N-Infinite Technology by livepanel

ESOMAR's AI 20 by Livepanel N-Infinite

A. Company Profile

1. **What experience does your company have in providing AI-based solutions for market research?**

   The genesis of our company's AI technology dates back to late 2017 when we decided to explore the implications of integrating machine learning resources to tackle challenges associated with managing large-scale field research projects, such as limited audiences, weighting, incomplete surveys, etc.

   To this end, the company onboarded professionals not only in technology related to data science but also from related disciplines such as Economics and Sociology. This multidisciplinary approach facilitated determining the conditions for the correct application of this technology.

   Furthermore, during this period, we have engaged in fruitful interactions with the Institute of Information Sciences under CONICET (National Science Committee) and the Department of Computer Science at the National University of the South (Argentina) being the latter the oldest and most extensive in terms of AI-related applied research in our country, aiming to validate our methods, applications, and results.

2. **Where do you think AI-based services can have a positive impact for research? What features and benefits does AI bring, and what problems does it address?**

   As a company, we are not only convinced of the potential for a positive impact, but we have also experienced it firsthand through successful product deliveries over the last four years. The benefits are clear: **Today, AI is fundamentally a tool that allows us to effectively combat the industry's quality crisis stemming from the use of unknown third party panels, river sample or social media surveys.. Additionally, it is a very effective way to reduce costs while easily meeting timelines and quotas**, while providing a more precise tool than weighting for addressing issues such as poor age, gender distribution, and even reaching challenging audiences.

3. **What practical problems and issues have you encountered in the use and deployment of AI? What has worked well and how, and what has worked less well and why?**

   AI-based processes are not immune to the generalities of computing, especially the principle of *garbage in, garbage out*. Synthetic responses cannot be generated if the quality of source data is not robustly addressed. This includes not only field survey data but also a solid set of profiling that ensures not only a prediction base but also that the algorithms comprehend the differences between one type of panelist and another.

In this regard, a significant challenge was adapting our panel profiling database so that each new panelist not only responds to current projects but also generates a sufficient knowledge base to be utilized in prediction projects.

Technologically, the greatest challenge was validating the legitimacy of the obtained results, not only using statistical tools that establish the confidence of each response but also comparing the results against fieldwork, both individually and collectively. This includes validating the compliance of questionnaire routing, the expected distribution of responses, and improving the pre and post-processing of data to ensure quality and simplify the analysts' work.

B. Is the AI capability/service trustworthy, ethical and transparent?

4. **Can you explain the role of AI in your service offer in simple, non-technical terms in a way that can be easily understood by researchers and stakeholders? What are the key functionalities?**

   We employ AI with a very specific function: To **create synthetic responses from real users**. This process involves starting from a pre-established base of profiles or even incomplete answers, seeking matches with real respondents and the best prediction candidates. From there, we generate synthetic responses that mimic the candidates' responses to the target questionnaire.

5. **What is the AI model used? Are your company's AI solutions primarily developed internally or do they integrate an existing AI system and/or involve a third party and if so, which?**
   We use many Machine Learning models including, but not limited to, Regressions, Random Forests and Neural networks. There is not an enhanced or repeatedly used model but rather, a new machine learning model is created for each questionnaire's question during the training and prediction process. The type of machine learning model is either chosen automatically or can be indicated by the analyst. In each training, between 10 and 255 machine learning models are evaluated, and the one demonstrating the best prediction capabilities is selected after being tested with a partial population of field cases. The end-to-end solution has been entirely developed in-house, using external libraries or services such as those provided by various MLaaS providers (Google, Amazon).
   We also use auxiliary LLM services like GPT-4 from OpenAI or Claude from Anthropic, although these do not constitute the central driver of the system's functionalities.

6. **How do the algorithms deployed deliver the desired results? Can you summarise the underlying data and the way in which it interacts with the model to train your AI service?**

   The algorithms deliver results by predicting with information from field responses and profiling questions regularly answered by panelists. These same questions are answered by

other panelists, and it is at this intersection where the raw material is generated to feed the trained models at the time of predictions.

Due to the system's nature, and to optimize processing times, the models are trained in parallel, considering all prior information within the questionnaire sequence before the target question. However, prediction is performed sequentially since the preceding prediction's result is relevant as a parameter for the next model.

Client information of any type is **NOT** used in general profiling or training or prediction towards other clients. In the case of projects with questionnaire interrelation, it could be used in projects for the same client. Typically we use fully encoded information (we don't know either the questionnaire contents or answers) and since we do not use cross-information between clients, it is not necessary to offer opt-outs in this case.

7. **What are the processes to verify and validate the output for accuracy, and are they documented? How do you measure and assess validity? Is there a process to identify and handle cases where the system yields unreliable, skewed or biased results? Do you use any specific techniques to fine-tune the output? How do you ensure that the results generated are 'fit for purpose'?**

   Validation and verification of integrity processes are carried out within the platform through pre and post-questionnaire logic analysis. Additionally, each valid prediction distributes results concerning the set used for training.

   Additionally, the predictions undergo a group analysis that optimizes the established distribution based on the training data, as long as both display similar characteristics regarding the user distribution. We call this technology SmartMatch, and it is available if the evaluation of distributions yields positive results.

   Regularly, we conduct comparative studies consisting of taking a part of the field sample and comparing the remaining part with the system's predictions to evaluate the quality and precision of the models generated. These studies are documented in the form of widely distributed papers by organizations including ESOMAR.

8. **What are the limitations of your AI models and how do you mitigate them?**

   The main limitation of our models is the absence of high-quality field data. To address this, we have developed an analysis tool based on historical experience that allows us to weigh the distribution of demographic data but also the richness of profiling, training, and prediction data available. In this way, we can set reliability limits before proceeding with a project that might have insufficient data.

   Also, within each project, the tool provides data engineering resources that filter anomalous cases that would degrade the overall quality of predictions. We prefer high-quality synthetic responses to a collection of poor data from panelists responding reluctantly.

9. **What considerations, if any, have you taken into account, to design your service with a duty of care to humans in mind?**

Our data originates entirely from current field research, preventing deviations or attitudinal biases and hallucinations typical of LLM systems. In our technology, these systems play an auxiliary role and are not responsible for generating synthetic responses.

How do you provide Human Oversight of your AI system?

10. **Transparency: How do you ensure that it is clear when AI technologies are being used in any part of the service?**

    All projects delivered by our company with AI involvement are communicated to the client, indicating the intervention of this technology. In the case of projects to third-party panels, field cases are also distinguished from those generated by AI. In all cases, they are synthetic responses unequivocally based on real cases.

11. **Do you have ethical principles explicitly defined for your AI-driven solution, and how in practice does that help to determine the AI's behaviour? How do you ensure that humandefined ethical principles are the governing force behind AI-driven solutions?**

    All projects for creating synthetic responses are carried out and supervised by an analyst who performs pre and post-processing, with fundamental activities at each stage:

    - Pre-processing
    - Selection of cases (Age, Gender, Social Grade, Regions)
    - Choice of training questions
    - Filtering of inconsistent cases
    - Normalization of open-text questions
    - Training and prediction parameters
    - Selection of age, gender, and Social Grade quotas over the total number of users, prioritizing field cases
    - Post-processing

    As mentioned earlier, every step of the process involves verification by an engineer and/or a data scientist. The fact that each case involves simple, non-reusable models configures a different scenario than the creation and perpetual maintenance of one or a set of large models. Working primarily with quantitative information and classification questions mostly avoids scenarios of hallucinations or biases typical of LLMs.

12. **Responsible Innovation: How does your AI solution integrate human oversight to ensure ethical compliance?**

    Our process primarily focuses on the input of training and prediction: The taxonomies we create and manage to understand user behavior - and above all, allow the models to understand them - are developed with this premise. Respecting the individuality of each panelist, their interests, tastes, and beliefs, without inducing them to speculate or alter their statements - which are anonymous in practice - to be included or achieve a better result.

This pursuit also produces predictions, synthetic data, that we understand to be honest, genuine, and without deviations.

What are the Data Governance protocols?

13. **Data quality: How do you assess if the training data used for AI models is accurate, complete, and relevant to the research objectives in the interests of reliable results and as required by some data privacy laws?**

    In all cases, we work only with anonymous information from our sources or anonymized data supplied by our clients (including both questionnaires and answers). The greatest effort to achieve accurate and quality information comes from the selection of training responses and subsequent analysis on the platform. In the case of clients, we encourage them to optimize their profiling information and provide support on this purpose.

14. **Data lineage: Do you track the origin and processing of data throughout its lifecycle, from collection to analysis and reporting and are these sources made available?**

    All the proprietary information we use comes from fieldwork with our panelists, whom we prefer to remain anonymous to us. We do not use navigation information, advertising, and we restrict the request for georeferencing to onboarding only to simplify the panelist's location in the terms and scopes required by our clients (city).

15. **Please provide the link to your privacy notice (sometimes referred to as a privacy policy). If your company uses different privacy notices for different products or services, please provide an example relevant to the products or services covered in your response to this question.**

    https://www.livepanel.ai/privacy-policy/ is mainly focused on panel's privacy, 100% GDPR compatible. Since we don't get PII from panelists, we think it's not necessary to include AI synthetic data generation as a privacy issue.

16. **What steps do you take to comply with data protection laws and implement measures to protect the privacy of research participants? Have you evaluated any risks to the individual as required by privacy legislation and ensured you have obtained consent for data processing where necessary or have another legal basis?**

    Our platform has been developed to respond to GDPR as a general framework for user data protection, ensuring the removal of data that could compromise the anonymity of the users. By using these same users in the training and prediction tasks, we ensure that these privacy guarantees are preserved.

17. **What steps do you follow to ensure AI systems are resilient to adversarial attacks, noise and other potential disruptions? Which information security frameworks and standards do you use?**

The process of creating models, as well as their use and deactivation, has a very short life cycle, as they are created for each questionnaire analyzed and discarded after the prediction process. They would even require a re-generation if a new wave is processed considering exactly the same questionnaire. In these terms, we ensure that cloud service providers, MLaaS comply with standards like COBIT, HITRUST or ISO 27001, but not maintaining trained models in the long term nor large dictionaries as those entailed, for example, by an LLM model guarantees this resilience.

18. **Data ownership: Do you clearly define and communicate the ownership of data, including intellectual property rights and usage permissions?**

    At Livepanel, we clearly divide data ownership into three large groups:

    - Livepanel's ownership corresponds to the generic (and anonymous) data of the different panelists, who only reveal contact information when requesting a reward, for example.
    - The profiling questions that are carried out with complete independence from the questionnaires of clients.
    - The ownership of the clients concerning the data collected in the different field projects. This information is not crossed on any occasion with that of other clients, not even to make predictions within the framework of a project, and the synthetic data generated in the process also limit themselves to the ownership of the client in each case.

19. **Data sovereignty: Do you restrict what can be done with the data?**

    Our company does not establish any type of restriction on the use of data, as it does not include any type of PII, both in the case of field data and those resulting from synthetic generation.

20. **Ownership: Are you clear about who owns the output?**

    Yes. The intellectual property of the processes and the models generated is exclusively owned by Livepanel, as well as the profiling data of our users. In the case of 100% third-party projects, the ownership of the input and output data is 100% of the clients of each project.