CONVERSATIONS WITH AI: PART II

Unveiling AI quality in qualitative workstreams

September 2023

AUTHORS

Jim Legg

Ajay Bangia

IPSOS VIEWS

AI SERIES

GAME CHANGERS

We're excited, as a rising tide of change has the power to lift all boats, making us more productive and creative, and providing access to new tools and forms of expression. ”

WAVES OF TRANSFORMATION

It's clear that the next major sea change hitting our shores is associated with how we work with generative artificial intelligence (AI). The potential use cases are no longer abstract concepts or distant possibilities on the horizon. The latest wave of AI tools is user-friendly and can perform tasks such as transcribing audio or video files, summarizing content, generating images and videos, and even writing code. A feeling of nervous excitement arises from the fact that generative AI applications possess the capability to formulate seemingly original content from basic prompts.

We're excited, as a rising tide of change has the power to lift all boats, making us more productive and creative, and providing access to new tools and forms of expression.

But we're also nervous, as crucial gaps persist in the realm of generative AI. These systems are trained on extensive datasets that inherently encode biases from the source material. They are susceptible to errors and hallucinations, some of which are apparent while others are subtle.

Sailing blindly into the waves is unwise. It is up to us to be cautious and thoughtful in how we use these technologies now and how we consider their implications for the future.

MEASURING THE WAVES

In the realm of qualitative research, the integration of AI holds the potential for a transformative wave, poised to revolutionize traditional workstreams.

Embracing the inevitable disruption, engaging in well-designed and strategic research-on-research allows us to skillfully navigate the waves, gauging their speed and intensity. A fitting metaphor for AI in market research is the art of timing, raising the sails at the opportune moment to harness the winds of progress. Through studying the impact of AI on qualitative research, we equip ourselves to embrace the forthcoming changes and devise strategies to harness the power of this technology to our advantage. While the storm may challenge us, with astute navigation and a comprehensive grasp of the risks and opportunities, we have the resilience to weather the storm and emerge stronger and more effective in our research endeavors.

To set the stage for this conversation, it is crucial to emphasize the importance of thoroughly assessing these tools within specific use cases to unlock their maximum value while mitigating potential risks. Existential questions about how these tools may change society and how we work are beyond the scope of this brief paper; we focus here on practical questions about how to evaluate their utility in practice. Although there are some unique new aspects, the dimensions overlap with how we have evaluated analytic AI & machine learning in the past. Ipsos classifies these into the domains of Truth, Beauty, and Justice.

At Ipsos we evaluate AI tools using the criteria of Truth, Beauty, and Justice:

TRUTH

This domain focuses on the accuracy of the models and their outputs; examining their quality and avoiding hallucinations or false fabrications.

BEAUTY

The most important aspect of Beauty in AI focuses on the explainability of its output. Some use cases also include a model's ability to surprise and generate new insights.

JUSTICE

This domain encompasses multiple important areas – AI ethics, algorithmic fairness, data security, privacy, alongside the rights and responsibilities of creators of data used for training and by users of the models.

We used three criteria to evaluate the risk associated with AI outputs:

1. How will a set of AI suppliers, using primarily qualitative research platforms, score in transcription, translation, and sentiment (human vs. machine)?
2. Will poorer quality in AI machine transcriptions lead to poorer AI-generated summaries or analysis?
3. How will the quality of prompts impact the AI-generated summaries or analysis?

We started with five countries: the US, Mexico, China, Thailand, and Denmark. We selected these countries based on language, choosing three mainstream and two less common languages to test. Each country delivered client reports (in PowerPoint) for two different categories, as well as the raw data collected for each report (video focus group videos or online community discussion board transcripts). For the videos, we engaged a third-party, Ipsos-approved transcription and translation supplier and had humans transcribe the first 5,000 words of each video. We also had them translate the first 5,000 words of all non-English language transcripts into English. We then evaluated how seven different AI or AI-assisted platforms (all leaders in the space) compared.[1]

All of our tests were conducted within Ipsos Facto: Ipsos' own internal generative AI platform. It was designed to enable Ipsos to work with generative AI in a secured environment, keeping our clients' and Ipsos' IP and data private.

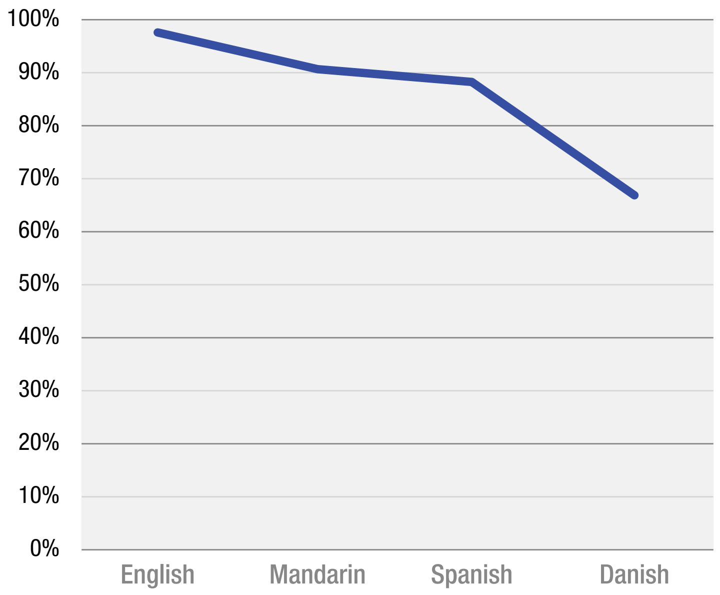
THE IMPACT ON TRANSCRIPTIONS

For **transcriptions**, we leveraged a credible Word Error Rate (WER)[2] Utility for our assessment. To compute WER, a machine-generated transcription is compared to a benchmark human-generated "ground truth" transcription. The WER is the number of errors found when comparing the ground truth to the machine transcription, divided by the total number of words in the ground truth. Consider the following example in Figure 1.

To better evaluate the impact of machine transcription errors, we took one additional step, creating a custom methodology that differentiates significant or insignificant errors. An example of an insignificant error in English would be: "gonna" vs. "going to." We then coded all errors and removed insignificant errors from the results.

Automatic speech recognition (ASR) systems performed well. That said, generative AI platforms and their association probability models exhibited remarkable multilingual capabilities, their performance varies significantly across different languages.[3] A discernible trendline emerged in the realm of transcription quality, revealing that machine transcriptions achieve higher accuracy with more widely used languages such as English. While certain languages are already approaching optimal levels e.g., Mandarin and Spanish, others may require additional time, potentially becoming ready for accurate transcriptions soon. Continuous testing and evaluation will play a crucial role in this process.

Figure 1: Lead transcription AI supplier's accuracy rate by language



Ground Truth: I went to the **...** store yesterday **and** bought apples, milk, **eggs**, and bacon.

Machine Transcription: I went to the **grocery** store yesterday, **...** bought apples, milk, **legs**, and bacon.

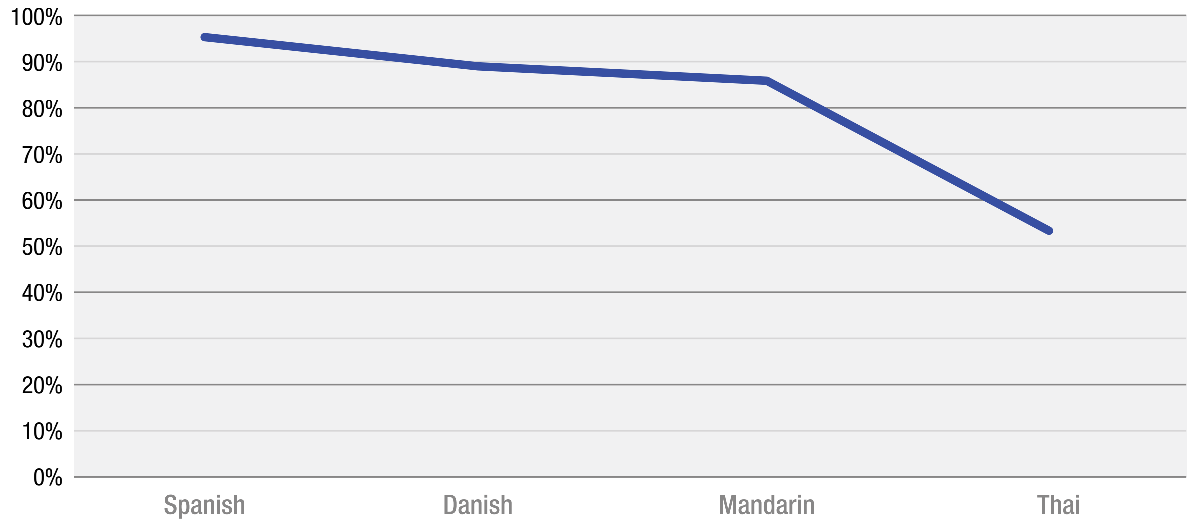
23%
Since the ground truth contains 13 words and there are 3 errors, the WER is approximately 23% (3 as a percentage of 13).

Note: Due to intermittent irregularities during data extraction, Thai was not included in the transcription assessment. Source: Ipsos UU

For **translations**, our Ipsos in-house translation team assessed critical error rates by applying the American Translation Association's (ATA) 'Meaning Transfer' framework.[4] The framework defined when meaning had changed and should be counted as an error.

We assessed: 1. audio-to-machine translation, 2. audio-to-machine transcription-to-machine translation, and 3. audio-to-human transcription-to-machine translation. The winning AI translation supplier performed with the highest accuracy when human transcriptions were used for the

Figure 2: Lead translation AI supplier's accuracy rate by language



Source: Ipsos UU

translation (see Figure 2). However, it's worth noting that one supplier achieved superior translation quality by directly translating from the audio, specifically for Spanish-to-English translations. In contrast, other languages showed better results when human transcriptions were used as the starting point. Therefore, except for Spanish, where the quality was sufficiently high for direct audio-to-text translation, human transcriptions are still necessary to achieve machine translation accuracy of 90% or higher in all tested languages.

Sentiment analysis consists of classifying the polarity of a given text. We chose to assess AI machine-coded sentiment at the sentence level. Custom data science techniques were deployed to split the transcripts into sentences for each of the languages. To evaluate the AI models, native speakers were asked to annotate each set of sentences (across all languages), categorizing each sentence as positive, negative, neutral, or unsure. Once the data was evaluated, the models were projected with a weighted F1-score per category, supplier, and language. The F1-score serves as a measure of the model's accuracy.

Some of the AI suppliers tested have generative Large Language Models (LLM) and are not specifically designed for sentiment analysis, so we compensated by employing iterative prompts to generate the desired outputs. Other suppliers tested are specifically trained for audio transcription and sentiment analysis. Surprisingly one LLM model not designed to code sentiment performed at almost the same level as one designed specifically for machine-coded sentiment. One perspective suggests that AI should make significant advancements in sentiment coding at the sentence level before it can operate independently (see Figure 3).

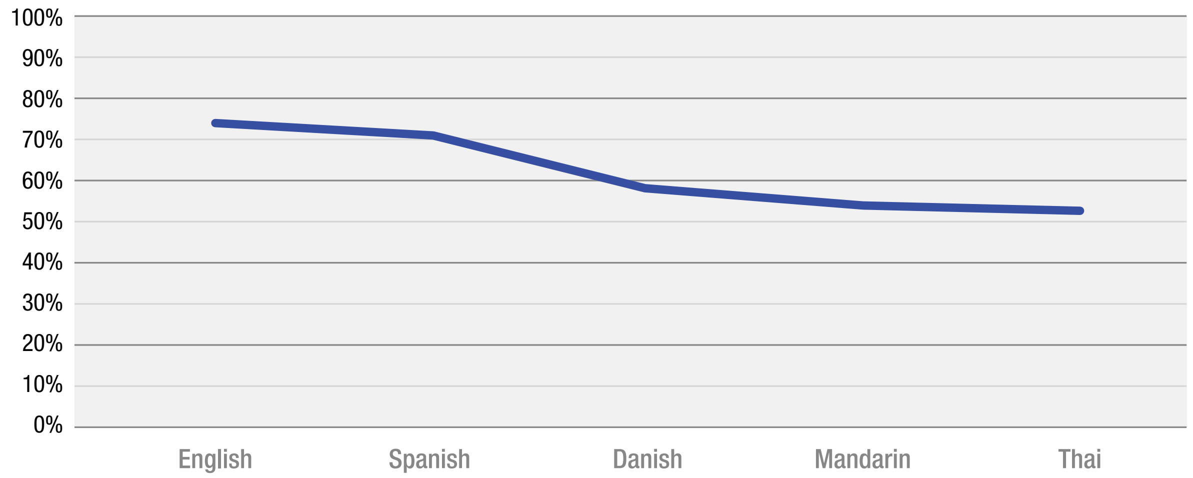
We tested the capability of generative AI in **thematic analysis** by sourcing reports from multiple categories and countries. An Ipsos expert prompt engineer with 20 years of research experience reverse-engineered the reports using custom prompts while working on AI-generated transcripts. The reports were compiled and evaluated by researchers who were not part of the original study, providing an unbiased perspective. Furthermore, a visual template and flow similar to the original report was used, and senior researchers evaluated blind copies of the original and AI-generated reports based on strategic parameters covering Business & Strategy, Insight & Storytelling, and Efficiency.

We found that generative AI provided a good topline summary, *akin to a novice researcher.* The AI suppliers tested lacked the ability to elevate insights to draw business implications. Even with expert-level prompts, the top AI supplier faced challenges with domain expertise related to research business questions and categories. While efficiencies are possible and undoubtedly coming, current token limitations require stitching micro-summaries together to produce quality, comprehensive AI outputs. Moreover, domain expertise gaps, possible biases, and hallucinations necessitate human intervention.

Overall, generative AI did better with:

- **Clear over ambiguous questions.** Our results indicated that AI performed well in answering clear and straightforward questions, such as how to improve the performance of a single ad analyzing qual research run on the advertising. However, when it came to broader questions, such as how to improve the overall campaign, the human moderator performed better by incorporating domain and communication expertise. Furthermore, AI demonstrated better performance in evaluating clear executional levers, like voiceover or music acceptance, but was less consistent when it came to more subjective dimensions,

Figure 3: Lead sentiment analysis AI supplier's accuracy rate by language



Note: The AI machine-coded sentiment assessment was applied to long-form transcripts. Results may vary when assessing more finite data sets like open ends from a survey.
Source: Ipsos UU

like relevance, which can be expressed differently by different consumers.

- **Summarizing vs. elevating insights.** We found that AI provided a strong summary of the research discussion against the prompt questions. However, the human researcher added significant value by making additional connections and elevating the insights. For example, the moderator was able to interpret the reactions to an ad and conclude that it inspired people to think of the brand in a different context, thereby opening a valuable new consumption occasion.
- **Writing.** The AI-generated summary was well-executed in terms of grammar and sentence structure, requiring no further proofreading. Furthermore, we did not observe any hallucinations or inaccurate conclusions in the summary within the finite qualitative research dataset organized by the researcher.

It goes without saying that the higher the quality of AI machine transcriptions, the tighter the AI summaries.

> When it came to broader questions, such as how to improve the overall campaign, the human moderator performed better by incorporating domain and communication expertise. ”

MASTERING THE WAVES, HERE AND ON THE HORIZON

Just as harvesting the energy of a storm can help sailors navigate turbulent waters, businesses can drive innovation and growth by adapting to and harnessing the potential of emerging technologies. AI is a prime example of such a technology that holds the power to fortify companies, making them more robust and adaptable in times of transformation.

HOW IS IPSOS MASTERING GENERATIVE AI?

- **Access for all:** Ipsos' continued investment in generative AI enables our teams to invent innovative applications within the privacy, security, and governance principles of Ipsos generative AI policy.
- **Prompt excellence:** We believe the secret is in the human-AI partnership, having the right iterative conversations and providing "coaching" via prompts. As prompt quality is one key driver of outcome quality, Ipsos has created prompt guidelines for optimal prompt engineering and will continue to evolve them.
- **Tuning with proprietary frameworks:** Based on our AI quality assessment results, Ipsos has already started developing a domain-specific prompt library. These will allow us to leverage our proprietary domain (research and category) knowledge and scientific framework intellectual property (IP) within our internal Ipsos chatbot.
- **Being trusted AI advisors to our clients:** Ipsos has been an active advocate of storytelling for a long time and has been providing training to its researchers on this skill for years. As AI automates basic tasks such as summarization, storytelling is becoming even more crucial. Generative AI would enable research teams to devote more time to crafting impactful stories, elevating insights, and activating findings to drive business outcomes.

Ipsos stands ready to help its clients **be sure** when navigating AI. We have a clear understanding of the quality from generative AI. We have assessed the risks and the timing around each wave coming towards us. Ipsos will continue to assess the waves on the horizon and is ready to raise the sails at the right moment so our clients can maximize the benefits generative AI delivers.

In our next Conversations with AI paper, we will be exploring the application of generative AI in augmenting creativity and divergent thinking in Ideation workshops.

AI QUALITY HEADLINES TO REMEMBER (AS OF SEPTEMBER 2023)

Initial Research Question	Ipsos Findings
How do AI suppliers and primarily qualitative research platforms score in transcription, translation, and sentiment (human vs. machine)?	• **Not all languages are created equal:** Automatic speech recognition (ASR) systems performed well. That said, generative AI platforms and their association probability models exhibited remarkable multilingual capabilities, their performance varies significantly across different languages. • **Fewer errors in widely used languages:** Machine transcriptions achieve higher accuracy with more widely used languages such as English and are approaching optimal levels with Mandarin and Spanish. Others - like Thai or Danish - may require additional time to reach acceptable accuracy levels for machine transcriptions. Continuous testing and evaluation will play a crucial role in this process. • **Different strokes for different folks:** Translation quality seems to be determined by how much each AI platform has been developed in each specific language. It's important to note that one AI supplier performed better in transcriptions while another had higher accuracy in translations. Ipsos Facto, Ipsos' own internal Generative AI chatbot, enables teams to leverage different foundational and fine-tuned models for different use cases (transcription vs. translation). • **Sentiment Decoding:** AI Still a Work in Progress. For AI machine-coded sentiment, AI has made significant strides, but still has some ground to cover to get closer to human-like performance.
Will poorer quality in AI machine transcriptions lead to poorer AI-generated summaries or analysis?	• **Poor roots, poor shoots:** We found that the more mainstream the language, the lower the Word Error Rates, the higher the summary quality.
How will the quality of prompts impact the AI-generated summaries or analysis?	• **Unlocking greater precision through targeted prompts:** Prompt quality impacted the quality of the output tremendously. A clear, superiorly crafted prompt, specifying the desired tone, had the ability to generate inspired responses. • **A novice in need of domain expertise:** There is also a clear gap of research and category domain expertise in the larger AI data corpus. Finally, the researcher who understands their clients' business and can connect insights to non-obvious opportunities will continue to be invaluable.

ENDNOTES

1. Ipsos's current efforts are focused on "universal" models or multi-lingual speech recognition and not on measuring the quality of monolingual models.

2. Worthy, B. "What Is Word Error Rate? Measuring the WER of Machine-Generated Transcripts and Its Limitations". Medium, 2 December, 2019. https://medium.com/@bethworthy/what-is-word-error-rate-measuring-the-wer-of-machine-generated-transcripts-and-its-limitations-1457be914f3b

3. Radford, A., Kim, J.W., Xu, T., Brockman, G., Mcleavey, C. and Sutskever, I. 2022. "Robust Speech Recognition via Large-Scale Weak Supervision." ArXiv, abs/2212.04356. https://cdn.openai.com/papers/whisper.pdf

4. American Translation Association. "Explanation of Error Categories: Section 2. Meaning Transfer." Accessed 8 September, 2023. https://www.atanet.org/certification/how-the-exam-is-graded/error-categories/

- **Cover page prompt:** Robot and human working together to raise the halyard on a sailboat (on *MidJourney*)

- **Anemometer prompt:** Closeup on a golden anemometer on a wooden sail ship in a dark storm, 4k, hyper realistic (on *Lexica*)

- **Waves 1 prompt:** large frothy waves on a rough sea. Lightning forking the stormy sky (on *MidJourney*)

- **Waves 2 prompt:** large frothy waves on a rough sea. Lightning forking the stormy sky viewed from a rocky beach (on *MidJourney*)

CONVERSATIONS WITH AI: PART II

Unveiling AI quality in qualitative workstreams

AUTHORS

Jim Legg, Global Operations Lead, UU, Ipsos

Ajay Bangia, Global Scale Lead, UU, Ipsos

CONTRIBUTORS

Dan Gahan, Candace Kozak, Manali Vivek, Karin O’Neill, Betsy Georgiton, Luana Lopez, Andres Martinez, Dominik Hugle, Daniel Lungu, Maria Mora, Aisya Adi, Jennifer Wade, Alex Tamvakis

The **Ipsos Views** white papers are produced by the **Ipsos Knowledge Centre.**

www.ipsos.com
@Ipsos

GAME CHANGERS