

Content of Interview Protocol:

- **Introduction:** You are invited to participate in evaluating Gherkin specifications generated by a Large Language Model (LLM) (ChatGPT) as part of a research study aimed at improving software artifacts related to food safety monitoring systems. Your feedback is crucial to understanding the usability and accuracy of these AI-generated outputs.
- **Objective:** Your task is to review the Gherkin specifications generated by (LLM) based on the regulatory provisions provided. We will use your review results to identify the strengths and weaknesses of LLMs in automatically generating Gherkin specifications. You are asked to provide a rating for each criterion per specification and provide feedback on identified issues to help us understand problems or room for improvement.
- **Background:** The regulatory provisions are extracted from food-safety regulations. The provisions have been colour-coded to help you more easily see the food-safety concepts that they relate to. Use this information to guide your review and ensure the specifications reflect the intended concepts accurately. Gherkin specifications are expected to follow the **Given-When-Then** format:
  1. Given describes the context or initial state of the system.
  2. When specifies the event or action that triggers the behaviour.
  3. Then defines the expected outcome or result of that action.
- **Evaluation Criteria:** Use the following criteria to assess the Gherkin specifications:
  1. **Relevance:** Specification matches the intended system behaviour as described in the regulatory text while focusing on what is directly pertinent.  
 Key Points: Only pertinent details for the specified behaviour are presented, i.e., nothing is said that is not relevant.  
 Scale:
    - (a) All irrelevant: The output does not address the requirement at all. The output is entirely off-topic or unnecessary.
    - (b) Mostly irrelevant: A small portion of the output addresses the requirement, while most content is off-topic or unnecessary.
    - (c) Somewhat relevant: A substantial portion of the output relates to the requirement, but an equally substantial portion remains off-topic or unnecessary.
    - (d) Mostly relevant: Most of the output addresses the requirement, with only a small portion being off-topic or unnecessary.
    - (e) All relevant: All content in the output addresses the requirement, with no off-topic or unnecessary information.
  2. **Completeness:** Specification covers all intended functionality.  
 Key Points: All functionalities implied by the regulatory requirement are included without omissions.  
 Scale:
    - (a) Fully incomplete: No required information is included. The specification omits all relevant information from the requirement.
    - (b) Mostly incomplete: Only a small fraction of the required information is included, while most of the relevant information is missing.
    - (c) Partially complete: A substantial portion of the required information is included, but an equally substantial portion is missing.
    - (d) Mostly complete: Most of the required information is included, with only a small portion of relevant information missing.
    - (e) Fully complete: All required information is included. The specification omits no relevant information.
  3. **Clarity:** Specification is written in a clear and unambiguous manner.  
 Key Points: 1) Specifications can be uniquely interpreted, i.e., not being vague or misleading. 2) Technical jargon commensurate with the need for specification.  
 Scale:
    - (a) Completely unclear: No part of the specification is understandable. The content is highly vague, ambiguous, or difficult to interpret.

- (b) Mostly unclear: Only a small portion of the specification is understandable, while most parts are vague, ambiguous, or difficult to interpret.
  - (c) Somewhat clear: A substantial portion of the specification is understandable, but an equally substantial portion is vague, ambiguous, or difficult to interpret.
  - (d) Mostly clear: Most parts of the specification are easy to understand, with only a small portion being vague, ambiguous, or difficult to interpret.
  - (e) Completely clear: All parts of the specification are easy to understand, with no vague, ambiguous, or difficult-to-interpret content.
4. Integrity (later removed): Specification follows the BDD structure of Given-When-Then, i.e., Each scenario has a non-empty context, event, and outcome.
- Key Points: If the regulatory requirement is not written with a specific event or lacks context, the specification might not include valid steps. 2) Context might be included in the background rather than scenarios.
- (a) No adherence: The specification does not follow the BDD structure at all.
  - (b) Minimal adherence: The specification follows the BDD structure in a few scenarios, while it does not follow it in most scenarios.
  - (c) Somewhat adherent: The specification follows the BDD structure in a substantial number of scenarios, but it does not follow it in a substantial number of scenarios.
  - (d) High adherence: The specification follows the BDD structure in most scenarios, while it does not follow it in a few scenarios.
  - (e) Fully adherent: The specification strictly follows BDD structure in all scenarios.
5. Singularity: Each scenario focuses on a single purpose i.e., context, event, and outcome are provided to describe that purpose.
- Key Points: 1) Specification avoids mixing multiple purposes within a single scenario. 2) Specification splits scenarios when multiple functions or events serve distinct purposes to improve testability and maintainability.
- Scale:
- (a) Completely mixed: The output combines multiple purposes in each scenario, making testing and maintenance difficult.
  - (b) Mostly mixed: The output separates purposes in a few scenarios but still combines multiple purposes in most, making testing and maintainability difficult.
  - (c) Somewhat singular: The output features a substantial number of scenarios with a single purpose, but an equally substantial number that combine multiple purposes.
  - (d) Mostly singular: The output ensures most scenarios focus on a single purpose, but a few still combine multiple purposes, making testing and maintainability somewhat difficult.
  - (e) Completely singular: The output ensures each scenario has a single purpose, making testing and maintenance easy.
6. Time Savings: The extent to which the specification reduces the time a participant would otherwise spend creating the specification from scratch.
- Key Points: Reflects how much of the specification can be reused or adapted instead of fully rewritten.
- (a) Completely unhelpful (no time saved) Does not reduce the workload at all. The specification must be entirely rewritten from scratch.
  - (b) Mostly unhelpful (little time saved) Only a small fraction of the specification can be reused. Most parts require to be rewritten from scratch.
  - (c) Somewhat helpful (moderate time saved) A substantial portion of the specification can be reused, but an equally substantial portion still needs to be rewritten from scratch.
  - (d) Mostly helpful (significant time saved) Most of the specification can be reused. Only a small portion that needs to be rewritten from scratch.
  - (e) Completely helpful (maximum time saved) Completely eliminates manual work. The specification can be adopted virtually as-is.

• **To do:**

1. Review best practices in writing BDD

2. Read the regulatory requirements carefully:
    - a) Understand the intent and scope of the input.
  3. Review each specification against the criteria:
    - a) Use the provided criteria (Relevance, Completeness, Clarity, and Singularity, Time Savings) and the scale per criterion.
  4. Provide constructive feedback:
    - a) Be precise about issues per criterion e.g., ‘ambiguous or unclear terms’, ‘specific functionality not being covered’, etc. to justify your ratings and present areas for improvement in your feedback.
  5. Write your own version of the specification if you think your version is more plausible based on common sense.
    - a) If you detect major issues with the given specification, it might be easier for you to rewrite it as an alternative to providing feedback (you still need to provide ratings for the given specification, though).
    - b) Even if you strongly agree with the definition of criteria for the given specification, please write down your observations when you notice something interesting or room for improvement. For instance, If the specification is ‘all relevant’ and ‘fully complete’ but could be improved by designing it differently, provide your feedback and suggest your version of the specification. Or, if the specification is ‘Mostly singular’, but you would design it differently for any specific reason, explain your feedback and include your version.
  6. Is the specification plausible for real-world implementation, given standard physics constraints and common sense (Yes/No)? If not, please elaborate on any concerns or issues that might arise.
  7. Ask for our feedback:

If you have any concerns/questions, please don’t hesitate to reach out to us at (s.hassani@uottawa.ca). The quality of your work is our top priority.
- **Final Note:** Your feedback is crucial to evaluating and improving the quality of LLM-generated specifications and ensuring alignment with best practices in BDD. Thank you for your participation!
  - **Observations and Insights:** Below are common patterns and suggestions based on our evaluations:

1. Assumptions: LLMs tend to include assumptions. We should decide to what degree their generated content is relevant. Content could be considered relevant and acceptable based on common-sense interpretation of input or potentially irrelevant, i.e., they contradict the input or are not supported by the input.

**Example:**

Regulatory Requirement: “(5) In the case of an inspection of a consumer prepackaged food that consists of a liquid, the net quantity of the food must be determined on the basis of the assumption that the liquid is at a temperature of 20°C.”

**Original Specification:**

“Feature: Liquid Food Product Inspection

Scenario: Measuring net quantity of prepackaged liquid food

Given a consumer prepackaged food product containing liquid

And the liquid temperature is stabilized at 20 degrees Celsius

When the inspector measures the net quantity

Then the system should record the measurement value

And store the measurement with the temperature condition

Scenario: Attempting measurement at incorrect temperature

Given a consumer prepackaged food product containing liquid

And the liquid temperature is not at 20 degrees Celsius

When the inspector attempts to measure the net quantity

Then the system should display a warning message

And prevent the measurement from being recorded

And prompt for temperature adjustment to 20 degrees Celsius”

**Improved Specification:**

“Feature: Liquid Food Product Inspection

As a food safety inspector, I need to measure the net quantity of prepackaged liquid food adjusted to 20°C so that I can ensure accurate quantity measurements for consumer protection

Background: Given the inspection system is calibrated for temperature-adjusted liquid measurements And the system has temperature monitoring and adjustment capabilities

Scenario: Measuring net quantity of prepackaged liquid food

Given a consumer prepackaged food product containing liquid

And the liquid temperature is measured at its current state

When the inspector measures the net quantity

Then the system should calculate the adjusted volume as if the liquid is at 20°C

And record the adjusted measurement value

And store the temperature used for adjustment”

Tip: The second scenario introduces a counter-argument with steps not pertinent to the input (e.g., the system displaying a warning message and preventing measurement recording) that could be avoided to improve relevance.

Also, the specification should cover intended functionalities that are plausible. For this regulatory requirement, measuring liquid net quantity at 20°C does not follow common practices. The reason is that liquids, conceivably, would never be at a specific temperature we want, i.e., liquids are not always temperature-controlled. In this case, it is more plausible that the liquid is volume adjusted as if it is at 20 degrees.

2. Details: Sometimes, details usually in the user story or background section and mostly as context, are provided which are not exact words of the given regulatory requirement, but they are pertinent and valid. These relevant details often help clarify the intent of the feature without compromising relevance. For instance, in a regulatory requirement that addresses the temperature measurement of food products having a background of “Given the inspection system is calibrated for liquid measurements And temperature monitoring capability is available” could be considered a bonus for the model. In these cases, you might want to write down your feedback about the specific details you found acceptable and relevant.
3. Single clear intention:

**Example:**

Regulatory Requirement: “Liquid Whole Egg, Dried Whole Egg and Frozen Whole Egg are the foods that meet the standard set out in section B.22.034 of the FDR and:

- (a) in the case of liquid whole egg and frozen whole egg, contain at least 23.5% egg solids by weight;
- (b) in the case of dried whole egg, contain not more than 5% water; and
- (c) contain not more than 50,000 viable bacteria per gram and 10 coliform organisms per gram.”

Original Specification:

“Feature: Egg Product Requirements

Scenario: Liquid Whole Egg requirements

Given a liquid whole egg product

When the product is analyzed

Then it must contain at least 23.5% egg solids by weight

And it must not contain more than 50,000 viable bacteria per gram

And it must not contain more than 10 coliform organisms per gram

Scenario: Frozen Whole Egg requirements

Given a frozen whole egg product

When the product is analyzed

Then it must contain at least 23.5% egg solids by weight

And it must not contain more than 50,000 viable bacteria per gram

And it must not contain more than 10 coliform organisms per gram

Scenario: Dried Whole Egg requirements

Given a dried whole egg product

When the product is analyzed

Then it must contain not more than 5% water

And it must not contain more than 50,000 viable bacteria per gram  
And it must not contain more than 10 coliform organisms per gram”

Improved Specification:

“Feature: Egg Product Requirements

Background:

Given an egg product ready for analysis

Rule: Egg products must meet the FDR standards for composition and bacteria limits

Scenario: Verify composition for Liquid and Frozen Whole Eggs

Given the product type is liquid whole egg or frozen whole egg

When the product is analyzed

Then it must contain at least 23.5% egg solids by weight

Scenario: Verify composition for Dried Whole Eggs

Given the product type is dried whole egg

When the product is analyzed

Then it must contain not more than 5% water

Scenario: Verify bacteria and coliform limits for all egg products

When the product is analyzed for bacteria content

Then it must not contain more than 50,000 viable bacteria per gram

And it must not contain more than 10 coliform organisms per gram”

Tip: The improved version separates different functions into distinct scenarios, maintaining distinct purposes for each scenario. This improves singularity.

Example:

Regulatory Requirement: “Heating requirements for item 3 a Liquid processed egg, Whole egg with no less than 24% and no more than 38% egg solids is Minimum temperature of the processed egg at the automatic diversion valve 61 (°C), Minimum temperature of the processed egg at the automatic diversion valve 142 (°F) and Minimum heating time 3.5 (minutes).

Heating requirements for item 4 b Liquid processed egg, Whole egg mix with less than 2% added salt or sweetening agent, or both is Minimum temperature of the processed egg at the automatic diversion valve 60 (°C), Minimum temperature of the processed egg at the automatic diversion valve 140 (°F) and Minimum heating time 6.2 (minutes).”

Original Specification:

“Feature: Heating Requirements for Liquid Processed Eggs

Scenario Outline: Verify heating requirements for liquid processed eggs

Given a liquid processed egg of type "<Egg Type>"

And the egg has "<Egg Solids>" egg solids

When it is heated at the automatic diversion valve

Then the minimum temperature should be "<Temperature (°C)>" degrees Celsius

And the minimum temperature should be "<Temperature (°F)>" degrees Fahrenheit

And the minimum heating time should be "<Heating Time>" minutes

Examples: | Egg Type | Egg Solids | Temperature (°C) | Temperature (°F) | Heating Time | | Whole egg | 24%-38% | 61 | 142 | 3.5 | | Whole egg mix (≤2% salt/sweetener) | N/A | 60 | 140 | 6.2 |”

Improved Specification:

“Feature: Heating Requirements for Liquid Processed Egg

The feature provides behavioral specifications for compliance with regulatory requirements on pasteurization/heating of various liquid processed egg mixtures.

@Requirement-3a

Scenario: Ensure heating requirements for Item 3(a) - Whole egg (24-38% solids)

Given the product is "Liquid processed egg, Whole egg

And the egg solid content is between 24% and 38%

When the system measures temperature at the automatic diversion valve

Then the minimum temperature shall be 61 °C (142 °F)

And the minimum heating time shall be 3.5 minutes

@Requirement-4b

Scenario: Ensure heating requirements for Item 4(b) - Whole egg mix (<2% salt/sweetener)

Given the product is "Liquid processed egg, Whole egg mix"

And less than 2% of salt or sweetening agent (or both) is added

When the system measures temperature at the automatic diversion valve

Then the minimum temperature shall be 60 °C (140 °F) And the minimum heating time shall be 6.2 minutes"

Tip: Splitting scenarios further enhances singularity and testability by isolating distinct functions.