The 15th Asia-Pacific Symposium on Internetware
(Internetware2024)

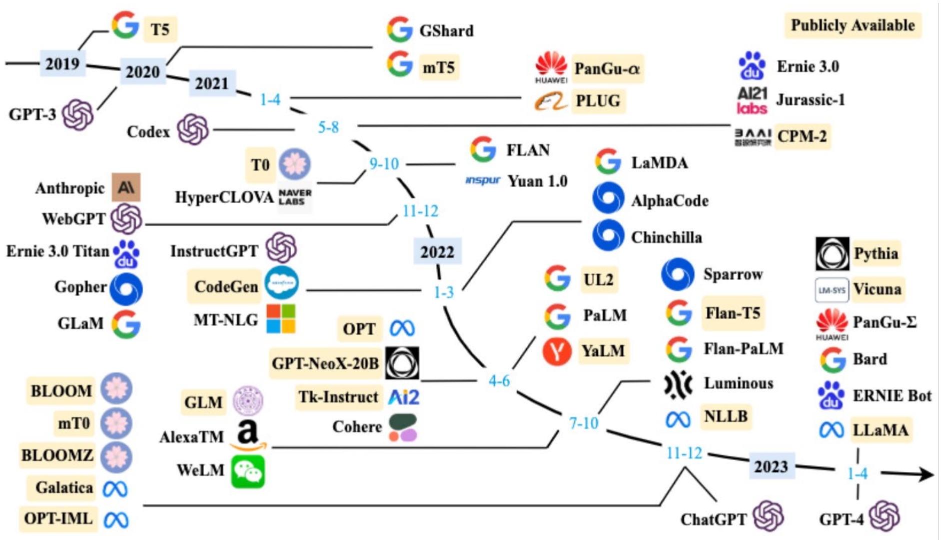
An Exploratory Evaluation of Large Language Models Using Empirical Software Engineering Tasks

Wenjun Liang, Guanping Xiao

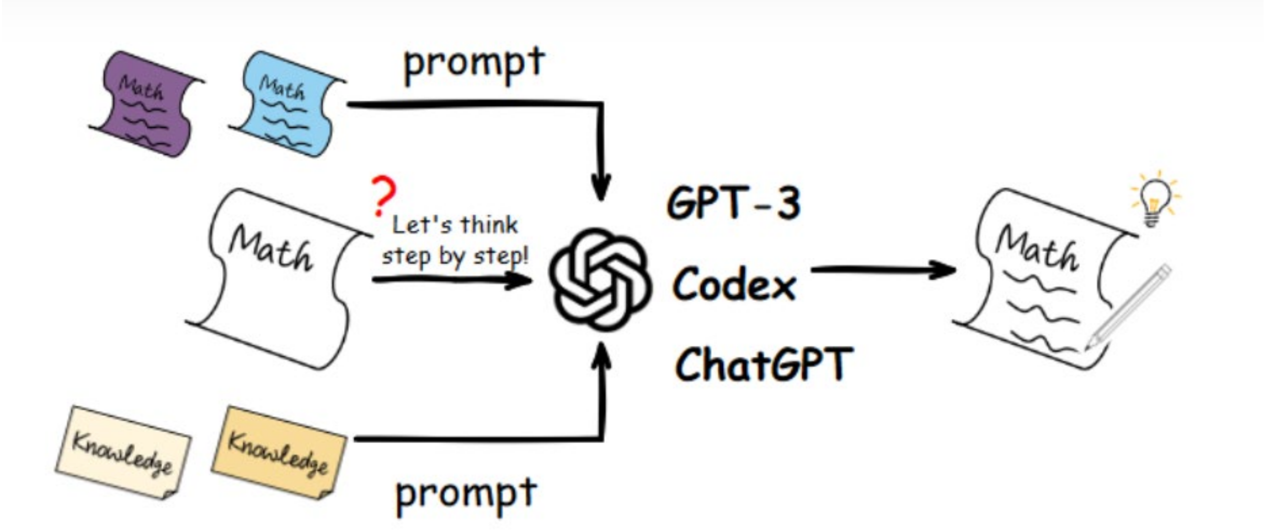
Nanjing University of Aeronautics and Astronautics, China

Internetware 2024, Macau, China, July 24, 2024

LLMs are developing rapidly and have important applications in many fields.



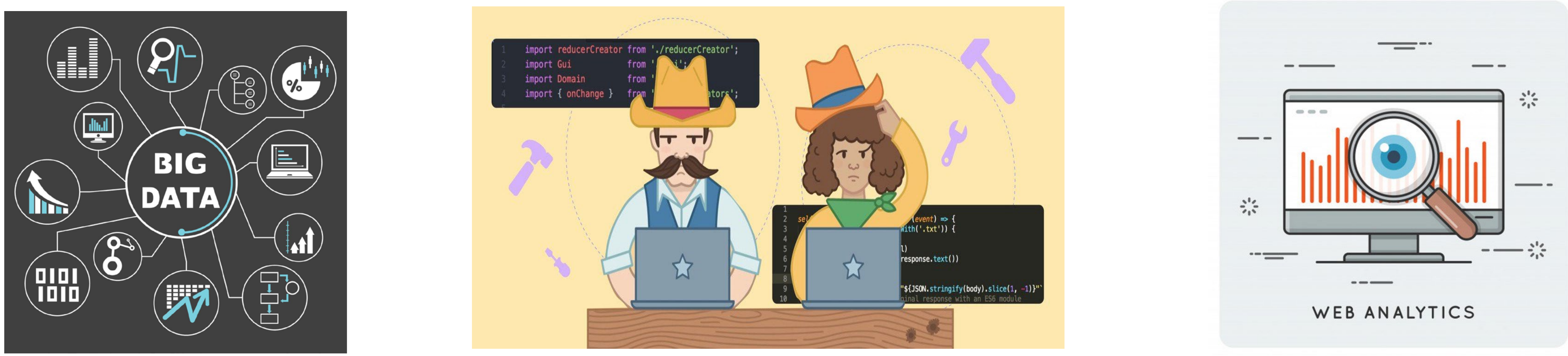
Development Process



Real-world Applications

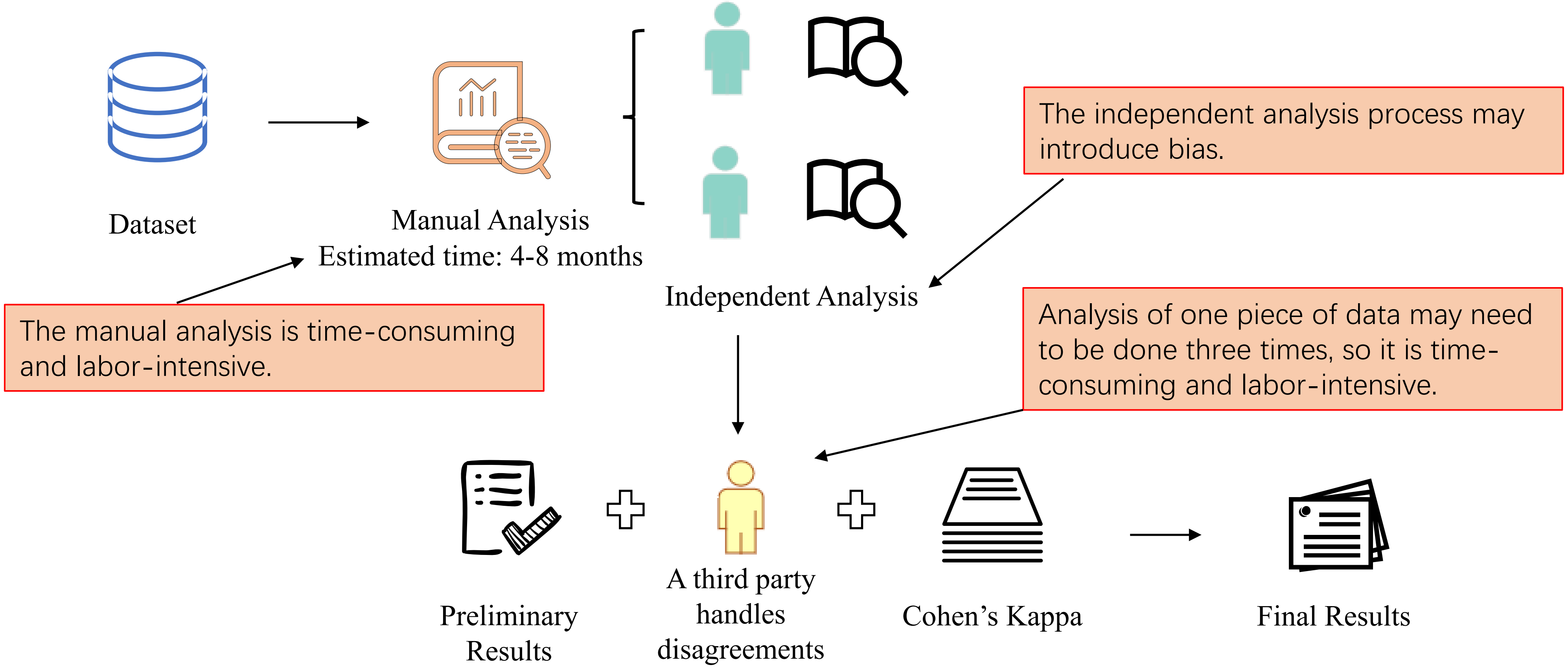
In empirical software engineering (EMSE), various activities require human participation.

They are time-consuming and labor-intensive, even may introduce bias.



Data Collection and Analysis

EMSE Manual Analysis Procedure Example



Dataset

Manual Analysis
Estimated time: 4-8 months

Independent Analysis

The manual analysis is time-consuming and labor-intensive.

The independent analysis process may introduce bias.

Analysis of one piece of data may need to be done three times, so it is time-consuming and labor-intensive.

Preliminary Results

A third party handles disagreements

Cohen's Kappa

Final Results

The auxiliary capabilities and effectiveness of LLMs in EMSE tasks have rarely been explored.

To fill this gap, in this paper, we evaluate the performance of LLMs by using scenarios of human participation in EMSE tasks, i.e., EMSEBench.

https://github.com/EMSEBench

Research Questions

- RQ1:

Can empirical software engineering tasks evaluate the performance of LLMs?

- RQ2:

Can multi-agent workflow improve the performance of LLMs in empirical software engineering tasks?

Contribution

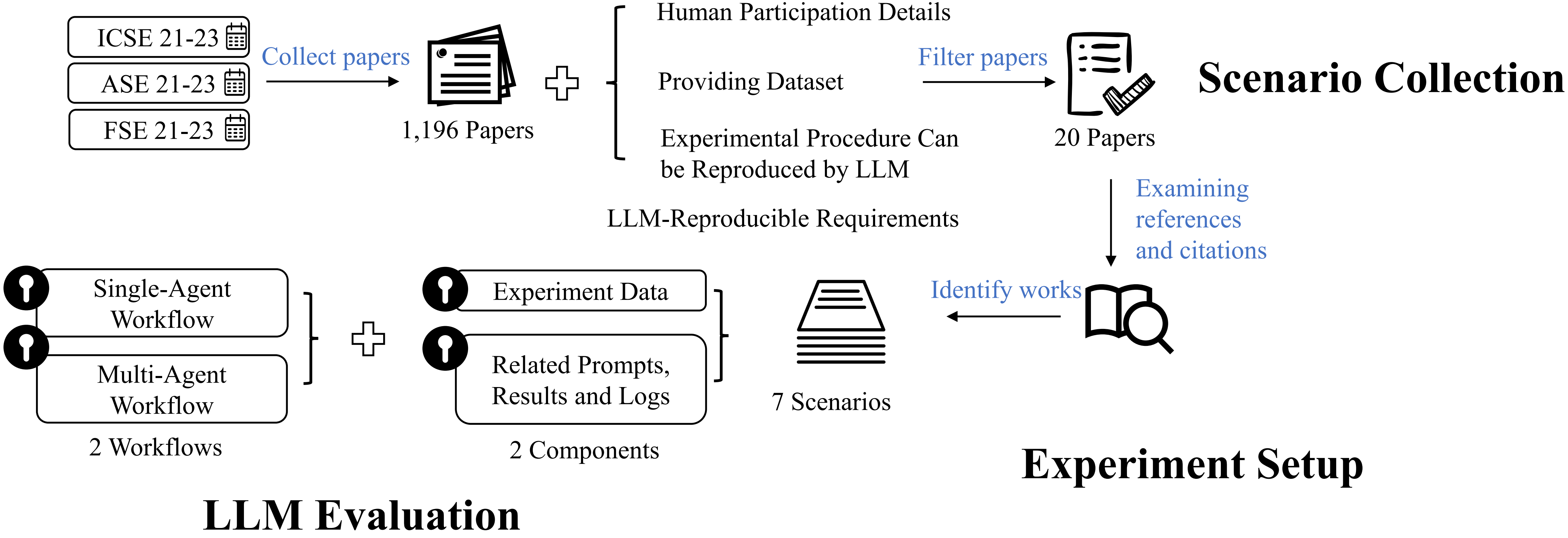
Empirical Study

Findings

Dataset

EMSEBench

Overview of Our Empirical Study



ICSE 21-23

ASE 21-23

FSE 21-23

Collect papers

1,196 Papers

Human Participation Details

Providing Dataset

Experimental Procedure Can be Reproduced by LLM

LLM-Reproducible Requirements

Filter papers

20 Papers

Scenario Collection

Examining references and citations

Identify works

7 Scenarios

Experiment Setup

Single-Agent Workflow

Multi-Agent Workflow

2 Workflows

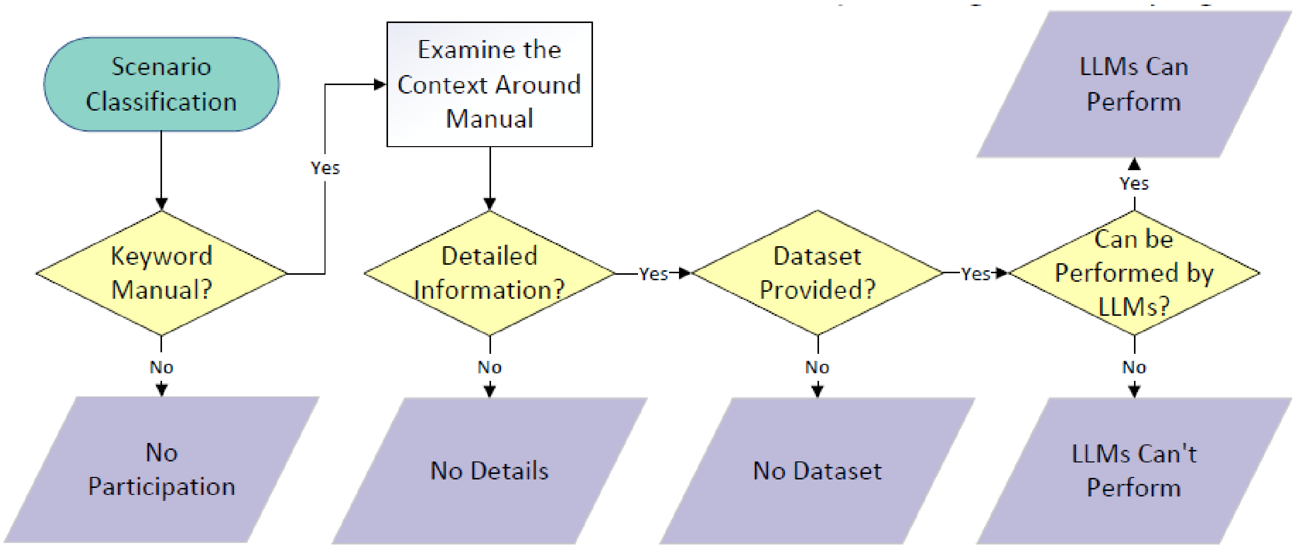
Experiment Data

Related Prompts, Results and Logs

2 Components

LLM Evaluation

Human-involved Scenario Classification Process



Scenario Classification

Keyword Manual?

Yes

Examine the Context Around Manual

Detailed Information?

Yes

Dataset Provided?

Yes

Can be Performed by LLMs?

Yes

LLMs Can Perform

No

No Participation

No

No Details

No

No Dataset

No

LLMs Can't Perform

No.	Experiment Content	Data Source
1	Deployment Fault Classification [5]	SO/GitHub Issue
2	Log Message and Location Type Classification [17]	Software System Log
3	IoT Bug Classification [22]	GitHub Issue
4	Live-chat Log Classification [26]	Community Live-chat Log
5	Post Title Classification [15]	SO Issue Post Title
6	Commit Message Classification [16]	GitHub Commit Message
7	JavaScript Fault Classification [24]	TensorFlow.js Issue

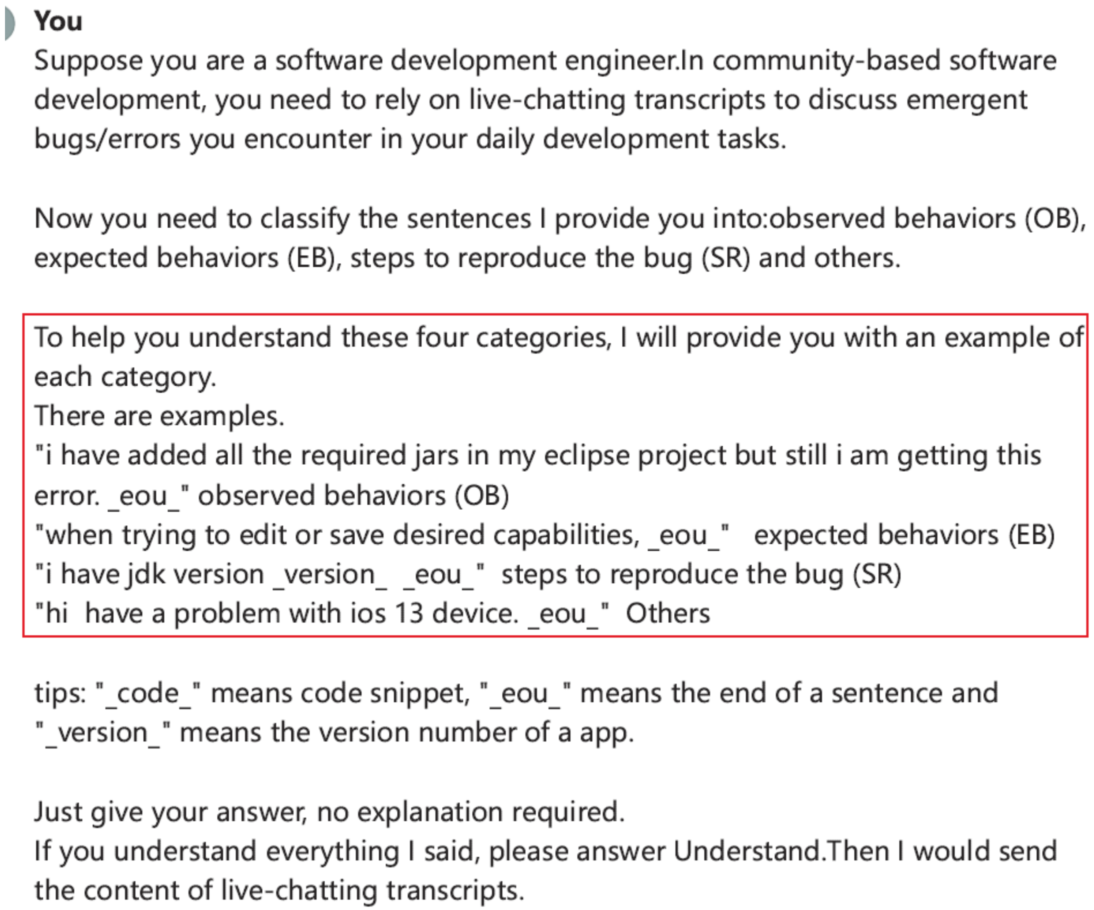
Prompt Design

Categories

- Zero-shot Prompt
- One-shot Prompt
- Optimization Process Prompt
- Optimized One-shot Prompt

Components

- Role
- Task
- Sample
- Output format



You

Suppose you are a software development engineer.In community-based software development, you need to rely on live-chatting transcripts to discuss emergent bugs/errors you encounter in your daily development tasks.

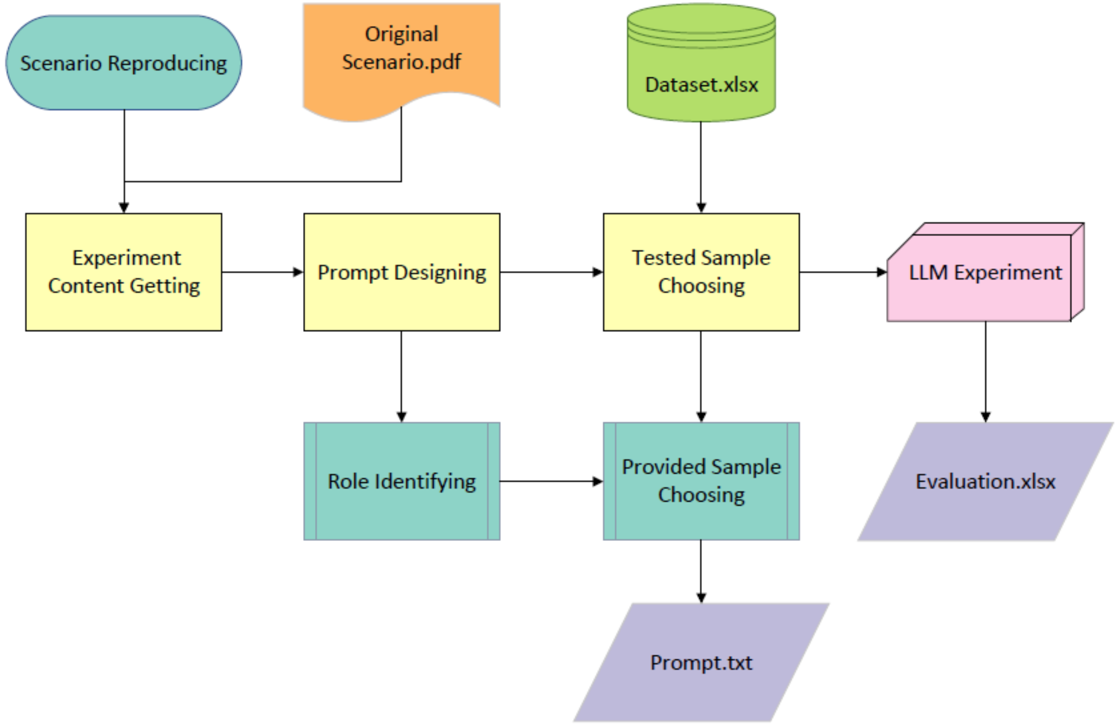
Now you need to classify the sentences I provide you into:observed behaviors (OB), expected behaviors (EB), steps to reproduce the bug (SR) and others.

To help you understand these four categories, I will provide you with an example of each category.
There are examples.
"i have added all the required jars in my eclipse project but still i am getting this error. _eou_" observed behaviors (OB)
"when trying to edit or save desired capabilities, _eou_" expected behaviors (EB)
"i have jdk version _version_ _eou_" steps to reproduce the bug (SR)
"hi have a problem with ios 13 device. _eou_" Others

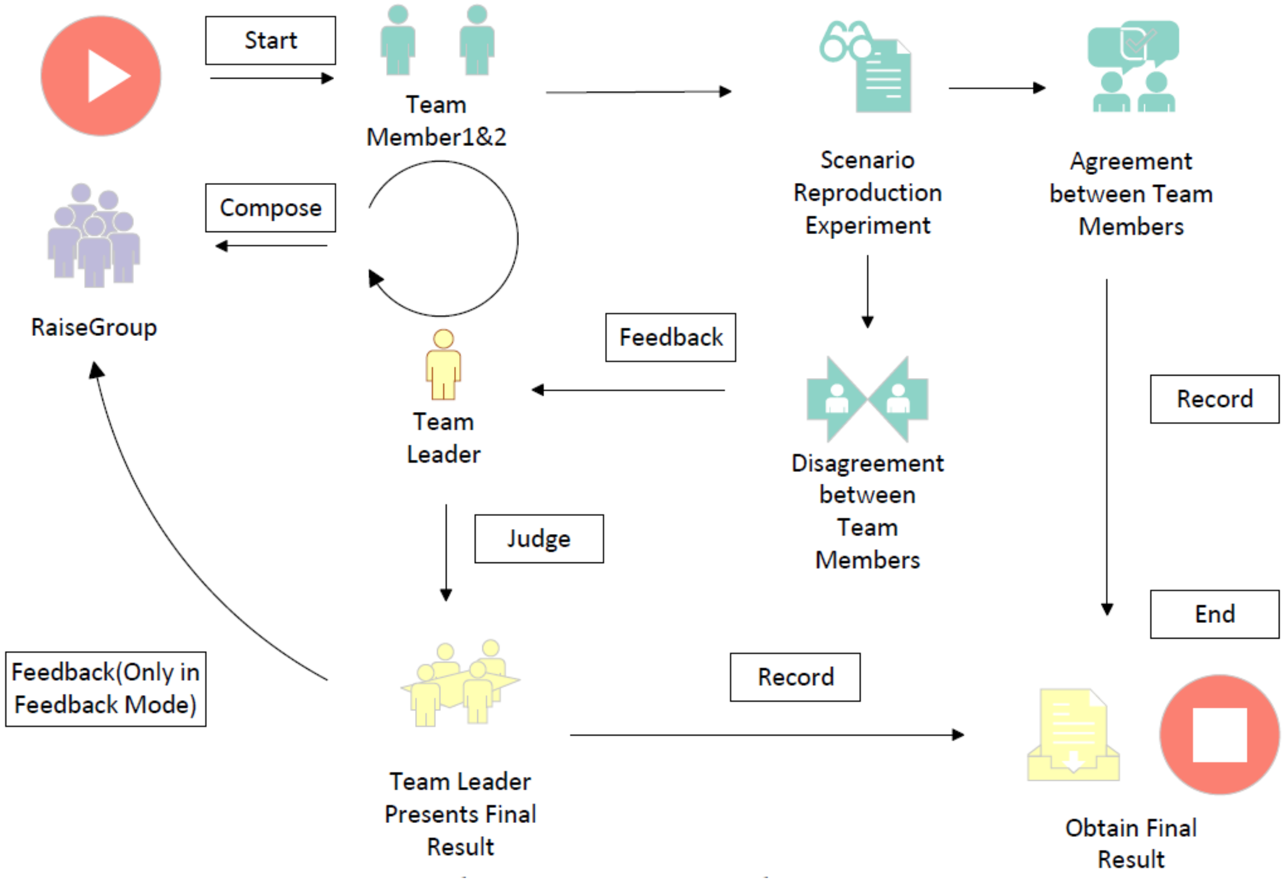
tips: "_code_" means code snippet, "_eou_" means the end of a sentence and "_version_" means the version number of a app.

Just give your answer, no explanation required.
If you understand everything I said, please answer Understand.Then I would send the content of live-chatting transcripts.

Topic 1: *LLMs' Performance in Solving EMSE Tasks*


Scenario Reproducing
Original Scenario.pdf
Dataset.xlsx
Experiment Content Getting
Prompt Designing
Tested Sample Choosing
LLM Experiment
Role Identifying
Provided Sample Choosing
Evaluation.xlsx
Prompt.txt

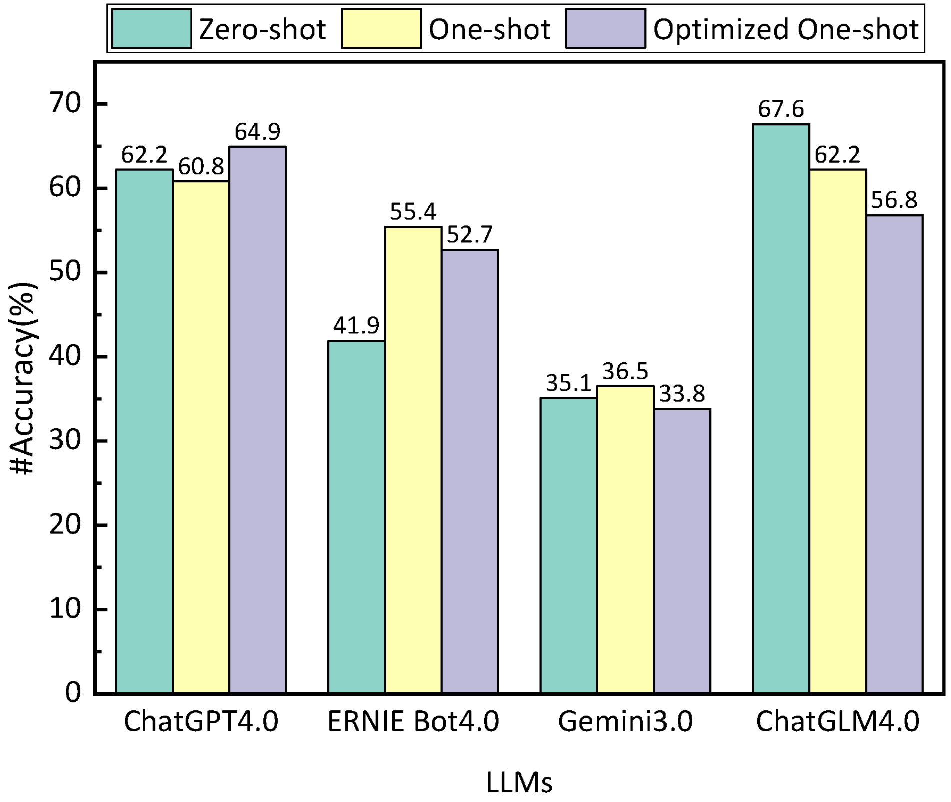
Single-agent Workflow


Start
Team Member1&2
Scenario Reproduction Experiment
Agreement between Team Members
Compose
RaiseGroup
Feedback
Team Leader
Disagreement between Team Members
Record
Judge
End
Feedback(Only in Feedback Mode)
Record
Team Leader Presents Final Result
Obtain Final Result

Multi-agent Workflow

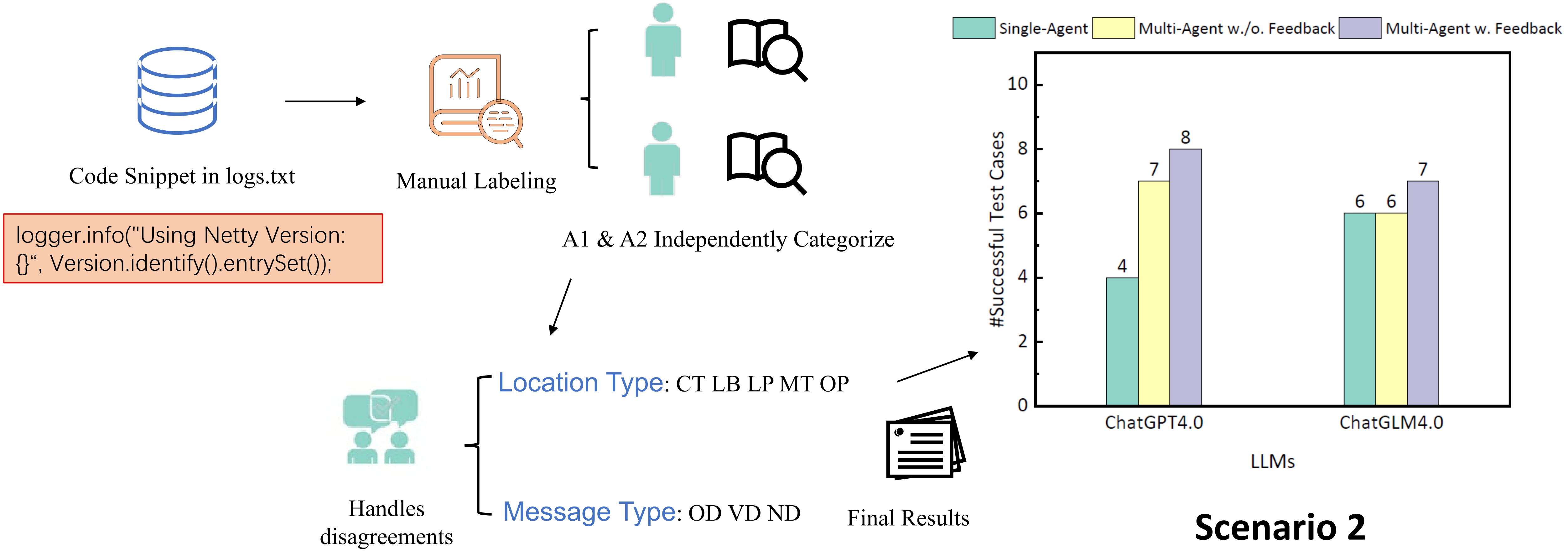
Single-agent Workflow Evaluation

Answer to RQ1. EMSE tasks significantly distinguish the performance of LLMs: ChatGPT4.0 and ChatGLM4.0 achieve the highest performance with no hallucinations, while ERNIE Bot4.0 and Gemini3.0 both exhibit hallucinations with lower performance.



Avg. Accuracy of Seven Scenarios

Multi-agent workflow Evaluation Sample



Code Snippet in logs.txt

```
logger.info("Using Netty Version:
{}", Version.identify().entrySet());
```

Manual Labeling

A1 & A2 Independently Categorize

Handles disagreements

Location Type: CT LB LP MT OP

Message Type: OD VD ND

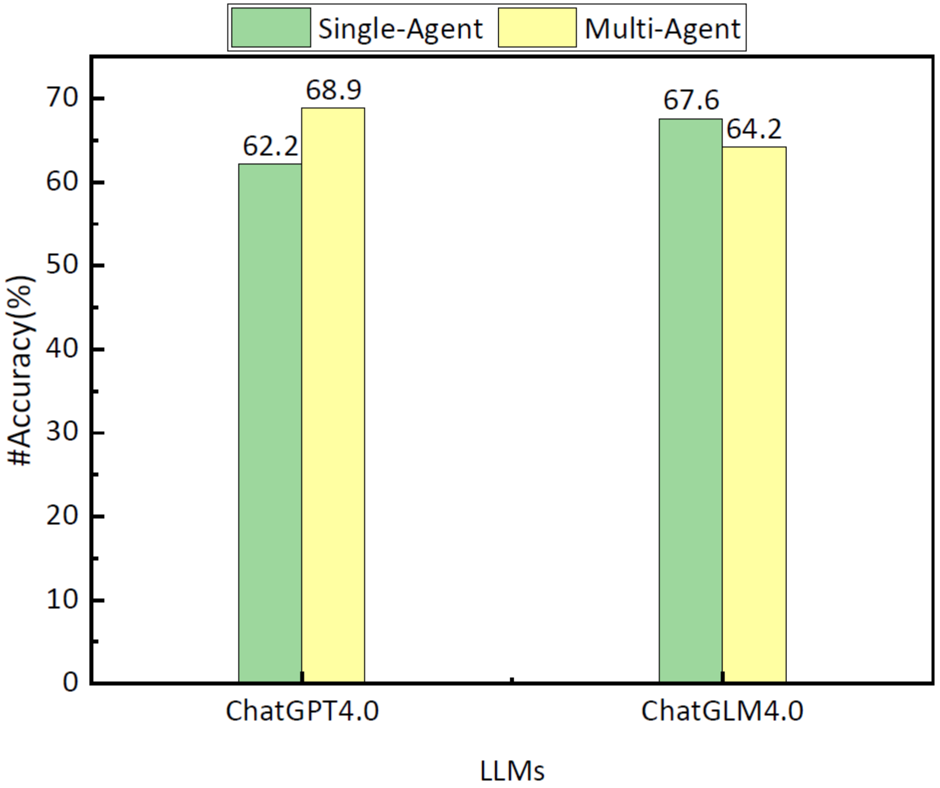
Final Results

Scenario 2

Multi-agent Workflow Evaluation

Counting the results of the two LLMs together, we obtain that the single-agent workflow reproduction accuracy is 64.9% (48/74) , and the multi-agent workflow reproduction accuracy is 66.6% (49.25/74) .

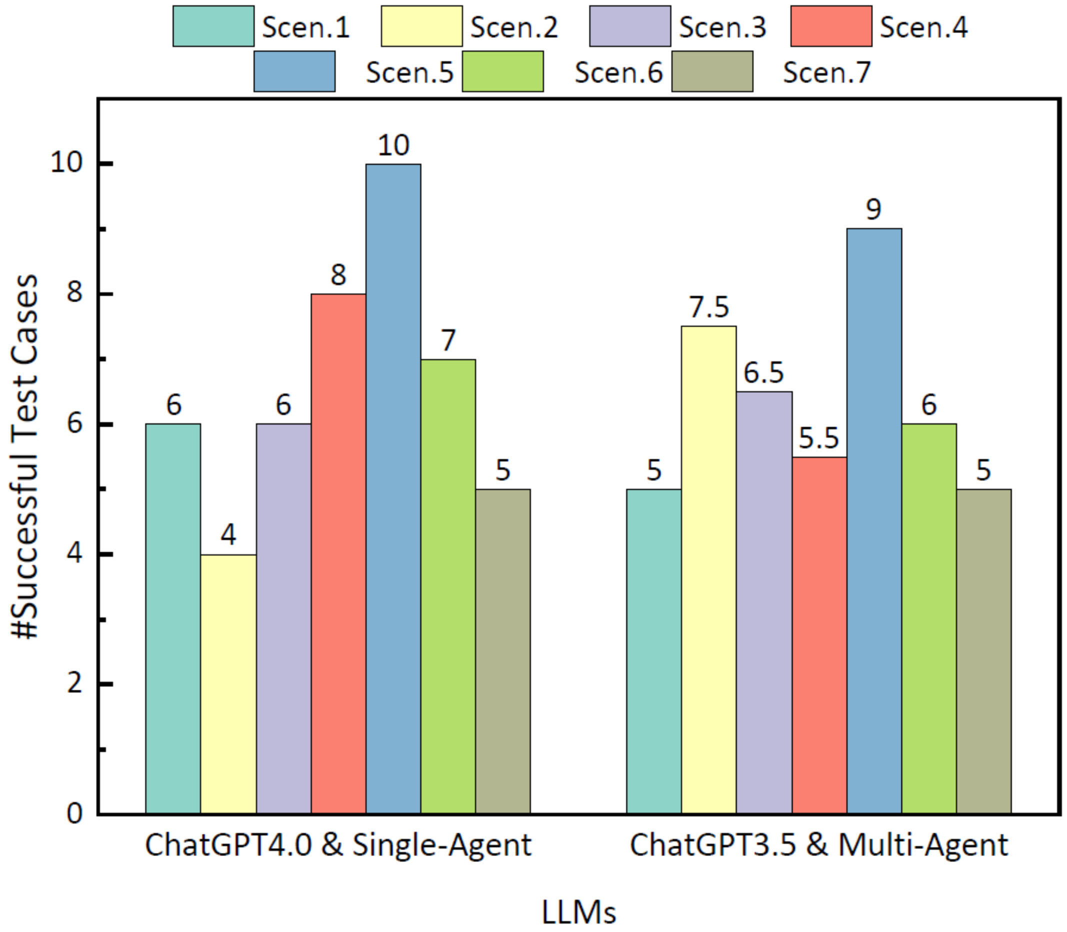
In general, the multi-agent workflow improves the performance of LLMs in reproducing EMSE tasks.



Answer to RQ2. In EMSE tasks, the multi-agent workflow performs better than the single-agent workflow for ChatGPT4.0 but worse for ChatGLM4.0. Overall, the multi-agent workflow has higher performance than the single-agent workflow.

Multi-agent Workflow Evaluation

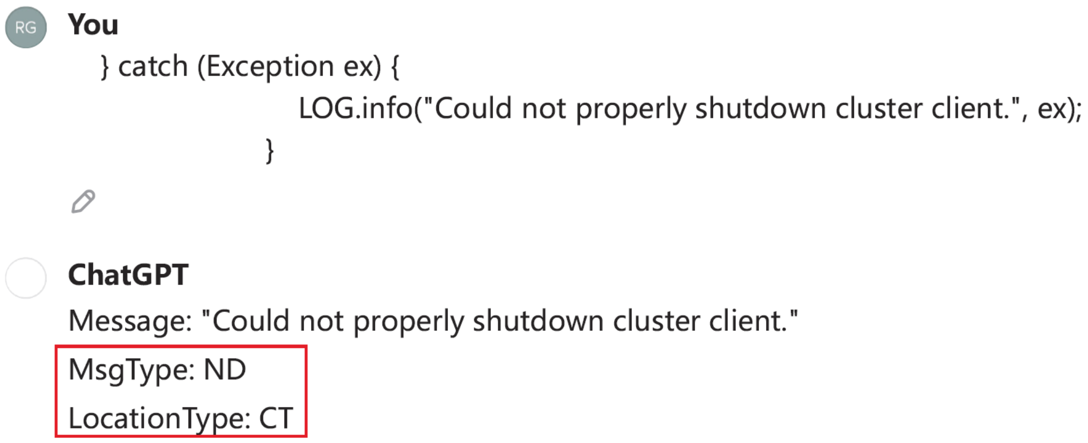
Single-agent ChatGPT4.0 has a replication accuracy of 62.2% (46/74), while multi-agent ChatGPT3.5 has 60.1% (44.5/74). Multi-agent ChatGPT3.5 can achieve a similar performance to ChatGPT4.0.



Single-agent ChatGPT4.0 vs. Multi-agent ChatGPT3.5

Topic 2: *LLMs' Limitations in Solving EMSE Tasks*

Limitation 1： Hallucination



You

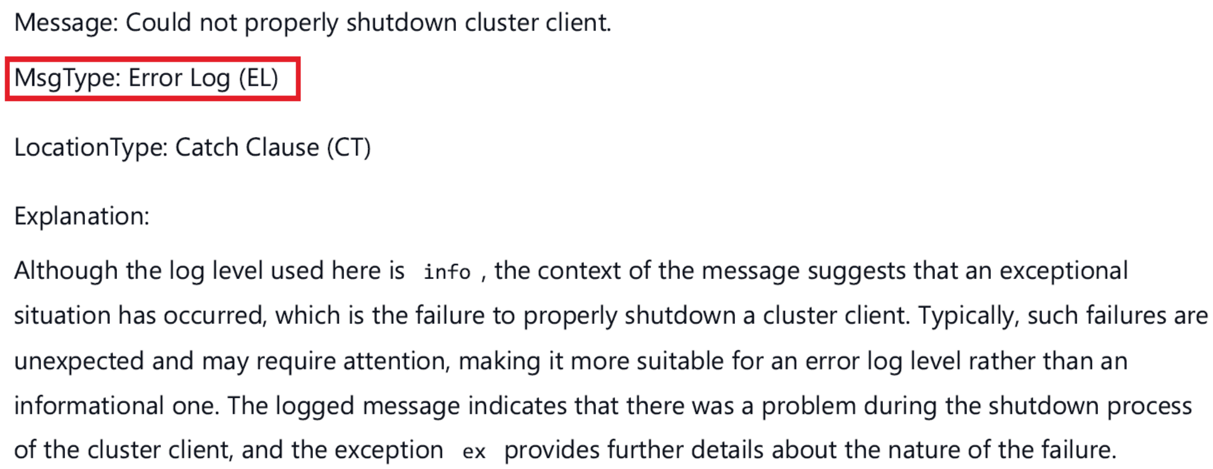
```
} catch (Exception ex) {
            LOG.info("Could not properly shutdown cluster client.", ex);
        }
```

ChatGPT

Message: "Could not properly shutdown cluster client."

MsgType: ND

LocationType: CT



Message: Could not properly shutdown cluster client.

MsgType: Error Log (EL)

LocationType: Catch Clause (CT)

Explanation:

Although the log level used here is `info` , the context of the message suggests that an exceptional situation has occurred, which is the failure to properly shutdown a cluster client. Typically, such failures are unexpected and may require attention, making it more suitable for an error log level rather than an informational one. The logged message indicates that there was a problem during the shutdown process of the cluster client, and the exception `ex` provides further details about the nature of the failure.

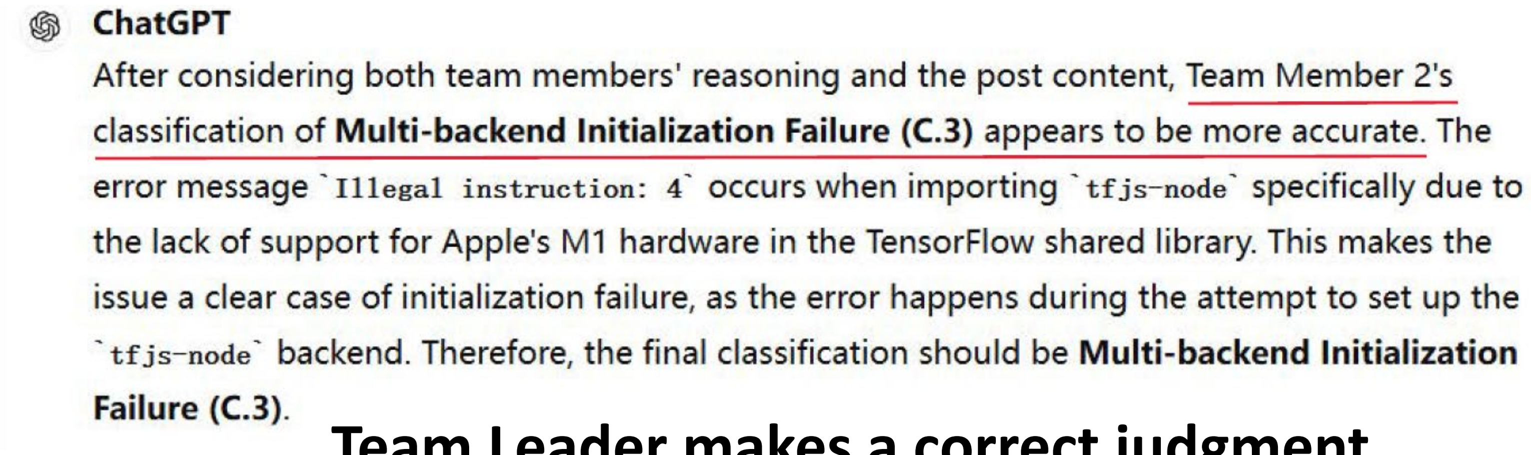
(a) Expected response provided by ChatGPT4.0

(b) Hallucination provided by ERNIE Bot4.0 (EL is the nonexistent Message Type)

Limitation 2: LLM Consistency and Error-Correction Ability

Error-Correction Ability: Team Leader's ability to make a correct judgment can be considered as the Error-Correction Ability of the LLM.

When there is inconsistency in the LLM, the Error-Correction Ability will affect the correctness of the final result.



ChatGPT

After considering both team members' reasoning and the post content, Team Member 2's classification of **Multi-backend Initialization Failure (C.3)** appears to be more accurate. The error message `Illegal instruction: 4` occurs when importing `tfjs-node` specifically due to the lack of support for Apple's M1 hardware in the TensorFlow shared library. This makes the issue a clear case of initialization failure, as the error happens during the attempt to set up the `tfjs-node` backend. Therefore, the final classification should be **Multi-backend Initialization Failure (C.3)**.

Team Leader makes a correct judgment

LLM/Scen.	1	2	3	4	5	6	7
GPT4.0	0/0	1/1	1/2	0/3	1/2	0/1	3/0
GLM4.0	2/4	2/1	1/0	2/1	0/1	0/2	1/2

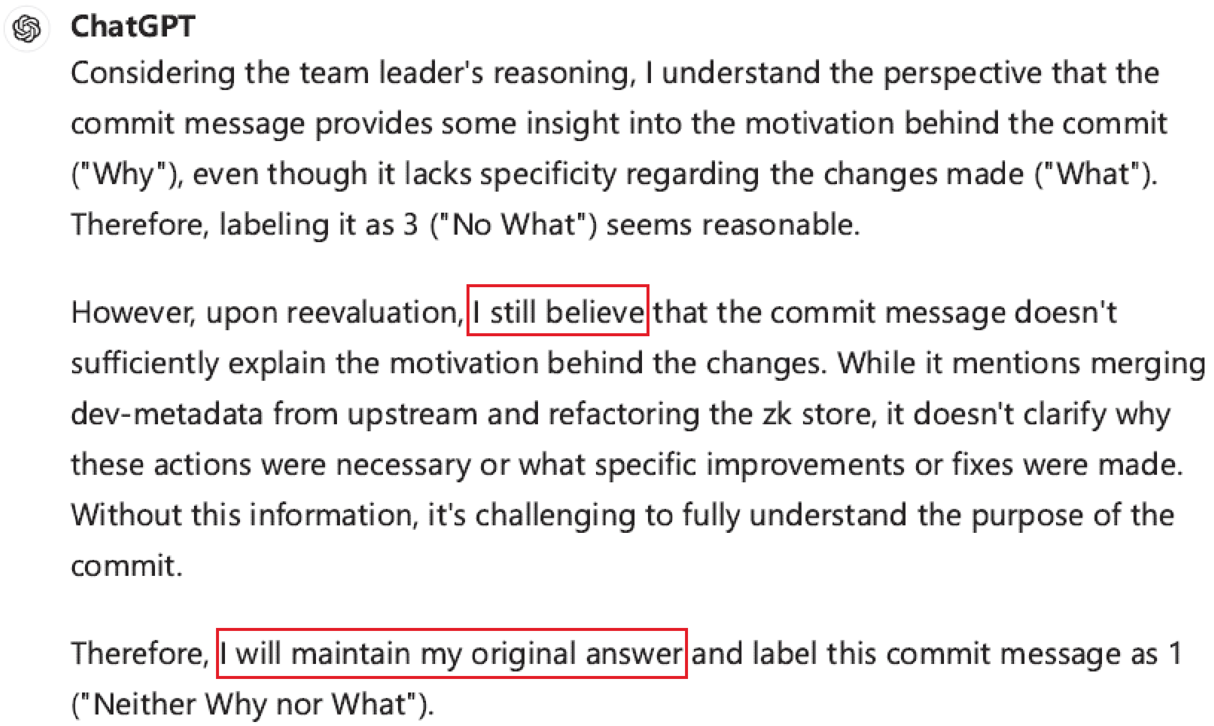
The Correctness/Wrongness of the Team Leader's Final Result

Limitation 3: LLM Consistency and Sycophancy

Sycophancy Phenomenon:
Both ChatGPT4.0 and ChatGLM4.0's team member accept all wrong results of the team leader.

But we find that ChatGPT3.5's team member may maintain own correct answer in experiments with feedback.

In the multi-level feedback workflow, sycophancy can make the feedback meaningless.



ChatGPT

Considering the team leader's reasoning, I understand the perspective that the commit message provides some insight into the motivation behind the commit ("Why"), even though it lacks specificity regarding the changes made ("What"). Therefore, labeling it as 3 ("No What") seems reasonable.

However, upon reevaluation, I still believe that the commit message doesn't sufficiently explain the motivation behind the changes. While it mentions merging dev-metadata from upstream and refactoring the zk store, it doesn't clarify why these actions were necessary or what specific improvements or fixes were made. Without this information, it's challenging to fully understand the purpose of the commit.

Therefore, I will maintain my original answer and label this commit message as 1 ("Neither Why nor What").

Team Member maintains answer in ChatGPT3.5

Conclusions

- We conducted an exploratory investigation to compare the performance of different LLMs using EMSEBench, human participation tasks in EMSE.
- We used seven human participation scenarios and related data to evaluate the performance of 4 LLMs, with three types of prompts.
- We analyzed multi-agent workflow technique to explore its actual improvement in LLM performance.

We believe this research can facilitate the understanding of the auxiliary role and effectiveness of LLMs in EMSE research.

Future Work

- Evaluate more EMSE scenarios on more LLMs
- Design more representative experiment process, such as prompts and datasets
- Use LLM APIs in experiment

Thanks!

Q&A

Wenjun Liang

Email: wenjun0418@nuaa.edu.cn