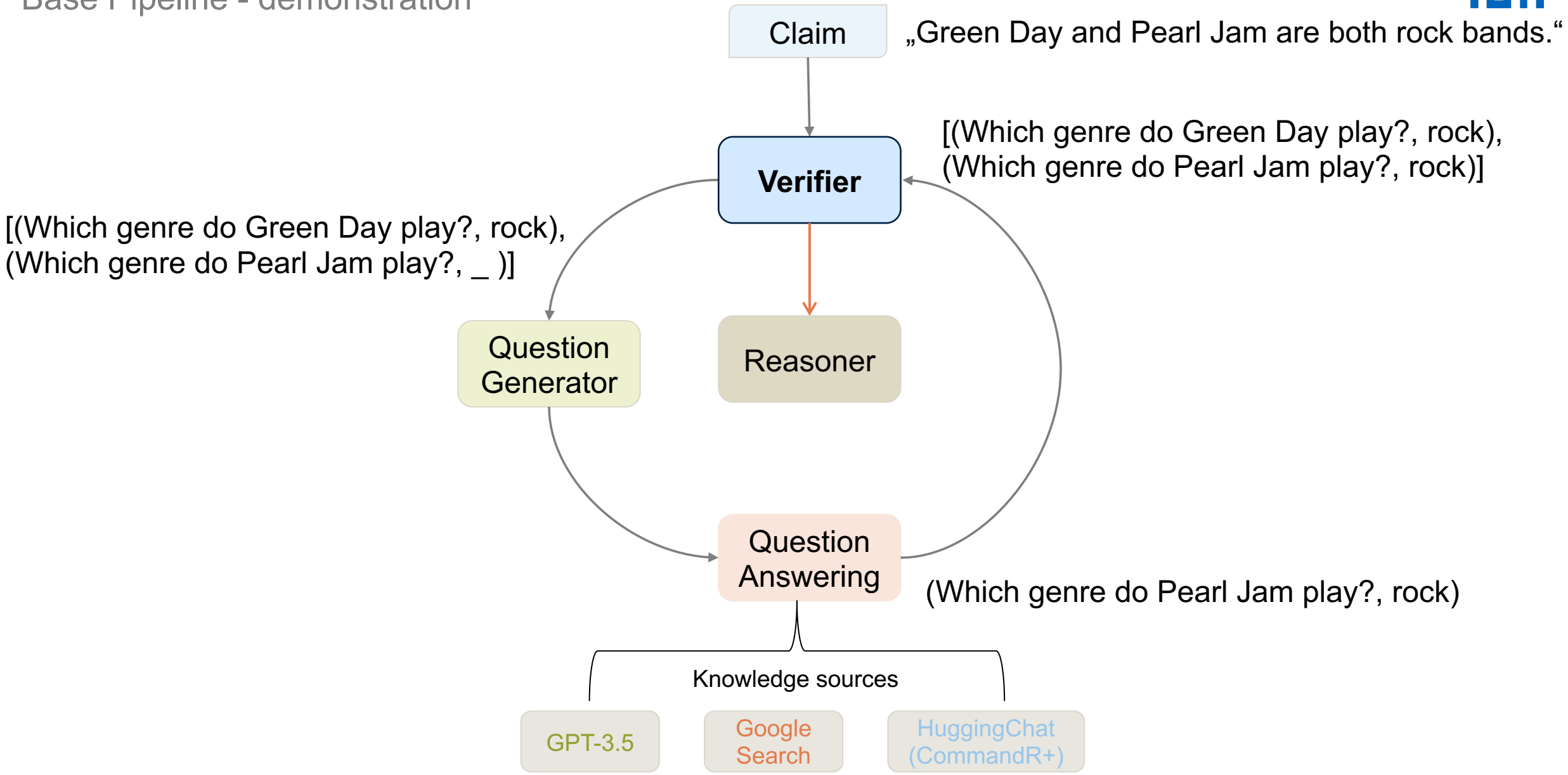
Methods

Base Pipeline - demonstration



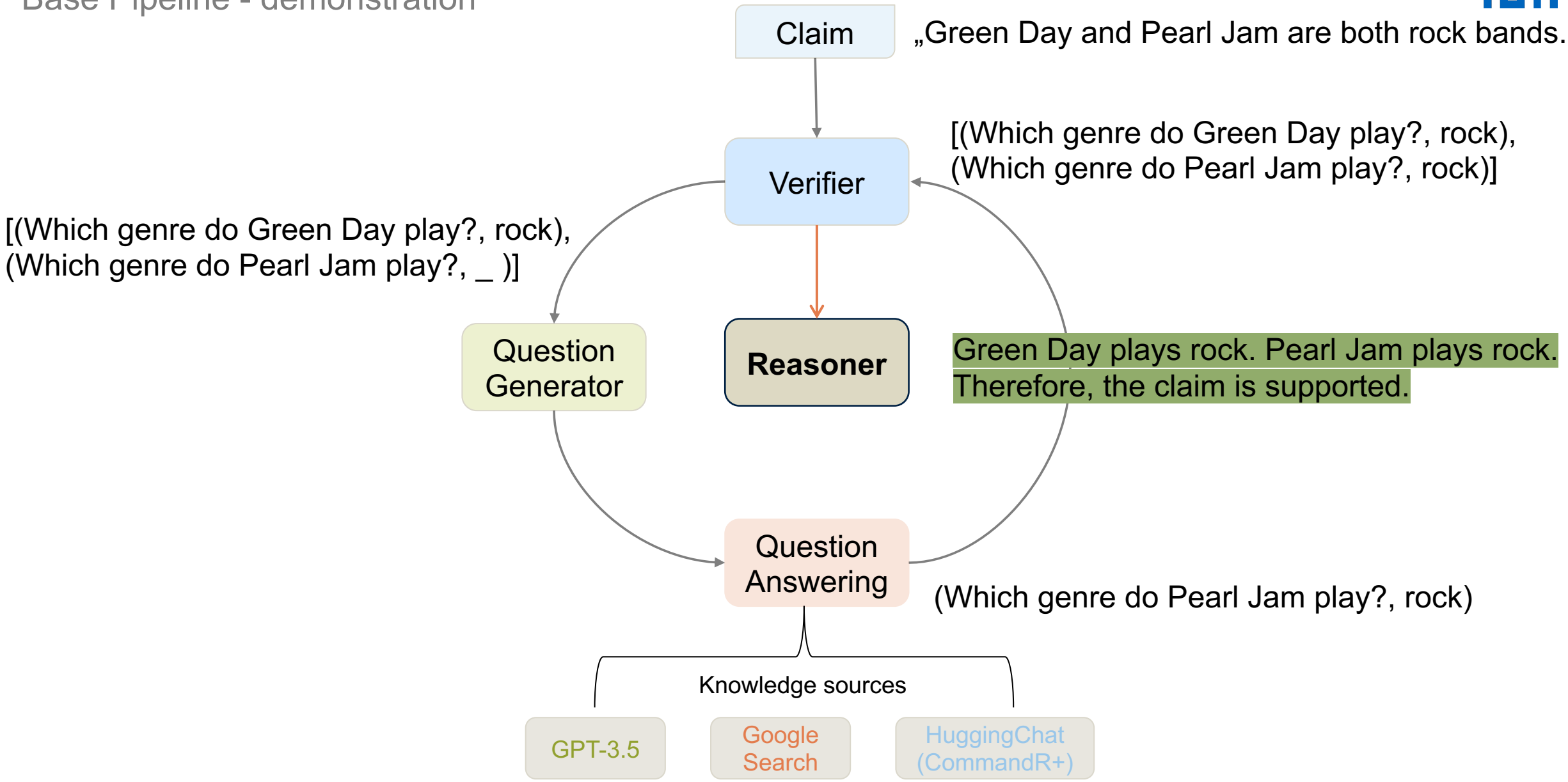
Methods

Base Pipeline - demonstration



Methods

Base Pipeline - demonstration



Methods

Example from the WebGUI

The Kentucky Department of Corrections is headquarted along the Kentucky River.

True

The headquarters of the Kentucky Department of Corrections is located in Frankfort, Kentucky. The Kentucky River runs along Frankfort, Kentucky. Therefore, the final answer is: True.

Where is the headquarters of the Kentucky Department of Corrections located?

Frankfort, Kentucky

The headquarters of the Kentucky Department of Corrections is located in Frankfort, Kentucky. For more information, you can refer to the Wikipedia article on the Kentucky Department of Corrections: https://en.wikipedia.org/wiki/Kentucky_Department_of_Corrections

Which river runs along Frankfort, Kentucky?

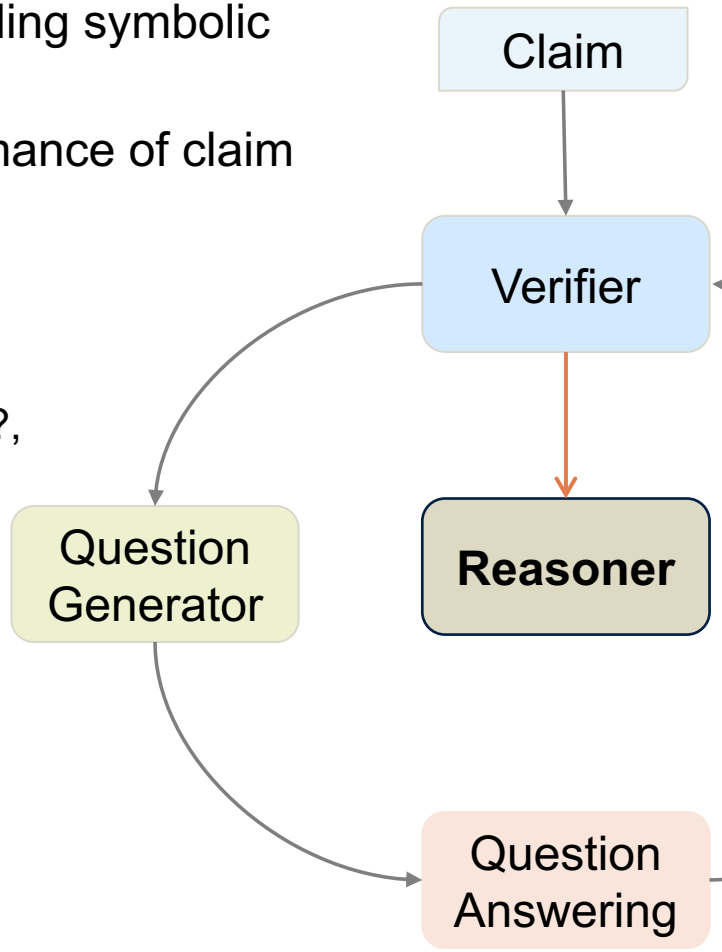
The Kentucky River.

The Kentucky River runs along Frankfort, Kentucky. Here is a relevant Wikipedia article: https://en.wikipedia.org/wiki/Kentucky_River

Methods

Predicate Pipeline

- Added predicate representation
- LLMs are good at understanding symbolic representations
- It has shown to boost performance of claim verification



„Green Day and Pearl Jam are both rock bands.“

[(Which genre do Green Day play?, **Genre(Green Day, rock),** rock), (Which genre do Pearl Jam play?, **Genre(Pearl Jam, rock),** rock)]

[(Which genre do Green Day play?, **Genre(Green Day, rock),** rock), (Which genre do Pearl Jam play?, **Genre(Pearl Jam, rock),** _)]

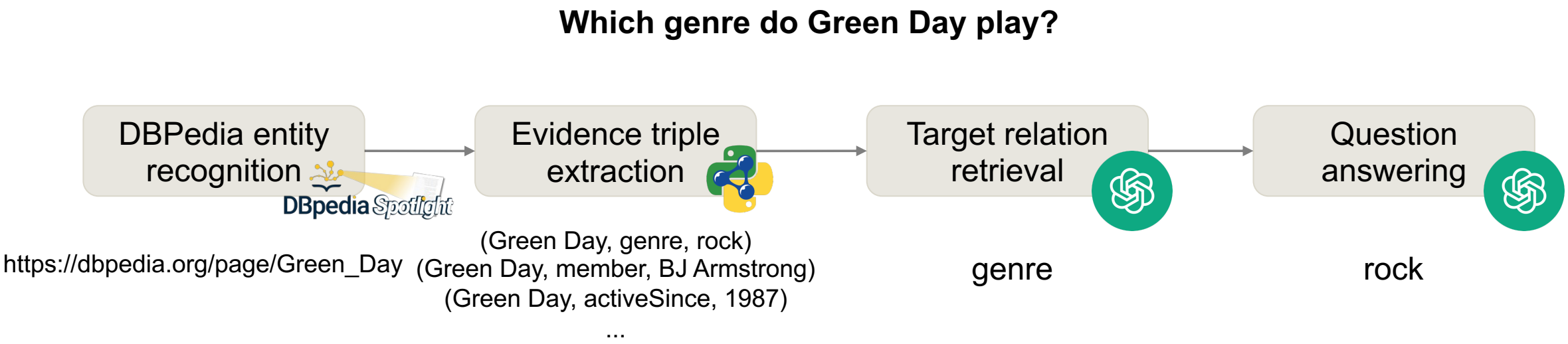
Genre(Green Day, rock) is True
Genre(Pearl Jam, rock) is True
Therefore, the claim is supported.

(Which genre do Pearl Jam play?, rock)

Methods

Knowledge Graphs Pipeline

- The source of evidence is a KG (here DBPedia)
- Same as the base pipeline, only the question answering module works differently
- LLM is used to process evidence from KG



Results

Perfomance of the base and predicate pipelines

Base pipeline outperformed many step-by step and direct claim verification models, except FOLK.

Predicate pipeline improved performance of the base pipeline only for the most complicated claims.

F1 scores of the base and predicate pipeline and of models from literature

Method	Reasoning	Evidence	HOVER		
			2-HOP	3-HOP	4-HOP
InstructGPT - direct	-	InstructGPT	56.51	51.75	49.68
Codex - direct	-	Codex	55.57	53.42	45.59
ProgramFC	Codex	FLAN-T5	54.27	54.18	52.88
FOLK	GPT-3.5	Google snippets	**66.26**	54.80	**60.35**
QACheck	InstructGPT	InstructGPT	55.67	54.67	52.35
Base pipeline	GPT-3.5	GPT-3.5	60.02	**55.66**	49.52
Predicate pipeline	GPT-3.5	GPT-3.5	58.55	54.39	51.9

Results

Performance of different evidence sources

Choice of the evidence source did not effect the performance much, when using GPT-3.5 for the reasoning.

For the most complicated claims it seems the reasoning capabilities are the bottle neck.

Table 5.3: F1 scores for the base pipeline and different evidence sources (GPT-3.5, Google snippets and CommandR+ with web search) on the HOVER dataset, separated by the number of hops.

Method	Reasoning	Evidence	HOVER		
			2-HOP	3-HOP	4-HOP
Base pipeline	GPT-3.5	GPT-3.5	60.02	**55.66**	49.52
		Google snippets	60.05	53.55	**50.21**
		CommandR+ w/ web search	**64.55**	52.47	49.89

Results

Performance of different LLMs

Huge performance boost when using Mixtral-8x7b instead of GPT-3.5.

Improvement in performance is also suggested by the number of generated questions, which goes consistently up with increasing number of hops.

Table 5.4: F1 scores for the base pipeline and different reasoning LLMs: GPT-3.5 and Mixtral-8x7b on the HOVER dataset, separated by the number of hops.

Method	Reasoning	Evidence	HOVER		
			2-HOP	3-HOP	4-HOP
Base pipeline	GPT-3.5	GPT-3.5	60.02	55.66	49.52
	Mixtral-8x7b	GPT-3.5	61.25	60.14	58.94

Table 5.5: Average number of generated questions for different pipelines on the HOVER dataset, separated by the number of hops.

Method	Reasoning	Evidence	HOVER		
			2-HOP	3-HOP	4-HOP
Base pipeline	GPT-3.5	GPT-3.5	2.30	2.28	2.34
		Google	2.24	2.28	2.28
		CommandR+ w/ web search	2.14	2.29	2.32
	Mixtral-8x7b	GPT-3.5	2.56	2.82	3.39
Predicate pipeline	GPT-3.5	GPT-3.5	2.75	2.86	2.89

Results

Different domain datasets

Claims from HOVER are much more difficult to verify than the domain specific, real-life claims.

The climate related claims have the best performance, then COVID-19 and the general health related claims turned out to be most difficult.

We managed to surpass baselines for binary and ternary classification (NotEnoughEvidence label), which were trained and fine tuned for the task and the domain.

Binary classification

Dataset	Model	Reasoning	Evidence	F1 score
HOVER	Base pipeline	GPT-3.5	GPT-3.5	60.02 55.66 49.52
	Predicate pipeline	GPT-3.5	GPT-3.5	58.55 54.39 51.9
HealthFC	Base pipeline	GPT-3.5	GPT-3.5	78.22
	Predicate pipeline	GPT-3.5	GPT-3.5	79.22
Climate-Fever	Base pipeline	GPT-3.5	GPT-3.5	84.85
	Predicate pipeline	GPT-3.5	GPT-3.5	75.8
CoVERT	Base pipeline	GPT-3.5	GPT-3.5	85.64
	Predicate pipeline	GPT-3.5	GPT-3.5	78.48

Ternary classification (GPT-3.5 for reasoning and evidence)

Dataset	Model	F1-macro	Refuted	Supported	NEI
HealthFC	Base pipeline	34.74	13.8	47.92	42.50
Climate-Fever	Base pipeline	48.74	46.21	67.54	32.49
CoVERT	Base pipeline	48.45	53.33	68.25	23.78

Results

Knowledge graphs pipeline

KG pipeline slightly outperformed the base pipeline on KG claims in terms of accuracy.

Using KGs has proved to be most beneficial for existence and negation claims.

The KG pipeline is lagging mostly for the multi-hop claims, this should be improved to increase the overall performance.

Results on the FactKG dataset of the base and KG pipeline and results from literature

Method	Accuracy % on FactKG					
	Total	One-hop	Conjunction	Existence	Multi-hop	Negation
BERT	65.20	69.64	63.31	61.84	**70.06**	63.62
GEAR	**77.65**	**83.23**	**77.68**	**81.61**	68.84	**79.41**
KG-GPT	72.68	-	-	-	-	-
Base pipeline	67.84	71.54	71.08	61.35	62.5	59.97
KG pipeline	68.95	72.22	71.02	75.59	57.74	69.93

Results

Question coverage survey

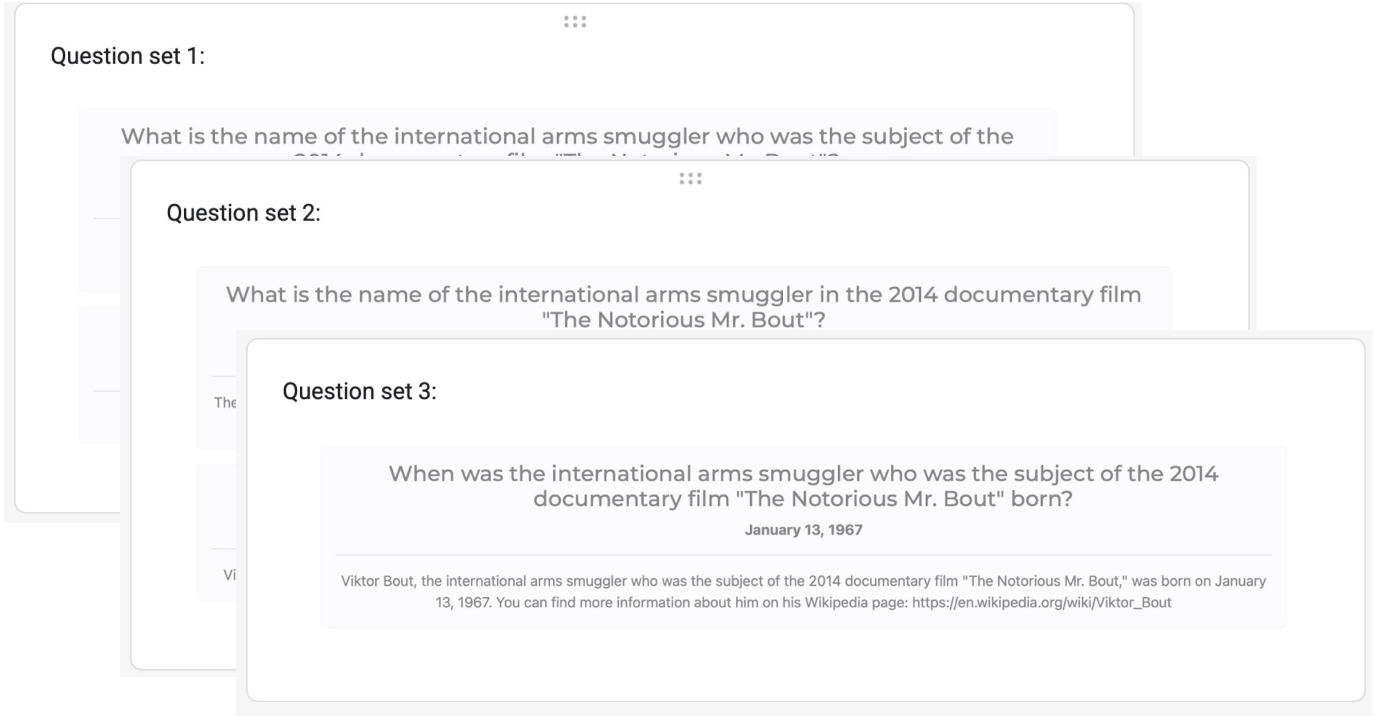
We conducted a survey on 3 pipelines: base pipeline with GPT-3.5 and Mixtral and the predicate pipeline with GPT-3.5.

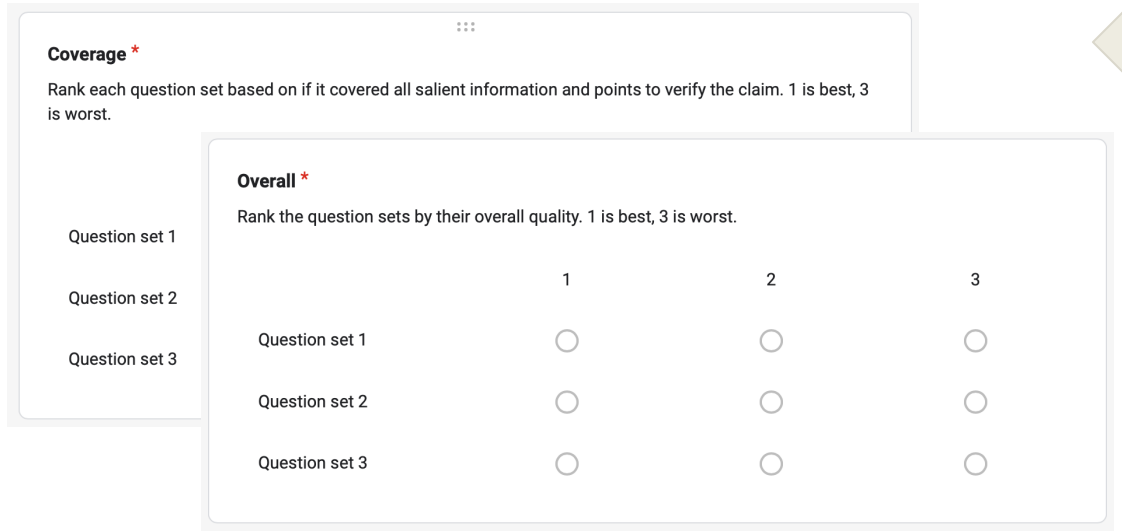
Evidence source is GPT-3.5.

Annotators ranked generated questions and answers for a claim based on 2 categories: **coverage** and **overall** quality.

Example claim from the survey:

The international arms smuggler who was the subject of the 2014 documentary film "The Notorious Mr. Bout" was born in the year 1967.





Results

Question coverage survey

The questionnaire had 6 claims (question sets) and was filled out by 12 respondents.

The pipeline using Mixtral as a reasoner was ranked highest in both categories, then the base pipeline with GPT-3.5 and the last was predicate pipeline.

	Base pipeline + GPT-3.5	Base pipeline + Mixtral-8x7b	Predicate pipeline + GPT-3.5
Coverage	1.85	**1.6**	2.13
Overall	1.81	**1.71**	2.03

Conclusion

RQ1: How can the use of LLMs help claim verification?

- Generating questions to verify a claim
- Reason over collected evidence
- Source of evidence
- Summarising multiple sources of evidence to find a relevant answer
- Explanation generation
- Reasoning over knowledge graphs

RQ2: Does leveraging knowledge from knowledge graphs and structured reasoning improve performance?

- Adding predicates did not improve the performance as expected
- It showed promising results for the most complicated claims
- Using Mixtral for the reasoning improved the performance much more
- Using KG as evidence source improved on the simpler types of claims
- Reasoning on KG could be improved for better results on the more challenging claims

Conclusion

RQ3: How do different domains compare in this task?

- Performance on real-life datasets was higher than on synthetic HOVER claims
- HealthFC was more challenging than CoVERT and Climate-Fever
- Model can reason over very informal claims from Tweets
- Showed good performance also for ternary classification on the domain specific tweets
- Future research could focus on analysing the verification outputs more into depth to find potential strengths and weaknesses of the reasoning mechanism

Bc.

Ivana Hacajová

Technical University of Munich (TUM)
TUM School of CIT
Department of Computer Science (CS)
Chair of Software Engineering for Business Information Systems (sebis)

Boltzmannstraße 3
85748 Garching bei München

+49.89.289.17132
matthes@in.tum.de
wwwmatthes.in.tum.de

Appendix

HealthFC results

Binary classification

Method	Reasoning	Evidence	HealthFC
DeBERTa-v3	-	Gold evidence	**81.9**
	-	Wikipedia	76.5
Base pipeline	GPT-3.5	GPT-3.5	78.22
		CommandR+ w/ web search	76.83
Predicate pipeline	GPT-3.5	GPT-3.5	79.22

Ternary classification

Method	Reasoning	Evidence	HealthFC			
			F1-macro	REFUTED	SUPPORTED	NEI
DeBERTa joint	-	Scientific papers	**67.5**	-	-	-
DeBERTa pipeline	-	Scientific papers	65.1	-	-	-
Base pipeline	GPT-3.5	GPT-3.5	34.74	13.8	47.92	42.50

Appendix

Climate-Fever results

Binary classification

Method	Reasoning	Evidence	Climate-Fever
Base pipeline	GPT-3.5	GPT-3.5	84.85
		CommandR+ w/ web search	**85.25**
Predicate pipeline	GPT-3.5	GPT-3.5	75.8

Ternary classification

Method	Reasoning	Evidence	Climate-Fever			
			F1-macro	REFUTED	SUPPORTED	NEI
ALBERT	-	Wikipedia	36.05	41.81	47.79	18.57
Base pipeline	GPT-3.5	GPT-3.5	**48.74**	**46.21**	**67.54**	**32.49**

Appendix

CoVERT results

Binary classification

Method	Reasoning	Evidence	CoVert
DeBERTa	-	Gold evidence	83.4
DeBERTa	-	Wikipedia	82.5
Base pipeline	GPT-3.5	GPT-3.5	**85.64**
		CommandR+ w/ web search	83.51
Predicate pipeline	GPT-3.5	GPT-3.5	78.48

Ternary classification

Method	Reasoning	Evidence	CoVert			
			F1-macro	REFUTED	SUPPORTED	NEI
MLP-FEVER	-	Gold evidence	46.00	-	-	-
MLP-Evidence	-	Gold evidence	**69.00**	-	-	-
Base pipeline	GPT-3.5	GPT-3.5	48.45	53.33	68.25	23.78

Appendix

Predicate Pipeline Example

The award-winning CD, that Jacqueline Horner-Kwiatek was a featured soloist on, includes the theme for the game \"Civilization IV,\" written by the man who released The Drop That Contained the Sea which is a classical crossover album. This album was released in 2014.

True

Explanation: Christopher Tin wrote the theme for the game "Civilization IV" and released the classical crossover album "The Drop That Contained the Sea" in 2014. Jacqueline Horner-Kwiatek was a featured soloist on the award-winning CD that includes the theme for "Civilization IV."

Who wrote the theme for the game "Civilization IV"?

Christopher Tin

The theme for the game "Civilization IV" was composed by Christopher Tin. You can find more information about him and his work on the Wikipedia pag https://en.wikipedia.org/wiki/Christopher_Tin

Who released The Drop That Contained the Sea?

Christopher Tin

The Drop That Contained the Sea was released by composer Christopher Tin. Here is the Wikipedia article for more information: https://en.wikipedia.org/wiki/The_Drop_That_Contained_the_Sea

```
PREDICATES:
Predicate 1 = WroteTheme("Civilization IV", writer) ::: Verify the writer of
the theme for the game "Civilization IV"
Predicate 2 = Released("The Drop That Contained the Sea", artist) ::: Verify
the artist who released The Drop That Contained the Sea.

PREDICTION:
"WroteTheme("Civilization IV", writer) is True because Christopher Tin wrote
the theme for the game "Civilization IV".
Released("The Drop That Contained
the Sea", artist) is True because Christopher Tin released The Drop That
Contained the Sea.
WroteTheme("Civilization IV", writer) && Released(\"The
Drop That Contained the Sea", artist) is True.
The claim is [SUPPORTED].

Explanation:
Christopher Tin wrote the theme for the game "Civilization
IV" and released the classical crossover album "The Drop That Contained the
Sea" in 2014. Jacqueline Horner-Kwiatek was a featured soloist on the
award-winning CD that includes the theme for "Civilization IV."
```

Appendix

KG Pipeline Example

Claim: **Well the Acura TLX has no V6 engine and was not assembled in Marysville Ohio.**	
Round 1	
Question	Does the Acura TLX have a V6 engine?
DBPedia URI	http://dbpedia.org/resource/V6_engine http://dbpedia.org/resource/Acura_TLX
Retrieved relations	manufacturer, class, modelYears, layout, width, aka, productionStartYear, designer, related, engine, predecessor, assembly, bodyStyle, transmission
Retrieved tuples	("Acura_TLX", "engine", 3.5) ("Acura_TLX", "engine", 2.4)
Rationale	The Acura TLX has engines available in 3.5 and 2.4.
Answer	Yes, the Acura TLX does have a V6 engine available in the 3.5 model.
Round 2	
Question	Was the Acura TLX assembled in Marysville, Ohio?
DBPedia URI	http://dbpedia.org/resource/Marysville,_Ohio http://dbpedia.org/resource/Marysville_Motorcycle_Plant http://dbpedia.org/resource/Acura_TLX
Retrieved relations	class, website, areaTotal, bodyStyle, populationDensity, subdivisionName, utcOffset, modelYears, areaCode, type, productionStartYear, areaWater, designer, related, predecessor, assembly, elevation, timeZone, populationTotal, engine, width, settlementType, nickname, transmission, manufacturer, layout, areaLand, location, leaderTitle, leaderName, postalCode, aka, governmentType
Retrieved tuples	("Acura_TLX", "assembly", "United States: Marysville, Ohio") ("Acura_TLX", "assembly", "Guangzhou, China")
Rationale	The Acura TLX is assembled in the United States in Marysville, Ohio and in Guangzhou, China.
Answer	Yes.
Prediction	The Acura TLX has a V6 engine available in the 3.5 model. The Acura TLX was assembled in Marysville, Ohio. Therefore, the final answer is: False.

Appendix

Dataset examples

HOVER

- Mike Leigh's second collaboration with Louis Marks, after BBC (1982 Film), starred the actor who was best known for the role of Captain Hans Geering on "Allo 'Allo!“
- The American Thoroughbred horse race facility in Elmont, New York opened on May 4, 1905. It has a $500,000 purse starting in 2016. It's where a horse race takes place every June. The horse race was won by Avatar in 1975.
- Before I Go to Sleep stars an Australian actress, producer and occasional singer.

HealthFC

- Can dementia be prevented?
- Can vitamin supplements protect against life-threatening diseases such as cancer or cardiovascular disease?
- Do static magnets help against pain?

Appendix

Dataset examples

Climate-Fever

- Global warming is driving polar bears toward extinction
- If CO2 was so terrible for the planet, then installing a CO2 generator in a greenhouse would kill the plants.
- Sea level rise has been slow and a constant, pre-dating industrialization

CoVERT

- @username The S1 spike protein used in the Covid vaccine may be toxic to endothelial cells. Likely the cause of the clotting problems we are seeing from people who get the vaccine.\n\nhttps://t.co/ZCepuEfWCD
- @username @username And the Covid vaccine can’t knock out all the variants either. \n\nSo once again, we are taking a vaccine with known risks that STILL doesn’t let us leave our basements or take off our double masks. But it does cause blood clots, so there’s that.
- @username @username @username That makes no sense. [NEWLINE][NEWLINE]If they have done so much work already why have they never got rid of the common cold. [NEWLINE][NEWLINE]Covid is noting like any other virus. Which other virus causes hair loss, skin irritations and lasts 6-12 months in people? Only one thing does that. Radiation!!!

Appendix

Prompt Examples from the Base Pipeline

Check if we have enough evidence

```
Claim = Superdrag and Collective Soul are both rock bands.
To validate the above claim, we have asked the following questions:
Question 1 =to explainAnswer 1 = Yes
Can we know whether the claim is true or false now?
Prediction = No, we cannot know.

Claim = Superdrag and Collective Soul are both rock bands.
To validate the above claim, we have asked the following questions:
Question 1 = Is Superdrag a rock band?
Answer 1 = Yes
Question 2 = Is Collective Soul a rock band?
Answer 2 = Yes
Can we know whether the claim is true or false now?
Prediction = Yes, we can know.
```

First question generation

```
Claim = Superdrag and Collective Soul are both rock bands.
To validate the above claim, the first simple question we need to ask is:
Question = Is Superdrag a rock band?

Claim = Jimmy Garcia lost by unanimous decision to a professional boxer that
challenged for the WBO lightweight title in 1995.
To validate the above claim, the first simple question we need to ask is:
Question = Who is the professional boxer that challenged for the WBO
lightweight title in 1995?
```

Follow-up question generation

```
Claim = Superdrag and Collective Soul are both rock bands.
To validate the above claim, we need to ask the following simple questions
sequentially:
Question 1 = Is Superdrag a rock band?
Answer 1 = Yes
Question 2 = Is Collective Soul a rock band?

Claim = Jimmy Garcia lost by unanimous decision to a professional boxer that
challenged for the WBO lightweight title in 1995.
To validate the above claim, we need to ask the following simple questions
sequentially:
Question 1 = Who is the professional boxer that challenged for the
WBO lightweight title in 1995?
Answer 1 = Orzubek Nazarov
Question 2 = Did Jimmy Garcia lose by unanimous decision to Orzubek Nazarov?
```