
Welcome

Hallucination Detection

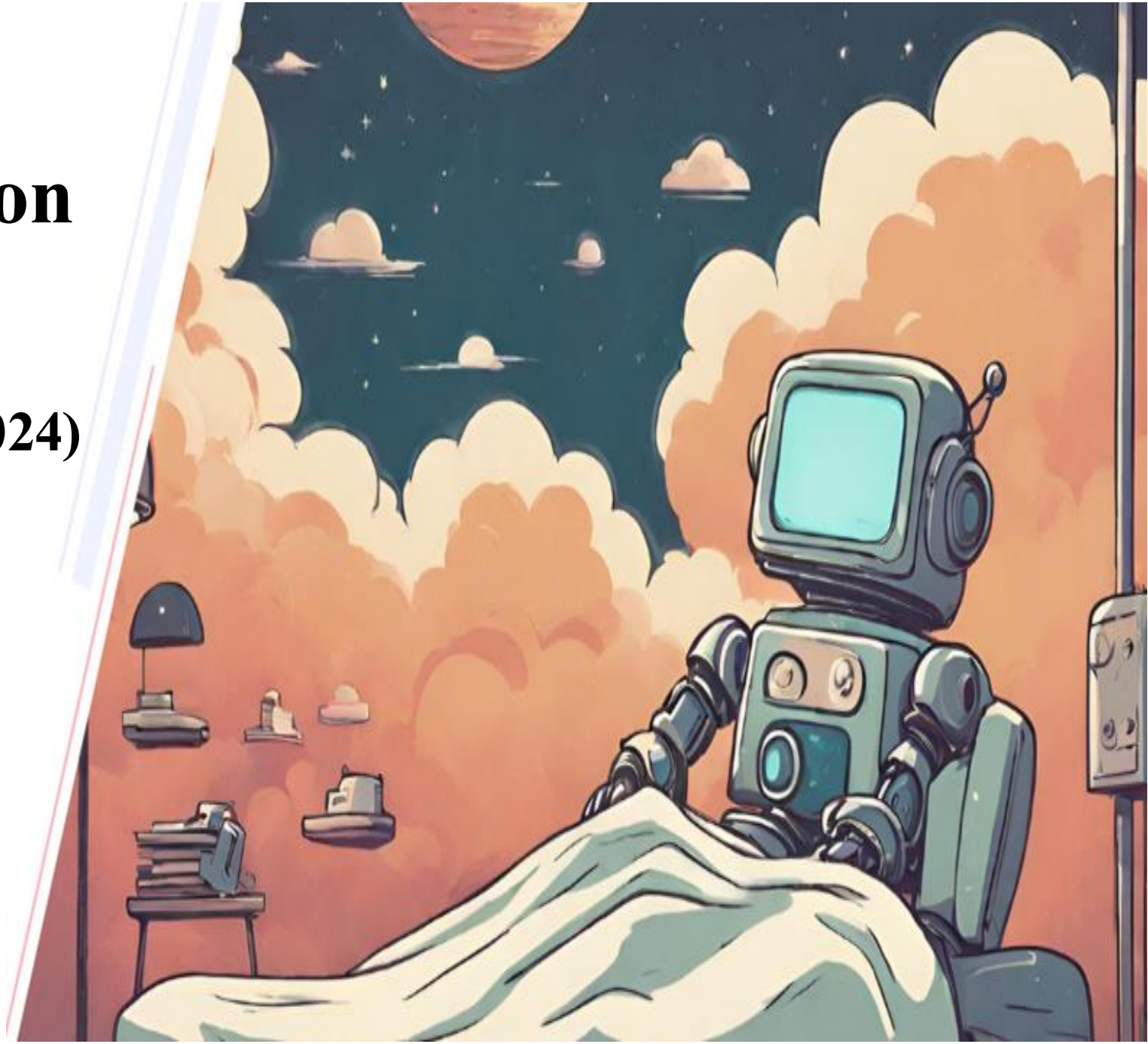
(ELOQUENT Lab @ CLEF 2024)

01 – Introduction

Philipp Schaer and Narjes Nikzad

08-04-24 – Cologne, Germany
https://ir.web.th-koeln.de

Technology
Arts Sciences
TH Köln



SPECIAL GUEST

CLEF 2024

- Know more about CLEF.
- CLEF stands for Conference and Labs of the Evaluation Forum.
- CLEF has been a leading annual conference since 2000 that focuses on evaluating information access systems.

ELOQUENT shared tasks for evaluation of generative language model quality

ELOQUENT provides a set a of tasks for evaluating the quality of generative language models.

1 - Topical competence
This task will test and verify a model's understanding of an application domain and specific topic of interest.

2 - Veracity and hallucination
This task will test how the truthfulness or veracity of automatically generated text can be assessed.

3 - Robustness
This task will test the capability of a model to handle input variation -- e.g. dialectal, sociolectal, and cross-cultural -- as represented by a set of equivalent but non-identical varieties of input prompts.

4 - Voight Kampff
This task will explore whether automatically-generated text can be distinguished from human-authored text. This task will be organised in collaboration with the PAN lab at CLEF.

In this project:

- We conduct **weekly or bi-weekly hybrid meetings**, with an **online focus**.

In this project:

- Finally, participants are required to **document their approach**.

Documentation

Task Title

Start date - End date

Task Description

What you have done including:

- the input, output and process.
- the explanation of why you did that and why you have such output.

Finally:

- The results of our work will **be submitted to ELOQUENT**, where the results will be part of the official evaluation campaign.

Schedule and Deadlines

- Begin of April: Start of the project, kick-off, and task assignment
- Mid of April to mid to May: Working on task, coding and evaluating
- End of May: Submission of the results to ELOQUENT
- June: Submission of workshop papers (optional).

Size	It's 6 credits

Grading process

- Each student gives a presentation.

Question

Problem Description

Hallucination

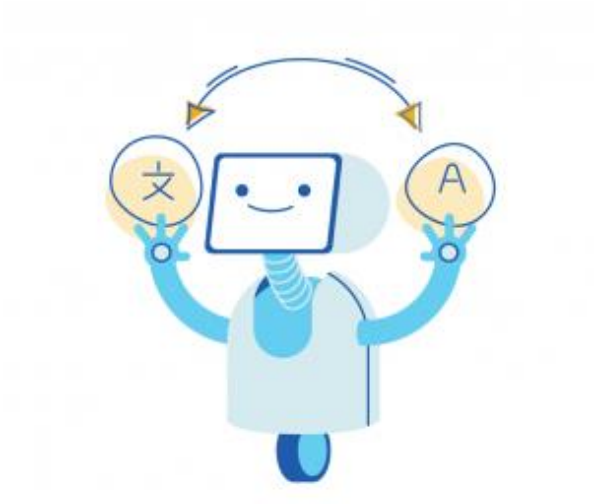
•A hallucination is when a LLM produces an output that is false, or that does not match the user's intent.

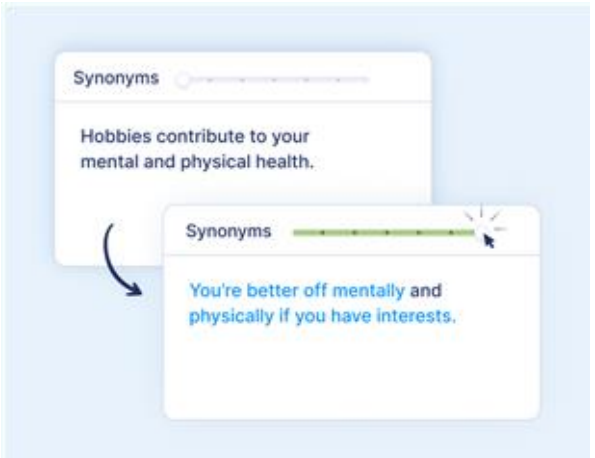
Hallucination Detection

•One possible approach for detecting hallucinated content, that we will use in this project, is **the use of LLMs to evaluate LLM output.**

Project Definition

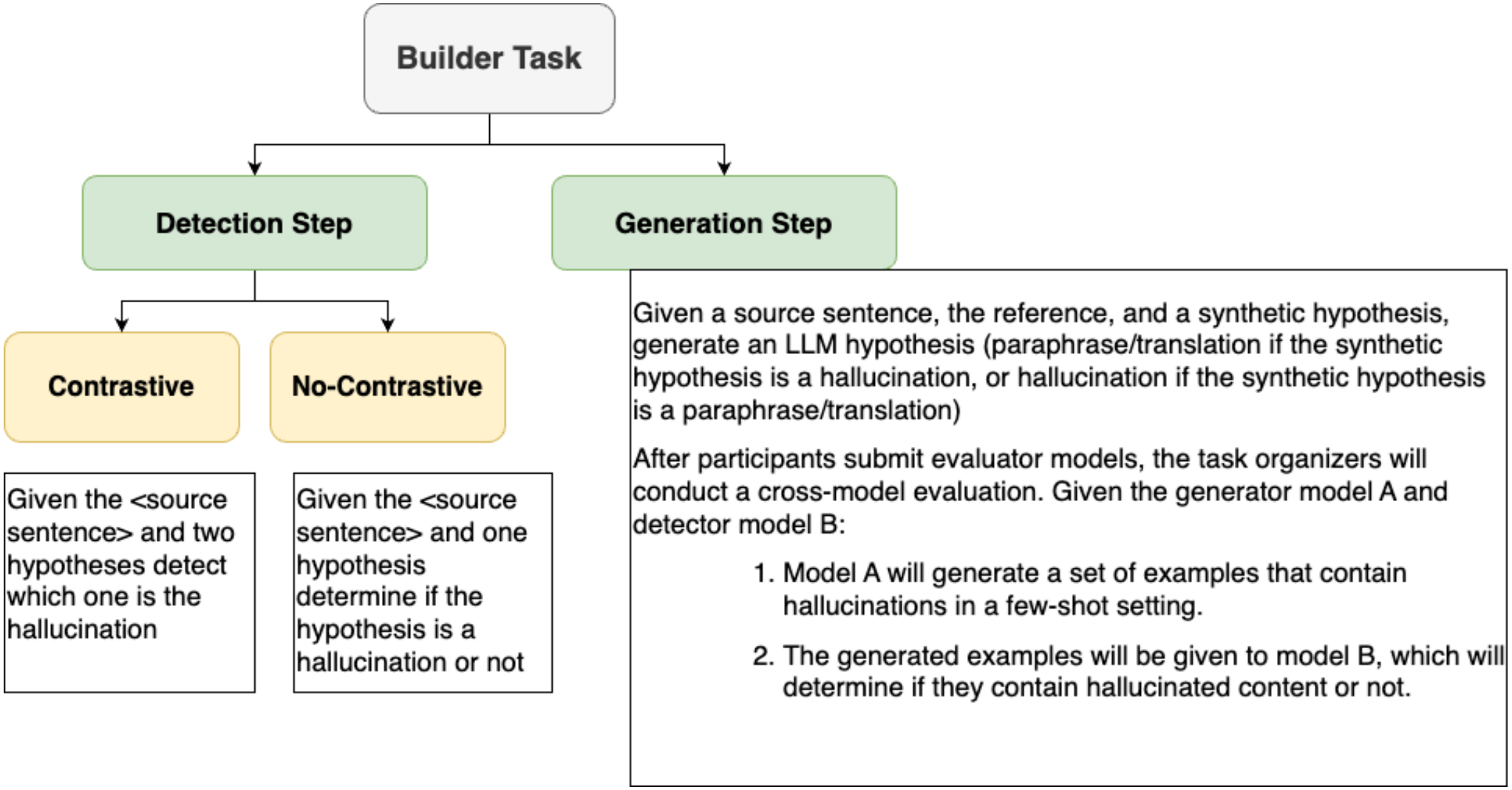
- In this project the goal is to focus on one of the ‘CLEF’ labs called ‘**ELOQUENT**’ that has a special ‘**HalluciGen Detection**’ task.
- HalluciGen is a hallucination detection and generation task that will be performed in two different scenarios:





- The data is available.

Task: builder task



Open-source LLMs:

1. MPT-7B

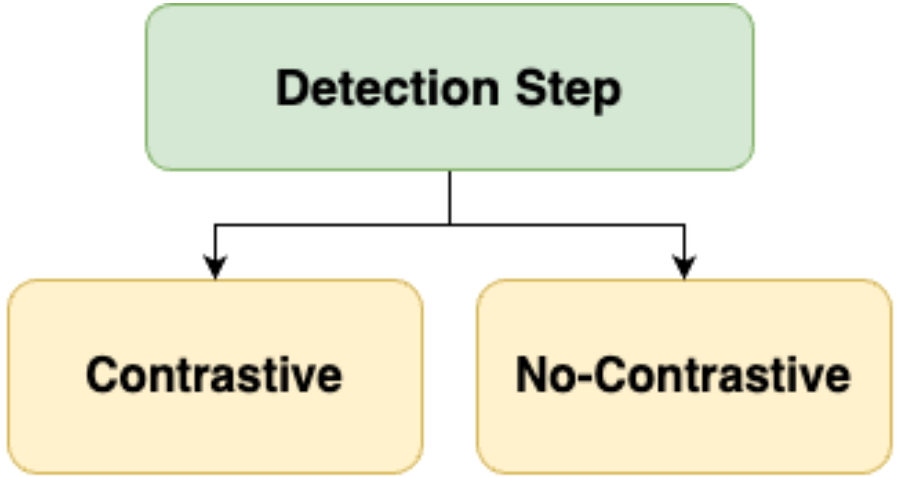
2. falcon-7b-instruct

3. 

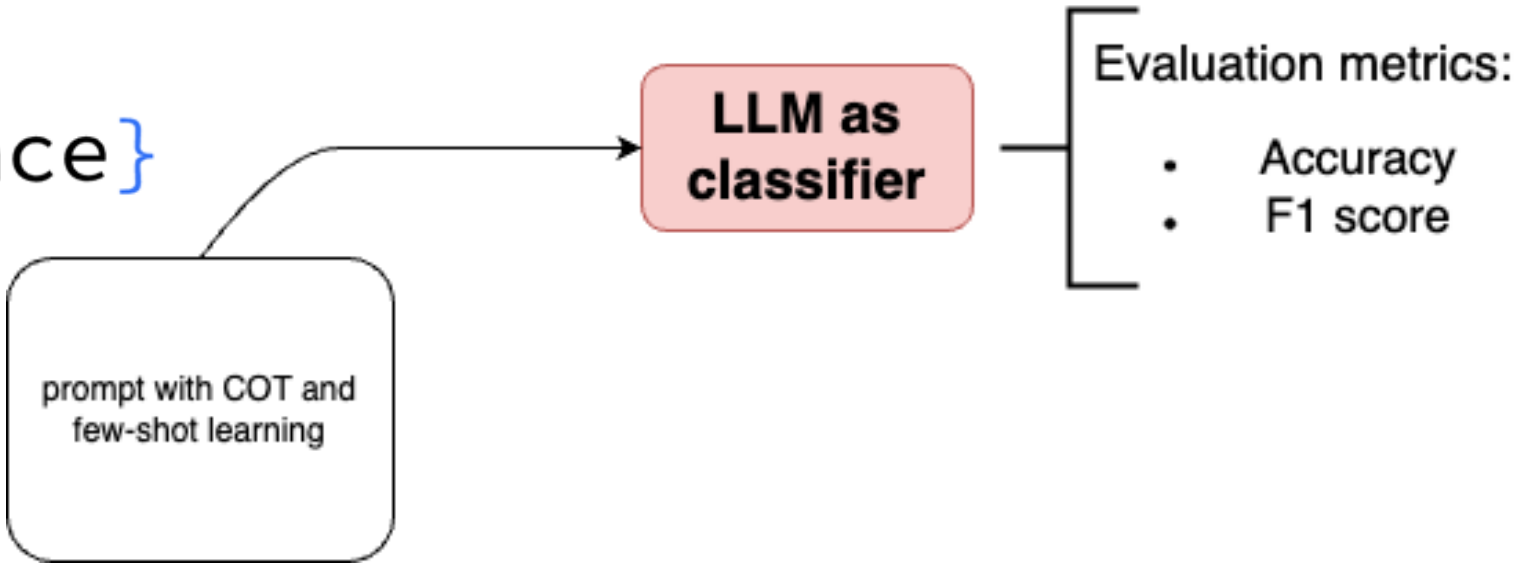
Close-source LLMs:

Gemini

Detection task

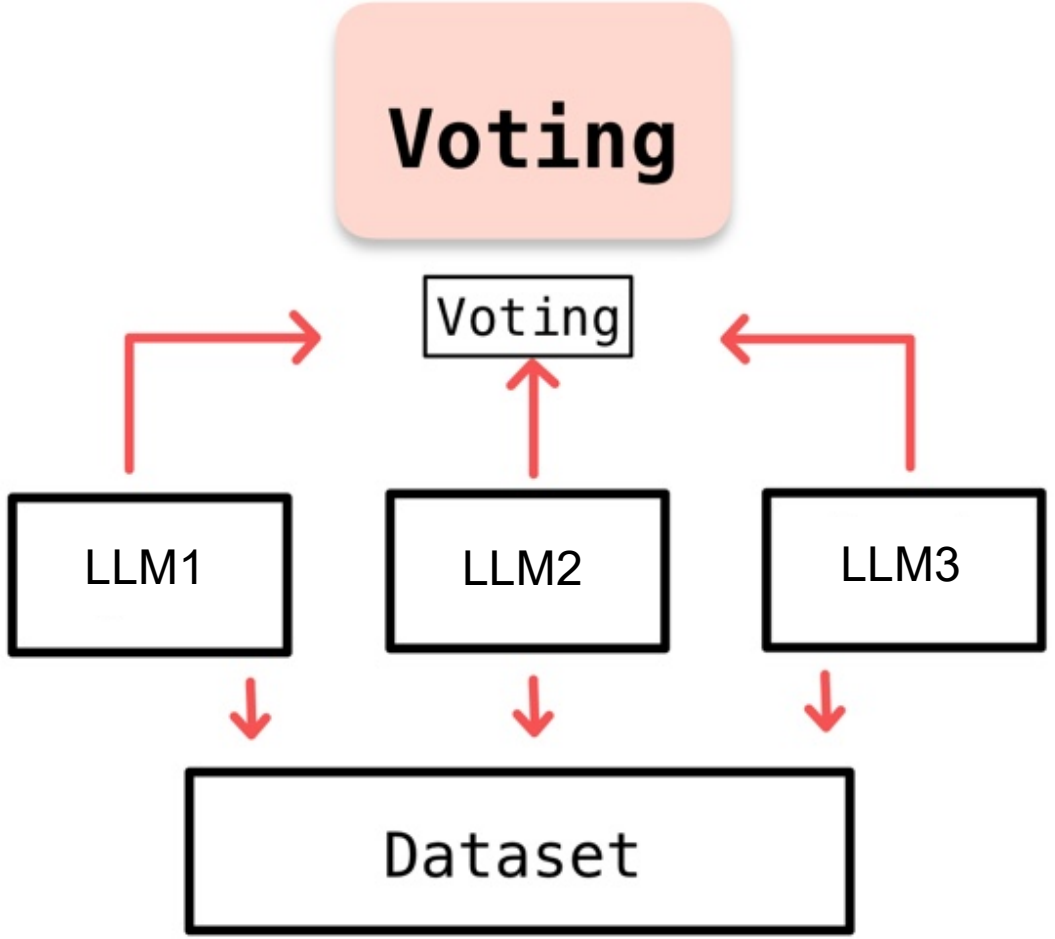


{guidance}

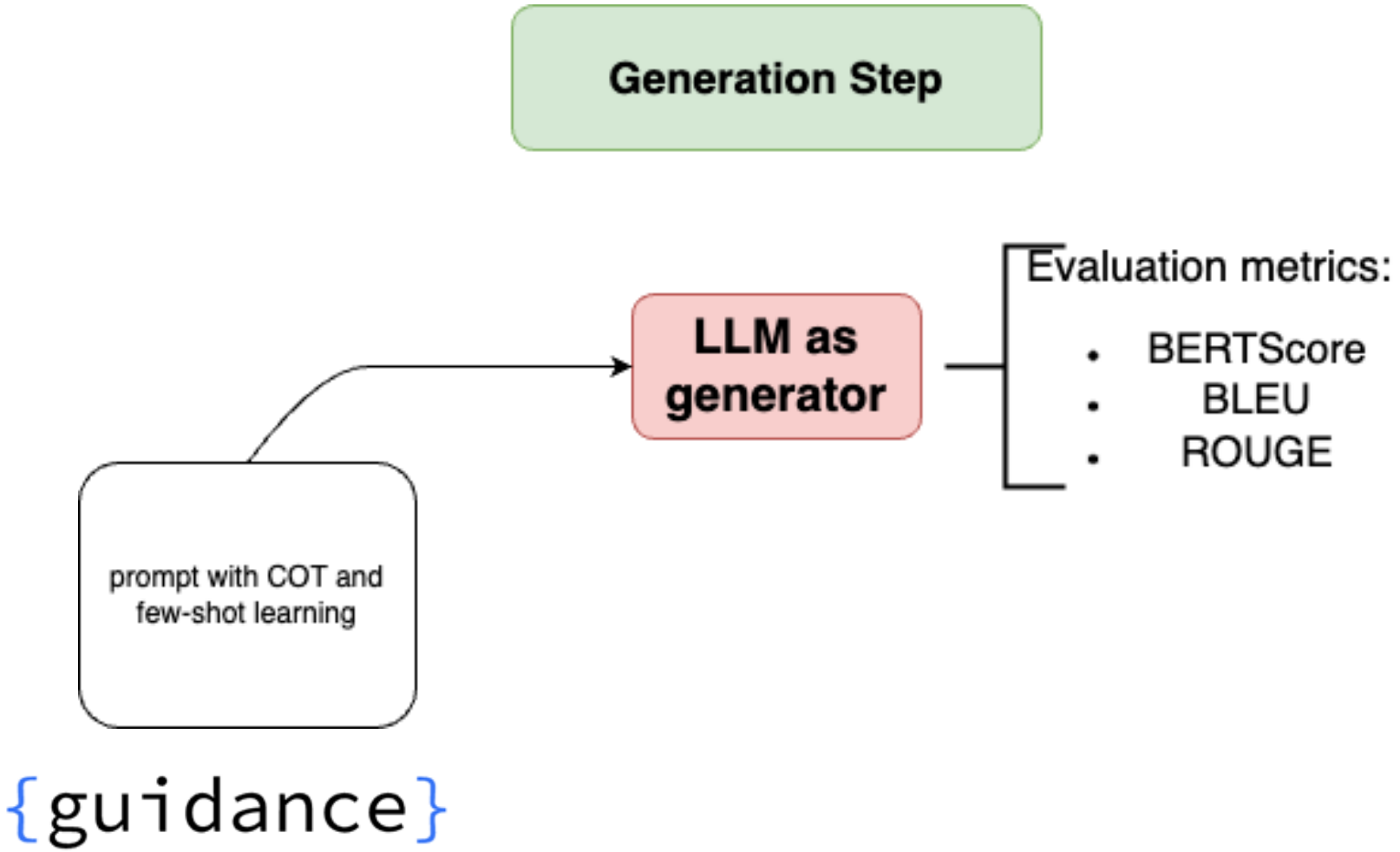


Voting

- Comparing the efficacy of LLMs as elevators using ensemble technique such as the voting approach.



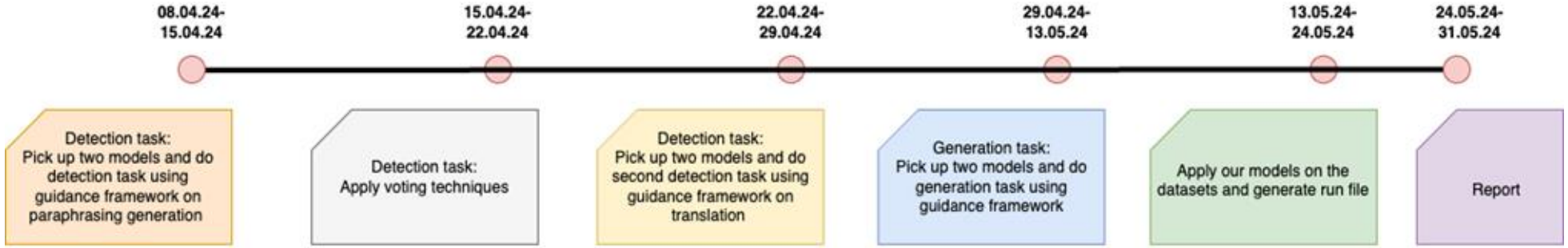
Generation Task

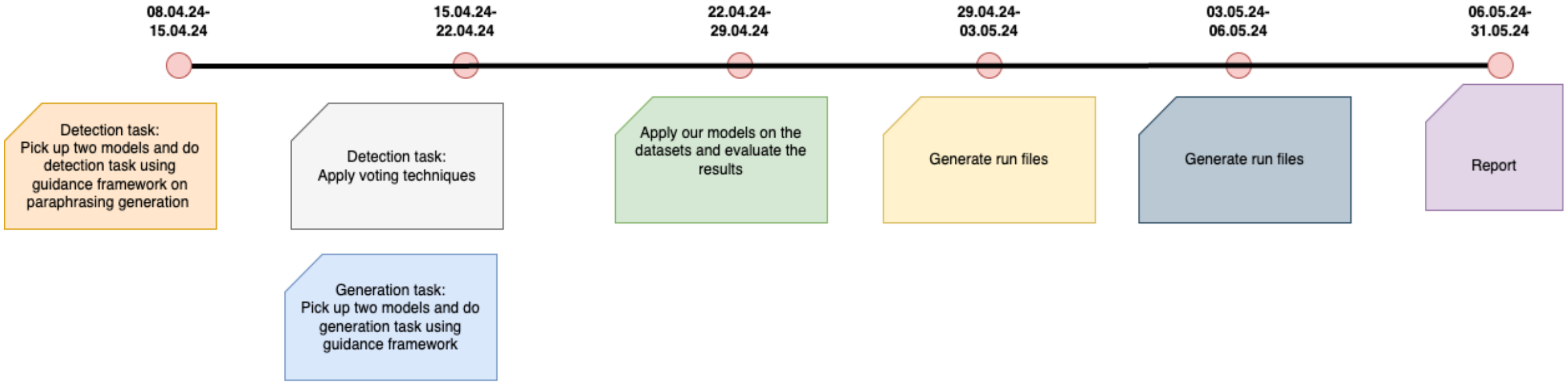


Timeline:

- May 6: submission of experimental run results
- May 24: results from evaluation are sent to participants
- May 31 (strict!): submission of report manuscripts from participants for workshop proceedings
- June 24: notification of acceptance for participant papers with reviewer comments
- July 8 (strict!): submission of camera-ready final manuscripts
- July 22-26: preview proceedings check by manuscript authors
- Sept 9-12: CLEF Conference and ELOQUENT workshop sessions in Grenoble!

Timeline and tasks

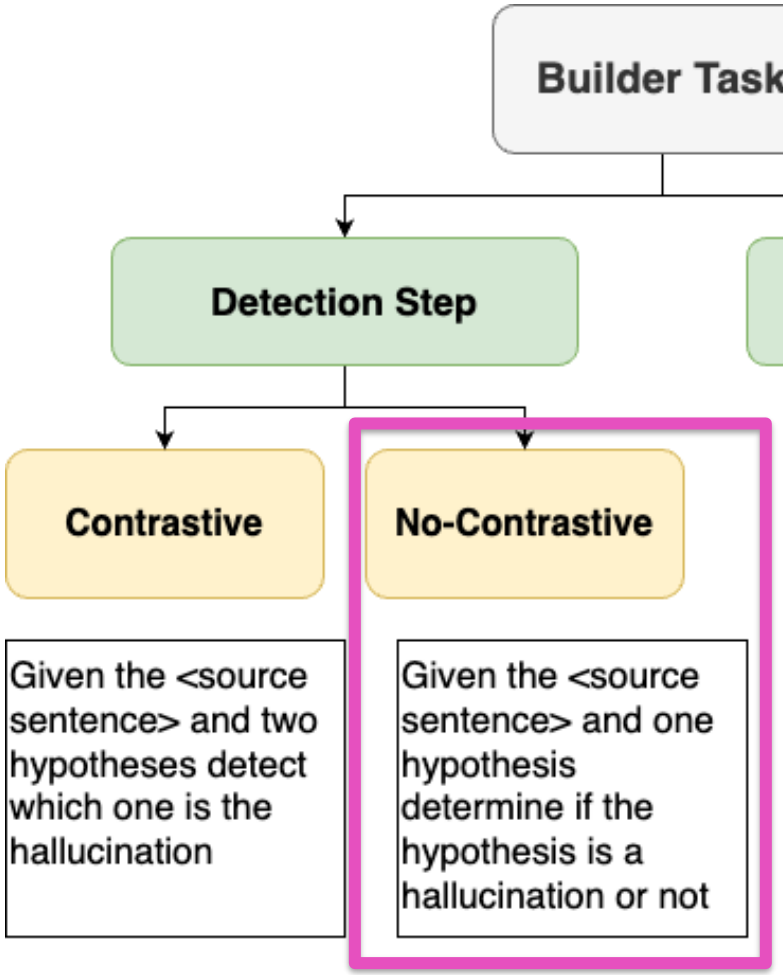




Task for next week: SHROOM Dataset

```
Hypothesis: Alternative form of scatter site
Reference: tgt
Source: It is designed for single - family use in subdivisions or <define> scatter sites </define> and availab
Target: An area of state-sponsored housing used as a shelter for homeless people; such housing is scattered ac
Label: Hallucination
Probability of Hallucination: 0.6666666666666666
--------------------
Hypothesis: The state or condition of being obsolescent.
Reference: tgt
Source: One way is to legislate extended warranties on products , so washing machines and refrigerators last f
Target: (uncountable) The state of being obsolete—no longer in use; gone into disuse; disused or neglected.
Label: Not Hallucination
Probability of Hallucination: 0.0
```

Task for next week:

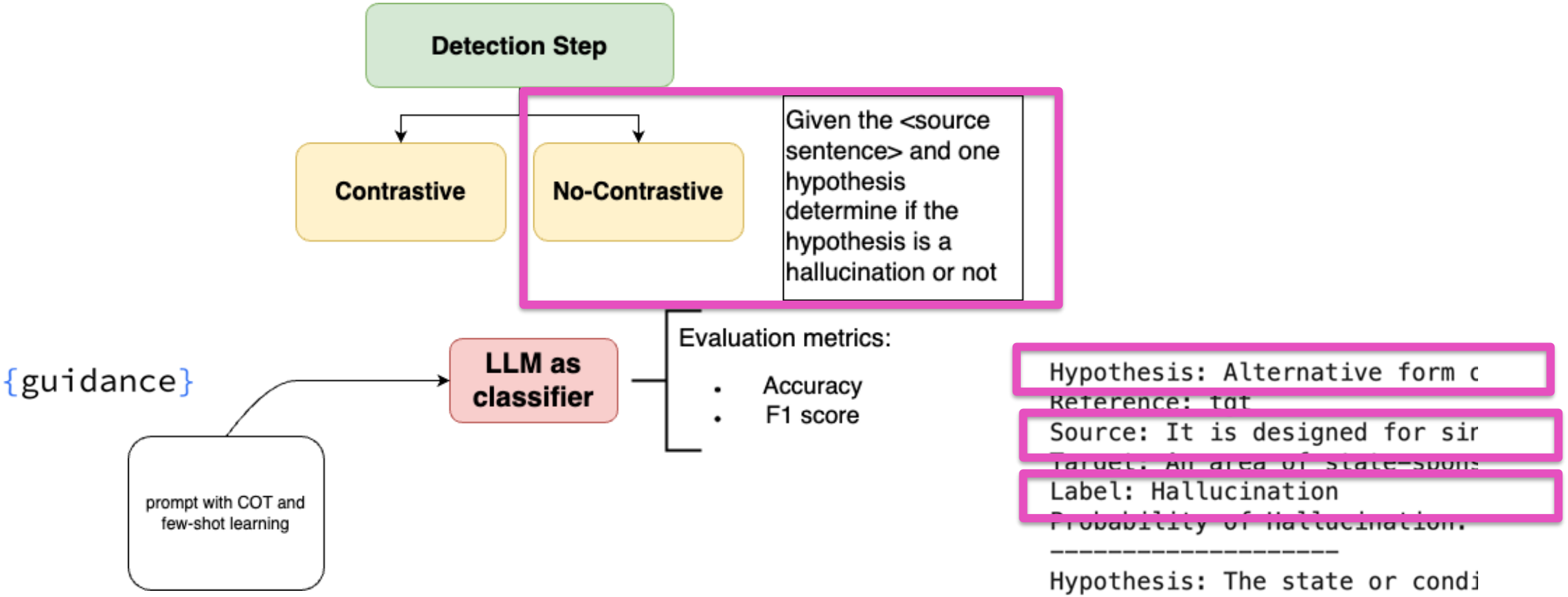


Given a source sentence, the reference, and a synthetic hypothesis, generate an LLM hypothesis (paraphrase/translation if the synthetic hypothesis is a hallucination, or hallucination if the synthetic hypothesis is a paraphrase/translation)

After participants submit evaluator models, the task organizers will conduct a cross-model evaluation. Given the generator model A and detector model B:

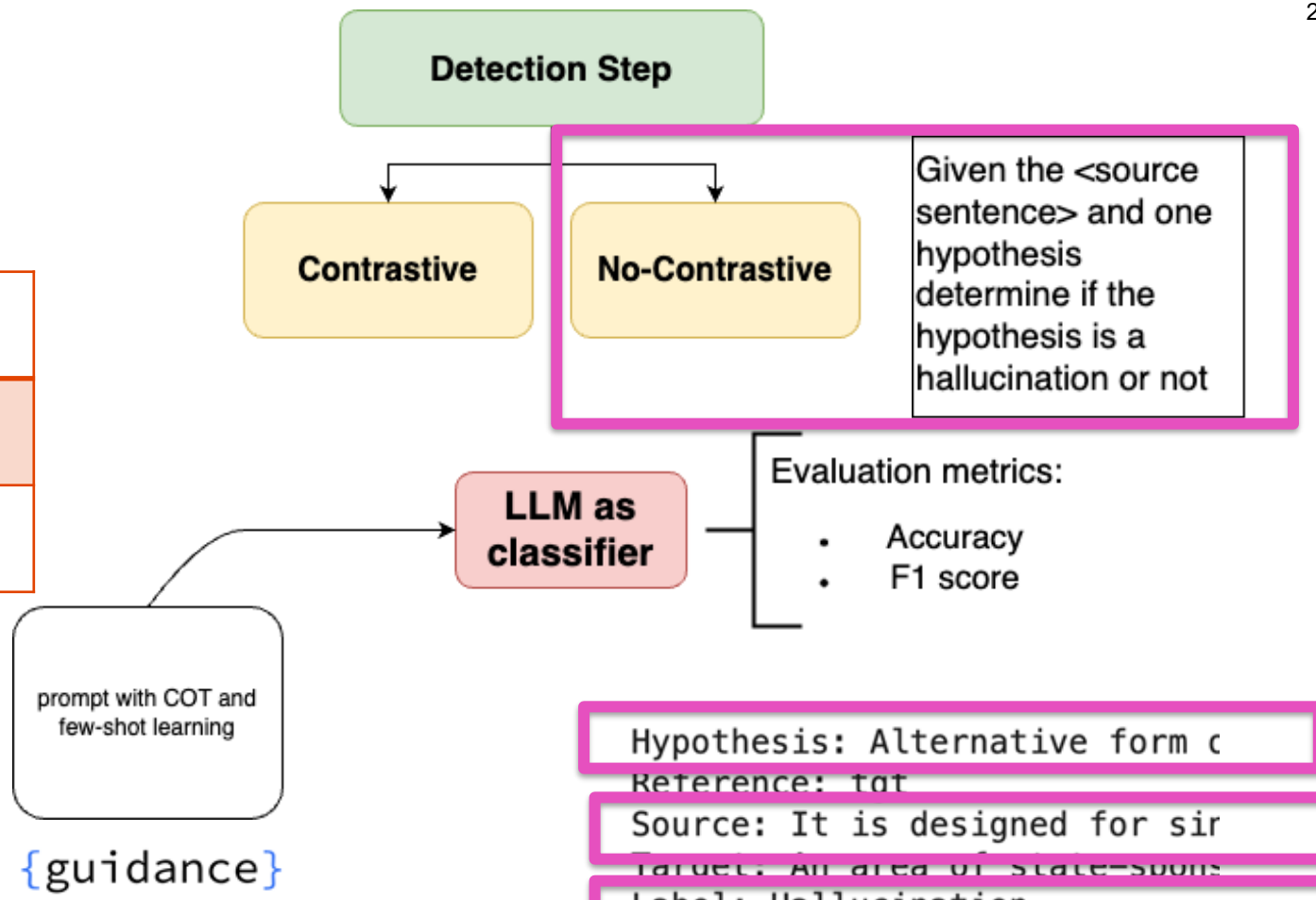
1. Model A will generate a set of examples that contain hallucinations in a few-shot setting.
2. The generated examples will be given to model B, which will determine if they contain hallucinated content or not.

Task for next week:



Task for next week:

Groups	Models




```
Hypothesis: Alternative form c
Reference: tgt
Source: It is designed for sir
Target: An area of state-owns
Label: Hallucination
Probability of Hallucination:
--------------------
Hypothesis: The state or condi
Reference: tgt
Source: One way is to legislat
Target: (uncountable) The stat
Label: Not Hallucination
Probability of Hallucination:
```

MPT-7B

falcon-7b-instruct

Sources you need for next week

- Colab google account:
  - ✓User: research.student.acc@gmail.com
  - ✓Pass: thkoln.nikzad@

- Link to the dataset:
  - ❖Find it on google drive with folder name ‘Master Project-Hallucination’
  - ❖Link

Sources you need for next week

- Link to the Models:
  - ❖ MPT model
    - ✓ Source1
    - ✓ Source2
  - ❖ Falcon Model
    - ✓ Source1
    - ✓ Source2
  - ❖ Llama 2
    - ✓ Source1
    - ✓ Source2
- Link to guidance framwork
  - ✓ Source

Sources you need for next week

- Link to evalution metrics:
  - ❖ Accuracy
    - ✓ Source1
  - ❖ F1 score
    - ✓ Source1

Link to important points

- BLEU
  - ✓ Source1
- Ensambling throught voting
  - ✓ Source1