Name : Nipurna Patel
Student ID : 202201061

PART C

[Q.1] How syntactically correct are LLM-generated user stories? Are they following INVEST framework? Provide examples from your problem specification.

LLM generated user stories vary with different LLMs, by analyzing all the results I observed Gemini performing the best and Llama performing the worst.

These are some common issues with LLM generated user stories:

1. Lack of independence, combine multiple functionalities.
2. Ambiguous acceptance criteria in some cases.
3. Not specifically and not always testable

Model	Syntax accuracy	INVEST criteria	Observations
Gemini	Good	Particularly follows	Structured but sometime over explained
Mistral	Decent	Particularly follows	Less independent
DeepSeek	Average	Does not fully follows	Combine too many details
Llama	Week	Rarely follows	Less clear

Example

Manually written :

"As a customer, I want to create my own insurance package so that I can personalize my coverage to the specific needs in my budget."

- Independent - Focuses on single functionality of custom package creation
- Negotiable - can be refined based on discussion.
- Valuable - Covers specific task
- Estimable and Small
- Tastable - can verify custom package creation

LLM generated - Llama

"As a user, I want a system that allows me to create, update, and review my package while comparing competitor pricing."

- Not Independent

- Not easily testable

[Q.2] How semantically correct are LLM-generated user stories?

LLMs understand general structure but fail in domain-specific knowledge. And they sometimes assumes common roles like “user” instead of “customer” or “insurance agent”

Model	Syntax accuracy	Observations
Gemini	Good	Captures general semantics but lacks deep insurance knowledge
Mistral	Decent	Some misinterpretation in of business terminology
DeepSeek	Average	Generate too common stories, sometimes missing business logic
Llama	Week	Sometimes introduces incorrect or incomplete user goals

LLM example - Deepseek
Issue :

- "As a user, I want to check my claim history."
- Who is "user"? Is it an insurance agent or a customer?
- What does "check" mean? Is it for approval or review?

[Q.3] Are LLMs capable of identifying the information about stakeholders and user stories from their own perspectives?

LLMs struggle with stakeholder differentiation, often considering everyone as "User."
Gemini and Mistral perform better in stakeholder identification.

Model	Syntax accuracy	Observations
Gemini	Good	Recognize basic stakeholders and user stories
Mistral	Decent	Confuses primary and secondary stakeholder
DeepSeek	Average	Often assumes common “user” roles instead of defining business roles

Llama	Week	Struggle with structure stakeholder differentiation and user stories

[Q.4] Are LLMs capable of identifying the acceptance criteria (both success and failure) for the user story?

LLMs mostly focus on success conditions but fail to define failure scenarios.
LLaMA and DeepSeek often do not consider failure cases entirely.

Model	Syntax accuracy	Observations
Gemini	Good	Misses some failure scenarios
Mistral	Decent	Lacks in clear failure
DeepSeek	Average	Incomplete failure condition
Llama	Week	Not handling failure

LLM example - DeepSeek

- "User uploads a document successfully." (No testable condition)
- What happens if upload fails? (Missing failure criteria)