DeepSeek

All items are rated on a **5-point Likert scale** unless otherwise indicated:
1 = Strongly Disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, 5 = Strongly Agree.
Section A. **Content Accuracy and Reliability**

1. The hazards identified by the system are **factually grounded in the scene description**. **5**

2. The analysis includes the **most critical hazards** relevant to the task and environment. **5**

3. The **severity and likelihood ratings** are reasonable and consistent with the hazards described. **3**

4. The analysis avoids mentioning **hazards not supported by the scene context** (hallucinations). **4**

5. The hazard list is **non-redundant** (few repeated or duplicate entries). **3**

Section B. **Explanation Quality**

6. Explanations are **specific to the scene context**, referencing concrete objects and spatial relations. **3**

7. Explanations provide a **causal or temporal account** (e.g., preconditions or sequences that could lead to the hazard). **3**

8. The outputs include **clear and actionable safeguards or instructions**. **4**

9. The hazard descriptions are **concise and free of irrelevant details**. **2**

10. The explanation structure makes it **easy to follow the reasoning process** of the system. **3**

Section C. **Trust and Usability**

11. I would find this hazard analysis **useful for supporting safety assessment** in assistive robotics. **3**

12. The outputs are **easy to interpret and understand** without further clarification. **2**

13. I would feel **confident relying on these results** in a real hazard analysis task. **3**

14. The system provides a **balanced level of detail**, neither overwhelming nor superficial. **2**

Section D. **Open Feedback**

(Free-text responses; supports qualitative analysis)

15. What did you find **most useful** about this hazard analysis output?

It recognised realistic hazards like tripping, burns, and obstacles around the wheelchair. The explanations made sense.

16. What did you find **least useful or problematic**?

The table: every item had its own line, and the text was packed tightly together. It felt more technical than practical, so reading it was tiring.

17. What **improvements** would make the outputs more reliable and usable for safety analysis?

Condense the table, group hazards by type, and give short summaries rather than full paragraphs.

Grok

All items are rated on a **5-point Likert scale** unless otherwise indicated:
1 = Strongly Disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, 5 = Strongly Agree.
Section A. **Content Accuracy and Reliability**

1. The hazards identified by the system are **factually grounded in the scene description**. **5**

2. The analysis includes the **most critical hazards** relevant to the task and environment. **5**

3. The **severity and likelihood ratings** are reasonable and consistent with the hazards described. **4**

4. The analysis avoids mentioning **hazards not supported by the scene context** (hallucinations). **4**

5. The hazard list is **non-redundant** (few repeated or duplicate entries). **4**

Section B. **Explanation Quality**

6. Explanations are **specific to the scene context**, referencing concrete objects and spatial relations. **4**

7. Explanations provide a **causal or temporal account** (e.g., preconditions or sequences that could lead to the hazard). **4**

8. The outputs include **clear and actionable safeguards or instructions**. **4**

9. The hazard descriptions are **concise and free of irrelevant details**. **3**

10. The explanation structure makes it **easy to follow the reasoning process** of the system. **3**

Section C. **Trust and Usability**

11. I would find this hazard analysis **useful for supporting safety assessment** in assistive robotics. **3**

12. The outputs are **easy to interpret and understand** without further clarification. **3**

13. I would feel **confident relying on these results** in a real hazard analysis task. **3**

14. The system provides a **balanced level of detail**, neither overwhelming nor superficial. **3**

Section D. **Open Feedback**

(Free-text responses; supports qualitative analysis)

15. What did you find **most useful** about this hazard analysis output?

Well organised, listed nearly everything relevant in the scene.

16. What did you find **least useful or problematic**?

he writing felt mechanical; lots of numbers and coordinate references that didn’t add meaning. After a few rows it started to feel repetitive.

17. What **improvements** would make the outputs more reliable and usable for safety analysis?

Keep the layout but trim unnecessary data. Add a short bullet summary for the top five risks.

GPT 5

All items are rated on a **5-point Likert scale** unless otherwise indicated:
1 = Strongly Disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, 5 = Strongly Agree.
Section A. **Content Accuracy and Reliability**

1. The hazards identified by the system are **factually grounded in the scene description**. **5**

2. The analysis includes the **most critical hazards** relevant to the task and environment. **5**

3. The **severity and likelihood ratings** are reasonable and consistent with the hazards described. **5**

4. The analysis avoids mentioning **hazards not supported by the scene context** (hallucinations). **3**

5. The hazard list is **non-redundant** (few repeated or duplicate entries). **4**

Section B. **Explanation Quality**

6. Explanations are **specific to the scene context**, referencing concrete objects and spatial relations. **5**

7. Explanations provide a **causal or temporal account** (e.g., preconditions or sequences that could lead to the hazard). **4**

8. The outputs include **clear and actionable safeguards or instructions**. **5**

9. The hazard descriptions are **concise and free of irrelevant details**. **4**

10. The explanation structure makes it **easy to follow the reasoning process** of the system. **5**

Section C. **Trust and Usability**

11. I would find this hazard analysis **useful for supporting safety assessment** in assistive robotics. **4**

12. The outputs are **easy to interpret and understand** without further clarification. **4**

13. I would feel **confident relying on these results** in a real hazard analysis task. **4**

14. The system provides a **balanced level of detail**, neither overwhelming nor superficial. **4**

Section D. **Open Feedback**

(Free-text responses; supports qualitative analysis)

15. What did you find **most useful** about this hazard analysis output?

This was the easiest to read (except the table). The explanations linked hazards to visible objects without many unnecessary data.

16. What did you find **least useful or problematic**?

Messy table. Some hazards were repeated, and the spacing made it difficult to see where one row ended.

17. What **improvements** would make the outputs more reliable and usable for safety analysis?

Fix the table. (Content is good)