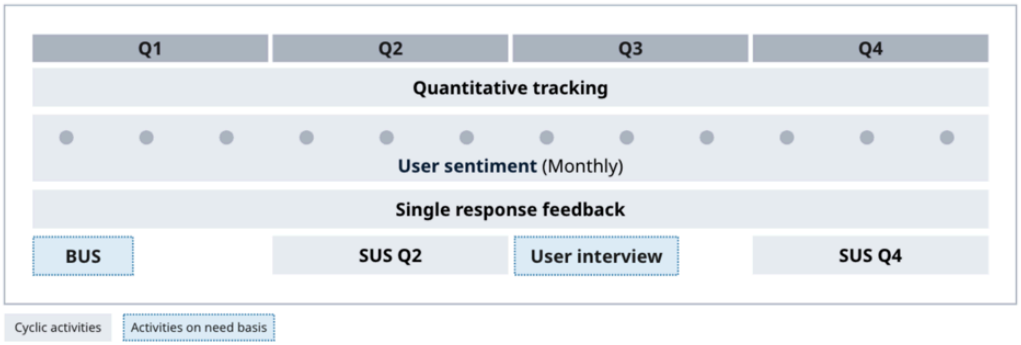
Conversational AI evaluation

For conversational AI we recommend different methods to measure the user experience, and user behaviour. It is not mandatory to use all but it is not recommended to add more or different methods then outlined in this chapter. This is to create a coherent way of measuring, establish baselines and benchmarks for our products.

Read the guidelines for conversational AI components here

Content in this section

This illustration shows an example of what a full year of data collection could look like for an application with conversational AI.



Quantitative user behavior data

Collecting quantitative data is essential for understanding how users engage with conversational AI and for identifying friction early. It helps us make informed design decisions, measure impact, and continuously improve the experience.

Below are key metrics to track usage and engagement. For general user behavior metrics, refer to the general guidelines here

Engagement & conversation metrics

- **Monthly active users:** Adoption baseline.

- **New vs. returning users:** Adoption vs. Retention.
- **Number of session & conversations:** Volume over time, and number of conversations pr session.
- **Average conversation length**: Number of messages back and fourth.
- **Drop-off Rate:** % of sessions ending prematurely (E.g. before AI responses)
- **Top 10 queries:** Most frequently asked questions/intents.
- **Clarification loops:** Frequency of follow-up/clarifying questions.
- **Links CTR:** Click-through rate when the bot suggests links (not sources).

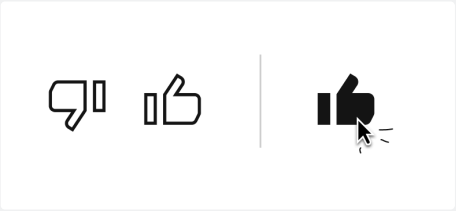
Bot performance & data quality

- **Fallback rate:** % of queries the bot couldn't answer (fallbacks/escalations)
- **Response latency:** Average reply speed (in seconds).
- **Error rate:** Failed API calls, broken flows, or system errors.

Single response feedback (Thumbs up/down)

Purpose

For conversation AI you can use the thumbs up/down to assess the effectiveness of a single interaction between the user and the AI—typically one message. It's a simple, low-effort way to capture sentiment. It's focused on individual exchanges, offering granular insight into each response.



Definition

👍Thumbs up: Positive sentiment and successful interaction.

👎Thumbs down: Flag for review.

💬Comments (Optional): Use for qualitative analysis and training improvements.

Guidelines

- Display thumbs up/down immediately after each AI reply.

- Feedback is optional
- Provide visual confirmation after feedback is submitted (Solid color).
- Allow the user to undo their choice, and deselect their feedback selection
- Prompt users to add a comment after a thumbs down, but let it be optional

Examples of predefined questions to ask for additional feedback

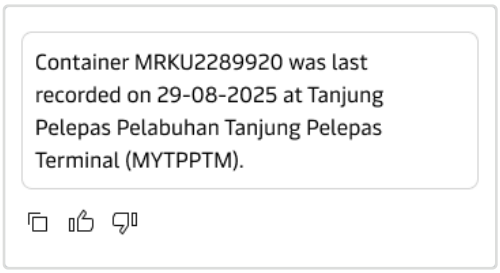
- The response was inaccurate or incomplete
- It lacked enough explanation or detail
- The wording was unclear or confusing
- It included a broken link or outdated information
- It wasn't clear what I should do next
- The response took too long
- Other (please describe)

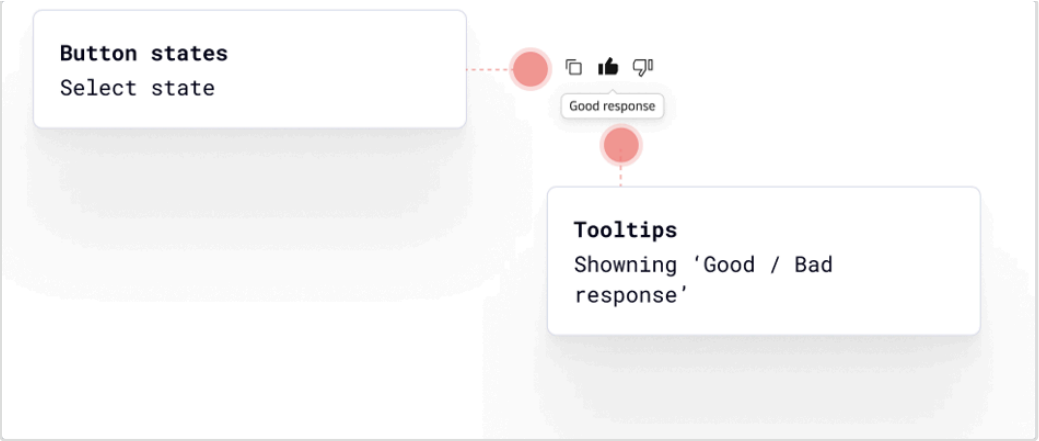
How to use it

- For conversational AI, it's useful in testing with SMEs and test groups. Enables lightweight, in-app feedback instead of using spreadsheets or surveys.
- Feedback should automatically be captured and used to improve LLM models.
- For open questions, make sure that a review takes place and the input are used (if its not used or analysed by the team, then remove the comment feature in the UI)
- Sharing improvements with testers builds trust and encourages continued involvement.

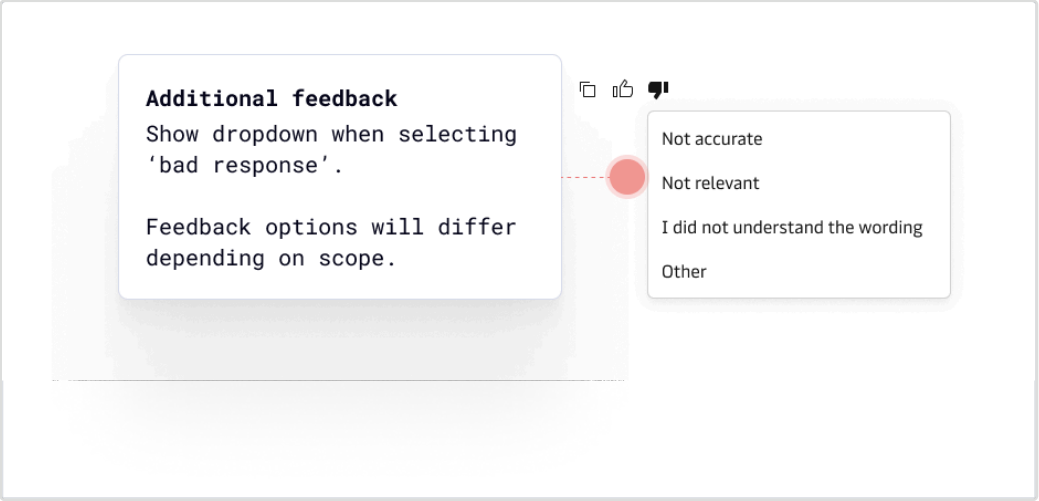
Design specification

Figma design file



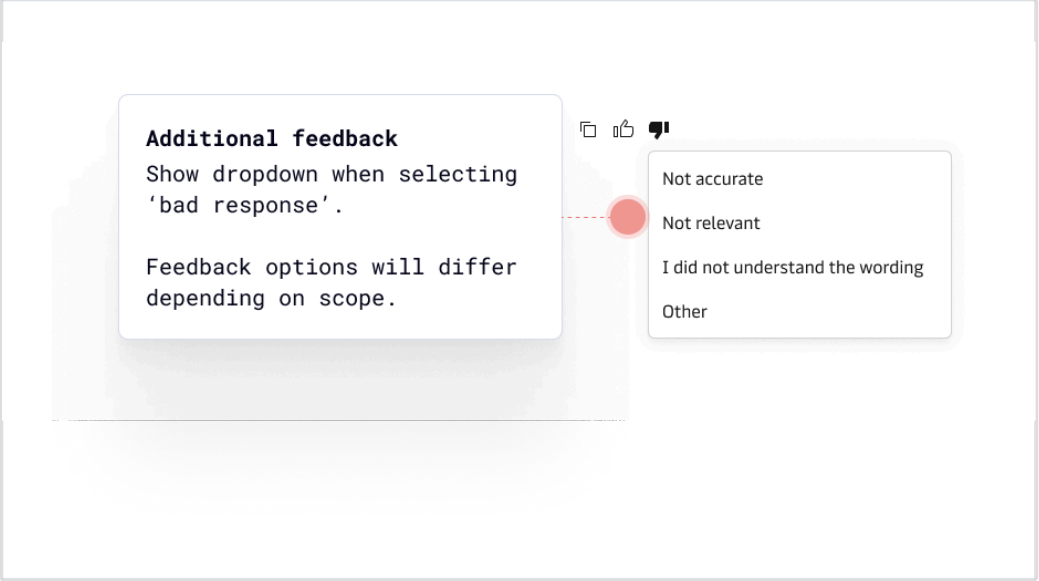


Interactions

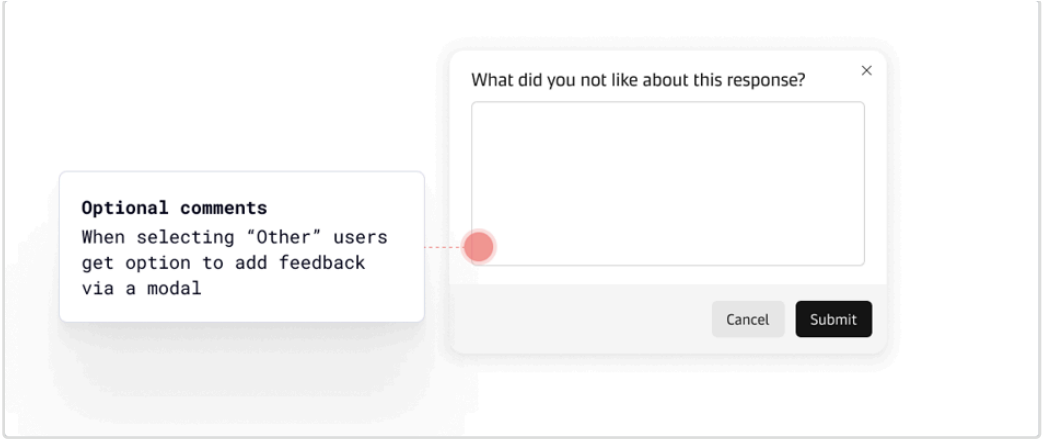


Predefined options (max 5)

Optional feedback on “Other” - two ways of showing



Comment in dropdown



Comments in modal

Umux-lite: Perceived usability & usefulness

Why we use UMUX-Lite

UMUX-Lite (Usability Metric for User EXperience-Lite) is a lightweight usability metric that enables us to continuously track user experience across new AI initiatives—without the overhead of larger surveys like SUS. It’s fast to deploy, easy to interpret, and ideal for capturing early signals on usefulness and ease of use.

By integrating UMUX-Lite into our feedback loops, we can:

- Monitor UX quality over time
- Identify friction early in development

Definition

Two fixed questions on a 5-point scale, 1 = Strongly disagree → 5 = Strongly agree

> **“[Application name]´s functionality meet my needs”**
>
> **“[Application name] is easy to use”**

How to use it

- Embed the two feedback questions directly into the chat interface—after a conversation or task is completed.
- To keep the experience lightweight, show them no more than once per month per user.
- Avoid redirecting users to a separate survey; keeping feedback in-flow ensures higher response rates and a smoother experience.

- Use conversational phrasing when embedding: *How much do you agree with the following?"*
- Optional: Follow-up with a free-text. Keep in mind that the goal is to keep feedback quick and lightweight for users. Only include it if there's a clear need for deeper insights.

Analysing the results

Once you've collected UMUX-Lite responses from users, you'll want to turn those into a score that's easy to interpret and share.

1. Add the two item scores together.
   Example: 4 + 4 = 8
2. Subtract 2 (to adjust for the minimum possible score).
   This shifts the scale so the lowest possible score (1 + 1 = 2) becomes 0.
   → 8 − 2 = 6
3. Multiply the result by 12.5 to scale it to a 0–100 range.
   → 6 × 12.5 = 75

This gives you a UMUX-Lite score of 75 out of 100.

Interpreting the score

UMUX-Lite scores reflect perceived usability. While there's no universal threshold for what counts as "good" or "bad," the most reliable way to interpret results is through benchmarking against your own previous scores.

Track how scores evolve over time or across releases. A rising score suggests improvements in usability, while a drop may signal new friction or unmet expectations. This approach helps you stay focused on trends and progress rather than isolated numbers.

Sources

- UMUX-Lite paper
- The original UMUX paper
- Simplifying wording

BUS-11: Conversational usability rating

Why use BUS-11

The ChatBot Usability Scale (BUS-11) helps us measure the usability of chatbot interactions in a structured way. Unlike quick sentiment checks, BUS-11 provides a detailed view across four factors: accessibility, conversation quality, privacy, and responsiveness.

By using BUS-11, we can:

- Identify weak points in conversation design
- Track improvements across releases
- Support decisions on automation and UX enhancements
- Benchmark chatbot usability against industry standards

Definition

BUS-11 is a validated questionnaire with **11 items**, grouped into four factors:

- Accessibility – How easy it is to find and use the chatbot
- Functional Interactive Conversation – How well the chatbot understands context and provides helpful responses
- Privacy – Whether the chatbot informs users about privacy concerns
- Responsiveness – How quickly the chatbot replies

Each item is rated on a **5-point scale** (1 = Strongly Disagree → 5 = Strongly Agree).

BUS-11 Items

Factor	Item
1. **Accessibility**	1. The chatbot function was easily detectable
	2. It was easy to find the chatbot
2. **Functional Interactive Conversation**	3. Communicating with the chatbot was clear
	4. The chatbot was able to keep track of context
	5. The chatbot's responses were easy to understand
	6. I find that the chatbot understands what I want and helps me achieve my goal
	7. The chatbot gives me the appropriate amount of information
	8. The chatbot only gives me the information I need

	9. I feel like the chatbot's responses were accurate
3. **Privacy**	10. I believe the chatbot informs me of any possible privacy issues
4. **Responsiveness**	11. My waiting time for a response from the chatbot was short

How to use it

Step 1: Setting up the survey

- Create a survey in InsightsHub or copy the BUS-11 template
- Add the 11 items with a 5-point Likert scale
- Include optional demographic or open-ended questions if needed

Ask users:
"How much do you agree with the following statements about your experience with the chatbot?"
(1 = Strongly Disagree, 5 = Strongly Agree)

Step 2: Collecting and cleaning data

- Aim for at least 40 responses or 30% of your user base
- Remove duplicates, invalid patterns (e.g., all identical ratings), incomplete entries, and users who never interacted with the chatbot

Step 3: Calculate and interpreting the scores

Overall BUS-11 score

Calculate

- Add all 11 item ratings for each respondent and divide by 11. This gives the overall BUS-11 score.

Convert

- To make the score easier to interpret, convert it to a percentage:

```
Percentage Score = (Overall Score ÷ 5) × 100
```

Example: If the overall score is 4.2 →

```
(4.2 ÷ 5) × 100 = 84%
```

Interpret

Percentage	Interpretation
< 67%	Very Poor
67–71%	Poor
71–74%	Average
74–80.1%	Good
≥ 80.1%	Very Good

Factor scores

Calculate

- Accessibility = Average of Items 1–2
- Functional Interactive Conversation = Average of Items 3–9
- Privacy = Item 10
- Responsiveness = Item 11

Convert

- Factor scores stay on the original 1–5 scale (no percentage conversion).

Interpret

Each factor has its own interpretation range:

Factor	Very Poor	Poor	Average	Good	Very Good
Accessibility	0–3.50	>3.50–3.70	>3.70–3.92	>3.92–4.19	>4.19–5
Functional Interactive Conversation	0–3.32	>3.32–3.53	>3.53–3.68	>3.68–4.05	>4.05–5
Privacy	0–2.52	>2.52–2.63	>2.63–2.77	>2.77–3.18	>3.18–5
Responsiveness	0–4.03	>4.03–4.23	>4.23–4.38	>4.38–4.58	>4.58–5

Sources

BUS-11