Human Factors in Design, Engineering, and Computing, Vol. 238, 2026, 1781–1788
https://doi.org/10.54941/ahfe1008406

Exposing Tool Use in AI Assistant Interfaces: Inspectable Evidence for Industrial Assistance

Andrea Alesani

VTT Technical Research Centre of Finland Ltd, Tampere, Finland

ABSTRACT

AI assistants use tools to obtain information and perform actions, but users may still need to inspect the basis of an answer or action. This paper explores an interface for inspecting document retrieval and PDF navigation, with tool requests and results shown alongside a conversation and source PDF. Three author walkthroughs established concrete evidence conditions, including incomplete retrieval, in which inspection could help a user check an answer. Three moderated expert sessions on Meta Quest 3 then explored how participants responded to this support. When invited to check an installation answer, all pursued a source check; two identified omitted or additional details, while one could not verify the answer. Participants used different evidence routes and suggested clearer links, optional detail, and task-relevant visuals. The findings motivate further development of accessible evidence interfaces that support users in checking assistant outputs and deciding what to do next.

Keywords: Human-computer interaction, Tool transparency, Artificial intelligence, Interface design, Industrial assistance

INTRODUCTION

Tool-using assistants connect users' requests to external information and actions. An assistant can retrieve passages, navigate a manual, check sensor data, and search databases, but a fluent answer may conceal what was consulted. Users may still need to check whether a surprising answer reflects a documented restriction, incomplete evidence, or an error.

The design objective is to help users relate an answer to the available evidence and direct the next interaction. For example, a user checking an installation procedure could compare the summary with a source passage, inspect omitted details in the manual, or request a focused search. Trust should relate to automation's capabilities and conditions of use (Lee and See, 2004). The interface provides observable points of reference: the tool request, its returned information, and the source document.

This paper presents a prototype (Figure 1) and a two-part exploratory study. Author walkthroughs establish concrete situations in which evidence is available, incomplete, or reused. Moderated expert sessions then examine how participants use and respond to the proposed inspection support. The contribution is a set of observations and design implications for

Received September 24, 2026; Revised October 30, 2026; Accepted November 9, 2026; Available online December 1, 2026

© 2026 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

connecting answers, observable tool activity, and accessible evidence. The study uses one model configuration; its focus is the support available for user participation.



Figure 1: The three interface components used in the study: conversation at left, tool panel in the centre with document-retrieval request and result cards collapsed, and source PDF at right. The Meta Quest screenshots in Figures 1, 2, and 4 were upscaled using AI to improve readability.

BACKGROUND AND DESIGN RATIONALE

Human–AI interaction guidance emphasizes communicating capability, supporting correction and helping users understand system behavior (Amershi et al., 2019). Liao, Gruen and Miller (2020) organize explainability around users' questions. Here, the inspection target is an external tool request and its result, rather than a generated explanation of the model's answer.

Situation awareness includes perceiving relevant elements and understanding their significance (Endsley, 1995). This motivates a practical goal: support reliable use by helping users check an answer against available evidence and decide what to inspect or request next. The walkthroughs establish concrete evidence limitations, while the expert sessions explore users' responses to activity cues, inspectable results, and source access.

Compact activity entries expand into details for checking returned evidence, while a tool-call sound provides an additional cue when attention is elsewhere in the XR workspace. The adjacent PDF offers another route to source information. Together, these surfaces let users choose the level and form of detail they need.

INTERFACE AND SYSTEM

The Unity client places the document, conversation, and tool information in a shared workspace (Figure 1). Author walkthroughs used the laptop implementation; the formative expert sessions used a Meta Quest headset with voice input.

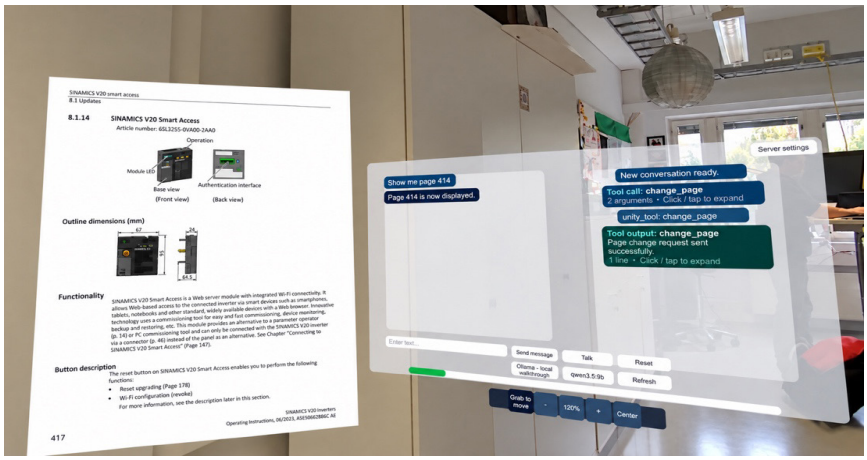


Figure 2: Interface illustration of PDF navigation: the conversation requests a specific page, and the tool panel shows the change_page call and returned status beside the displayed PDF.

Compact entries identify observable tool activity as events arrive. An expandable side panel provides the tool name, supplied arguments, and returned output. For document retrieval, this includes page-labelled passages that can be compared with the answer. A subsequent request can navigate the PDF to inspect the source in context. The panel exposes the change_page call, its page argument, and the returned status, allowing users to check whether the requested action was invoked (Figure 2). These routes serve related purposes: retrieved passages show what the tool supplied to the assistant, while the full document may contain further details beyond those passages.

Figure 3 summarizes the event flow. The backend sends the request and result to the client as structured events. At tool-call start, the client also plays an audio cue to indicate activity; the returned evidence remains available in the visual record. The renderer consumes generic tool-name, argument, and output fields, allowing different tools to use the same event presentation. This provides a basis for extending inspection support while adapting the visible evidence to each task and information source.

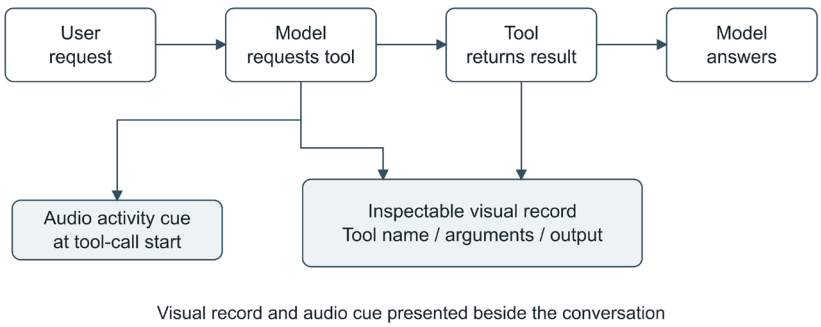


Figure 3: Requests and results populate the visual inspection record; an audio cue signals the start of tool activity.Tool-event visibility does not expose hidden reasoning.

A secondary context indicator reports token occupancy against the configured window. It describes resource use for the reported model invocation, not confidence or correctness, and was not evaluated in the expert sessions.

METHOD

Three structured author walkthroughs used the SINAMICS V20 operating manual (Siemens AG, 2022), the local qwen3.5:9b model served through Ollama, and an 8,192-token context window. The assistant had access to a documentation retrieval tool and a PDF navigation tool. Each scenario began with a fresh conversation. Short, fixed utterances standardized the intended evidence condition and could plausibly be spoken.

The retrieval pipeline used lightweight lexical matching and fixed chunking to create inspectable, repeatable evidence conditions. The backend set how many passages to return and how long each passage could be. The broad overview case applied 20 results with a maximum of 2,500 characters per snippet; the normal Flat Plate passage was limited to 700 characters. These settings produced inspectable evidence conditions without requiring users to specify retrieval parameters. The study concerns inspection under these conditions, rather than the performance of an optimized retrieval system.

The technical baseline was recorded through the actual WebSocket interface. Saved events include responses, tool requests and arguments, returned text, navigation calls, and token statistics. Automated checks verified the intended scenario conditions. The walkthroughs establish what happened in these retained interactions, including limits in the retrieved evidence; the participant sessions examine how people respond to the available inspection support. Later runs could produce different outputs.

On 23 September 2026, three experts completed moderated Quest sessions lasting 18–28 minutes. They followed the same three task goals in order using natural voice prompts. The moderator invited them to identify document use, check answer support, try a next action, and distinguish a follow-up from fresh retrieval. Prompts, tool events, and audio were retained under anonymous participant IDs. Initial coding distinguished behavior demonstrated without extra help, after a hint, not demonstrated, or not assessable. The analysis also considered participants' chosen evidence routes, the information they found, and their design suggestions. The protocol initially emphasized the tool panel, and the moderator sometimes redirected participants toward it after they chose the PDF. These sessions therefore examine responses during prompted exploration. Participants received no standardized tutorial. The amount of moderator help differed between sessions, and speech recognition sometimes altered participants' intended questions.

AUTHOR WALKTHROUGHS OF EVIDENCE CONDITIONS

Supported surprise. The first request asked whether the Smart Access module, an accessory used to set up the motor controller, could remain connected after setup was complete. The assistant answered no. One document query

returned a passage on PDF page 414 stating that the module is intended only for initial setup and cannot remain in permanent use with the converter. Inspection therefore revealed documentary support for a potentially surprising restriction. The follow-up, "Open PDF page 414 for me," issued a separate change_page call with page_number=414 and did not repeat retrieval. The trace made the distinction between searching for evidence and navigating to its location observable.

Incomplete evidence. The second request asked how to install the Flat Plate version, including screw torque, clearance, and a temperature check. Retrieval returned part of the procedure without every requested numerical value, and the response acknowledged these gaps. The case establishes an observable limitation in the available evidence. In a separate author rehearsal on 22 September, a shorter installation question returned a passage cut off during step 4 (illustrated in Figure 4). The author noticed the cutoff and proposed another search or checking the source, without performing either action. These interactions illustrate why a user may need to examine evidence and seek further information. They motivate the inspection support explored with participants.

Evidence reuse. The broad commissioning and fault-information request returned manual passages. A follow-up asked which three sections to start with, explicitly requesting no further lookup. No additional search occurred. The trace made this distinction observable and retained the earlier result for inspection. This case provides a further condition for user awareness: an answer may follow earlier retrieval without a new search.

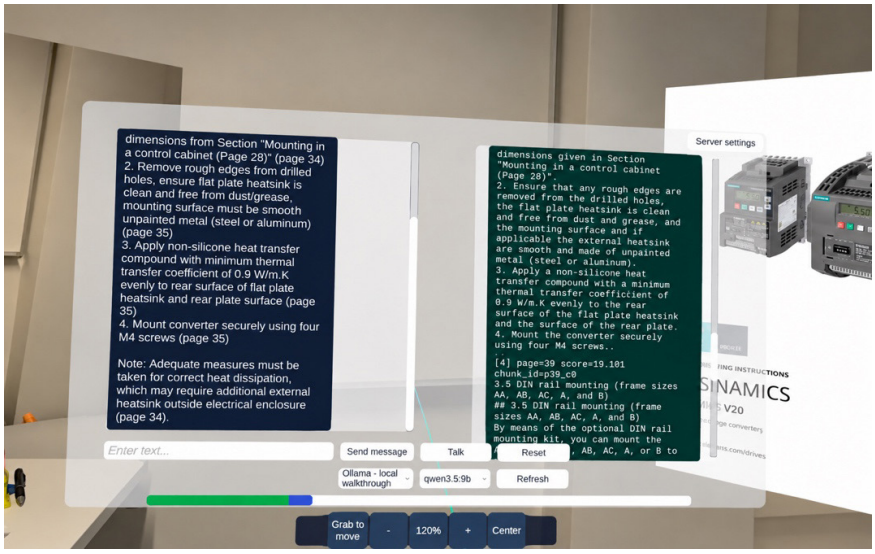


Figure 4: Interface illustration of the incomplete-evidence case (walkthrough 2): the expanded tool-output panel on the right shows the retrieved procedure cut off during step 4, "Mount the converter securely using four M4 screws…". The assistant's answer appears on the left.

FORMATIVE EXPLORATION OF EVIDENCE USE

Participants followed the same three task goals as the author walkthroughs: checking the Smart Access restriction and navigating to its source, inspecting the Flat Plate installation procedure, and asking a follow-up to a broad overview without a new search. They used natural voice prompts rather than fixed wording.

The expert sessions explored participants' responses to the prototype's evidence support. When invited to check the installation answer, all three pursued a source check. They chose different routes through the PDF and retrieved passages; two identified omitted or additional details. Table 1 summarizes these routes and outcomes. The observations show how participants used the available support and where its presentation impeded their checks.

P01 preferred the visible PDF and independently identified missing installation details after opening the source page. P02 compared retrieved chunks with the answer after navigation help and identified a 50 Hz condition omitted from the response. Later PDF inspection revealed further screw details, although P02 had not noticed the retrieved procedure ending mid-step. P03 attempted PDF navigation and then inspected tool output, but could not establish support for the answer. The unsuccessful check identifies a gap in the prototype's support for reaching and interpreting relevant evidence.

Table 1: Evidence routes and outcomes in the prompted installation task.

ID	Evidence Route and Outcome	Presentation Needs
P01	Chose the PDF; identified omitted installation details independently.	Direct links to relevant pages or sections; optional logs.
P02	Compared retrieved passages with help; found an omitted condition, then further detail in the PDF.	Readable excerpts linked to the source; concise steps with access to detail.
P03	Tried PDF navigation and tool output; could not verify the answer.	Clearer source navigation and readable evidence; selected installation images.

Activity awareness and evidence inspection served different purposes. P03 recognized retrieval from the tool-call sound. All three correctly reported no new search in the final follow-up, although P01 did not name a cue and P03 relied on absent sound. In the first task, none demonstrated locating the search arguments or the intended passage in the card; P02's changed question made the intended passage unavailable. Page references also acted as credibility cues before content was checked. A few prototype shortcomings affected inspection, such as a mismatch between the PDF page numbering and the page labels printed in the document footers, and speech playback that competed with reading for P01 and P02. The assistant and retrieval system used PDF page positions, whereas participants could see the page numbers printed on the displayed document pages. These numbering systems can differ; the interface needed to make their correspondence clear. Their suggestions addressed how evidence should be presented: P01

proposed an optional activity icon linked to the relevant page or section; P02 preferred concise steps with the PDF for documentation, where the source viewer offered a more usable check than the raw results panel, but suggested that the tool call card in the panel could be useful for inspecting selected records from – for example – a large database; P03 valued the audio cue and proposed selected installation images in place of a prominent raw-text panel.

DISCUSSION AND LIMITATIONS

The sessions motivate further investment in evidence interfaces that support active user participation. P01 checked the manual, P02 compared retrieved passages with the answer, and P03 attempted both routes while identifying problems in their presentation. These responses suggest that inspection support should connect the answer with accessible evidence across the workspace. The practical goal is to give users usable points of reference for checking what information is available, what the assistant has done, and what to inspect or request next.

The prototype's rough presentation exposed some opportunities for improvement. P03's request for selected installation images, together with a preference for the activity sound, suggests that awareness and detailed inspection need different forms of support. Further refinements include a compact activity cue with optional full logs, readable excerpts with clear boundaries and cutoff indicators, and direct links from citations to the matching source page and passage. Controls to pause or mute speech could support reading. Task-relevant images could complement text, while preserving access to the underlying evidence. These proposals follow participants' behavior and feedback and remain to be evaluated.

The broader design argument concerns user participation as assistant capabilities develop. More capable models and improved retrieval may reduce some errors or omissions, while users still need ways to inspect task-specific conditions and direct further interaction. Evidence access is therefore worth developing alongside model capability. The full document and the retrieved excerpt offer different information; other tools may call for different representations. When prompted about sensor data, P03 suggested nominal ranges and relevant trends. This suggestion illustrates a direction for future interfaces, rather than a tested result beyond documentation tasks.

This exploratory study used three participants, one manual, one model configuration, and a rough prototype, with no comparison interface or field installation. Lexical retrieval and fixed chunk limits shaped the available evidence. Voice recognition changed prompts, and moderator hints, redirection toward the panel, descriptions of its formatting, and page-navigation help may have influenced behavior and feedback. The observations concern prompted exploration and design opportunities. The sessions do not establish spontaneous verification habits or improvements in trust, accuracy, safety, or productivity, and they do not compare models. A further study should test whether the proposed refinements help users find relevant evidence, interpret it, and choose a useful next action.

CONCLUSION

The prototype makes tool requests, returned evidence, and source navigation available for inspection. The author walkthroughs established concrete evidence conditions, including incomplete retrieval, that motivate support for checking assistant answers. The three moderated Quest sessions then showed how participants responded to the prototype: all pursued evidence when invited to check an installation answer, two found omitted or additional details, and one could not verify it. Their behavior and suggestions motivate clearer connections between concise answers, readable excerpts, source documents, and task-relevant visuals. The next design step is to improve these connections so that users can inspect the basis of an answer and direct further interaction as assistant capabilities develop.

REFERENCES

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R. and Horvitz, E. (2019). Guidelines for Human-AI Interaction. Proceedings of CHI 2019, pp. 1–13. https://doi.org/10.1145/3290605.3300233

Endsley, M. R. (1995). Toward a Theory of Situation Awareness in Dynamic Systems. Human Factors, 37(1), pp. 32–64. https://doi.org/10.1518/001872095779049543

Lee, J. D. and See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. Human Factors, 46(1), pp. 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Liao, Q. V., Gruen, D. and Miller, S. (2020). Questioning the AI: Informing Design Practices for Explainable AI User Experiences. Proceedings of CHI 2020, pp. 1–15. https://doi.org/10.1145/3313831.3376590

Siemens AG (2022). SINAMICS V20 Converter Operating Instructions. May 2022 edition.