Appendix of Supplementary Findings

1 Reliability Analysis of Results using Cohen's K

In the following, we present the results of the reliability analysis for all four case studies.

Case 1: Industrial Disassembly Robot (IDR)

Table 1 presents inter-rater reliability results obtained using Cohen's K measure. The results reveal unanimous agreement between participants on both *agree* and *disagree* decisions. Therefore, no false positives or false negatives were observed.

Table 1: IDR case study results for inter-rater agreement analysis with Cohen's K

		Rater 2	
		Agree	Disagree
Rater 1	Agree	126	0
	Disagree	0	58

Case 2: Warehouse Robotic Swarm (WRS)

Table 2 shows inter-rater reliability results calculated using Cohen's K measure. The results indicate a unanimous agreement between both participants on *agree* and *disagree*. Hence, there are no false positives or false negatives in this case.

Table 2: WRS case study results for inter-rater agreement analysis with Cohen's K

		Rater 2	
		Agree	Disagree
Rater 1	Agree	200	0
	Disagree	0	20

Case 3: Prolonged Hull of an Autonomous Vessel (PHAV)

Table 3 presents inter-rater reliability results obtained using Cohen's K measure. The results indicate unanimous agreement between participants on both *agree* and *disagree* choices. Thus, no false positives or false negatives were observed.

Table 3: PHAV case study results for inter-rater agreement analysis with Cohen's K

		Rater 2	
		Agree	Disagree
Rater 1	Agree	218	0
	Disagree	0	41

Case 4: Human-Robotic Interaction (HRI)

Table 4 shows inter-rater reliability results obtained using Cohen's K measure. The results demonstrate that both participants unanimously *agreed* and *disagreed*. Consequently, no false positives or false negatives exist.

Table 4: HRI case study results for inter-rater agreement analysis with Cohen's K

		Rater 2	
		Agree	Disagree
Rater 1	Agree	160	0
	Disagree	0	56

2 Results for Confidence Intervals with Bootstrap Method

In the following, we present the results of non-parametric confidence intervals calculated using the bootstrap method for all four case studies.

Case 1: Industrial Disassembly Robot (IDR)

The results in Table 5 provide 95% bootstrap confidence intervals (CI) for each agreement category across the evaluated models. The **Strongly Disagree** category shows a relatively narrow confidence interval (2.40–3.40), indicating consistent responses across models. Similarly, the **Agree** and **Strongly Agree** categories show higher and stable agreement levels, with confidence intervals of 5.60–6.40 and 6.00–7.00, respectively. The **Neutral** category has the smallest confidence interval (0.00–0.40), reflecting minimal neutral responses across the models, while **Disagree** exhibits a narrow range (2.80–3.00), suggesting low disagreement levels. Overall, the results highlight that most models are rated positively, with strong agreement dominating the responses.

Table 5: Confidence Intervals for IDR Results

Category	**Lower Bound (95% CI)**	**Upper Bound (95% CI)**
Strongly Disagree	2.40	3.40
Disagree	2.80	3.00
Neutral	0.00	0.40
Agree	5.60	6.40
Strongly Agree	6.00	7.00

Case 2: Warehouse Robotic Swarm (WRS)

Table 6 presents 95% bootstrap confidence intervals for the mean responses across five Likert-scale categories in the WRS results. The **Strongly Disagree** and **Disagree** categories show narrow confidence intervals ([0.00, 0.00] and [2.00, 2.00], respectively), indicating consistent responses across the models with minimal variation. In contrast, the **Neutral** category has a slightly broader interval ([1.10, 1.90]), reflecting moderate variability in responses. The **Agree** and **Strongly Agree** categories exhibit higher average values with intervals of [7.60, 10.20] and [10.10, 11.90], respectively, highlighting a strong positive agreement among the models. These results suggest that the majority of the responses are skewed toward agreement, with minimal disagreement or neutrality.

Table 6: Confidence Intervals for WRS Results

Category	**Lower Bound (95% CI)**	**Upper Bound (95% CI)**
Strongly Disagree	0.00	0.00
Disagree	2.00	2.00
Neutral	1.10	1.90
Agree	7.60	10.20
Strongly Agree	10.10	11.90

Case 3: Prolonged Hull of an Autonomous Vessel (PHAV)

Table 7 shows 95% bootstrap confidence intervals for the five Likert-scale categories in the PHAV results. The **Strongly Disagree** and **Strongly Agree** categories have narrow confidence intervals ([1.00, 1.00] and [0.00, 0.00], respectively), indicating highly consistent responses across models with minimal variation. The **Agree** category exhibits the widest interval ([20.20, 23.20]), reflecting a strong positive agreement but with moderate variability in the responses. Meanwhile, the **Neutral** category has a relatively small interval ([0.70, 1.10]), suggesting limited neutral responses across the models. Overall, the results indicate that most responses are skewed towards agreement, with little disagreement or neutrality observed.

Table 7: Confidence Intervals for PHAV Results

Category	**Lower Bound (95% CI)**	**Upper Bound (95% CI)**
Strongly Disagree	1.00	1.00
Disagree	3.00	3.10
Neutral	0.70	1.10
Agree	20.20	23.20
Strongly Agree	0.00	0.00

Case 4: Human-Robotic Interaction (HRI)

Table 8 presents 95% bootstrap confidence intervals for the five Likert-scale categories in the HRI results. The **Strongly Disagree** category shows a very narrow confidence interval ([1.00, 1.00]), indicating consistent responses across all models. The **Disagree** and **Neutral** categories exhibit slightly broader intervals ([4.30, 5.30] and [1.30, 2.00], respectively), reflecting moderate variability in responses. The **Agree** category shows a wider interval ([4.40, 6.60]), suggesting variability in agreement levels across models. Finally, the **Strongly Agree** category has a high interval ([9.60, 11.20]), indicating strong agreement among the models with some variability. Overall, the results highlight consistent disagreement and strong agreement trends across most models.

Table 8: Confidence Intervals for HRI Results

Category	**Lower Bound (95% CI)**	**Upper Bound (95% CI)**
Strongly Disagree	1.00	1.00
Disagree	4.30	5.30
Neutral	1.30	2.00
Agree	4.40	6.60
Strongly Agree	9.60	11.20