The Viability of Large Language Models for Conjoint and MaxDiff Analysis

In Market Research

The advent of Large Language Models (LLMs) such as GPT-4 has sparked a significant shift in the landscape of data analytics within the market research industry. These advanced AI-based tools have the potential to emulate complex human decision-making processes, offering new avenues for understanding consumer behaviour and preferences. Early explorations have investigated the capabilities and limitations of LLMs in executing more sophisticated tasks such as making choices between products, but the rapidly evolving nature of these models necessitates further comprehensive research to fully comprehend their impact.

The potential of LLMs to accurately predict consumer choices and quantify trade-offs presents an opportunity to streamline market research practices, offering insights without the need for exhaustive surveys. However, the emergent nature of generative AI demands a rigorous examination of its predictive reliability, biases, and limitations in capturing the nuanced aspects of human cognition.

Ipsos has undertaken one of the largest research exercises in this field, eliciting over 250,000 AI generated responses to Conjoint and MaxDiff choice tasks, evaluating a range of LLMs across a diverse set of scenarios, comparing their performance against real-world data. This research offers a comprehensive insight into the transformative potential of LLMs in answering choice experiments and the strategic implications for businesses.

Background

Generative AI exhibits promise in replicating human language, presenting innovative pathways to automate and enhance market research processes. These models, pre-trained on extensive volumes of data, generate content based on statistical probabilities, enabling responses to diverse stimuli such as product features or pricing strategies. Despite these advantages, potential challenges and ethical concerns exist such as the risk of replicating human biases ingrained in training data and a lack of nuanced understanding of human cognition and emotion.

Conjoint analysis is a statistical technique frequently used in market research that helps understand how consumers value different features of a product or service. It requires respondents to choose or rank hypothetical products, each with specific combinations of features. By analysing these responses, the relative importance of each feature and preferences for new combinations of features can be ascertained. MaxDiff, or Maximum Difference Scaling, is a technique used to measure the preference of many items from a list. It involves presenting a set of items and asking respondents to identify the most and least preferred/important items, enabling a ranking of each item.

Early research papers in the area, such as "Using GPT for Market Research" by Brand et al, laid the groundwork for understanding how generative AI could align with fundamental economic theories. Their research underscored key economic principles like the downward-sloping demand curve, which posits that as a product's price increases, consumer demand decreases. Building on this foundation, Ipsos has expanded the knowledge of research in this area, examining the performance of different LLMs, the impact of LLM parameters, such as 'Temperature', and the influence of prompt text and prompt execution on both Choice Based Conjoint and MaxDiff studies.

The research was structured around several hypotheses, each intended to evaluate the potential, as well as constraints of LLMs when working with choice designs:

1. LLMs are capable of handling complex choice designs
2. There is an optimal LLM for generating choice responses
3. The temperature setting impacts the performance of LLMs
4. How prompts are executed can enhance the results generated by LLMs
5. Positional bias is a fundamental issue in LLMs
6. LLMs can achieve differentiation at the respondent/persona level
7. Training LLMs with external data will enhance results
8. Results derived from LLMs provide the same commercial insights as studies with real respondents

Research Design

The research incorporated five commercial studies: three Choice Based Conjoint (CBC) and two MaxDiff (Figure 1). These designs, which ranged from low to mid-complexity, encompassed diverse service sectors and different treatments of price. The initial research phase used the first three of the data sets and tested the language models GPT-3.5, GPT-4, Claude 2.1, and Gemini Pro, employing different LLM temperature settings (0.2, 0.5, and 0.8). The 'temperature' setting controls the randomness or variability of the model's responses, with higher values leading to more diverse outputs and lower values resulting in more deterministic responses.

Figure 1

	CBC	CBC (SKU)	MaxDiff	CBC 2	MaxDiff 2
# Attributes / Items	7	9	11	6	20
# Levels	33	17	-	26	-
Fieldwork location	UK	UK	UK	France	UK
Design complexity	Mid	Mid	Low	Low	Low
Price attribute	No price	Linear	-	Part-worth	-

The research benchmarked the analysis of the LLMs against a sample of N=500 respondents drawn from the real study. To generate synthetic responses, a persona was constructed incorporating demographic and behavioural information based on the real respondent sample information, including Gender, Age, Region, and study-specific behavioural data, e.g., frequency of travel, frequency of product purchase, etc.

Prior research has indicated the influence of prompt text on LLM outcomes. Utilising GPT-4, multiple queries were generated to identify a structure that the LLMs could comprehend and contained all the necessary information for answering the choice tasks. After experimentation and repeated querying of the LLMs on its task comprehension and decision-making processes, the final structure used is illustrated in Figure 2. In addition to providing the text prompt, all choice tasks were submitted in a single prompt and the LLMs were asked to select the preferred concept from each of the choice tasks for the CBC exercises, or most and least preferred items in the MaxDiff exercises.

Figure 2

Final prompt consisted of the following sections:

1. Background e.g., client business question	2. Description of Factors (attributes)	3. Persona Information
4. Structure of Task e.g., scenarios, options (tasks)	5. Expected Response and Format	6. Summary

In the second phase of the research, the most promising LLM and parameter settings were carried forward and further experimentation, aimed to enhance the accuracy of the models was conducted. The focus of the experiments in phase 2 was three-fold: refining prompt execution, training the LLM with external data and investigating positional bias further (Figure 3). The experiments included refinements such as changing the persona tense, simplifying the prompt text, adding practice tasks, adding information about when to select certain options, submitting choice tasks one at a time, and conversational prompts. Experiments on training the LLM focussed on training it with actual choices made from a different set of real respondents or providing the LLM with incomplete choices from respondents to then impute responses to the remaining choice tasks. To investigate positional bias further, experiments included re-running the analysis where the original concept position was randomised or reversed and running a Dual response None methodology which first forces the LLM to make a choice from one of the concepts, then in a second stage to state whether it would purchase the product or not.

Figure 3

Refining Prompt Execution

- Change persona tense
- Add practice tasks with clear winners
- Simplifying prompt description
- Promoting LLM one task at a time
- Conversational prompts

LLM Training

- Add complete real choice tasks (using previous respondents)
- Add incomplete real choice tasks (to impute additional choice tasks)

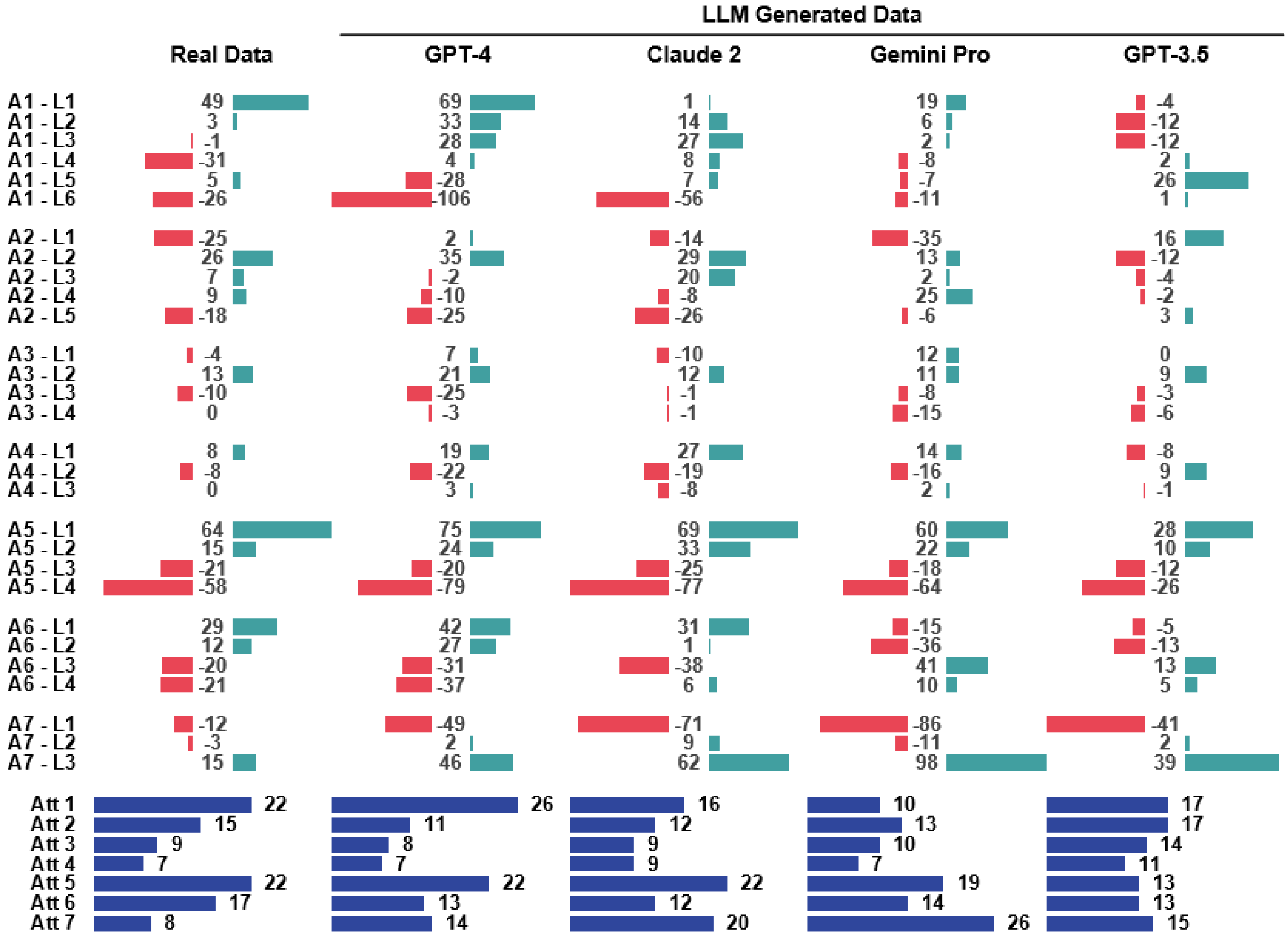
Positional Bias

- Reverse concept positions
- Combinatorial analysis
- Additional instructions about None option
- Dual response None

Selected Results

The outcomes of the experiments highlight the potential of LLMs, particularly GPT-4, in navigating more complex choice designs. Figure 4 compares the preference structure of different LLMs versus actual data from one of the CBC studies. The figure displays standardised utility scores for every level within each attribute (A1-A6), where higher scores denote a higher level of preference. The attribute importance in the decision-making process is also illustrated at the bottom of the figure. Where attributes had a clear preference structure the performance of the LLMs were mostly accurate, but results were inconsistent when dealing with non-ordered categorical attributes or where there were interactions between attributes.

Figure 4



In the case of MaxDiff, GPT-4 was a clear winner, where other models often produced illogical answers and non-sensical results. A comparison between the ranking of items between the real survey and those generated by GPT-4 suggested that the LLM was generally adept at identifying items ranked at the top and bottom, as shown in Figure 5.

Training, or fine-tuning the LLM with additional data led to a significant improvement in result accuracy. This was particularly noticeable when the LLM was trained on responses from a separate sample of 500 real respondents that had already gone through the choice exercise. Improvements included an increase in variability among respondent/persona utility scores, elimination of positional bias, and increase in the correct ranking of level preference within attributes, as shown in Figure 6.

Figure 5

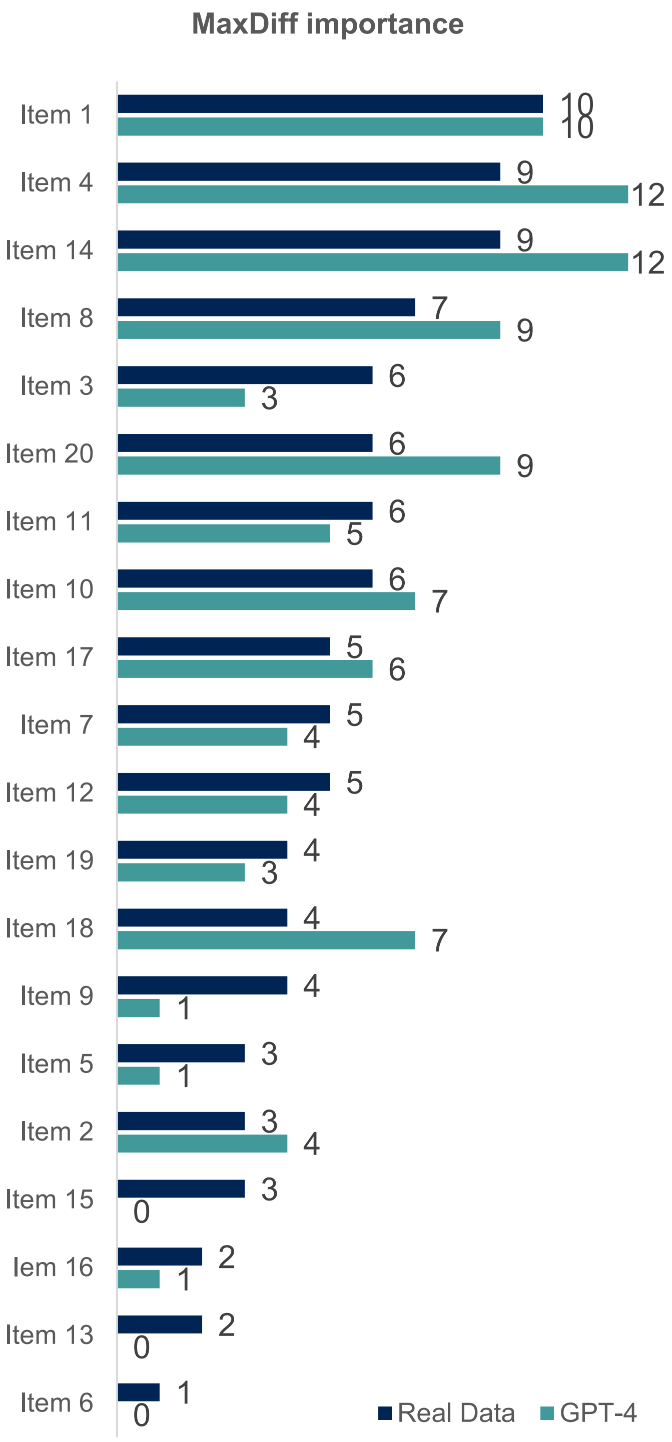
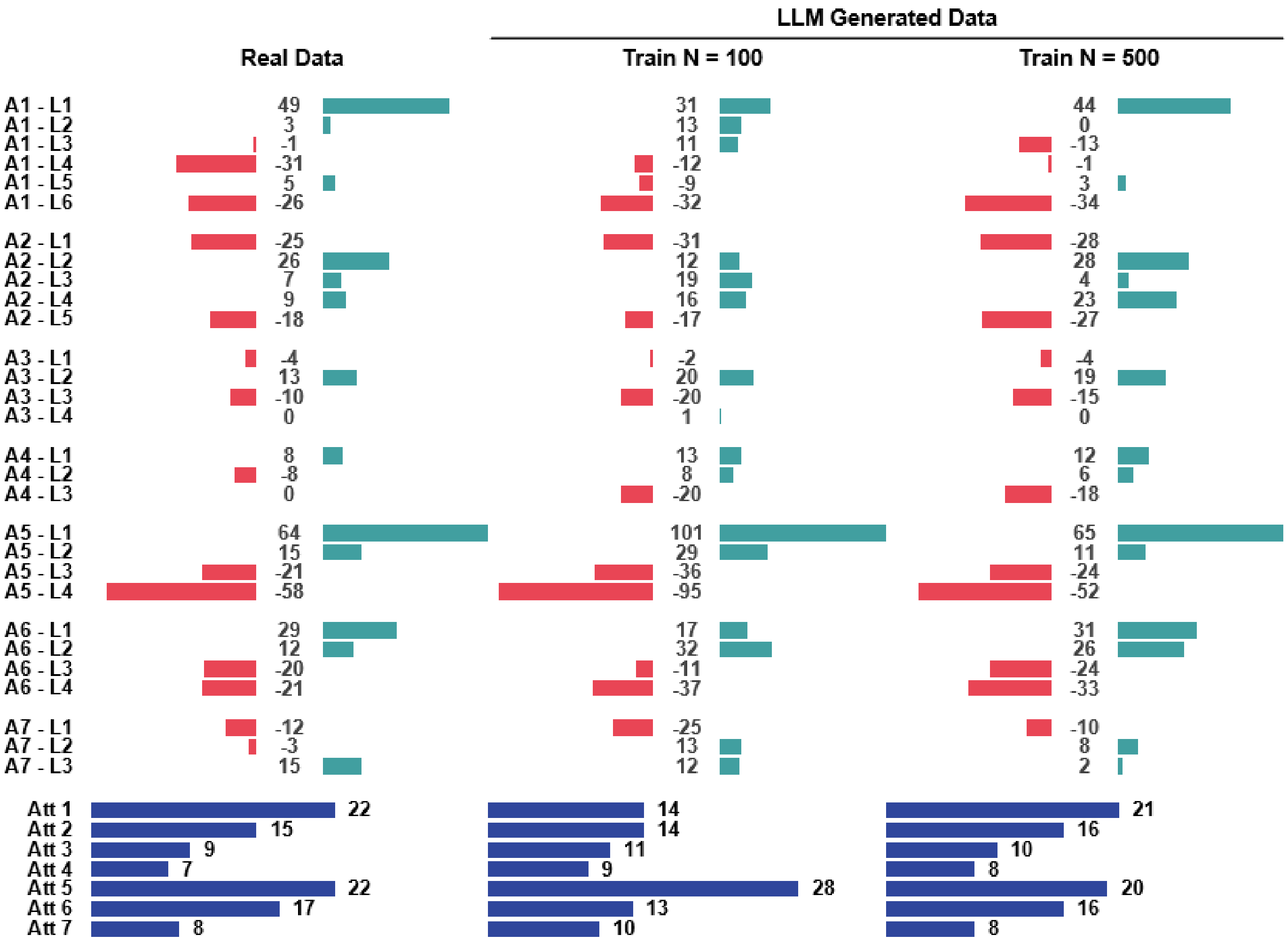


Figure 6


LLM Generated Data
Real Data
Train N = 100
Train N = 500
A1 - L1 49 31 44
A1 - L2 3 13 0
A1 - L3 -1 11 -13
A1 - L4 -31 -12 -1
A1 - L5 5 -9 3
A1 - L6 -26 -32 -34
A2 - L1 -25 -31 -28
A2 - L2 26 12 28
A2 - L3 7 19 4
A2 - L4 9 16 23
A2 - L5 -18 -17 -27
A3 - L1 -4 -2 -4
A3 - L2 13 20 19
A3 - L3 -10 -20 -15
A3 - L4 0 1 0
A4 - L1 8 13 12
A4 - L2 -8 8 6
A4 - L3 0 -20 -18
A5 - L1 64 101 65
A5 - L2 15 29 11
A5 - L3 -21 -36 -24
A5 - L4 -58 -95 -52
A6 - L1 29 17 31
A6 - L2 12 32 26
A6 - L3 -20 -11 -24
A6 - L4 -21 -37 -33
A7 - L1 -12 -25 -10
A7 - L2 -3 13 8
A7 - L3 15 12 2
Att 1 22 14 21
Att 2 15 14 16
Att 3 9 11 10
Att 4 7 9 8
Att 5 22 28 20
Att 6 17 13 16
Att 7 8 10 8

Key Findings

The research, comprising c.50 experiments, and over 250,000 AI generated responses has advanced the understanding of the use of LLMs in the field of choice modelling. It not only provides insights into the capabilities of LLMs but also paves the way for future explorations. The outcomes of the experiments have allowed Ipsos to address the key hypotheses set out at the start of the research.

LLMs are capable of handling complex choice designs

LLMs demonstrated the potential to process a larger number of attributes and levels than previous research tested. LLMs can understand the ranking of levels within attributes that have a clear order, but their performance in considering interactions and the accuracy of level preference in categorical attributes, without additional training, is inconsistent. While LLMs can handle complex choice models to a degree, they require training for handling more sophisticated tasks.

There is an optimal LLM for generating choice responses

Among the LLMs tested, GPT-4 outperformed others in most accurately reflecting the utility and importance structure of the real responses. For MaxDiff, GPT-4 emerged as a clear winner over other models, which produced illogical answers, such as coding the same item as both best and worst.

The temperature setting impacts the performance of LLMs

The temperature setting, which controls the stochasticity or randomness of LLMs responses, had minimal impact on the results, indicating that for numerical selections within the confines of a choice task, its effect is negligible.

How prompts are executed can enhance the results generated by LLMs

Different prompt engineering can improve the performance of LLMs in choice modelling tasks. Asking choice tasks one at a time and having a conversation with the LLM between tasks enhanced the accuracy of the results. However, the inclusion of certain behavioural information, e.g., ‘most often purchased product’ led to spurious results, highlighting the importance of careful prompt design.

Positional bias is a fundamental issue in LLMs

Positional bias was detected in responses from GPT-4, with a tendency to select option one over other options. However, when the order of options was reversed, GPT-4 adapted its preferences, demonstrating that it could make consistent choices based on the concepts shown to it. Positional bias was less apparent in other LLM but came at the expense of accuracy. Despite specific instructions being included in the prompt, the LLMs rarely select the ‘None’ option. In the MaxDiff exercises, positional bias was detected in the selection of the ‘Worst’ item with option one being selected most frequently in GPT-4.

LLMs can achieve differentiation at the respondent level

While LLMs stated that they considered the persona information, there was limited differentiation in utility scores at the respondent/persona level. It suggests that the persona information provided to the LLMs in these experiments may not have been sufficient to generate the differentiation seen in real human responses.

Training LLMs with external data will enhance results

When responses from real respondents were utilised in training the LLMs, a marked improvement was observed in the results. This included increased variability in utility scores across respondent / persona's, eradication of positional bias and increase in the correct ranking of level preference within attributes. Due to constraints in prompt size that the LLM could handle, the potential for integrating further external information was limited. In experiments where the LLM was presented with incomplete responses from real respondents and tasked with answering the remaining tasks, while the output bore resemblance to the real data, the accuracy of the concept selected in each task by the LLM was found to be poor.

Results derived from LLMs provide the same commercial insights as studies with real respondents.

The comparison of results derived from LLM models with real studies is nuanced. While LLMs, particularly GPT-4, have shown the ability to replicate certain aspects of consumer choice behaviour, they have limitations in handling interactions between attributes and non-ordered attributes. In the CBC studies, comparing simulations with real data against LLM generated data in many instances provided different commercial insight. However, the results from the MaxDiff studies are more encouraging. In the two data sets tested, without any training, GPT-4 was able to identify most of the top and bottom items, albeit with a different ranking.

Use Cases and Further Research

LLMs present a variety of potential future use cases. LLMs could be employed to screen items for quantitative studies, thereby providing a more efficient method for survey preparation.

LLMs could in the future supplement quantitative studies by generating additional synthetic respondents, or aid in reducing questionnaire length by only asking respondents to answer a small number of choice tasks. However, these applications should be pursued with an understanding of the current limitations of LLMs, particularly their current inability to fully capture the complexity and variability of human behaviour.

More research is required to fully harness the capabilities of LLMs. A deeper understanding of how to improve respondent or persona information could allow for more nuanced decision-making processes. Another area to investigate, as LLMs become more powerful and allow more information, involves enhancing the LLM learning capabilities by incorporating additional external data.

Conclusions

The exploration of LLMs in Conjoint and MaxDiff analysis signals a transformative shift in data analytics. While the potential to emulate complex human decision-making processes is clear, the journey to fully harness these models' capabilities has only just begun.

The Ipsos research has shown that while LLMs can replicate certain aspects of human choice behaviour, a large gap remains. They do not yet provide sufficiently similar commercial insights as studies with real human respondents, indicating an inherent limitation in capturing the entirety of human cognitive complexity and variability. As models are pre-trained, their ability to accurately predict choices from experiments that contain new and/or innovative features will be limited. In addition, given the sources that LLM are trained upon, they may not be representative of the specific research sector, be outdated and/or inconsistent in their responses based on the geographical region due to training data predominantly coming from Western, educated, and democratic societies (Atari et al).

Generative AI is not ready to take over the choice modelling industry. It requires significant human intervention to overcome the biases that exist within its training corpus. That said, as our understanding of LLMs deepens and the technology evolves, generative AI has the potential to become a powerful complement to choice modelling methodologies. The horizon is vast and full of possibilities, and the future of LLMs in choice modelling promises to be both challenging and transformative.

References

James Brand, Ayelet Israeli, Donald Ngwe, “Using GPT for Market research”, 2023, Harvard Business School Marketing Unit Working Paper No. 23-062

Mohammad Atari, Mona J. Xue, Peter S. Park, Damián E. Blasi, Joseph Henrich, 2023, “Which Humans?”, Department of Human Evolutionary Biology, Harvard University

Contact us to find out more:

Chris Moore

UK Head of Data Science and
Advanced Analytics Transversal,
Ipsos

chris.moore@ipsos.com