AI-moderated interviews (AIMI) in practice: early evidence and open questions

A practitioner's guide to AI-moderated interviews (AIMI)

1. Abstract

Glaut is pioneering a new research methodology: AI-Moderated Interviews, or AIMIs. Positioned between surveys and traditional in-depth interviews, AIMIs use conversational AI to conduct one-on-one, open-ended interviews at scale adapting to each participant's response in real-time.

This whitepaper introduces AIMIs as a distinct hybrid quantitative and qualitative approach: explaining how they work, what makes them different, and where we believe they hold particular value. It also shares early findings from our first comparative study between AIMIs and "static" surveys, (or CAWI), which suggests that AIMIs can produce richer, cleaner, and more engaging data than surveys.

Please note that this paper is not a validation of the method, it's a first articulation of the logic behind it, the evidence supporting our hypotheses so far, and the open questions we believe still need rigorous exploration.

2. Introduction

Surveys often fall short on depth. In-depth interviews deliver richness but are hard to scale.

At Glaut, we've been developing a third approach: in 2024, we coined the term **AIMI - AI-Moderated Interview** to define a new methodology that combines qualitative depth with quantitative scalability.

AIMIs are one-on-one, open-ended, AI-moderated interviews. During the interview, the Moderation Agent makes questions, listens and understands answers, and probes in real time without requiring a human moderator. Like IDIs, AIMIs adapt to each participant, like surveys they can scale to hundreds or thousands of sessions.

This paper introduces the AIMI methodology, explains how it works, and shares early findings from our first comparative study. It is not a validation, but an invitation: a first step toward exploring where this method adds value and where its limits lie.

3. Methodological approach

AI-Moderated Interviews (AIMIs) represent a new methodological category combining the adaptive depth of qualitative interviews with the scale and consistency of survey infrastructure. Unlike human-moderated interviews or static forms, AIMIs are built from the ground up to operate through conversational AI.

This section outlines how AIMIs work, what defines them, and how the underlying technologies support their reliability and scalability, grounded in Glaut's current design, safeguards, and deployment protocols.

3.1 What are AI-moderated interviews (AIMIs)?

AI-moderated interviews (AIMIs) are 1-o-1 interviews led by a conversational agent powered by large language models (LLMs) and natural language processing (NLP). Designed to simulate the flow of live qualitative interviews, they ask open-ended questions, listen actively, adapt to the respondent, and generate real-time follow-ups all without human intervention during the session.

The interviews can take three formats:

- **Voice-based** (voice answers via microphone)
- **Text-based** (written answers via keyboard)
- **Hybrid** (voice answers + scaled inputs, e.g., checkboxes)

AIMIs are not just surveys with better UX. They create an interview-like experience: asking, listening, and clarifying in a semi-structured but adaptive way. This allows researchers to gather rich, theme-dense responses from large samples, including harder-to-reach groups like children, low-literacy populations, or multilingual audiences.

3.2 How do they work?

Every AIMI follows a modular structure that allows it to operate seamlessly from respondent onboarding to insight delivery. Below, we unpack the three key operational layers: conversation design, pre-interview safeguards, live interview dynamics, and post-interview interpretation.

1. **Conversation design:** Every AIMI begins with a researcher-led conversation design process. This includes 3 to 15 core open-ended questions, often supplemented by optional scaled items. The backbone is deliberately minimal, typically 3 to 5 well-crafted prompts, so that the AI can maximize space for respondent spontaneity.

   Each question is paired with logic that guides the moderator Agent's follow-up behavior. Researchers define how the AI should respond to vague, incomplete, or off-topic answers, and set tone boundaries to ensure a respectful, on-brand interaction. Importantly, this design phase defines not just *what* is asked, but *how the interview adapts*—creating a semi-structured yet responsive conversation flow.

2. **Interview safeguards for higher data quality:** before a single question is asked, researchers can set up the AIMI to apply a set of automated controls to ensure the quality and consistency of the sample. These include:

    - Participant pre-screening (e.g., by demographic, language, prior exposure)
    - Question randomization
    - "Voice-only" enforcement for studies where verbal expression is key

    Built-in agents such as:

    - Interpretative score: estimates the expected insight richness from a respondent
    - Consistency checker: flags contradictions in responses and redirect respondents out
    - Uncooperative detector: screens out participants unlikely to engage meaningfully

    These safeguards ensure that interviews begin with the right participants, continue with engaged participants and reduce the volume of unusable data before it enters the system.

3. **Live interview dynamics**: once the session begins, the Moderation Agent takes over. It introduces the study, asks the first open-ended question, and listens. As the respondent answers, the AI evaluates their input in real time:
    - If the response is clear and complete, it moves to the next question.
    - If the response is vague ("I like it"), the AI prompts for depth ("What do you like about it?").
    - If the respondent becomes unresponsive, contradictory, or nonsensical, the moderator agent may rephrase, prompt again, or terminate the session based on predefined thresholds.

    This live moderation behavior is powered by a combination of natural language processing (to parse meaning) and large language models (to generate follow-ups with contextual relevance). The result is a conversation that feels fluid, natural, and tailored without needing a human moderator.

4. **Automated data processing and analysis**: the final stage is where AIMIs deliver their most strategic value: transforming raw verbatim into interpretable insights. This process includes:
    - **Thematic clustering**: surfacing key narratives, subtopics, and emergent issues.
    - **Sentiment and emotion analysis**: parsing tone, intensity, and attitudinal signals.
    - **Interpretative scoring**: assessing depth, clarity, and nuance of each response.
    - **Metadata tagging**: capturing duration, number of probes, and engagement markers.

- **Structured output generation**: full transcripts (including downloadable original audio files), theme-indexed datasets, and editable reports with verbatim highlights

Every theme, summary, and insight is linked back to the original participant response ensuring full traceability and transparency. Rather than simply processing data, AIMIs provide researchers with a layered, explorable insight asset.

AIMIs operate with real-time logic and fallback conditions, for instance, handling off-topic responses or activating a consistency check agent if participant answers seem contradictory. The result is a structured yet flexible conversation that mirrors qualitative interviewing while scaling far beyond what is typically possible in manual formats.

3.3 The technology behind: Large language models (LLM) and Natural Language Processing (NLP)

AIMIs are powered by a combination of natural language processing (NLP) and large language models (LLMs). These technologies enable the system not only to understand participant input, but also to respond intelligently, generate tailored follow-ups, and support downstream interpretation at scale. During the interview, NLP systems parse each response to assess relevance, depth, sentiment, and structure. This first layer ensures that responses are coherent and actionable and allows the AI to decide whether a follow-up is needed or if the interview can proceed.

LLMs, layered on top of NLP, generate the actual follow-up questions. They do so by considering the participant's answer, the research objective, and predefined tone and guardrails. This dynamic interaction allows AIMIs to simulate many of the core behaviors of a skilled moderator, clarifying vague responses, probing for examples, or gently pushing for specificity.

After the interview, the same LLMs support the second core function: interpretation. They help extract themes, cluster sentiment, and surface underlying motivations or tensions. This is where the AI augments - not replaces - human insight. it's easy to fall into the *"illusion of insight*" when AI output is treated as conclusive. Glaut's approach emphasizes interpretability, auditability, and transparency: all themes are traceable, all summaries are editable, and researchers retain full control over the narrative being built.

While LLMs are powerful, they're not infallible. Their performance can vary by language, cultural context, and input quality. That's why every AIMI is built with prompt boundaries, fallback conditions, and human-in-the-loop checkpoints during design and analysis. Glaut wants to be a structured tool for scaling both depth and reliability.

4. The hypotheses: our working assumptions

We hypothesize that AIMIs offer three key advantages:

- **Richer responses:** more complex answers due to voice interaction and dynamic probing.

- **Cleaner data:** lower incidence of gibberish or low-effort replies.

- **Higher engagement:** more natural, satisfying experience for respondents.

These are hypotheses, not conclusions. To explore them, we ran a first comparative test.

4.1 AIMI vs. Surveys: a controlled comparative study

To explore the potential of AIMIs as a scalable alternative to traditional surveys, Glaut conducted a controlled experimental study, designed and authored by Glaut and G. M. Occhipinti *(Enhancing Qualitative Market Research with Conversational AI-Agents, 2024)*. The study compared AI-moderated interviews and static online surveys using the same topics, questions, and thematic focus allowing for a like-for-like evaluation across two balanced groups of 100 Italian participants.

All open-ended responses - whether from AIMIs or surveys - were analyzed using the same protocols for theme extraction and quality evaluation, ensuring a fair and consistent comparison across methodologies.

Performance comparison: key metrics

The study evaluated six key metrics:

Metric	AIMI	Survey	% Change
Avg. Experience Rating (1–10)	8.48	8.03	+5.6%
Avg. Words per Respondent	71.97	31.42	+129.1%
Avg. Themes per Respondent	8.23	6.94	+18.6%
Preferred Transcripts (LLM eval)	66%	34%	+94.1%

Gibberish Transcript Rate	26%	56%	-53.6%
Valid Completion Rate	61%	39%	+56.4%

(*Source: Glaut, Occhipinti, 2024*)

The most striking difference was the total word count: AIMIs elicited responses that were, on average, over twice as long as those from the survey group. Thematic richness also increased meaningfully, with AIMIs surfacing more unique ideas per participant. Perhaps most importantly, the incidence of low-quality or "gibberish" responses was more than halved in the AIMI group. These cleaner, more elaborate transcripts were also rated higher in quality: in head-to-head comparisons, AIMI transcripts were judged superior in two out of three cases.

Why did AIMIs outperform surveys?

Two features likely account for AIMIs' improved performance:

- **Voice interaction** made the experience feel more like a natural conversation than a task.
- **Contextual follow-ups** allowed the AI to probe vague answers in real time—something no static survey can do.

For example, a respondent answering "I like the brand" would be asked, "What specifically do you like about it?", a prompt absent in the static format. These micro-adjustments increased linguistic and thematic depth without introducing bias or fatigue.

Additionally, the AI's real-time detection of non-responses (e.g., "idk," irrelevant replies) helped filter out disengaged participants before they completed the session, leading to a higher ratio of meaningful completes.

Statistical significance

To ensure the differences were robust, the researchers used non-parametric and categorical tests:

- **Mann-Whitney U tests** confirmed significant differences in word count and theme count ($p < 0.001$).

- **Chi-square tests** validated transcript quality and gibberish rates as significantly associated with completion mode ($p < 0.01$), after Bonferroni correction.

This reinforces the central hypothesis: AIMIs - when powered by voice and dynamic follow-ups - produce richer, cleaner qualitative data and better respondent engagement than traditional surveys.

However, as the original study authors note, these findings should be treated as first evidence, not final validation. Larger-scale replications across diverse populations will be necessary to confirm how broadly these results generalize.

4.2 AIMI vs. In-depth interviews (IDIs)

AIMIs share much in common with traditional in-depth interviews. They replicate key elements such as contextual probing, adaptive follow-ups, and respondent engagement while introducing the possibility of scaling that experience across hundreds or even thousands of participants. In many respects, AIMIs can be seen as a scalable form of qualitative depth: able to listen, clarify, and explore meaningfully in a one-on-one format without requiring a human moderator for each session.

That said, we don't assume equivalence. While early results are encouraging, especially for exploratory research and product feedback, we recognize that highly nuanced or emotionally charged topics may still benefit from the sensitivity and intuitive judgment of a trained qualitative researcher. To understand where the line truly lies, Glaut plans to run direct comparative studies between AIMIs and IDIs, exploring where the two approaches align and where they differ most.

4.3 Signals from early field applications

While structured experiments like the previous Occhipinti study help test specific hypotheses, they don't fully reflect the messy, real-world conditions where research is deployed. That's why we've also been attentive to early signals emerging from client projects. These are not designed to validate AIMIs per se, but they do offer contextual clues about where the methodology may hold particular value.

- In a recent project with **a Media and Entertainment group**, AIMIs were used to explore trust and preferences among **children aged 3 to 13**. The interviews were conducted entirely through voice, with no parental mediation required. Despite the young and typically hard-to-engage target, the approach yielded a **96% completion rate**. The AI's ability to adapt tone and pace to each child's input appeared to play a key role in sustaining engagement.
- Separately, an **international research firm** used AIMIs to conduct a **large-scale qualitative project for a leading e-commerce company**. Rather than treating AIMIs as a post-collection tool, they designed and ran the full study through Glaut's AI-moderated interviews capturing open-ended responses at scale. By combining this methodology with AI-powered analysis, including automated theme extraction and real-time quality controls, they reported a **95% reduction in the time required for coding and interpretation**. Workflows that would typically span several days were compressed into a matter of hours without sacrificing traceability or thematic depth.

These examples are not conclusive evidence. They represent early, contextual applications in which AIMIs appear to unlock new forms of access, efficiency, or inclusion. More controlled and comparative research will be needed to isolate the specific contributions of the methodology, but these initial signals encourage us to explore further.

5. How to Decide

Choosing a research methodology is never just about features, it's about fit. Each approach carries strengths, limitations, and trade-offs that become more or less relevant depending on your objectives, your audience, and the type of insight you need.

AIMIs introduce a new third methodology option that blends elements of both qualitative and quantitative. They're structured like surveys, but feel like conversations. The AI Agent probe like moderators, but operate at scale. So when should you consider using one?

Use AIMIs when:

- You're exploring a topic that benefits from narrative responses rather than single-word answers.
- You want to ask open-ended questions and ensure respondents actually engage with them.
- You need scale and consistency but still want depth (e.g., hundreds of interviews across markets or audiences).
- You're researching participants who may be easily distracted or demotivated by static formats (e.g., younger users, low-literacy audiences, mobile-first contexts).
- You need qualitative insights on a faster timeline, and your team can benefit from AI-assisted coding, cleaning, or summarizing.

In these cases, AIMIs can unlock new possibilities. The ability to probe in real time, adapt to each respondent, and collect high-quality open ends without a human moderator opens a middle ground between survey scale and interview depth.

Stick with traditional methods when:

- Your stakeholders expect or require live moderation (e.g., in-person ethnographic work, IDIs, etc.).
- You're working in contexts where AI-led tools might not yet be culturally or technically appropriate.

6. Open questions and next steps

The introduction of a new methodology is never the end of inquiry, it's the beginning of a new set of questions. AIMIs, while promising in structure and early performance, raise important considerations about context, bias, applicability, and long-term reliability.

We see their potential, but we also know what we don't yet know. This section lays out the core open questions we're actively investigating, and the directions we plan to explore as this methodology evolves.

- **Benchmarking vs. standard quantitative methods (e.g.: surveys):** our first comparative study offered a compelling first look at how AIMIs outperform static surveys on basic metrics: word count, thematic richness, completion rate, and gibberish reduction. We now aim to expand this benchmark in three key directions:

    - **New KPIs:** going beyond linguistic volume to measure brand narrative density, clarity of preference, motivation quality, or idea originality.
    - **Wider contexts:** running comparative studies across different markets, age groups, and topics especially in more emotional or conceptual domains.
    - **Mixed-mode comparisons:** testing AIMIs against surveys with rich open-ends.

    The goal is to identify when AIMIs provide not just more data, but better insight.

- **Benchmarking vs. standard qualitative methods (e.g.: IDIs)**: if AIMIs have demonstrated strength in replacing open-ended surveys, the next and more complex question is whether they can approximate the value of a live, human-moderated in-depth interview. We do not assume equivalence, but we believe it's worth testing systematically, rigorously, and transparently.
    We are designing a new phase of comparative studies focused on:

    - **Engagement quality**: how do participants describe the experience? Do they feel "heard" by the AI?
    - **Interpretative richness**: how many layers of meaning emerge in the transcripts? Are responses more descriptive, more conceptual, more emotionally complex?
    - **Disclosure depth**: do participants open up as much to an AI as they would to a skilled moderator?
    - **Coherence of insight**: can AIMIs surface the same actionable themes that human interviewers identify?

    That's why our approach is not about proving that AIMIs are "better" than IDIs, it's about identifying where they align, where they diverge, and where a hybrid model might unlock new value.

- **Voice vs. text formats**: initial signals suggest that voice interviews elicit richer, more natural responses, especially in younger or low-literacy audiences. Text, on the other hand, may perform better in sensitive topics or time-constrained environments. We aim to run systematic comparisons to understand where each format fits best.

As we continue this work, we're not just looking to validate what we've built: we're looking to challenge it. We welcome feedback from researchers, clients, and practitioners on what else AIMIs should be tested against, where they may fall short, and which questions we should be asking next.