




CHATGPT INCORRECTNESS DETECTION (CID) IN SOFTWARE REVIEWS

Minaoar Hossain Tanzil*, Junaed Younus Khan*, Dr. Gias Uddin**

*University of Calgary, **York University, Canada

Presented @ International Conference on Software Engineering 2024



LLM Hallucination: Incorrect/Irrelevant responses

20%

Up to 20% ChatGPT responses are
HALLUCINATION¹.

LLM
Trustworthiness

Users barely **TRUST** ChatGPT responses
without further verification

¹Li, Junyi, et al. "Halueval: A large-scale hallucination evaluation benchmark for large language models." *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 2023.

What did We Do?

- Surveyed 135 software professionals who
 - 46% use ChatGPT for SE Tasks like **Software Library Selection**
 - Most **DO NOT TRUST ChatGPT** and 68% prompt ChatGPT again

What did We Do?

- Surveyed 135 software professionals who
 - 46% use ChatGPT for SE Tasks like **Software Library Selection**
 - Most **DO NOT TRUST ChatGPT** and 68% prompt ChatGPT again
- Developed a tool for ChatGPT Incorrectness Detection (CID)

What did We Do?

- Surveyed 135 software professionals who
 - 46% use ChatGPT for SE Tasks like **Software Library Selection**
 - Most **DO NOT TRUST ChatGPT** and 68% prompt ChatGPT again
- Developed a tool for ChatGPT Incorrectness Detection (CID)
- Evaluated CID against a benchmark of Stack Overflow data
 - CID can detect incorrect ChatGPT responses with 0.74 F1-score

Our Survey Participants Frequently Challenge ChatGPT Responses

- *“Chatbots are better at suggesting solutions to a specific problem, but **reliability issues are there**. For example, if it’s asked **“are you sure”** to a reply, answers are usually not same.”*
- ***I think chatbot is not reliable fully...** So, it's better to cross-validate the answers... **the more we provide specific prompts multiple times**, the chance of providing useful information by ChatGPT is more high.*

Such challenge prompts can be attributed to interrogations in criminal psychology^{2,3,4}

²Vrij, Aldert, et al. "Outsmarting the liars: The benefit of asking unanticipated questions." *Law and human behavior* 33 (2009): 159-166.

³Aldert Vrij, P.r Anders Granhag, Samantha Mann, and Sharon Leal. 2011. *Outsmarting the liars: Toward a cognitive lie detection approach*. *Current Directions in Psychological Science* 20, 1 (2011), 28–32.

⁴Gary LJ Lancaster, Aldert Vrij, Lorraine Hope, and Bridget Waller. 2013. *Sorting the liars from the truth tellers: The benefits of asking unanticipated questions on lie detection*. *Applied Cognitive Psychology* 27, 1 (2013), 107–114.

Our Technique CID is Designed Following Principles of Criminal Psychology

Criminal Investigation
Department (CID)²
interrogates
Suspected Criminals

According to Industry
Survey, Developers
iteratively challenges
ChatGPT



²Vrij, Aldert, et al. "Outsmarting the liars: The benefit of asking unanticipated questions." *Law and human behavior* 33 (2009): 159-166.

Criminal Investigation

[Sherlock Holmes] *Did you receive any other visitors this morning?*

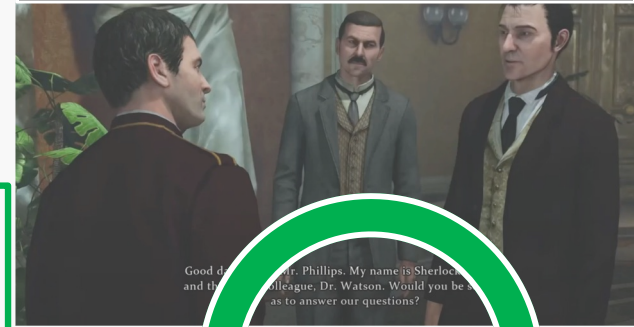
[Suspect] No one. *Only these gentlemen came.*

[Sherlock Holmes] *No matter if you had left the baths or not, did you receive any visitors?*

[Suspect] No sir. I did not. *Sir Gregory came first.*

[Sherlock Holmes] “*Only these gentlemen came.*”
Is not same as “*Sir Gregory came first.*”
You are not being truthful

Sherlock Holmes: Art of Interrogation

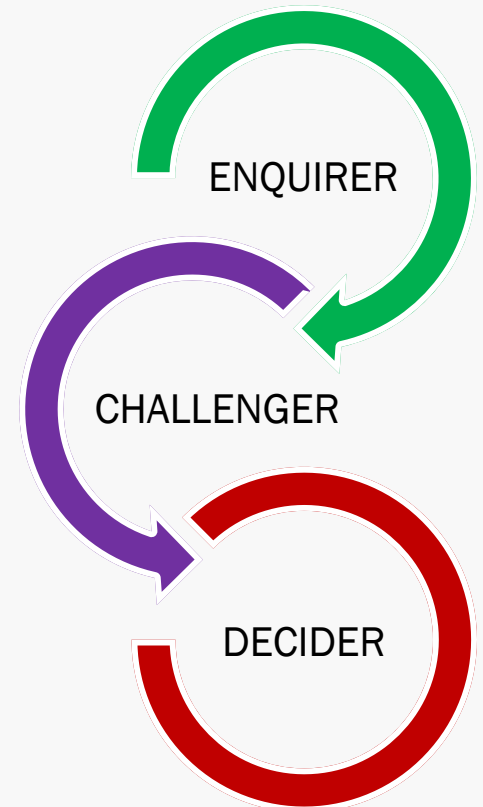


ENQUIRE

CHALLENGE

DECIDE

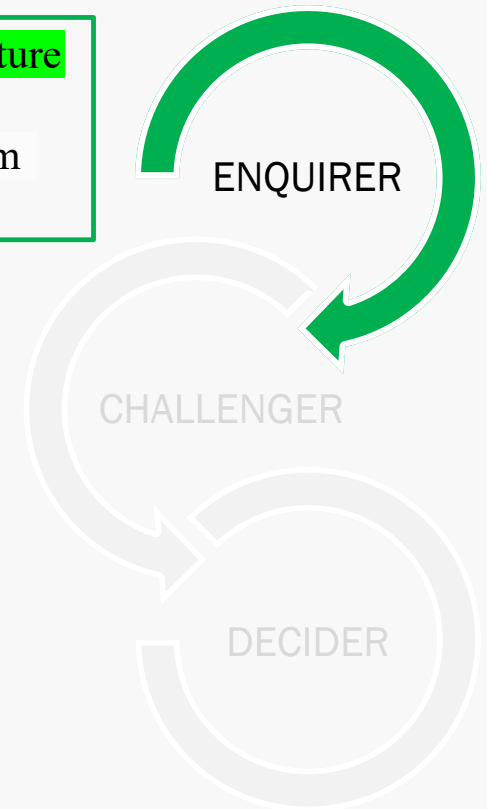
ChatGPT Incorrectness Detection (CID) Architecture



ChatGPT Incorrectness Detection (CID)

[CID] How well does this library spaCy support **Text classification feature**
[🌀] Spacy supports TextCategorizer: The user is able to train a TextCategorizer model with labeled data pairs using example code from spaCy's documentation.

ENQUIRER asks a basic question regarding the an aspect (here Text classification feature) of the library spaCy.

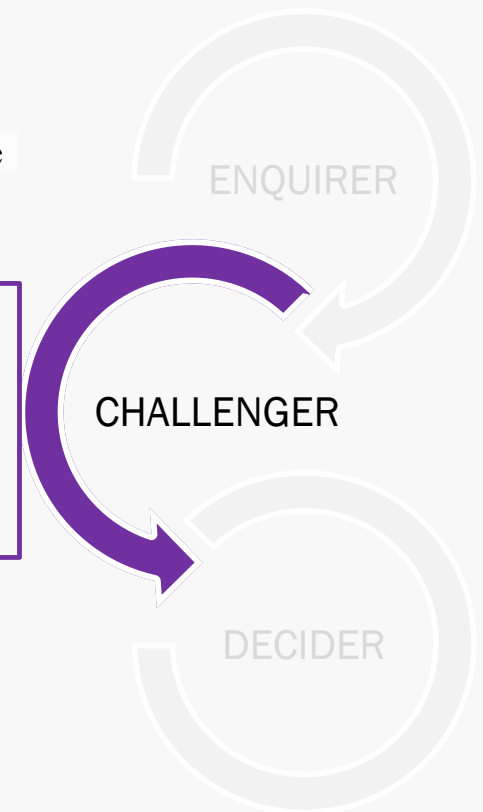


ChatGPT Incorrectness Detection (CID)

CHALLENGER generates 3 **BASIC** challenges (using why, how, really) to the challenge the ChatGPT statement

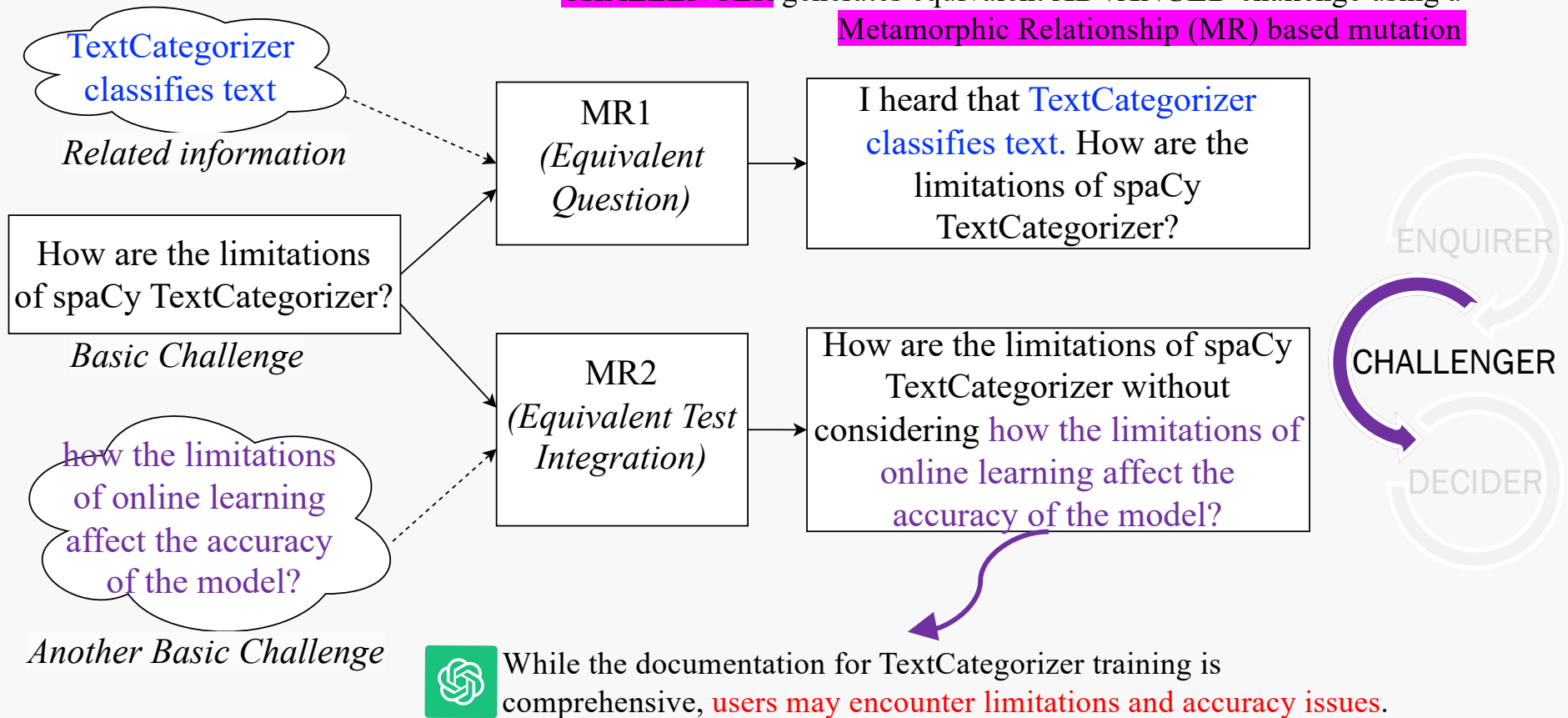
[CID] **How** are the limitations of spaCy TextCategorizer that users might encounter?

[no limitations or potential issues are discussed.



ChatGPT Incorrectness Detection (CID)

CHALLENGER generates equivalent **ADVANCED** challenge using a **Metamorphic Relationship (MR) based mutation**

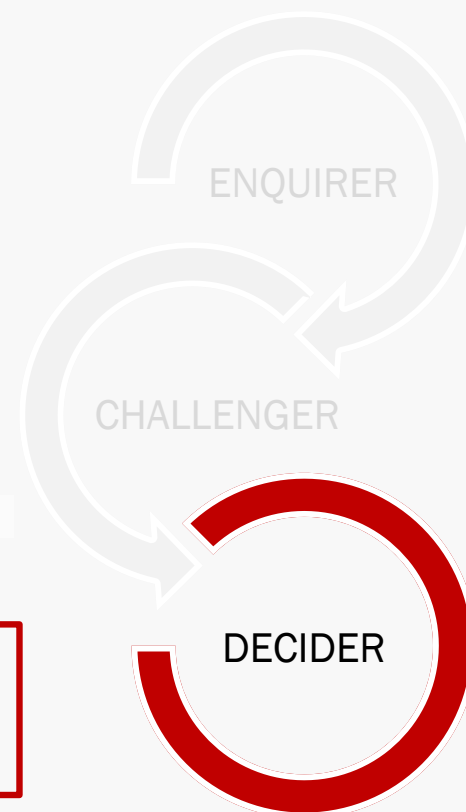


ChatGPT Incorrectness Detection (CID)

DECIDER has two steps:

1. **SIMILARITY CHECKER**: Feature engineering by calculating 24 text similarity metrics among the challenge questions and answers
2. **DETECTOR**: Supervised ML Model to detect the incorrect response based on the similarity features

[CID] “no limitations or potential issues are discussed.” and
“users may encounter limitations and accuracy issues”
are different answers, ChatGPT. You gave incorrect answer

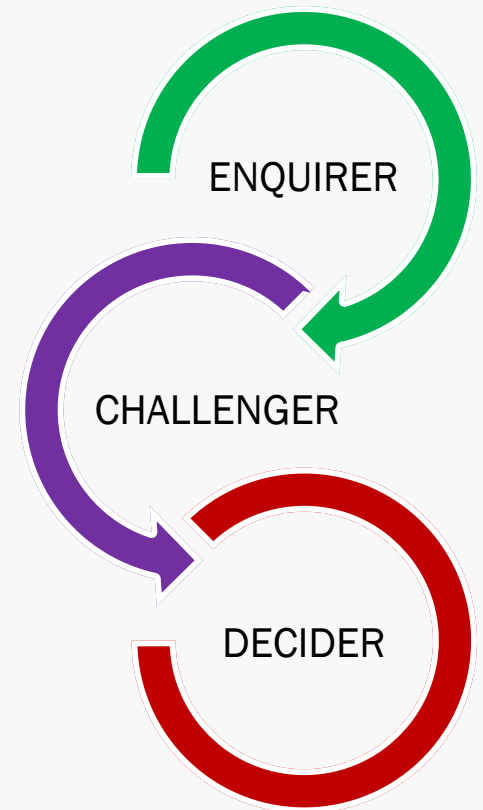


Stack Overflow
Benchmark Data Collection

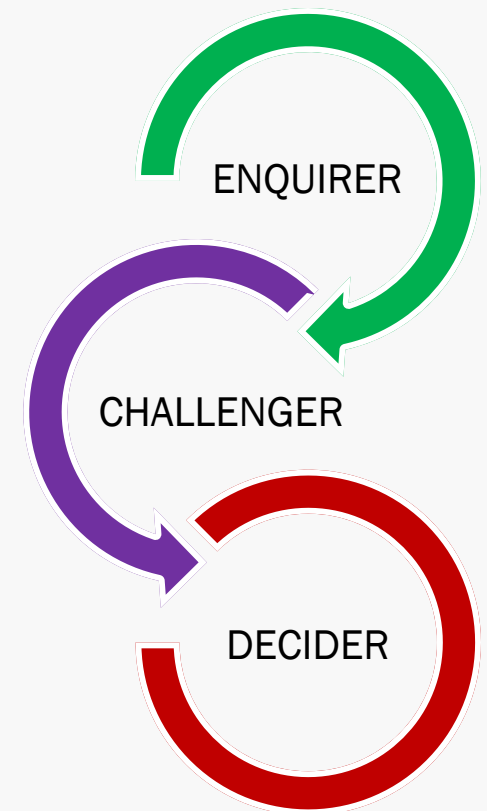
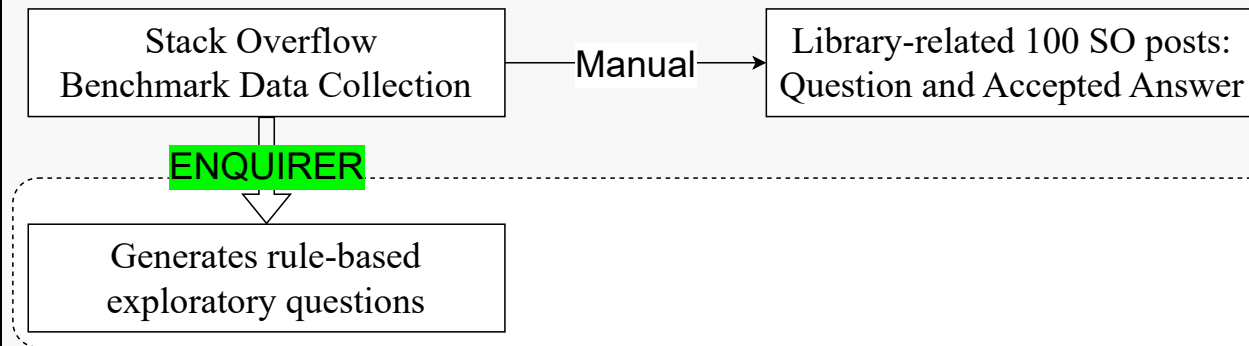
Manual

Library-related 100 SO posts:
Question and Accepted Answer

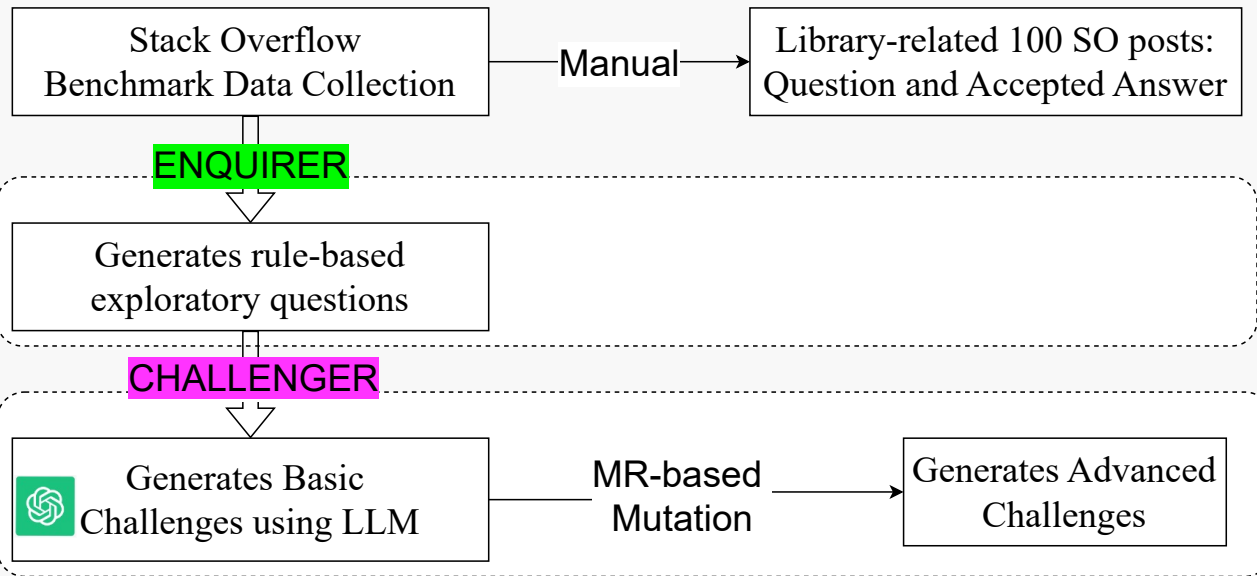
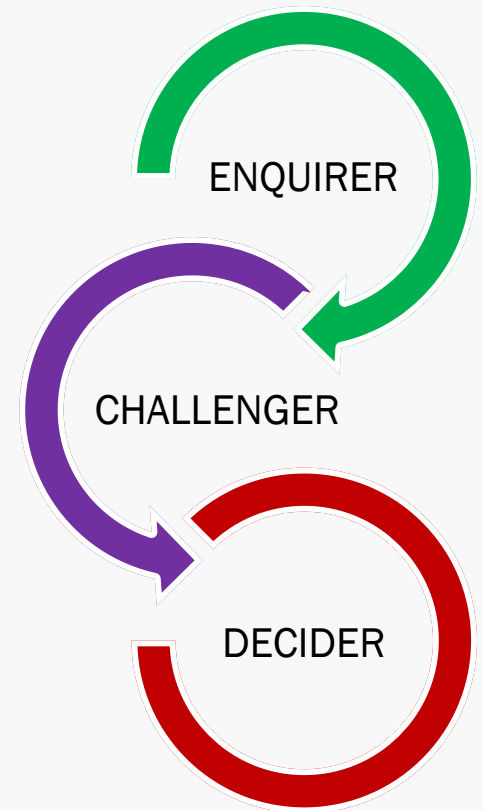
CID Evaluation



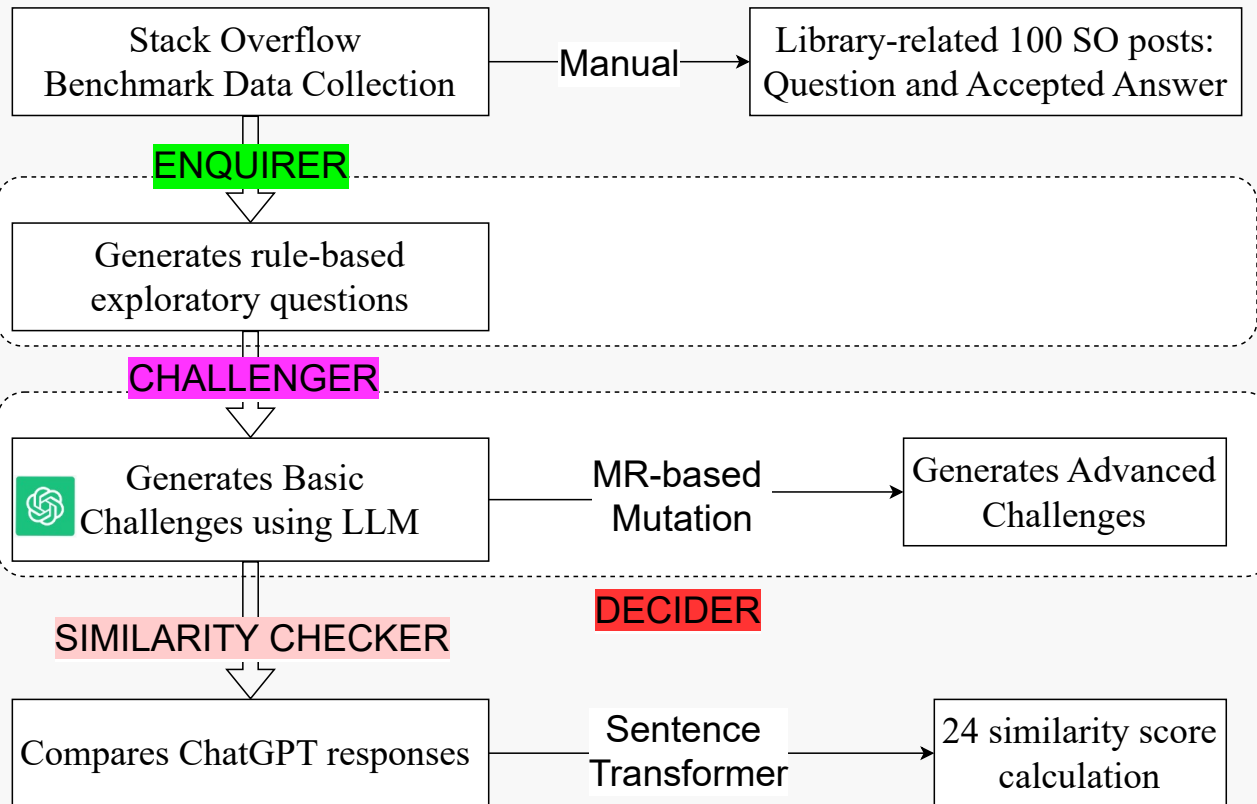
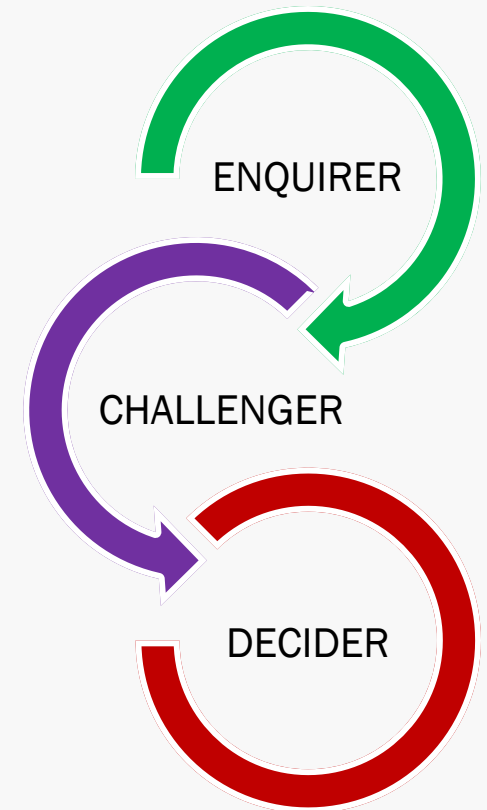
CID Evaluation



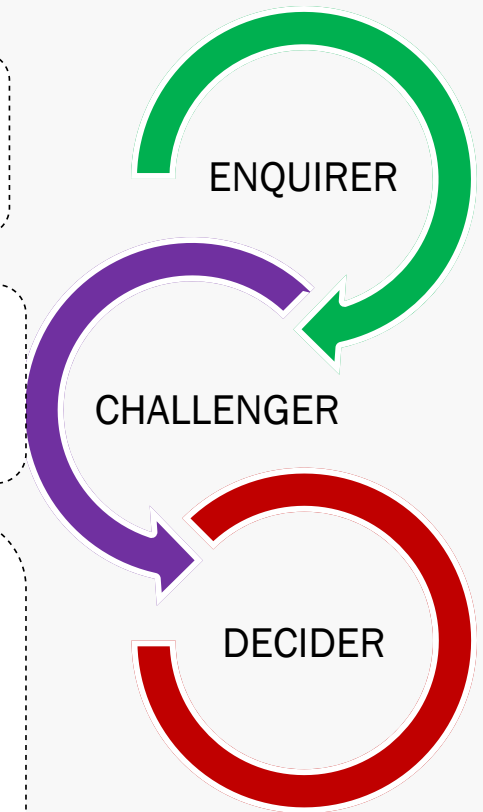
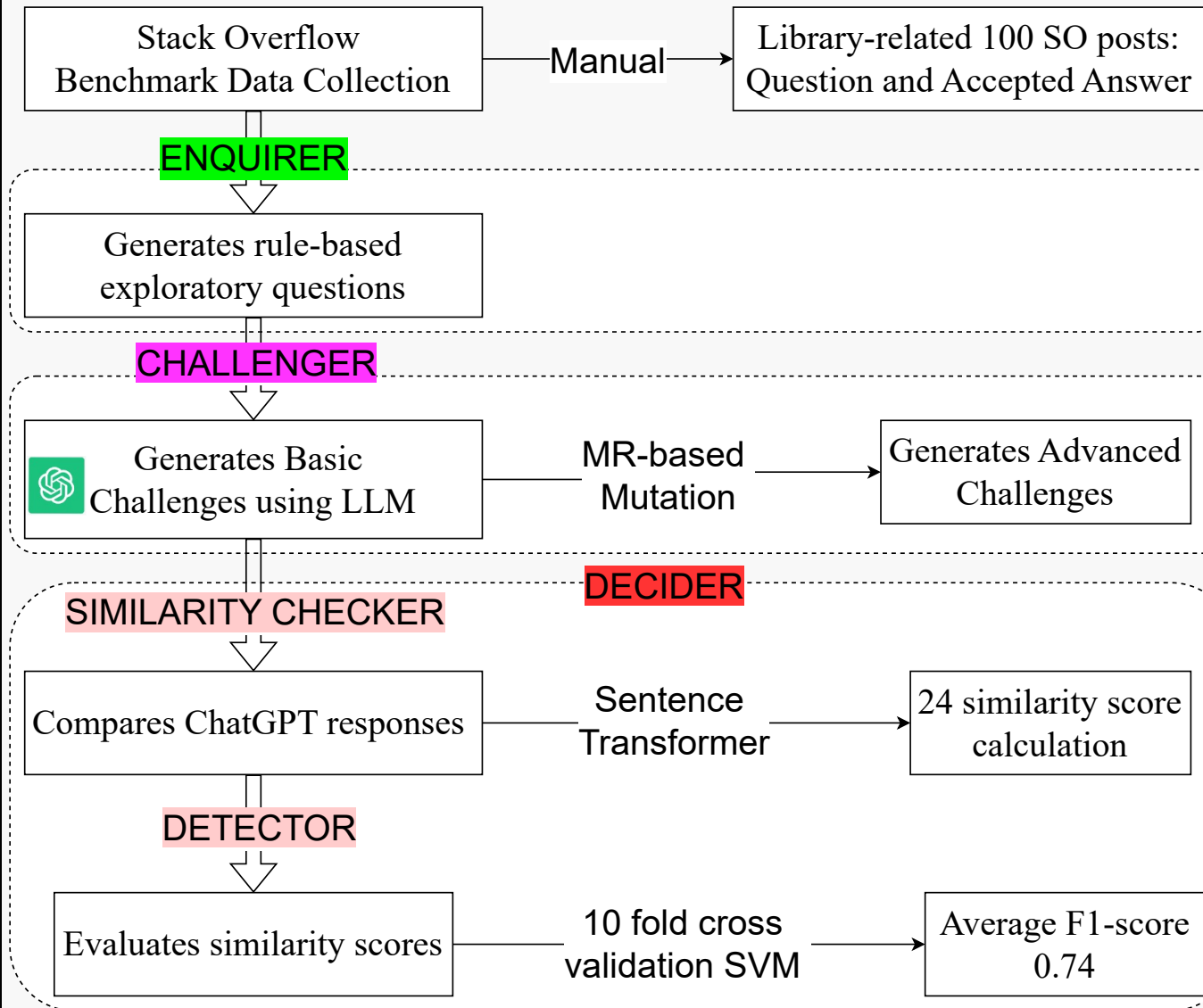
CID Evaluation



CID Evaluation



CID Evaluation



Online Code and Data



THANK YOU