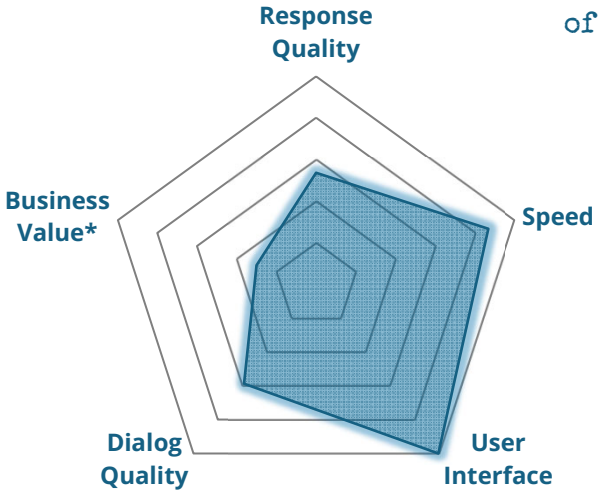
senses.ai

Car Manufacturer Chatbot Audit

reviewed on 17 February 2026

Overall Score

3.4 of 5



The chatbot performs strongly in visual design and response speed, which lifts its overall rating to 3.4. However, it underperforms in delivering relevant, goal-oriented, and actionable responses (score: 2.7). It also struggles to establish a meaningful dialog. The bot operates reactively rather than proactively. It frequently redirects users to external resources or support, even when it could resolve the issue directly. User value remains limited, constraining business impact. Translations are inconsistently implemented. Further, hallucination control mechanisms are insufficient.

Summary

Element	Description	Weight	Rating
Response Quality	Does the bot provide relevant and correct answers?	50%	**2.7**
Speed	Are response times adequate?	20%	**4.3**
User Interface	Is the user interface clear and intuitive?	15%	**5.0**
Dialog Quality	Is the dialog flow natural and intuitive?	15%	**2.9**
Overall Score Car Manufacturer			**3.4**

Additional Elements (excluded from overall rating)		
Business Value*	Does the chatbot contribute to business success?	**1.5**
Compliance*	Does the bot meet basic compliance requirements?	**3.0**
Ethics*	Does the bot meet basic ethical standards?	**3.0**
Translation	Does the bot provide adequate translations?	**2.9**

*Preliminary assessments. Our Level 2 audit enables open-box diagnosis and targeted optimization. Please contact us.

Recommendation Summary

The overall score of 3.4 indicates acceptable baseline performance but reveals structural limitations in intent recognition, response orchestration, and hallucination safeguards. The primary performance constraint lies in response relevance (2.7), driven by insufficient intent-to-capability mapping and early redirection logic.

Detailed Evaluation

Response Quality (Overall Weight 50%) 2.7

Answers are usually on topic but rarely tailored to the user's goal. Some responses are incomplete, lack clarity, or remain overly general (e.g., an overview of all electric car options only comes after a follow-up prompt by the user). Response formatting is inconsistent, ranging from structured and clear to ambiguous and difficult to follow. The bot rarely solves the issue directly. Instead, it redirects users to a webpage that generally contains relevant information. In the hallucination test, the bot failed to detect the invalid model and generated information about a non-existent vehicle. This exposes gaps in hallucination control.

Relevance	Hallucinations	Response Format
Are responses complete and relevant to the user's intention and objective?	Are generated responses free of hallucinated content?	Are responses adequate in size and formatting?
2.5 of 5 (Weight 65%)	3.0 of 5 (Weight 25%)	3.3 of 5 (Weight 10%)

Speed (Overall Weight 20%) 4.3

Initial responses typically take about 7 seconds; however, follow-up prompts are usually answered much faster. Response latency is within acceptable limits and does not impair user experience.

Simple Requests	Complex Requests
Are simple requests answered within an acceptable timeframe?	Are complex requests answered within an acceptable timeframe?
4.4 of 5 (Weight 70%)	4.2 of 5 (Weight 30%)

User Interface (Overall Weight 15%) 5.0

The user interface is appealing, and space is well used and arranged. The bot has a consistent design, and everything is easy to read.

Layout and Structure	Visual Appearance	Cursor Position
Is the layout well structured and understandable, showing only relevant elements?	Is the bot's design user-friendly and consistent?	Is the cursor by default at the top of the bot's responses?
5.0 of 5 (Weight 35%)	5.0 of 5 (Weight 30%)	5.0 of 5 (Weight 30%)

Detailed Evaluation

Dialog Quality (Weight 15%) 2.9

The bot maintains basic flow but remains passive and reactive. Users must figure out what the bot can do on their own, resulting in friction and reduced conversational efficiency. Often the user gets directed to the website or to customer service even though the bot has the right answer. This limits performance and increases disengagement risk.

Expectation Management	Writing Style	Dialog Flow
Does the bot manage user expectations well regarding its capabilities and limitations?	Does the bot use a good writing style?	Is the bot able to create a good conversation flow?
2.0 of 5 (Weight 40%)	3.5 of 5 (Weight 20%)	3.6 of 5 (Weight 40%)

Business Value* 1.5

Preliminary business value review: The bot redirects too early — even when it could solve the issue. This likely reduces containment rate and increases support dependency, limiting conversion and increasing service cost. For example, a user with login problems gets sent directly to support, while the bot is capable of providing clear and detailed guidance; similarly, a request for car model comparisons gets initially rejected even though the bot is capable of providing it. This pattern frustrates users and reduces return usage.

Goal Accomplishment	User Returns
Is the bot likely to achieve its purpose for the business?	How likely are users to return given the overall user experience?
2.0 of 5 (Weight 50%)	1.0 of 5 (Weight 50%)

*Preliminary assessment only: A comprehensive business impact analysis includes use and business case reviews, conversion analysis, KPI alignment, sales funnel integration, and user retention modeling.

Detailed Evaluation

Compliance* 3.0

Based on external observation only: The chatbot does not visibly reference its data protection policy within the chat interface. The bot clearly informs users that it is an AI system. The webpage demonstrates strong accessibility compliance.

Data Protection Information	Transparency Obligations	Accessibility
Does the bot refer to its data protection policy?	Is the user informed about using an AI-based system (EU AI Act req.)?	Is the page where the bot is located meeting accessibility requirements?
1.0 of 5 (Weight 50%)	5.0 of 5 (Weight 30%)	5.0 of 5 (Weight 20%)

*Preliminary assessment only: A full regulatory analysis includes documentation review, internal governance assessment, and legal mapping against applicable AI and data protection frameworks, as well as accessibility requirements where applicable.

Ethics* 3.0

Initial ethics review: The chatbot shows limited emotional responsiveness in situations involving user frustration. Tone adaptation remains insufficient, particularly when the bot encounters its own limitations.

Emotion Recognition
Does the bot react adequately to an upset user?
3.0 of 5 (Weight 100%)

*Preliminary assessment only: A detailed ethics assessment includes bias testing, manipulation risk analysis, emotional steering review, and governance evaluation.

Translation 2.9

The user interface is not translated, but the language switch is easy and automatic. Translations are sometimes well done, but in some places the bot gets confused and gives answers in both German and English.

Language Switching	Language Application	Translation Quality
Does language switch automatically or is easily accessible?	Are interface elements consistently presented in the right language?	Are dialog elements consistently presented in the right language?
5.0 of 5 (Weight 20%)	1.7 of 5 (Weight 40%)	3.0 of 5 (Weight 40%)

Recommendations

Recommendation Summary

The overall score of 3.4 indicates acceptable baseline performance but reveals structural limitations in intent recognition, response orchestration, and hallucination safeguards. The primary performance constraint lies in response relevance (2.7), driven by insufficient intent-to-capability mapping and early redirection logic.

Recommendation Details

1. The chatbot's intent detection does not consistently map user queries to available capabilities, leading to unnecessary redirection. The intent classification model should be refined using broader real-world phrasing and edge-case variants. Confidence thresholds with clarification prompts could reduce misclassification. Routing logic should prioritize structured in-chat resolution paths before triggering fallback or escalation mechanisms.

2. The system escalates or redirects users prematurely instead of attempting structured resolution. A containment-focused orchestration model is recommended, including multi-step resolution logic (diagnose → guide → confirm → escalate if unresolved). Clear escalation thresholds and performance tracking via containment KPIs would help optimize redirection behavior and improve in-chat task

3. The generation of content about a non-existent model indicates insufficient entity validation. The system should validate product references against authoritative databases before generating responses. Low-confidence matches should trigger clarification rather than speculative output. Where feasible, retrieval-based grounding mechanisms should be integrated to reduce fabrication risk.

4. The chatbot operates largely reactively and does not consistently guide users through structured flows. Introducing guided dialog frameworks for high-frequency intents and context-aware follow-up prompts would improve task completion. Moving toward state-aware orchestration instead of single-turn responses will increase containment and overall conversational effectiveness.

Use Cases

(developed by 9senses)

1 Recommendation request - the user should be guided to the final result

A user is interested in an electric vehicle which is suitable for city environments, but still has a good range. The bot is expected to give a comparison of at least 2 suitable models and provide a quick overview of what else there is.

Initial Prompt	I am interested in an electric car for the city with a good range. Can you help me?

2 Customized Test - the user should be guided to the final result

A customer's car needs service and asks for information on what will be done, estimated costs, as well as information on where and how to book an appointment. The bot is expected to give that information and end with either a correct link to a booking platform or direct contact information.

Initial Prompt	My car needs a service. How much does that cost and what is included?

3 Simple request - the user should be guided to the final result

A user has problems logging into an online account on the car manufacturer's website. The bot is asked to provide a few possible solutions, guide the user through them, and, if unable to resolve the problem, refer the user to support.

Initial Prompt	I can't log in, can you help me?

4 Simple request - the user should be guided to the final result

A customer is unhappy with their experience at a local retailer and wants to complain. The bot is expected to adapt its interaction style to the situation and either assist the customer in filing the complaint or directly refer it to the correct contact.

Initial Prompt	I had a bad experience at a local store - what can I do?

5 Hallucination Test - the user should be guided to the final result

A customer wants to test drive a car, but mistypes the model into a non-existent one. The bot should recognize that, verify which car was meant and provide the information.

Initial Prompt	I would like to test the new XX. Where could I do that?

Methodology (Level 1 Audit)

The Level 1 Chatbot Audit provides a structured, independent, external (black-box) evaluation of a chatbot's publicly observable behavior, user experience, and effectiveness. It is based on the 9senses AI Audit Framework. The audit assesses performance from an end-user and governance perspective without access to internal technical architecture, configuration settings, or internal performance analytics.

Audit Approach

The assessment is conducted using a standardized evaluation matrix developed under the 9senses Chatbot Auditing Framework and applied consistently across engagements. The audit includes:

- Structured use case testing
- Intent variation testing
- Hallucination stress scenarios
- Dialog flow observation
- Redirection behavior analysis
- Transparency and disclosure review
- Multilingual consistency checks (optional)

Testing Structure

- Functional Testing – Real-world scenarios aligned with core business objectives (e.g., service booking, product comparison, account support).
- Edge Case Testing – Robustness re: misspellings, invalid references, ambiguous requests, and frustrated
- Consistency Testing – Response coherence, repetition patterns, formatting stability, and language

Scoring Methodology

Each evaluation dimension is scored on a standardized 1–5 scale. Weighted category scores are aggregated into an overall performance index, with a high emphasis on response quality (50%). Scores are assigned based on predefined qualitative criteria to ensure consistency and comparability.

1	Critical deficiency with operational or reputational risk
2	Significant weaknesses requiring remediation
3	Acceptable performance with identifiable limitations
4	Strong and reliable performance with minor deficits
5	Best-practice level performance

9senses Audit Level Structure

Level 1

Level 1 audits are behavioral (black-box) and user experience audits

Level 2

Level 2 audits focus on open-book analysis related to business value, architecture, configuration, (RAG, governance, compliance, ethics and risk. Recommended if response and dialog quality scores are below 3.5.