# Enhancing Qualitative Market Research with Conversational AI-agents: Comparative Study of AI-Moderated Interviews (aka “AIMIs”) and Traditional Static Surveys

Giovanni Maria Occhipinti, Glaut

## Abstract

This study investigates the efficacy of an AI-moderated interviewer in eliciting richer and more insightful qualitative data compared to traditional web-based static surveys. We conducted a comparative experiment with two balanced groups of 100 participants, employing Glaut, a novel AI-moderated voice interview platform, and a widely used survey builder platform. The analysis, leveraging a large language model for thematic analysis and transcript quality assessment, reveals that AI-moderated interviews (aka “AIMI”), incorporating voice interaction and dynamic follow-up questions, yield significantly higher word counts, a greater number of extracted themes, and are overwhelmingly preferred for transcript quality. The results firmly indicate that voice interaction and the inclusion of follow-ups significantly enhance the richness and depth of the data collected, without diminishing user satisfaction. Furthermore, the AI-moderated interview approach demonstrates a markedly higher completion rate, particularly when considering the proportion of valid and meaningful responses relative to the total number of initiated surveys. These findings highlight the potential of AI-moderated interviews in revolutionizing and scaling qualitative data collection, though further research with larger, more diverse samples across various research contexts is warranted to confirm the generalizability of these results.

**Keywords:** AIMI, AI-moderated interview, Conversational AI, HCI, Qualitative Research,, Large Language Models

## 1. Introduction

In today's rapidly evolving digital landscape, artificial intelligence (AI) is transforming a wide array of industries, fundamentally altering how businesses operate and make decisions. One area where AI's impact is profoundly felt is in market research. Traditional methodologies often struggle to keep pace with the increasing need for nuanced, high-quality insights that can be obtained at scale.

Glaut mission is to understand people beyond the numbers. Glaut AI-native market research software is designed to act as an experienced qualitative researcher, collecting deeper and more detailed responses from respondents and analyzing them accurately and efficiently.

This research aims to explore the efficacy of AI-moderated interviews (aka “AIMI”), built upon large language models (LLMs), in enhancing market research processes. It serves as a critical step towards validating our approach and understanding the potential benefits and limitations of using conversational AI in market research. By comparing the results of AIMIs with traditional static surveys, this research aims to uncover whether AI-moderated research can truly enhance the quality and depth of insights obtained, thereby offering a superior alternative to conventional methods.

The core hypothesis driving this research is that an AI agent capable of interacting with respondents through spoken language, understanding open-ended responses, and providing contextual follow-ups when responses are incomplete, will stimulate richer and more detailed answers compared to traditional static surveys thought for qualitative research, such as those conducted via traditional survey providers. Thus, this study aligns with the comparative research domain explored by Villalba et al. (2023).

To test this hypothesis, this research conducted an experiment involving two equal and balanced samples of Italian participants. One group of 100 individuals completed a traditional static survey hosted by a globally known survey builder platform, featuring both closed and open-ended questions. The other group of 100 individuals participated in an AIMI, which included the same closed and open-ended questions, created and executed with Glaut AIMI.

The collected responses were compiled into two distinct datasets. Each row was enriched with multiple metrics, including total word count, number of extracted themes, completion time, user experience rating, voice usage annotation, and a response quality score. Additionally, this research included the results of a comparison between the interview transcripts generated by the two methodologies and a classification of transcripts as either "gibberish" or "not gibberish."

To maintain consistency in thematic analysis, both research types (AIMIs and surveys) were processed using the same language model (GPT4o) with standardized prompts for theme extraction and entity coding. Another model (Gemini 1.5 fast) was then used to compare interview transcripts concerning their overall quality and to categorize them as either gibberish or meaningful.

The following sections of this paper will detail the experimental design, including hypotheses, survey questions and topics’ rationale, sample quotas, settings, and tools. We will also discuss the market research topics addressed, the datasets collected, and the results of our analysis. Ultimately, we aim to provide a comprehensive understanding of how conversational AI can be harnessed to transform market research and deliver unparalleled qualitative insights at scale.

## 2. Experiment Design

### 2.1 Hypotheses

The primary hypothesis is that an AI agent, capable of conversing, understanding open-ended responses, and making contextual follow-ups, will extract richer, more detailed, and higher-quality answers compared to a traditional static survey.

Specifically, this research wants to prove whether the AI-moderated interview conducted through Glaut outperforms or not the traditional survey platform in several key areas:

- Generating a higher word count.
- Identifying a greater number of themes.
- Achieving a higher user experience rating due to its interactive nature.
- Earning preference from the LLM in the comparison of interview transcript quality.
- Resulting in fewer gibberish answers and transcripts.

In these cases, we hypothesize that higher performances in the AIMI methodology would be due to the combined effects of using voice interactions and the presence of contextual follow-ups. The ability of the AI agent to engage respondents through spoken dialogue would likely make the interaction more engaging and natural, encouraging respondents to provide more detailed and comprehensive answers. Additionally, AIMI capacity to offer contextual follow-ups for incomplete or vague responses ensures that valuable insights are not lost, further enhancing the depth and richness of the data collected. These factors, absent in the traditional static survey, are believed to significantly contribute to the superior performance of the AI-moderated interviews.

### 2.2 Chosen performance evaluation metrics.

To evaluate the performance of Glaut and traditional methodologies, we selected several key metrics that provide insight into different aspects of survey effectiveness and user experience.

1. **Completion Rate:** This metric measures the percentage of valid, non-gibberish completed interviews relative to the total number of surveys initiated for each methodology. It excludes transcripts containing gibberish responses, which are indicative of attempts by respondents to quickly complete the survey and secure a prize without providing meaningful answers. A higher completion rate could suggest a more engaging and user-friendly experience, as respondents could be less prone to cheat and more inclined to complete the survey.
2. **Number of Words per Respondent**: We assessed the number of words per each completed survey, only counting the words in open-ended responses, for each methodology. While a greater word count does not necessarily correlate with more information, it helps determine which methodology yields richer linguistic data.
3. **Number of Themes per Respondent**: This metric involves thematic analysis of open-ended responses to identify the number of distinct themes or codes extracted.

A higher number of themes suggests a richer, more detailed, and more comprehensive extraction of information. The thematic analysis was performed for both data samples using the Glaut encoder for themes, operating upon the Large Language Model GPT4o, with consistent instructions across both methodologies to ensure comparability.

4. **User Engagement Rating per Respondent**: We measured user satisfaction through a direct engagement rating question, asking respondents to rate their experience on a scale from 1 to 10. This metric provides an assessment of how engaging and satisfactory each survey experience was perceived to be by the participants.
5. **Comparison of Transcript Quality:** The analysis involved comparing the quality of transcripts from interviews and surveys using an instructed large language model (*Gemini-1.5-fast*). The LLM was guided by carefully crafted prompts that specified the criteria for evaluation, including adequacy with respect to context, meaningfulness of answers, and the depth and richness of responses. To ensure a fair and unbiased comparison, transcripts were ranked by length, and the quality comparison was conducted pairwise between transcripts of the same rank. This approach minimized the risk of biased judgments that could arise from randomly pairing transcripts of significantly different lengths.
6. **Percentage of Gibberish Transcripts:** This metric reflects the proportion of transcripts with at least one gibberish response. An instructed LLM (*Gemini-1.5-fast*) evaluated each answer, classifying the entire transcript as gibberish if any response was flagged. This stringent approach reflects the researchers' need for reliable data, as even one nonsensical answer can compromise the entire interaction, undermining the integrity of the research findings. This metric is mainly required to normalize the completion rate in accordance with the updated numbers of completed and valid interviews and surveys.

In particular, these three last metrics are crucial for addressing a common challenge in market research: poor-quality responses, especially prevalent when panelists are compensated. Researchers frequently encounter nonsensical answers (e.g., random letters), serial responses, off-topic replies, or evasive responses like "I don't know." These last three metrics are crucial in evaluating which methodology effectively mitigates this issue, encouraging respondents to engage more thoughtfully and seriously with survey questions.

In summary, the completion rate and user engagement rating serve as measures of user experience and effectiveness in retaining respondents, while the number of words, the number of themes extracted, and the "quality metrics" reflect the adequacy, significance, depth, and richness of the data collected, indicating how well each methodology stimulates detailed answers. This comprehensive evaluation allowed us to assess both the quality of the insights gathered and the overall respondent experience.

## 2.3 Addressing concerns and substantiating methodology: leveraging LLMs for consistent thematic analysis and evaluating responses' quality

Using a Large Language Model (LLM) for thematic analysis and responses' quality evaluation might raise concerns about relying on a "black box" and losing control over variable conceptualization and computation. Critics might argue that this approach sacrifices transparency and precision. However, we counter this by emphasizing that, while an LLM may introduce bias in its linguistic judgments, this bias would be consistently applied across all responses, ensuring a uniform and balanced analysis. Furthermore, given the advanced cognitive capabilities of state-of-the-art language models, entrusting an LLM with these tasks is almost comparable to relying on a human researcher's evaluation based on predefined frameworks.

In particular, studies support that thematic analysis performed by Large Language Models (LLMs) is comparable with human-generated TA because the LLMs achieve coding quality similar to human coders, while reducing labor and time demands (Dai, Shih-Chieh et al., 2023). The implementation of LLMs for thematic analysis is defended also by Paoli, who underlines the ability of Large Language Models in effectively perform an inductive Thematic Analysis on semi-structured interviews, inferring most of the main themes from previous research, thus demonstrating the viability and a good degree of validity of using this approach in qualitative research (Paoli, Stefano De., 2023).

Analog evidence surrounds the open-text quality evaluation task. As it is stated in the work of Chiang (Chiang et al., 2023) LLMs can effectively evaluate texts by presenting them with the same instructions, samples, and questions used in human evaluation.

Finally, GPT-based methodologies outperforming other approaches in short text classification tasks (such as sentiment analysis) through prompt engineering, encourages us to implement generative LLMs to categorize interviews and surveys responses as "gibberish" or "not gibberish" (Kheiri, K., & Karimi, H., 2023).

As a result, we have defined "theme", "gibberish response", and "transcript quality" primarily as textual phenomena and features emerging from text and labeled examples, according to the LLM internalized linguistic competence and judgment.

Therefore, regarding the concept of "theme", we allowed the LLM significant freedom to apply its own implicit conceptualization of the phenomena. This approach was guided by a straightforward prompt, which instructed the LLM to analyze survey responses and identify key themes. The process involved careful reading of each response, considering the underlying message, emotions, and context. The LLM was then asked to identify and list the main themes, summarizing each in a few words or a short phrase, and to provide a brief explanation for each theme, detailing why it was categorized as such and what aspects of the responses led to this conclusion. Incorporating an explanation step in our methodology is essential for enhancing the model's mindfulness about its own output, and to reinforce its accountability. The output was formatted to include the theme title, a brief explanation, and example responses that fit the theme, ensuring a structured and consistent analysis.

Regarding the ‘responses’ quality’ definition, the rationale communicated to the LLM for scoring answers was based on three primary criteria:

1. **Adequacy to Context:** This criterion assesses whether the response directly addresses the question and remains relevant and on-topic. A response that diverges from the context or fails to engage meaningfully with the prompt receives a lower score.
2. **Depth/Richness:** This measures the extent to which the response provides detailed, thoughtful insights and demonstrates a deep understanding of the topic. Responses that include multiple perspectives, examples, or in-depth reasoning are ranked higher.
3. **Presence of Serial Cheating Responses:** This criterion identifies instances where respondents provide serial nonsensical, out-of-context, or purposely evasive answers, often as an attempt to quickly complete the survey without genuine engagement. The presence of responses of this type reduces the overall quality score of the transcript.

These criteria were used to guide the LLM selection during the pairwise transcripts comparison. In this case, each AIMI transcript originated on Glaut was evaluated in the same run, paired with the corresponding traditional survey transcript.

For the evaluation of gibberish answers, the prompt defined the category of “gibberish” as responses that are “nonsensical,” “meaningless,” or “out of context”. Examples for each sub-category are provided to guide the model in alignment with a few-shot learning approach. This approach leveraged the LLM’s internalized linguistic competence, allowing it to assess responses within the provided semantic framework and context.

## 2.4 Research questions and topics’ rationale

To ensure comparability between the AIMI and the static survey, both questionnaires were designed identically, or at least as similar as possible. All follow-up questions that a researcher might have included in a traditional survey *a priori* were embedded within the static survey’s questions. This design ensures that any additional information gleaned by Glaut’s AI agent is due to its real-time, context-sensitive interaction rather than the presence of additional questions. The rationale behind this design is to evaluate the fundamental benefit of the conversational AI approach over traditional static surveys.

The survey focused on the themes of trust and loyalty towards brands. These topics were chosen because they elicit rich, qualitative data that can be challenging to capture with static survey methods. Plus, this kind of research represents a very common use case in the market research industry. The survey questions were the following. Consider that open-ended questions are identified with their dataset ID (e.g. “*Q1*”).

1. Let's start: you are…?

- A male
- A female
- Other/non-binary or prefer not to answer

2. And how old are you?

- [age: number]

3. In which region do you live?

- [options: Italian regions]

4. How many inhabitants does the city where you live have?

- Up to 10.000
- From 10.001 to 30.000
- From 30.001 to 100.000
- From 100,001 to 500.000
- Over 500,001

5. Let's talk about trust in general. What is trust for you? (Q1)

- [open-ended response]

6. Now think about trust in a brand. When is a brand reliable? (Q2)

- [open-ended response]

7. Among the products you buy at the supermarket, which brand do you think is very reliable? (TRUSTED)

- [open-ended response]

8. Why do you think this brand is reliable? (Q3)

- [open-ended response]

9. Which brand have you lost trust in recently? (UNTRUSTED)

- [open-ended response]

10. And why did you lose trust in this brand? (Q4)

- [open-ended response]

11. In general, how much trust do you have from 1 to 10 in companies that produce animal-based products (like cold cuts, meat, etc.)?

- [rating: number]

12. Why did you assign this value? (Q5)

- [open-ended response]

13. And how much trust do you have from 1 to 10 in companies that produce coffee?

- [rating: number]

14. Why did you assign this value? (Q6)

15. What is your profession?

- Entrepreneur/freelancer
- Artisan/shopkeeper
- Manager/Executive
- Employee
- Worker
- Housewife
- Student
- Retired
- Unemployed/looking for a job

16. And your level of education? (only one answer possible)

- None/Elementary school
- Middle school
- High school diploma
- University/Postgraduate degree

## 2.5 Sample quotas

Two equivalent samples of 100 respondents each were selected to participate in the study. These participants were balanced in terms of demographics to ensure the comparability of results. One group completed the AIMI on Glaut, while the other group completed the traditional survey. This equal sample size and demographic balance are critical for minimizing bias and ensuring the validity of our comparative analysis.

In particular, we balanced the samples according to the following variables and quotas:

| GENDER | quota |
|---|---|
| Female | 49.9% |
| Male | 50.1% |

| AGE | quota |
|---|---|
| 18-24 | 9.7% |
| 25-34 | 18.5% |
| 35-44 | 22.6% |
| 45-54 | 20.7% |
| 55-64 | 15.3% |
| over 64 | 13.2% |

| EDUCATION LV | quota |
|---|---|
| Middle/High school | 60% |
| University | 40% |

| JOB | quota |
|---|---|
| Employed | 73% |
| Unemployed | 27% |

| GRG AREA (Italy) | quota |
|---|---|
| North-west | 26.9% |
| Nord-east | 19.6% |
| Centre | 19.9% |
| South and Islands | 33.7% |

| N RESIDENTS | quota |
|---|---|
| 1-10.000 | 26.9% |
| 10.001-30.001 | 20.8% |
| 30.001-100.000 | 21.4% |
| 100.001-500.000 | 14.7% |
| >500.000 | 16.2% |

## 2.6 Settings and Tools

The AIMI was conducted using the Glaut AI voice-enabled interview platform, capable of natural language processing and contextual follow-ups when the responses are vague, incomplete or inadequate. Instead, the traditional static survey was implemented through a widely used survey builder application. Both surveys contained the same closed and open-ended questions. Both surveys were designed to be user-friendly, ensuring that respondents could complete them without undue difficulty.

The AIMI on Glaut was configured in-app to perform contextual follow-ups (*“what loops”*) only if needed, thus if the response was inadequate, out of context, or vague. Plus, the AIMI was instructed to maintain a correct, inclusive, and tolerant behavior when interacting with users.

Respondents for both samples were contacted by a panel service provider and received the link to the survey and to the AIMI. We periodically monitored the progress toward our pre-fixed sample quotas through redirect links activated after the completion of the surveys.

Both the survey and the AIMI were completed remotely, using smartphones or personal computers. The AIMIs were hosted on the Glaut platform, while the traditional static survey was hosted on the survey builder platform.

## 2.7 Datasets

The dataset comprises survey responses collected from participants across various demographic and regional categories. Each entry includes multiple variables, such as gender, age, region, city size, occupation, and education level, providing a comprehensive view of the respondent's background.

Each entry captures, also, the key metrics mentioned above for all the open-ended questions (Q1 to Q6, TRUSTED, UNTRUSTED): the number of words in the

response (N_W_[question ID]), the number of themes identified in the response (# T_[question ID]), the number of follow-ups required (# follow-ups [question ID]), and the classification as “gibberish” or “not gibberish”. Plus, for each respondent, the user experience rating was registered. Finally, for each survey entry, we computed the total number of words, the total number of themes, and the completion time in seconds.

Each row is further enriched with categorical annotations indicating key variables: the mode of completion used (“Glaut” vs. “Traditional Survey”), the classification of the transcript as “gibberish” or “not gibberish,” and the outcome of the quality-based comparison (“better” if the transcript was deemed superior, “worst” if it was not).

## 3. Results and Analysis

### 3.1 Computing Metrics and Preliminary Analysis

As we stated above, to evaluate the performance of the AIMIs on Glaut against traditional surveys, we computed several key metrics (‘*User Experience Rating’*, *‘Total Words per Respondent’*, ‘*Total Themes per Respondent’*), summarized through their averages and their medians in the tables below. Additionally, we recorded the completion rate for both Glaut and traditional methodologies and adjusted it subtracting the number of gibberish transcripts from the total count of completed interactions. Finally, we included in the table also *the ‘Percentage of Preferred Transcripts’* and the *‘Percentage of gibberish transcripts’* for both methodologies over the total of transcripts. These metrics provided an initial overview of the performance differences between the two methodologies. However, to determine whether these differences are statistically significant or simply due to chance, we proceeded with a formal hypothesis testing process in the next step.

| Performance metric | AIMI | Survey | Δ (%) |
|---|---|---|---|
| **Avg ‘Rating’** | 8.48 | 8.03 | 5.6 |
| **Mdn ‘Rating’** | 9.0 | 9.0 | 0.00 |
| **Avg ‘no. words/ respondent’** | 71.97 | 31.42 | 129.1 |
| **Mdn ‘no. words/ respondent’** | 63.0 | 31.0 | 103.2 |
| **Avg ‘no. themes/ respondent’** | 8.23 | 6.94 | 18.6 |
| **Mdn ‘no. themes/ respondent’** | 8.0 | 7.0 | 14.3 |
| **‘Percentage of Preferred Transcripts’ (%)** | 66 | 34 | 94.1 |
| **‘Percentage of Gibberish Transcripts’ (%)** | 26 | 56 | -53.6 |
| **‘Completion Rate’ (%)** | 61 | 39 | 56.4 |

**Table 1**: Performance Comparison of AI-moderated Interviews and Traditional Surveys

### 3.2 Preliminary Results Analysis and Interpretation

The preliminary overview of performance metrics presents that the AI-moderated interviews (AIMIs) performed on Glaut platform outperform the static surveys in terms of stimulating detailed and rich

responses while ensuring an overall better user experience and capacity of retaining engaged and thoughtful respondents.

User Experience and Engagement:

- **Average Rating**: The AIMI received a higher average rating (8.5) compared to the static survey (8.0), with a percentage difference of 5.6%. This could suggest that participants found the AI-moderated survey more engaging or satisfactory.
- **Completion Rate**: The static survey has a lower completion rate, considering the amount of invalid transcripts (38%) compared to the AIMI (61%), with a difference of 56.4%. Despite the AIMI approach being entirely novel to users compared to the well-established and familiar static approach, the Glaut methodology demonstrated a superior completion rate while effectively securing high-quality data.

Quality, Depth, and Richness of Data Collected:

- **Average Number of Words per Respondent**: Glaut participants produced an average of 71.97 words per interview compared to 31.42 words in the traditional surveys. This significant percentage difference of 129.1% could indicate that the AIMI elicited longer responses and, thus, it produced way more linguistic data.
- **Average Number of Themes per Respondent**: The AIMI extracted an average of 8.23 themes per interview, compared to 6.94 for the static survey, with an 18.6% difference. This could suggest that Glaut's ability to interact through speech and provide contextual follow-ups results in a deeper exploration of topics and in richer qualitative data.
- **Percentage of Preferred Transcripts**: The AIMI were considered better based on their quality in 66% of cases, securing a performance increment of 94.1%.
- **Percentage of Gibberish Transcripts:** Among the AIMI, only 26% were categorized as "gibberish," whereas 56% of the traditional surveys were deemed "gibberish." This metric highlights that the AI-moderated approach produced gibberish interactions 53.6% less frequently than the static surveys.

To establish the statistical significance of the influence of voice and follow-ups on performance metrics, it is necessary to implement a hypothesis testing procedure. This final analytical step will be detailed in the following section.

## 3.3 Hypothesis testing: variables and methodologies

To evaluate our hypotheses, we conducted a series of statistical tests to analyze the impact of the Mode of Completion, which contrasts a voice-based conversational experience against a static, traditional survey format. The dependent variables assessed were:

- User Experience Rating (continuous)

- Total Number of Themes per Respondent (continuous)
- Total Number of Words per Respondent (continuous)
- Transcripts' Comparison Result (categorical, levels: "better", "worst")
- Transcripts' Classification Result (categorical, levels: "gibberish", "not gibberish").

The Mode of Completion served as the independent variable, classified as "Glaut" (AI-moderated interviews) or "Traditional Survey" (common surveys). In this way, we tested whether the conversational approach, characterized by AI-driven follow-ups to deepen vague or hurried responses, significantly influenced performance metrics compared to the traditional static method.

The analysis began by exploring data distributions and central tendencies. To determine the appropriate statistical tests, we first assessed the normality of the data using the Shapiro-Wilk test and checked for homogeneity of variances using Levene's test. Where assumptions for parametric testing were violated, we opted for non-parametric Mann-Whitney U tests to compare the groups. For the categorical dependent variables, we applied the chi-square test to examine differences between modes in the absolute counts of results concerning transcripts' quality comparison and transcripts' classification as gibberish or not. Since multiple comparisons were made, we employed the conservative Bonferroni correction to adjust the significance level, ensuring robust evaluation of the Mode's impact on the performance metrics. Therefore, we interpreted results comparing the obtained p-values with the corrected significance threshold of $\alpha = 0.01$ to draw conclusions about the hypotheses. This approach allowed us to rigorously test whether any observed differences were truly statistically significant and not due to random chance.

## 3.4 Hypothesis testing findings

First, we examined the influence of our independent variable, 'Completion Mode' on the continuous performance metrics. In the graphs, Glaut AIMIs are showed in pink, and traditional surveys in blue.

1. For the relationship between the mode of completion and the 'User Experience Rating' the non-parametric Mann-Whitney U test yielded a p-value of 0.2218. Since this result exceeds the corrected threshold of 0.01, we reject the alternative hypothesis and conclude that the mode of completion does not negatively impact the user experience rating.


Figure 1: Distribution of User Experience Ratings, Grouped by Mode of Completion

2. Concerning the relationship between the mode of completion and the 'Total Number of Themes per

Respondent' the Mann-Whitney U test produced a p-value of 0.0002, indicating a statistically significant difference between the groups. This result allows us to accept the alternative hypothesis, demonstrating that the AIMI leads to a significantly higher number of themes per respondent compared to the static mode.

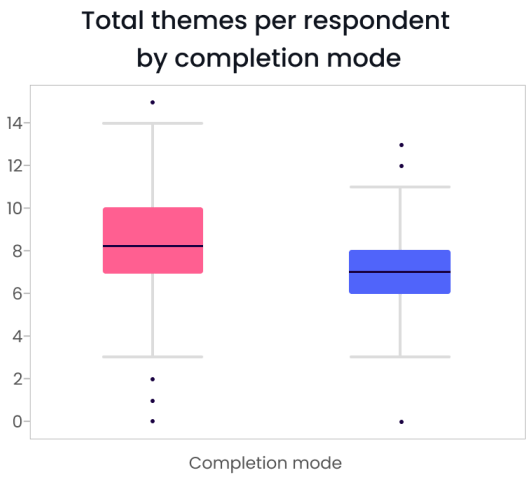


Figure 2: Distribution of Theme Counts per Respondent, Grouped by Completion Mode

3. Regarding the influence of the mode of completion on the 'Total Number of Words per Respondent', the Mann-Whitney U test returned a p-value of 5.3e-16, well below the significance threshold. This strongly supports the alternative hypothesis, showing that respondents using the AIMI provided a significantly higher word count than those using the static survey.

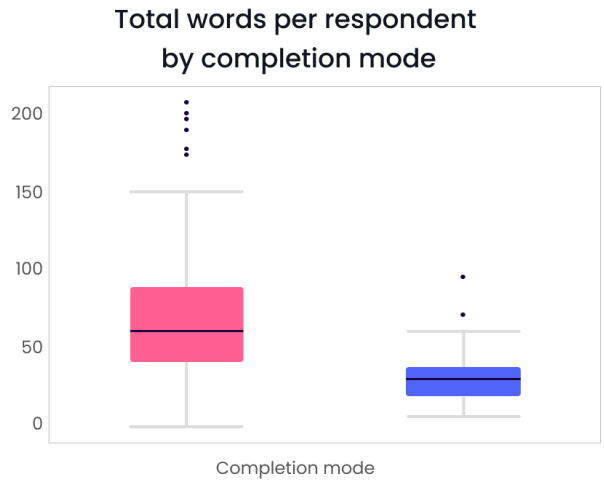


Figure 3: Distribution of Word Counts per Respondent, Grouped by Completion Mode

Next, we assessed the impact of the mode of completion on our categorical dependent variables.

4. For the 'Transcripts' Comparison Result', the Chi-Square test of independence resulted in a p-value of 0.00001, which is less than the corrected threshold of 0.01. This significant result leads us to accept the alternative hypothesis, confirming that the AIMI significantly influences the likelihood of producing better-quality transcripts.

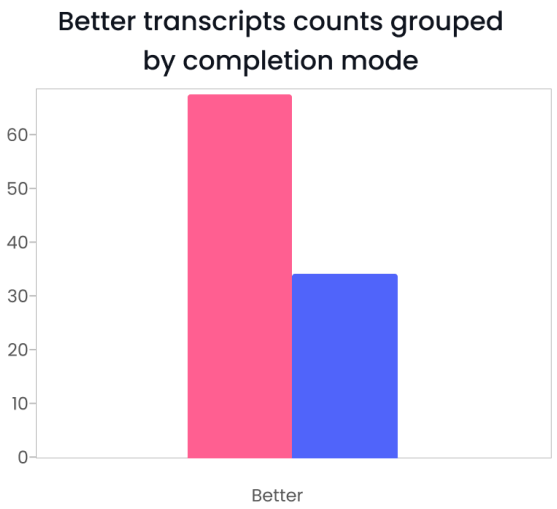


Figure 4: Better Transcripts by Quality Absolute Counts, Grouped by Completion Mode

5. When evaluating ‘Transcript Classification Result’ the Chi-Square test yielded a p-value of 0.00003. This result supports a statistically significant association between the mode of completion and the likelihood of producing non-gibberish transcripts, indicating that the AIMI lead to significantly cleaner and more meaningful responses compared to the static survey.

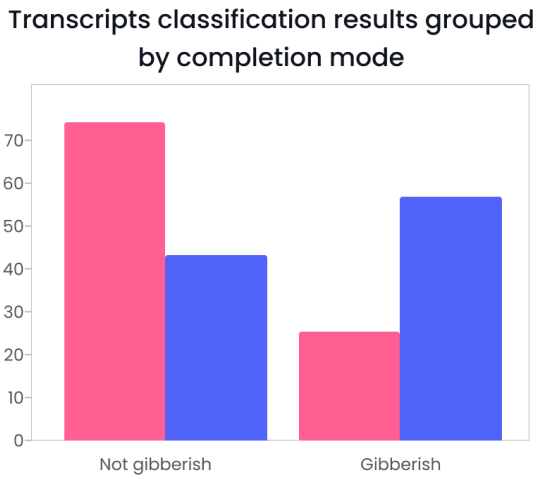


Figure 5: Transcript Classification Absolute Counts, Grouped by Completion Mode

These findings demonstrate that the AIMI approach positively influences the richness and quality of the data collected without compromising user experience ratings.

## 4. Conclusion and Future Directions

This research provides evidence that AI-moderated interviews, or AIMIs, leveraging both voice interaction and dynamic follow-up questions, offer significant advantages over traditional static survey methods. Our findings demonstrate that the Glaut AI-native platform, despite its novelty, successfully elicits richer, more detailed responses and facilitates deeper exploration of thematic areas compared to established survey builders.

Specifically, the use of voice interaction and follow-ups significantly increases the number of words and themes extracted from participants’ responses, highlighting its potential to unlock deeper layers of qualitative data. Furthermore, without compromising user experience ratings, the incorporation of voice and follow-up questions significantly elevates the quality of the interaction, leading to a clear preference for AI-moderated transcripts.

However, to solidify the generalizability of these findings, future research should prioritize:

- Increased sample sizes and diverse research settings: evaluating the performance of AIMIs with larger and more diverse participant groups across a range of research contexts beyond brand research. This would involve replicating this study with different research goals, potentially including volunteers from various demographics and backgrounds.

- Cross-comparison across multiple domains: directly comparing the efficacy of AIMIs and static surveys in diverse fields such as healthcare, education, and social sciences. This would provide a more comprehensive understanding of the strengths and limitations of each approach across different data collection needs and participant profiles.
- Diversifying respondent pools: expanding participant recruitment beyond paid panel providers to include individuals with diverse motivations and levels of engagement. This will offer insights into how different populations interact with AIMIs and ensure the generalizability of findings across broader user groups.

Furthermore, by implementing different and tailored experimental configurations, we could isolate and measure the specific impact of the key features defining the AI-moderated conversational methodology - namely, the use of voice interaction and the ability to deepen responses through contextual follow-ups. Future studies could, therefore, provide a more precise evaluation of how these elements contribute individually to the overall effectiveness of the AI-driven survey experience.

Finally, by broadening the scope of the investigation, accumulating evidence across diverse contexts, and measuring the specific contributions of AI-moderation on a research project outcome, we can establish a more robust foundation for understanding the full potential of AI-moderated interview methodologies. This will pave the way for their widespread adoption and optimization across a variety of research and data collection endeavors.

## 5. References


1. Villalba, A.C., Brown, E.M., Scurrell, J.V., Entenmann, J., & Daepp, M.I. (2023). Automated Interviewer or Augmented Survey? Collecting Social Data with Large Language Models. *ArXiv, abs/2309.10187*.
2. Chiang, C., & Lee, H. (2023). Can Large Language Models Be an Alternative to Human Evaluations? *Annual Meeting of the Association for Computational Linguistics*.
3. Paoli, S.D. (2023). Can Large Language Models emulate an inductive Thematic Analysis of semi-structured interviews? An exploration and provocation on the limits of the approach and the model. *ArXiv, abs/2305.13014*.
4. Dai, S., Xiong, A., & Ku, L. (2023). LLM-in-the-loop: Leveraging Large Language Model for Thematic Analysis. *Conference on Empirical Methods in Natural Language Processing*.
5. Kheiri, K., & Karimi, H. (2023). SentimentGPT: Exploiting GPT for Advanced Sentiment Analysis and its Departure from Current Machine Learning. *ArXiv, abs/2307.10234*.