Accuracy and Satisfaction in Multi-Turn LLM Dialogues for NFR Assessment

University of Maine · University of Notre Dame

Ali Pourghasemi Fatideh
Wilder Baldwin
Maria Dhakal
Collin McMillan
Sepideh Ghanavati

SIGDIAL 2026 · Emory University, Atlanta

LLM assistants are now standard tools for software developers

1. But our benchmarks measure functional correctness

They don’t evaluate NFRs:

- which are vague (no unit test for “ensure the integrity of ePHI”)
- a property of the whole system rather than one function
- and often need domain knowledge like a regulation

2. And they evaluate only a single turn prompt

They don’t evaluate the multi-turn interaction: how developers actually use these tools

Related Work in Evaluating LLMs

Work	MT	SM	Evaluation target
Chen et al. (2021)	–	–	Code generation
Hendrycks et al. (2021)	–	–	Code generation
Yin et al. (2018)	–	–	Code generation
Jimenez et al. (2023)	–	–	Bug fixing
Jain et al. (2024)	–	–	Code generation
Aider (2024)	–	–	Code editing
Wu and Fard (2025)	–	–	Clarification ability
Laban et al. (2025)	✓	–	MT degradation
Vaithilingam et al. (2022)	✓	–	User expectations
Kumar et al. (2025)	✓	–	Agent collaboration
Liu et al. (2024b)	✓	–	Code quality
Sarkar et al. (2022)	✓	–	User experience
Liang et al. (2024)	✓	–	Usability
This work	✓	✓	**NFR assessment**

MT = Multi-turn · SM = Satisfaction Modeling

Research Questions

RQ1 How accurately do LLM-based dialogue assistants assess NFRs, and how do users perceive the response quality?

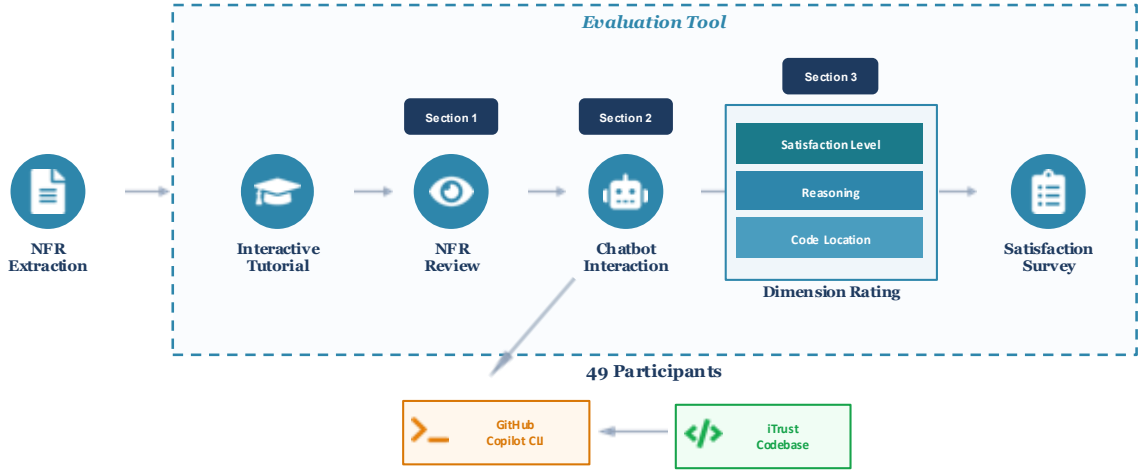
RQ2 What dialogue characteristics significantly correlate with user satisfaction?

Methodology

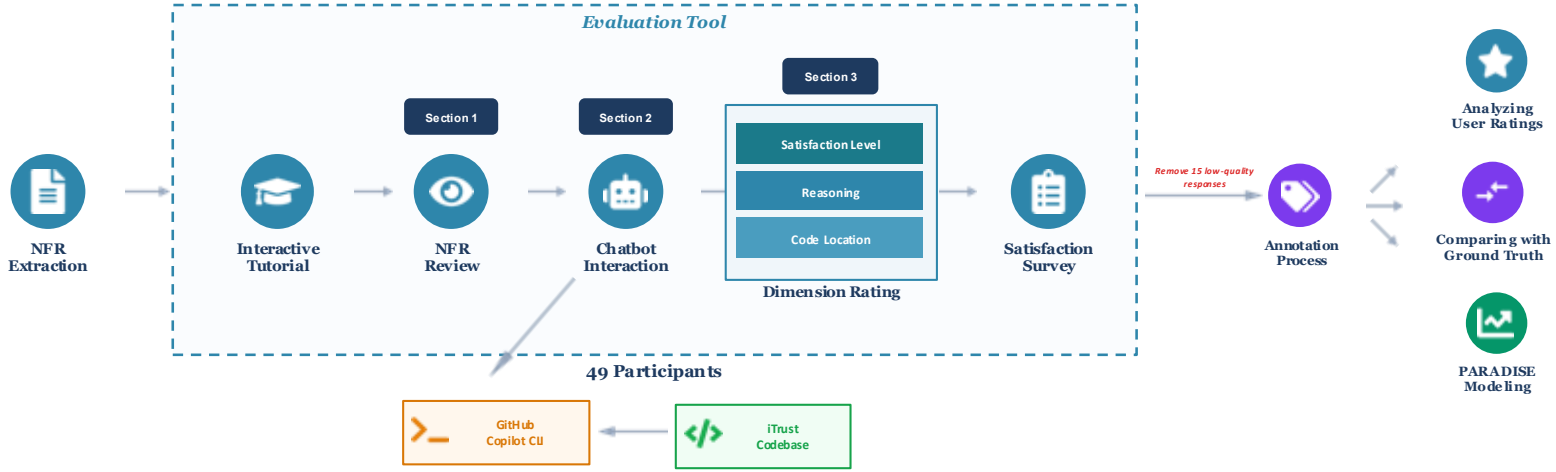
Study Design

NFR Extraction

Study Design



Study Design



Building the NFR Corpus

140 NFRs

extracted from HIPAA by two privacy & SE experts

HIPAA

HIPAA is a U.S. federal regulation that establishes standards for protecting the privacy and security of individually identifiable health information, known as electronic protected health information (ePHI). We use HIPAA as a representative domain of regulatory NFRs.

e.g., "The system shall ensure the integrity of ePHI that it creates."

Codebase & Ground Truth

iTrust: an EHR system built as a testbed for HIPAA security & privacy requirements

WHAT

Satisfaction level

Satisfied / Weakly Satisfied /
Weakly Denied / Denied

Codebase & Ground Truth

iTrust: an EHR system built as a testbed for HIPAA security & privacy requirements

WHAT	WHY
Satisfaction level Satisfied / Weakly Satisfied / Weakly Denied / Denied	**Reasoning** justification for the level

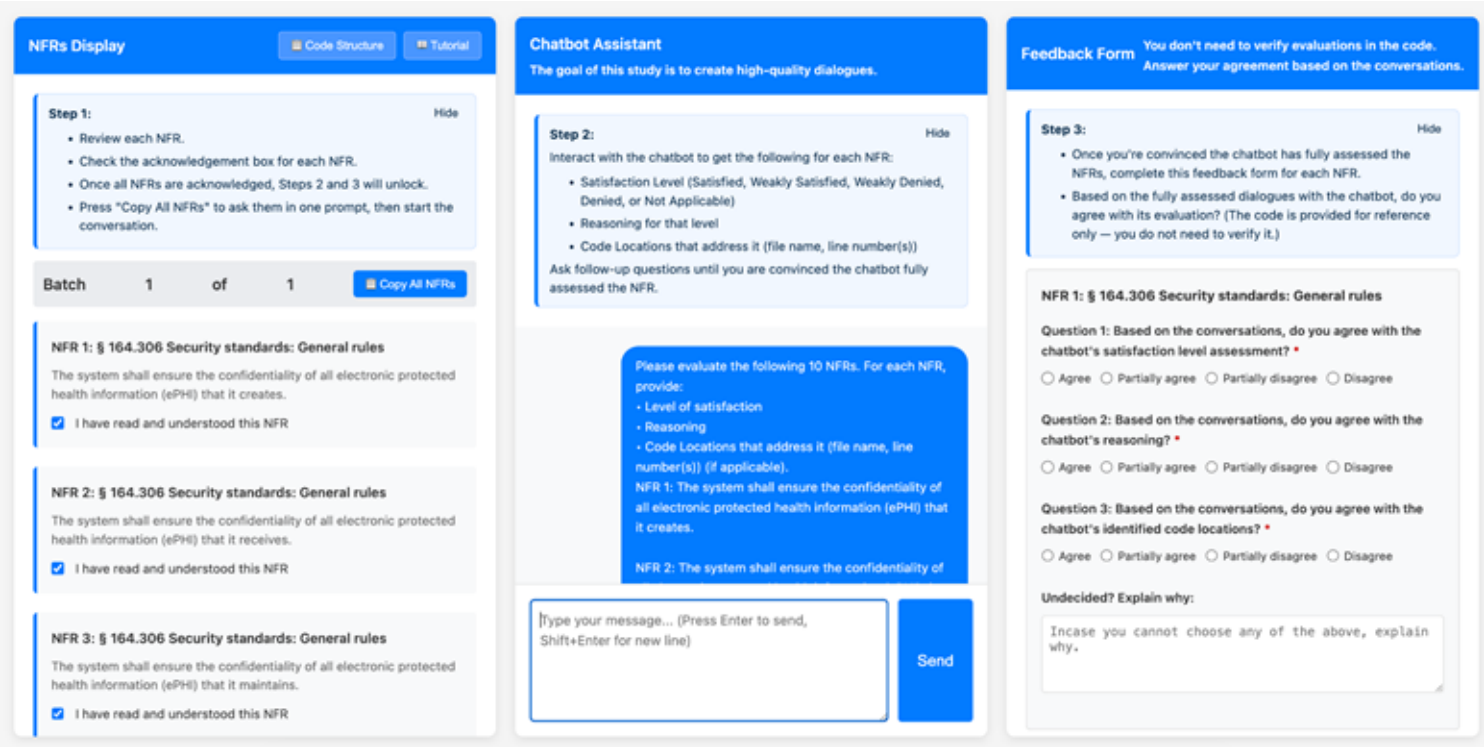
Codebase & Ground Truth

iTrust: an EHR system built as a testbed for HIPAA security & privacy requirements

WHAT	WHY	WHERE
Satisfaction level	**Reasoning**	**Code location**
Satisfied / Weakly Satisfied / Weakly Denied / Denied	Justification for the level	File names and line numbers

Two experts · independent review · disagreements discussed to consensus

The Study Tool



Turn Annotation

Turn labels drawn from PARADISE + MT-Eval + MT-Bench-101

Label	Count
Given Data	370
Follow-up	195
Recollections	156
Expansion	152
Refinement	53
User Initiatives	37
Proactive Interaction	36
Content Rephrasing	29
Check Move (CM)	27
Self-affirmation	23
Self-correction	16

Label	Count
Separate Input	11
Instruction Clarification	10
Anaphora Resolution	9
Unsuitable Request	3
Abandoned Requests	2
Help Messages	2
Format Rephrasing	2
Inappropriate Responses	1
Error	1
Content Confusion	0
Relevant Data	0

Results

RQ1: Accuracy vs. Perceived Quality

	Actual accuracy (vs. ground truth)		User Rating
Satisfaction level	**0.381** F1	*vs.*	**3.27/4**

RQ1: Accuracy vs. Perceived Quality

	Actual accuracy (vs. ground truth)		User Rating
Satisfaction level	**0.381** F1	*vs.*	**3.27/4**
Reasoning	**0.520** BERTScore F1	*vs.*	**3.31/4**

RQ1: Accuracy vs. Perceived Quality

	Actual accuracy (vs. ground truth)		User Rating
Satisfaction level	**0.381** F1	*vs.*	**3.27/4**
Reasoning	**0.520** BERTScore F1	*vs.*	**3.31/4**
Code location	**0.203** F1	*vs.*	**3.34/4**

Low accuracy on every row, yet user rating stays significantly above “agree” (t-test, $p < .001$)

91-94%

developer rating

F1 = 0.203 – 0.520

actual accuracy

Why It Happens: One Case

THE NFR

“The system shall implement procedures for creating passwords.”

Why It Happens: One Case

THE NFR

“The system shall implement procedures for creating passwords.”

THE SYSTEM SAID

Cites ResetPasswordAction.java, conflating password reset with password creation

Why It Happens: One Case

THE NFR

"The system shall implement procedures for creating passwords."

THE SYSTEM SAID

Cites ResetPasswordAction.java: conflating password reset with password creation

THE PARTICIPANT SAID

Strongly Agree

Semantically adjacent. Functionally distinct. Plausible, and wrong.

RQ2: Modeling User Satisfaction: The PARADISE Framework

Premise: a dialogue system's primary objective is to maximize user satisfaction.

$$\text{Performance} = \alpha \cdot N(\kappa) - \Sigma\, w_i \cdot N(c_i)$$

κ = task success (from a confusion matrix) · c_i = dialogue costs · w_i = learned weights · N = Z-score normalization

Dialogue costs	Dialogue interaction patterns
e,g,. number of turns, elapsed time, mean words per response	Number of follow up turns, number of refinement turns

RQ2: Modeling User Satisfaction

PARADISE · 21 candidate predictors · 32 dialogues · backward elimination → 3 survivors

RQ2: Modeling User Satisfaction

PARADISE · 21 candidate predictors · 32 dialogues · backward elimination → 3 survivors

$$\hat{N}(US) = -0.583 \cdot N(\text{Mean Words per Response})$$

Verbosity hurts

RQ2: Modeling User Satisfaction

PARADISE · 21 candidate predictors · 32 dialogues · backward elimination → 3 survivors

$$\hat{N}(US) = -0.583 \cdot N(\text{Mean Words per Response}) \quad \textit{Verbosity hurts}$$

$$-0.485 \cdot N(\text{Given Data}) \quad \textit{Information-dumping hurts}$$

RQ2: Modeling User Satisfaction

PARADISE · 21 candidate predictors · 32 dialogues · backward elimination → 3 survivors

$$\hat{N}(US) = -0.583 \cdot N(\text{Mean Words per Response}) \quad \textit{Verbosity hurts}$$

$$-0.485 \cdot N(\text{Given Data}) \quad \textit{Information-dumping hurts}$$

$$+0.422 \cdot N(\text{Proactive Interaction}) \quad \textit{Taking initiative helps}$$

Not in the model: task success. Whether the system was right did not predict satisfaction.

Summary

1. 91–94% agreement, F1 of 0.203–0.520 against expert ground truth

Summary

1. 91–94% agreement, F1 of 0.203–0.520 against expert ground truth
2. Satisfaction is driven by number of given data turns, mean word per turn and proactive interaction not by being correct

Summary

1. 91–94% agreement, F1 of 0.203–0.520 against expert ground truth
2. Satisfaction is driven by number of given data turns, mean word per turn and proactive interaction not by being correct
3. Corpus + annotated dialogues open-source: github.com/PERC-Lab/DialogueNFR

Thanks!

Repository Link

Any questions?

You can contact me at

- ali.pourghasemi@maine.edu

Paper Link

References

- Amyot, D., et al. “Evaluating goal models within the goal-oriented requirement language.” IJIS, 2010.
- Bai, G., et al. “MT-Bench-101: A fine-grained benchmark for evaluating LLMs in multi-turn dialogues.” ACL 2024.
- Burke, M. J., & Dunlap, W. P. “Estimating interrater agreement with the average deviation index.” ORM, 2002.
- Hajdinjak, M., & Mihelič, F. “The PARADISE evaluation framework: Issues and findings.” Comput. Ling., 2006.
- Kwan, W.-C., et al. “MT-Eval: A multi-turn capabilities evaluation benchmark for LLMs.” EMNLP 2024.
- Laban, P., et al. “LLMs get lost in multi-turn conversation.” arXiv:2505.06120, 2025.
- Meneely, A., Smith, B., & Williams, L. “iTrust electronic health care system case study.” 2012.
- Walker, M., et al. “PARADISE: A framework for evaluating spoken dialogue agents.” ACL 1997.
- Zhang, T., et al. “BERTScore: Evaluating text generation with BERT.” arXiv:1904.09675, 2019.

Building the NFR Corpus

650

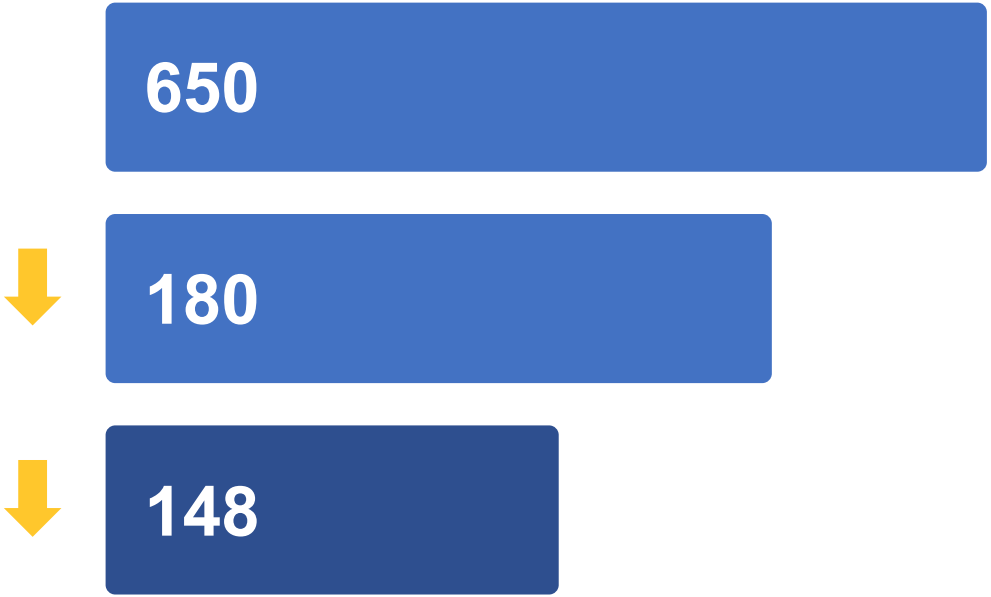
statements extracted from HIPAA by a privacy & SE expert

Building the NFR Corpus

statements extracted from HIPAA by a privacy & SE expert

after two filters: SWEBOK software-construction scope + within iTrust scope

Building the NFR Corpus



statements extracted from HIPAA by a privacy & SE expert

after two filters: SWEBOK software-construction scope + within iTrust scope

after second-expert review

e.g., "The system shall ensure the integrity of ePHI that it creates."

Participants

1. 49 programmers recruited: 8 pilot + 41 main study

Participants

1. 49 programmers recruited: 8 pilot + 41 main study
2. Screened: CS/Math degree, software engineering role, 95–100% approval

Participants

1. 49 programmers recruited: 8 pilot + 41 main study
2. Screened: CS/Math degree, software engineering role, 95–100% approval
3. 34 dialogues after attention-check and quality filtering (406 turns)

1 Turn labels drawn from PARADISE + MT-Eval + MT-Bench-101

Label	Count
Follow-up	195
Refinement	53
User Initiatives	37
Proactive Interaction	36
Content Rephrasing	29
Self-affirmation	23
…	…

The PARADISE Framework

Premise: a dialogue system's primary objective is to maximize user satisfaction.

$$\text{Performance} = \alpha \cdot N(\kappa) - \Sigma\, w_i \cdot N(c_i)$$

κ = task success (from a confusion matrix) · c_i = dialogue costs · w_i = learned weights · N = Z-score normalization

Efficiency costs

how much effort the task took

e.g., number of turns, elapsed time

Quality costs

what shapes the user's perception

e.g., response delay, errors, unsuitable responses

Walker et al. "PARADISE: A general framework for evaluating spoken dialogue agents." ACL 1997

Deriving the Performance Function

Fit multivariate linear regression: user satisfaction ~ task success + dialogue costs

1

Normalize Z-score every variable so the weights are directly comparable ($|w| \leq 1$)

Deriving the Performance Function

Fit multivariate linear regression: user satisfaction ~ task success + dialogue costs

1. **Normalize** Z-score every variable so the weights are directly comparable ($|w| \leq 1$)

2. **De-correlate** Drop predictors with $|r| > 0.7$ — keep the more significant one of each pair

Walker et al. ACL 1997 · Hajdinjak & Mihelič. "The PARADISE evaluation framework: Issues and findings." Computational Linguistics 2006

Deriving the Performance Function

Fit multivariate linear regression: user satisfaction ~ task success + dialogue costs

1. **Normalize** Z-score every variable so the weights are directly comparable ($|w| \leq 1$)
2. **De-correlate** Drop predictors with $|r| > 0.7$ — keep the more significant one of each pair
3. **Backward elimination** Iteratively remove predictors that are not significant ($p > 0.05$)

The surviving weights — and their signs — reveal what drives user satisfaction.

Walker et al. ACL 1997 · Hajdinjak & Mihelič. "The PARADISE evaluation framework: Issues and findings." Computational Linguistics 2006

Figure 1 — Methodology Overview

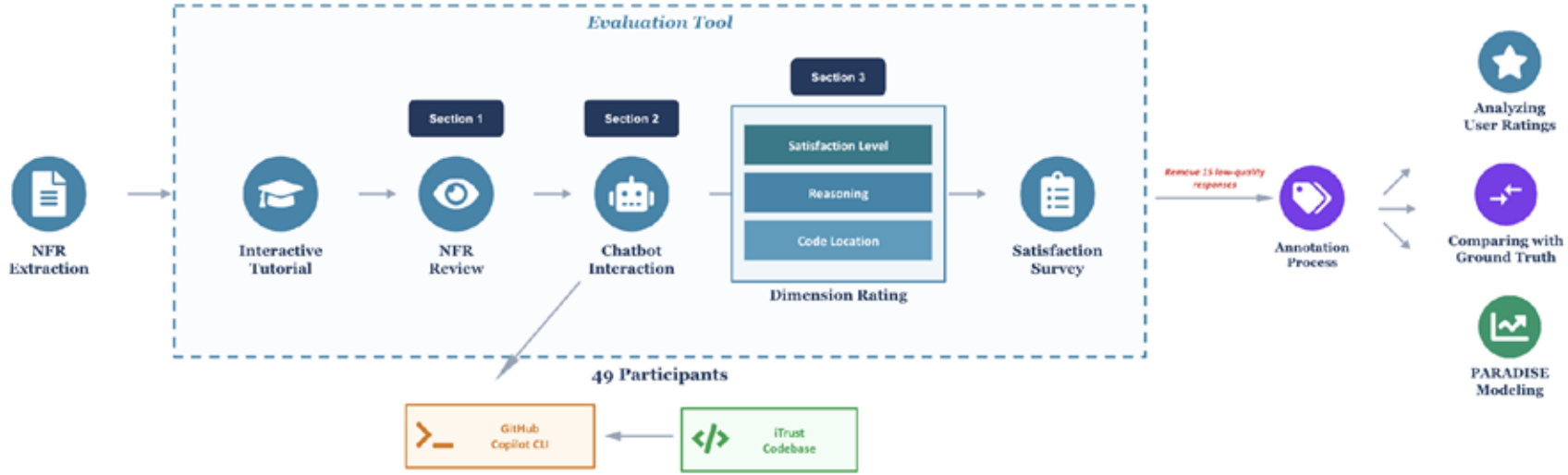


Figure 2 — Survey Tool Interface

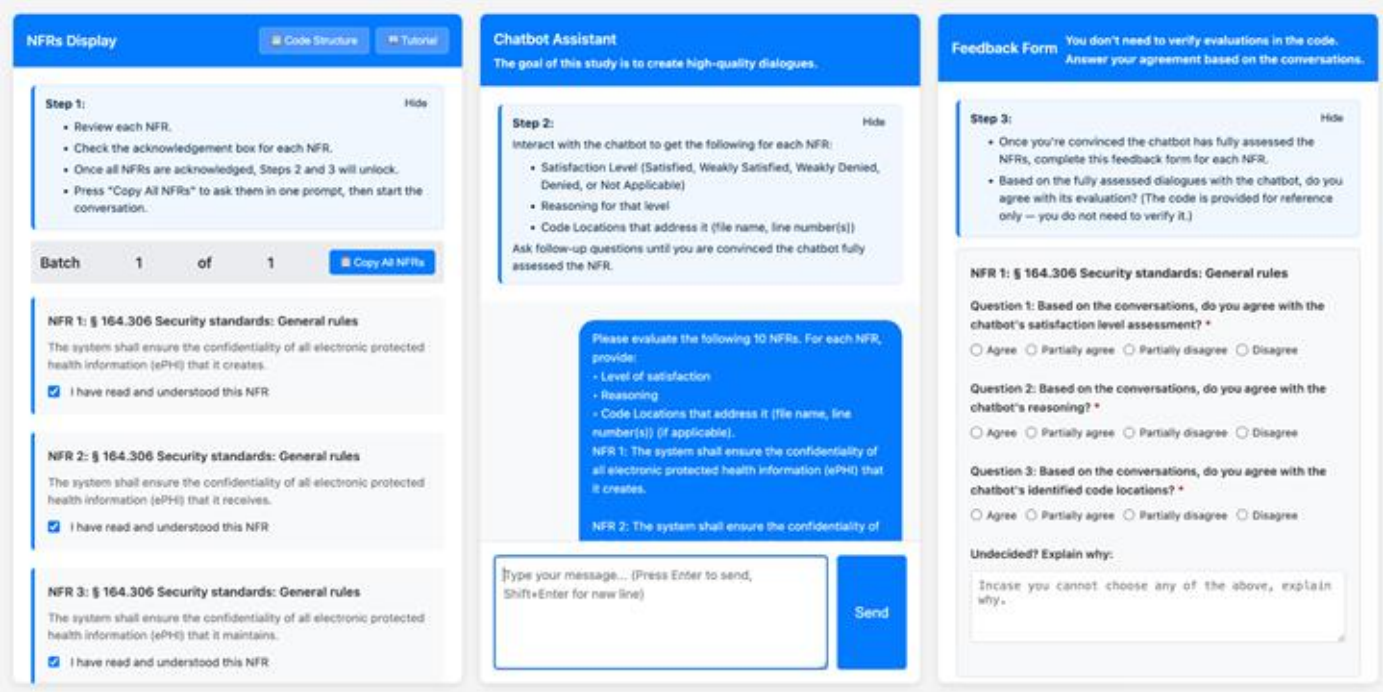


Table 1 — Related Work in Evaluating LLMs

Work	MT	SM	Evaluation target
Chen et al. (2021)	–	–	Code generation
Hendrycks et al. (2021)	–	–	Code generation
Yin et al. (2018)	–	–	Code generation
Jimenez et al. (2023)	–	–	Bug fixing
Jain et al. (2024)	–	–	Code generation
Aider (2024)	–	–	Code editing
Wu and Fard (2025)	–	–	Clarification ability
Laban et al. (2025)	✓	–	MT degradation
Vaithilingam et al. (2022)	✓	–	User expectations
Kumar et al. (2025)	✓	–	Agent collaboration
Liu et al. (2024b)	✓	–	Code quality
Sarkar et al. (2022)	✓	–	User experience
Liang et al. (2024)	✓	–	Usability
This work	✓	✓	**NFR assessment**

MT = Multi-turn · SM = Satisfaction Modeling

Table 3 — Key Statistics of Collected Dialogues

Statistic	Value
Total # Dialogues	34
Total # Turns	406
Avg. # Turns per Dialogue	11.94
Avg. # Words in Prompt	60.90
Max. # Words in Prompt	496
Avg. # Words in Response	129.46
Max. # Words in Response	877
Avg. # Words per Turn	190.36
Max. # Words per Turn	1206

Table 2 — Inter-Rater Reliability (Cohen's κ)

Label	R1	R1-redo	R2
Recollections	0.000	0.640	0.772
Expansion	0.129	0.602	0.666
Refinement	0.176	0.729	0.615
Follow-up	0.289	0.796	0.884
User initiatives	0.668	0.898	1.000
Unsuitable-request	–	–	–
Inappropriate resp.	–	–	–
Error	–	–	–
Help messages	0.000	–	–
Given-data	0.408	1.000	–
Check move	0.369	0.638	0.791

Label	R1	R1-redo	R2
Relevant data	–	–	–
Abandoned requests	0.000	1.000	–
Anaphora Resolution	0.000	1.000	1.000
Separate Input	–	–	–
Content Confusion	–	–	–
Content Rephrasing	0.370	0.912	0.813
Format Rephrasing	0.000	–	–
Self-correction	0.558	0.811	–
Self-affirmation	0.728	0.878	0.769
Instruction Clarif.	0.479	0.790	–
Proactive Interaction	1.000	1.000	–

Size: 50 / 50 / 54 **Average κ: 0.323 / 0.836 / 0.812**

– = insufficient variation to compute κ

Figure 3 — Participant Agreement Distribution (RQ1)

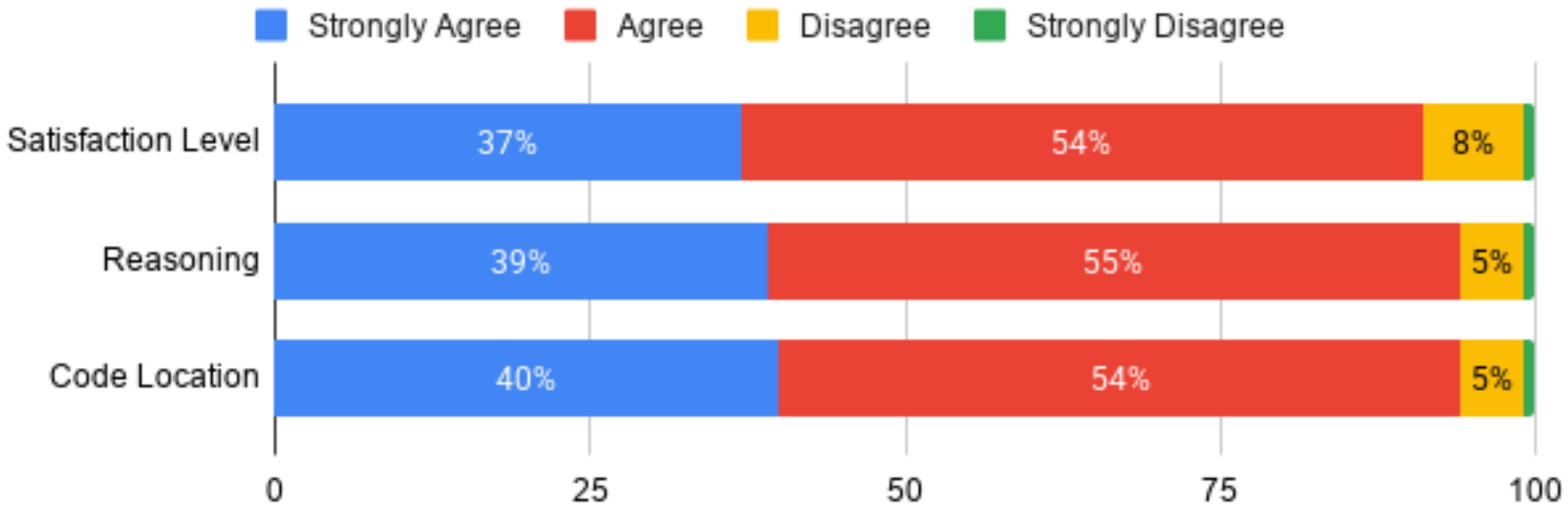


Table 4 — Accuracy of LLM vs. Ground Truth (RQ1)

Dimension	Metric	Score
Satisfaction Level	Precision	0.438
	Recall	0.418
	F1	**0.381**
Reasoning	BERTScore Precision	0.521
	BERTScore Recall	0.529
	BERTScore F1	**0.520**
	Cosine Similarity	0.774
	ROUGE-1	0.204
	ROUGE-2	0.037
	ROUGE-L	0.156
Code Location	Precision	0.151
	Recall	0.142
	F1	**0.203**

Table 8 — One-Sample t-Tests for Agreement (RQ1)

Dimension	Mean	SD	t	p
Satisfaction Level	3.27	0.66	7.49	< .001
Reasoning	3.31	0.63	9.05	< .001
Code Location	3.34	0.60	10.41	< .001

H_0: $\mu \leq 3$ (the “agree” threshold)

Table 5 — Mean Agreement & Rating Divergence (RQ1)

Dimension	Mean Agr.	High AD
Satisfaction Level	3.27	19
Reasoning	3.31	13
Code Location	3.34	15

Of the 148 NFRs: 6 diverged on all three dimensions, 6 on two. Single-dimension divergences: code location 9, satisfaction level 7, reasoning 1.

High AD = # NFRs with Average-Deviation index > 0.67 (lack of consensus)

Figure 4 — Participants' PARADISE Ratings (RQ2)

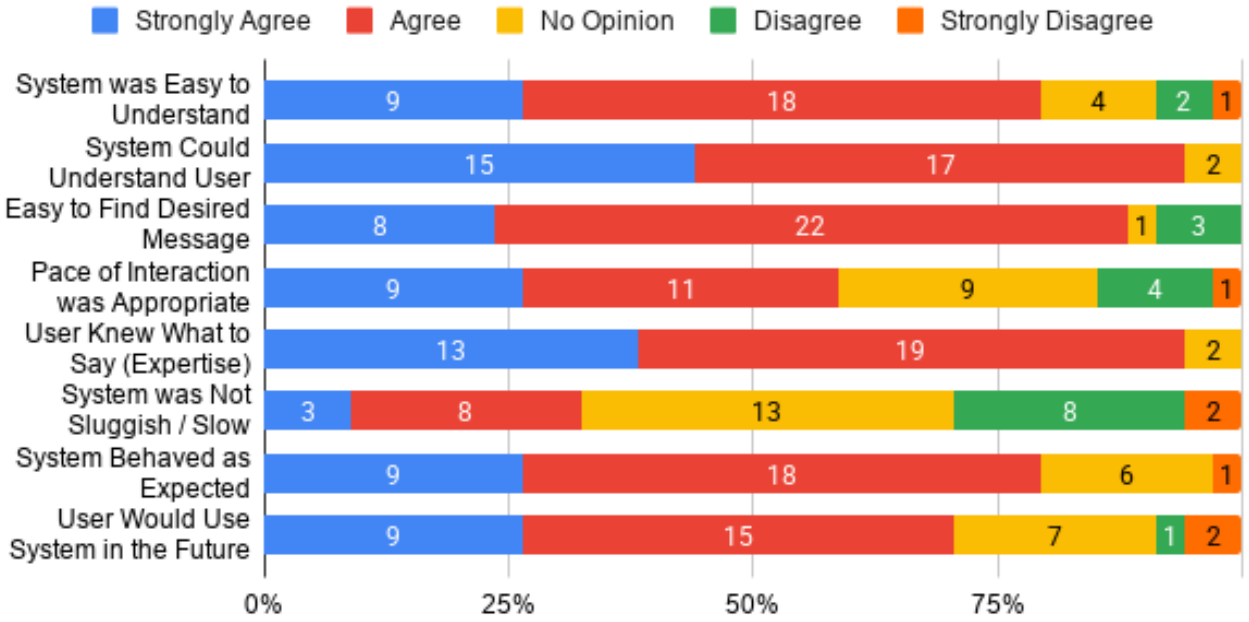


Table 6 — Total Counts of Dialogue-Act Labels (RQ2)

Label	Count
Given Data	370
Follow-up	195
Recollections	156
Expansion	152
Refinement	53
User Initiatives	37
Proactive Interaction	36
Content Rephrasing	29
Check Move (CM)	27
Self-affirmation	23
Self-correction	16

Label	Count
Separate Input	11
Instruction Clarification	10
Anaphora Resolution	9
Unsuitable Request	3
Abandoned Requests	2
Help Messages	2
Format Rephrasing	2
Inappropriate Responses	1
Error	1
Content Confusion	0
Relevant Data	0

Table 11 — Descriptive Statistics for All Predictors (RQ2)

Variable	Mean	SD	Min	Max
User Satisfaction (US)	31.26	3.65	20.00	37.00
Number of Turns	11.94	7.95	2.00	27.00
Mean Response Time (s)	45.80	33.05	12.59	157.77
Elapsed Time (s)	3496.4	9230.8	312.9	55369
Mean Words / Request	60.90	42.13	9.41	229.50
Mean Words / Response	129.46	105.49	50.91	548.00
User Initiatives	1.09	0.38	1.00	3.00
Unsuitable Requests	0.09	0.29	0.00	1.00
Inappropriate Responses	0.03	0.17	0.00	1.00
Errors	0.03	0.17	0.00	1.00
Help Messages	0.06	0.24	0.00	1.00
Given data	10.88	7.47	2.00	26.00
Check Moves	0.79	1.09	0.00	4.00
Relevant Data	0.00	0.00	0.00	0.00
Abandoned Requests	0.06	0.24	0.00	1.00
Unsuitable Request Ratio	0.009	0.036	0.000	0.200
Inappropriate Resp. Ratio	0.004	0.021	0.000	0.125

Variable	Mean	SD	Min	Max
Help Message Ratio	0.003	0.014	0.000	0.063
Given data Ratio	0.907	0.121	0.583	1.000
Check Move Ratio	0.064	0.105	0.000	0.500
Relevant Data Ratio	0.000	0.000	0.000	0.000
Abandoned Request Ratio	0.002	0.009	0.000	0.040
Recollections	4.59	3.64	0.00	11.00
Expansion	4.47	4.01	0.00	12.00
Refinement	1.56	2.27	0.00	10.00
Follow-up	5.74	5.48	0.00	18.00
Anaphora Resolution	0.26	0.57	0.00	2.00
Content Confusion	0.00	0.00	0.00	0.00
Self-correction	0.47	1.11	0.00	5.00
Self-affirmation	0.68	1.39	0.00	5.00
Instruction Clarification	0.29	0.52	0.00	2.00
Proactive Interaction	1.06	2.17	0.00	12.00
Task Success	0.95	0.26	0.55	1.80
Task Completion	0.90	0.20	0.10	1.00

Table 9 — Predictors Removed for High Correlation (RQ2)

Removed	Retained	r
Errors	Inappropriate Resp.	1.000
Inappropriate Resp. Ratio	Inappropriate Resp.	1.000
Abandoned Requests	Abandoned Req. Ratio	0.999
Help Message Ratio	Help Messages	0.996
Number of Turns	**Given data**	**0.981**
Expansion	Given data	0.867
Follow-up	Given data	0.860
Recollections	Given data	0.853
Unsuitable Req. Ratio	Unsuitable Requests	0.806

|r| > 0.7; for each pair the predictor with the higher p-value against US was removed

Table 10 — Backward-Elimination Sequence (RQ2)

#	Removed predictor	p	R^2	#	Removed predictor	p	R^2
1	Mean Words Request	.853	.728	10	Inappropriate Responses	.484	.705
2	Given data Ratio	.802	.727	11	Anaphora Resolution	.446	.697
3	Mean Response Time	.865	.726	12	User Initiatives	.138	.688
4	Task Completion	.803	.725	13	Elapsed Time	.074	.652
5	Instruction Clarification	.725	.724	14	Refinement	.111	.597
6	Self-correction	.681	.721	15	Help Messages	.150	.549
7	Check Move Ratio	.652	.718	16	Self-affirmation	.157	.507
8	Check Moves	.688	.714	17	Abandoned Request Ratio	.166	.465
9	Unsuitable Requests	.534	.711	18	Task Success	.134	.423

Highest-p predictor removed at each step; stopped when all remaining were significant ($p < 0.05$)

Table 7 — PARADISE Performance-Function Coefficients (RQ2)

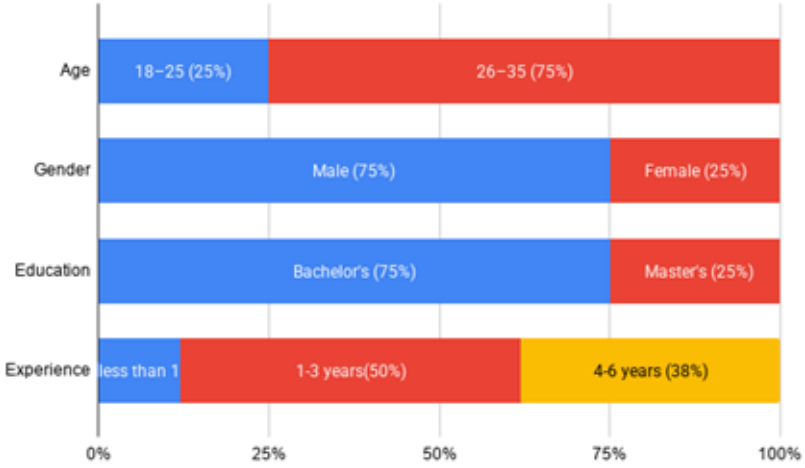
Predictor	β	SE	t	p
Mean Words per Response	**−0.583**	0.174	−3.35	.002
Given data	**−0.485**	0.162	−2.99	.006
Proactive Interaction	**0.422**	0.163	2.60	.015

$$\hat{N}(US) = -0.583 \cdot N(\text{Mean Words/Response}) - 0.485 \cdot N(\text{Given Data}) + 0.422 \cdot N(\text{Proactive Interaction})$$

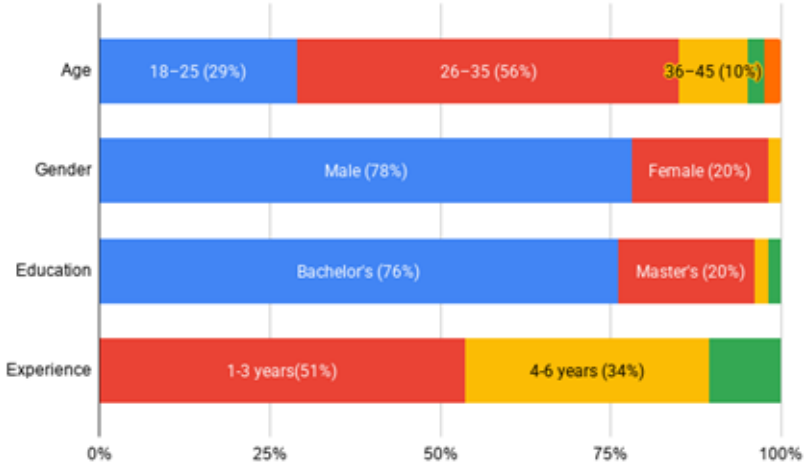
β: standardized coefficient · SE: standard error · t: test statistic

Figure 5 — Participant Demographics

Pilot study (n = 8)



Main study (n = 41)



Be concise. Be proactive. Don’t dump information.

What causes high satisfaction and a Warning

Be concise. Be proactive. Don’t dump information.

Do not optimize an agent against this function.

It would not produce a more accurate system, it would produce a system whose inaccurate responses are more convincing, which may cause the gap we report.