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. **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. **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**. **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?

It identified sensible hazards for the domestic setting – like the hot coffee, clutter, and wheelchair positioning – and gave good explanations of what could happen and why. The reasoning about the frail subject was particularly thoughtful.

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

The table is too long.

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

Improve the table; use clearer spacing or headings.

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?

It captured nearly every relevant hazard, including clutter and heat sources, and explained each one with clear ratings. The column structure looked tidy and consistent, which made scanning easier than DeepSeek's.

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

Too much of technical details; makes the user lost.

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

Simplify the wording.

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. **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**. **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?

The explanations were easiest to understand.

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

Messy table. Very confusing and hard to understand.

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

Fix the table.