APPENDIX A
USER STUDY QUESTIONS

In Table IV, we provide the specific questions and statements used to assess participant responses across various metrics in our user study. This table categorizes the survey items under the dependent variables they are intended to measure, including *Trust*, *Workload*, *Satisfaction*, *Clarity*, and *Actionability*. Each item is accompanied by a corresponding 7-point Likert scale.

TABLE IV
OVERVIEW OF QUESTIONS AND STATEMENTS ASSOCIATED WITH EACH DEPENDENT VARIABLE.

Dependent Variables	User Study Questions/Statements	7-point Scale
Satisfaction	I am satisfied with the answer to my question provided by the robot.	1-Low, 7-High
Clarity	I found the robot's answer regarding its behavior to be clear.	1-Low, 7-High
Actionability	I understand how to make this robot perform this task in the way I wanted.	1-Low, 7-High
Trust	To what extent can the robot's behavior be predicted from moment to moment?	1-Low, 7-High
	To what extent can you count on the robot to do its job?	1-Low, 7-High
	What degree of faith do you have that the robot will be able to cope with similar situations in the future?	1-Low, 7-High
	Overall, how much do you trust the robot?	1-Low, 7-High
Workload	How much mental and perceptual activity was required? Was the task easy or demanding, simple or complex, exacting or forgiving?	1-Low, 7-High
	How much physical activity was required? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?	1-Low, 7-High
	How much time pressure did you feel due to the pace at which tasks occurred? Was the pace slow and leisurely or rapid and frantic?	1-Low, 7-High
	How successful were you in accomplishing the goals of the task? How satisfied were you with your performance?	1-Good, 7-Poor
	How hard did you have to work to accomplish your level of performance?	1-Low, 7-High
	How did you feel during the task?	1-Low, 7-High

APPENDIX B
DEMOGRAPHICS BY CONDITIONS

Table V presents detailed breakdown of the demographic distribution of participants according to the response condition they were assigned to: Excuse-only, Explanation-only, and ARE. Each condition column, denoted by 'n=', represents the total number of participants in that specific group, providing a basis for the subsequent percentage calculations within each demographic category. These categories encompass Gender, Age, Education of participants, and whether participants have Computer Science (CS) Background or not, with their respective percentages summing up to 100% for each condition. The 'Overall' column, marked as 'N=90', aggregates the data across all conditions, offering a comprehensive view of the entire study's demographic landscape.

TABLE V
PARTICIPANT DEMOGRAPHIC DISTRIBUTION BY CONDITIONS

Demographics	Categories	Excuse (n=29)	Explanation (n=31)	ARE (n=30)	Overall (N=90)
Gender	Woman	27.59%	48.39%	46.67%	41.11%
	Man	65.52%	48.39%	50%	54.44%
	Non-binary	6.90%	3.23%	3.33%	4.44%
Age	18 -24	13.79%	16.13%	16.67%	15.56%
	25 -34	41.38%	41.94%	40%	41.11%
	35 - 44	17.24%	25.81%	10%	17.78%
	45 - 54	13.79%	9.68%	23.33%	15.56%
	55 - 64	10.34%	6.45%	10%	8.89%
	65+	3.45%	0%	0%	1.11%
Education	High school	34.48%	32.26%	33.33%	33.33%
	College	51.72%	48.39%	53.33%	51.11%
	Graduate	13.79%	19.35%	13.33%	15.56%
CS Background	No	82.76%	90.32%	83.33%	85.56%
	Yes	17.24%	9.68%	16.67%	14.44%

APPENDIX C
DESCRIPTIVE STATISTICS FOR DEPENDENT VARIABLES BY CONDITIONS

In Table VI, we present the mean and standard deviation (SD) values for various dependent variables under three different conditions: Excuse-only, Explanation-only, and ARE.

TABLE VI
MEAN AND STD FOR DEPENDENT VARIABLES BY CONDITIONS

Dependent Variable	Excuse	Explanation	ARE
Satisfaction	0.302 ± 0.330	0.312 ± 0.311	0.414 ± 0.362
Clarity	0.460 ± 0.342	0.538 ± 0.324	0.678 ± 0.339
Actionability	0.474 ± 0.336	0.336 ± 0.301	0.653 ± 0.325
Trust	0.270 ± 0.236	0.220 ± 0.201	0.271 ± 0.212
Workload	0.321 ± 0.109	0.272 ± 0.107	0.288 ± 0.118

APPENDIX D
TWO-TAILED T-TEST AND TWO ONE-SIDED T-TESTS (TOST) RESULTS

In Table VII, we present detailed statistical analyses to assess the equivalence of trust and workload across conditions.

TABLE VII
STATISTICAL ANALYSES INCLUDING THE TWO-TAILED T-TEST RESULTS (DISPLAYED IN THE 'STATISTIC' COLUMN UNDER 'T-TEST') AND THE TWO ONE-SIDED T-TESTS (TOST) RESULTS (PRESENTED IN THE 'STATISTIC' COLUMN, UNDER 'UPPER BOUND' AND 'LOWER BOUND' SECTIONS). P-VALUES MARKED WITH "*" DENOTE SIGNIFICANCE AT $p < 0.05$

Dependent Variable	Statistic	Excuse vs Explanation			Excuse vs Action			Excuse vs Explanation		
		t	df	p	t	df	p	t	df	p
Trust	T-Test	0.880	58	0.382	0.012	57	0.990	0.953	59	0.345
	Upper bound	1.074	58	0.144	0.204	57	0.419	1.148	59	0.128
	Lower bound	0.686	58	0.752	-0.180	57	0.429	0.758	59	0.774
Workload	T-Test	1.761	58	0.083	-1.109	57	0.272	0.563	59	0.576
	Upper bound	1.955	58	0.028*	-0.917	57	0.819	0.758	59	0.226
	Lower bound	1.568	58	0.939	-1.301	57	0.099	0.367	59	0.643

APPENDIX E
PRE-AGGREGATION RESULTS FOR DOMAIN-PROBLEM INSTANCE PAIRS

Table VIII and Table IX provide detailed breakdowns of the aggregated results presented in Table I and Table II, respectively.

TABLE VIII

COMPARISON OF PERFORMANCE METRICS ACROSS THREE APPROACHES: EXCUSE, EXPLANATION, AND ARE. FOR EACH DOMAIN AND PROBLEM INSTANCE PAIRS. $|\pi^R|$ AND $|\pi^H|$ REPRESENT THE AVERAGE LENGTHS OF ROBOT AND HUMAN PLANS, RESPECTIVELY. $|\zeta|$, $|\mathcal{E}|$, AND $|\zeta + \mathcal{E}|$ DENOTE THE LENGTHS OF THE EXCUSE, EXPLANATION, AND ARE. THE 'TIME(S)' COLUMN INDICATES THE COMPUTATION TIME (IN SECONDS). SUMMARIZED DATA CAN BE FOUND IN TABLE I

Domains	Problem	$\|\pi^R\|$	$\|\pi^H\|$	Excuse		Explanation		ARE	
				$\|\zeta\|$	Time(s)	$\|\mathcal{E}\|$	Time(s)	$\|\zeta + \mathcal{E}\|$	Time(s)
Logistics	p1	20	10	1	0.14	2	0.46	3	0.99
Logistics	p2	19	11	2	0.19	2	0.48	4	1.02
Logistics	p3	36	17	4	11.83	2	1.62	6	27.66
Logistics	p4	36	16	3	3.11	2	2.74	5	12.41
Logistics	p5	44	21	4	10.6	2	105.1	6	38.85
Depots	p1	10	4	2	2.71	4	3.60	6	4.71
Depots	p2	15	6	3	0.69	4	4.32	7	5.34
Depots	p3	27	12	5	2.47	4	96.94	9	110.42
Depots	p4	16	7	3	13.78	3	3.52	6	24.67
Depots	p5	27	12	5	2.68	4	576.68	9	620.11
Freecell	p1	8	5	1	0.24	2	3.33	3	3.04
Freecell	p2	14	6	2	1.72	3	26.04	5	34.13
Freecell	p3	8	4	2	1.02	3	7.08	5	8.73
Freecell	p4	21	7	6	78.01	3	49.2	9	173.37
Freecell	p5	9	3	2	1.09	3	7.12	5	8.87
Rovers	p1	10	5	2	0.55	2	0.49	4	1.47
Rovers	p2	14	8	3	3.52	2	0.52	5	4.53
Rovers	p3	11	7	2	0.51	2	0.49	4	1.18
Rovers	p4	8	6	2	1.34	2	0.47	4	2.55
Rovers	p5	22	13	4	27.1	2	5.06	6	30.73
Satellite	p1	17	16	1	0.16	1	0.54	2	1.27
Satellite	p2	15	13	2	1.06	2	2.48	4	5.98
Satellite	p3	20	17	4	7.40	3	41.31	7	62.1
Satellite	p4	15	12	3	33.14	2	13.71	5	64.53
Satellite	p5	14	13	3	4.23	1	1.49	4	4.67

TABLE IX

COMPARISON OF PERFORMANCE METRICS ACROSS THREE APPROACHES: EXCUSE, EXPLANATION, AND ARE. $|\pi^R|$ AND $|\pi^H|$ REPRESENT THE AVERAGE LENGTHS OF ROBOT AND HUMAN PLANS, RESPECTIVELY. $|\zeta|$, $|\mathcal{E}|$, AND $|\zeta + \mathcal{E}|$ DENOTE THE LENGTHS OF THE EXCUSE, EXPLANATION, AND ARE. PLEASE NOTE THAT THE LENGTH OF ARE IS GREATER THAN THE TOTAL LENGTH OF EXCUSE AND EXPLANATION WHEN COMPUTED INDIVIDUALLY. THE 'TIME(S)' COLUMN INDICATES THE COMPUTATION TIME (IN SECONDS). SUMMARIZED DATA CAN BE FOUND IN TABLE II

Domains	Problem	$\|\pi^R\|$	$\|\pi^H\|$	Excuse		Explanation		ARE	
				$\|\zeta\|$	Time(s)	$\|\mathcal{E}\|$	Time(s)	$\|\zeta + \mathcal{E}\|$	Time(s)
Blocks	p1	10	8	2	0.89	1	0.35	4	4.00
Blocks	p2	16	10	4	10.22	1	0.36	6	9.55
Blocks	p3	12	10	3	1.70	1	0.35	5	5.00
Blocks	p4	10	8	2	0.27	1	0.33	3	1.38
Blocks	p5	12	8	3	1.56	1	0.34	5	6.27
Zenotravel	p1	11	9	2	0.20	1	0.29	4	2.02
Zenotravel	p2	14	12	2	1.20	1	7.18	4	14.39
Zenotravel	p3	10	9	2	0.19	1	0.29	4	1.94
Zenotravel	p4	15	14	2	2.69	1	32.08	4	44.76
Zenotravel	p5	18	16	2	4.34	1	27.43	4	60.78
Logistics	p1	20	8	2	0.40	2	0.93	5	8.79
Logistics	p2	19	8	2	0.47	2	0.88	5	8.59
Logistics	p3	36	14	2	0.72	2	3.06	5	39.72
Logistics	p4	36	14	2	0.48	2	7.05	5	100.39
Logistics	p5	44	21	4	15.12	2	97.64	6	36.47

APPENDIX F
LIMITATION

Like all research, our study is subject to certain limitations. One of them lies in our capacity to interpret the findings. Despite our concerted efforts to engage in thorough discussions among co-authors and to draw insights from related studies in the field, the complexity of data interpretation remains a challenge.

Additionally, while we endeavored to include a broad spectrum of participants, achieving a perfectly representative sample is inherently difficult. There remains the possibility that not all participant groups were adequately represented in our study. Consequently, we advocate for future research to prioritize inclusivity and diversity among participant groups to enhance the generalizability of findings.

Our investigation was also confined to a limited selection of IPC domains, which may not encompass the full range of potential applications. Expanding the scope to include a wider variety of domains could provide a more comprehensive evaluation of the phenomena under study.

In our human subject experiments, we did not ask users to make the changes required to achieve the expected plans. Instead, we used a metric to assess the perceived actionability of our method. We did not require a separate test to measure actual actionability since our algorithm guarantees the soundness of the excuses generated. We also did not include additional experiments for explainability metrics, such as simulability, as the utility of model reconciliation as an explanation framework has been extensively studied (cf. [16]). However, future studies could explore the metrics like simulability and the ability of the users to carry out the changes recommended by the method.

Lastly, the nature of our study does not allow for the assessment of long-term implications associated with the types of information we examined. To gain a deeper understanding of the enduring effects, conducting longitudinal studies in real-world settings would be invaluable. Such research could offer significant insights into how these information types influence interactions and perceptions over extended periods.