Validation white paper

Executive Summary

Traditional surveys are time-consuming and frequently produce low-quality data. Simulated populations can deliver actionable insights in minutes rather than months.

- **The problem:** Traditional surveys are slow, costly, and often deliver uncertain data quality—limiting timely, evidence-based decision-making.
- **The approach:** Synthetic populations use large language models to simulate survey responses based on realistic population structures, complementing—not replacing—real-world surveys.
- **The validation:** Across 19 diverse survey items, simulated responses closely matched real-world benchmarks, with an average agreement of ~**89%**.
- **Key insight:** Highest alignment was observed for demographic and behavioral items; greater variation for attitudinal questions mirrors known survey sensitivities.
- **The value:** Synthetic surveys enable rapid hypothesis testing, better question design, and earlier insight generation—reducing cost and friction before fieldwork.
- **The takeaway:** Used alongside traditional surveys, synthetic populations support faster learning, better-designed research, and more informed decisions.

Why Traditional Surveys Fail Modern Decision-Making?

Thoughts, attitudes, and opinions shape human behavior across domains ranging from everyday consumer choices to consequential life decisions and political processes. Understanding the patterns underlying these mental states enables us to design products, services, and public policies that are not only effective, but genuinely desirable—and capable of driving meaningful change.

Because we cannot directly observe people's thoughts or cognitive processes, surveys remain the primary method for accessing them at scale. Yet designing high-quality surveys is both time-consuming and methodologically demanding. Poorly worded questions, response biases, low engagement, and sampling limitations can all compromise data quality. Moreover, traditional data collection methods often struggle to capture nuance, context, and change over time.

As a result, despite the central role surveys play in decision-making across industries and societies, there are strong reasons to question the reliability and validity of much of the data we routinely collect. Organizations today have to make decisions faster than traditional surveys can deliver actionable insights.

Despite decades of methodological development, core challenges in survey research remain largely unresolved by existing tools. Together, these three structural limitations create substantial hidden costs—through inefficiency, delayed insights, and compromised data quality—for organizations that depend on surveys for decision-making (Table 1):

1. The complexity and expertise required to design high-quality surveys
2. The slow and resource-intensive nature of data collection
3. The inconsistent and often low quality of data produced

Table 1: The three-fold problem of conventional survey practices and their impact on of decision-making quality and costs.

Problem	Description	Quality impact	Typical cost drivers
1. Difficulty of survey design	Designing valid and reliable survey instruments requires methodological expertise, careful wording, and iterative testing. In practice, surveys are often created under time pressure, using ad hoc questions or recycled items that are poorly aligned with the underlying construct of interest.	Measurement error, ambiguous results, and reduced interpretability of findings, leading to decisions based on incomplete or misleading information.	Expert time for question development, piloting, revisions, and validation; iterative stakeholder reviews; redesign due to unclear results
2. Time-consuming data collection	Traditional surveys rely on static questionnaires and discrete data collection cycles, often requiring weeks or months to design, deploy, and gather sufficient responses. Response rates are declining, further extending timelines.	Insights arrive too late to inform fast-moving decisions, limiting organizations' ability to respond to emerging trends or changing user needs.	Extended field periods, reminder campaigns, incentives to boost response rates, repeated survey waves due to insufficient data
3. Poor data quality	Survey responses are affected by low engagement, satisficing, social desirability bias, and lack of contextual understanding. These issues are difficult to detect and correct using conventional survey tools.	Reduced confidence in results, increased need for data cleaning or post hoc adjustments, and weakened trust in survey-based evidence across the organization.	Data cleaning, exclusion of low-quality responses, follow-up analyses, need for additional studies to confirm findings

Total estimated cost per traditional survey is €13,000–€65,000, excluding downstream costs of incorrect or delayed decisions. Beyond these direct and indirect costs, low-quality survey data can lead to strategic misalignment, ineffective product decisions, or poorly targeted policies—costs that often exceed the survey budget itself but remain largely invisible and unaccounted for.

Synthetic populations based on large language models offer a practical response to the limitations of traditional survey research. By modeling population structures and simulating how different individuals might respond, they provide decision-makers, researchers, and product teams with faster, lower-cost ways to generate high-quality insights.

Rather than replacing real-world data collection, synthetic populations enable early, directional insights in minutes instead of weeks. This allows organizations to test ideas, iterate rapidly, and enter fieldwork with sharper hypotheses—improving both efficiency and data quality.

Solution: kansa.io and synthetic populations

How kansa.io works?

Kansa.io generates synthetic survey data through a multi-stage process that combines population-level modeling with large language model–based simulation and empirical validation.

- First, a **population model** is constructed using a combination of open and proprietary datasets describing the demographic and structural characteristics of the target population (e.g., the Finnish population). The model captures key population-level distributions and relationships relevant for survey research and can be adapted to different geographic or demographic contexts.
- Second, **personas** are generated by sampling from this population model. Each persona represents an individual with a coherent set of demographic and structural attributes drawn to reflect realistic population heterogeneity.
- Third, during **survey prompting**, a large language model is instructed to respond to survey items from the perspective of a specific persona. The model is constrained to answer consistently with the persona's attributes

and the survey's response formats.

- Fourth, individual responses are **aggregated** to form a synthetic dataset that mirrors the structure of a conventional survey dataset, enabling standard analytical workflows.
- Finally, the synthetic data undergo **validation** by comparing response distributions and key relationships against established real-world survey benchmarks. This step is used to assess alignment with observed population patterns and to identify systematic deviations

The potential benefits for different user groups are presented in Table 2

Table 2: Potential benefits of synthetic populations and surveys across user groups

User group	Primary use case	Key benefits
Decision-makers (public and private)	Rapid exploration of policy, strategy, and market scenarios	Fast, directional insights delivered in minutes rather than weeks; ability to examine hard-to-reach or underpowered subgroups without additional cost; support for earlier, better-informed decisions under time pressure.
Researchers and analysts	Survey design, hypothesis testing, and methodological development	Iterative testing of question wording, formats, and framings; improved measurement quality before field deployment; reduced reliance on one-shot survey designs; integration into existing analytical workflows as a prototyping tool.
Product and service teams	Early-stage concept testing and prioritization	Low-cost validation of ideas, messages, and assumptions before development; rapid iteration without respondent fatigue; clearer signals on which concepts warrant real-world testing.
Policy and public-sector organizations	Policy design, evaluation, and stakeholder analysis	Ability to explore population-level responses to proposed interventions; improved visibility into minority or vulnerable subpopulations; more robust policy hypotheses prior to commissioning large-scale surveys.
Market research and insight teams	Continuous learning and insight generation	Expanded testing capacity without escalating fieldwork costs; consistent baselines across repeated simulations; shift from episodic surveys toward ongoing insight generation.

Does it work? Validation of kansa.io

To provide a preliminary understanding of our model's capabilities, we present comparative information from carefully selected benchmark surveys across various topics and industries. The process of survey selection and comparison, as well as the findings and their interpretation, is presented below.

4.1 Methdos

Selecting benchmark surveys

The validation survey comprised a deliberately heterogeneous set of items covering attitudes, beliefs, behaviors, and sociodemographic characteristics. Items were selected to maximize conceptual diversity across substantive domains—including institutional trust, political attitudes, perceived safety and wellbeing, economic conditions, media use, technology attitudes, and consumer behavior—rather than to measure a single latent construct.

Wherever possible, items were adapted from well-established national and international survey instruments (e.g., large-scale social, political, and wellbeing surveys) or closely aligned with question formats that have demonstrated validity and widespread use in population-based research. This approach was chosen to ensure that the item set reflected realistic survey content and response structures commonly encountered in applied research and policy contexts.

The final item pool included a mix of binary, ordinal, and continuous response formats, as well as both attitudinal and behavioral questions. Several items were intentionally included in parallel versions with and without explicit “don’t know” response options to evaluate robustness across common survey design choices. Collectively, the items were designed to span varying levels of abstraction, sensitivity, and cognitive demand, thereby providing a stringent test bed for validation analyses.

A complete list of survey items and response scales is provided in Table 2.

Table 2: Survey items included in the validation study

Topic	Item	Response scale
Trust in institutions & Democracy	To what extent do you think that the police are able to maintain public order and safety in Finland	1=Completely 2=Mostly 3=Not sure 4=To some extent 5=Hardly at all
Perceived local safety	How safe do you feel walking alone in your local area after dark?	1=Very safe 2=Safe 3=Unsafe 4=Very unsafe
Perceived happiness	All things considered, how happy do you feel?	0=Extremely unhappy 10=Extremely happy
Native language	Is Finnish your native language?	0=No 1=Yes
Investing in education	Investment in basic education must be increased significantly, even if it means reducing funding elsewhere.	1=Strongly agree 2=Somewhat agree 3=Hard to say 4=Somewhat disagree 5=Strongly disagree
Taxpaying	I am personally willing to pay more taxes to the state if wellbeing services counties need additional funding to ensure good care.	1=Strongly agree 2=Somewhat agree 3=Hard to say 4=Somewhat disagree 5=Strongly disagree
Outsourcing public services	A large share of our country’s public services should be outsourced to private providers to make service production more efficient.	1=Strongly agree 2=Somewhat agree 3=Hard to say 4=Somewhat disagree 5=Strongly disagree
Teachers authority	Teachers’ authority and right to maintain order in schools should be significantly increased.	1=Strongly agree 2=Somewhat agree 3=Hard to say 4=Somewhat disagree 5=Strongly disagree
Left-right political self-placement	Political attitudes are often described along a left-right scale. Where would you place yourself on this scale?	1=Left 2=Slightly left 3=Slightly right 4=Right
Financial coping	How would you describe your and your households financial situation and disposable income at the moment?	1=Must cut back on almost everything 2=Sometimes must cut back 3=Manageable with careful spending 4=Comfortably managing 5=Managing very well

Belonging (rural/urban- identity)	Do you see yourself as a rural, urban or both?	1=Urban 2=Rural 3=Both
Worry about finances	How often are you worried about whether your money will cover everything you need?	1=Daily 2=Weekly 3=Monthly 4=Every few months 5=Twice a year 6=Less than twice a year 7=Hardly ever 8=Not sure
Social media use: Facebook	How often do you use or follow Facebook?	1=Several times per day 2=1-2 times per day 3=3-7 times per day 4=1-2 times per week 5=Less than once per week 6=Not at all
AI attitudes	How much do you agree with the following statement: I believe AI will bring more positive than negative changes	1=Strongly agree 2=Somewhat agree 3=Neither agree or disagree 4=Somewhat disagree 5=Strongly disagree 6=Not sure
Fake news & Disinformation	How much do you agree with the following statement: Fake news and disinformation have made it harder to identify reliable information	1=Strongly agree 2=Somewhat agree 3=Neither agree or disagree 4=Somewhat disagree 5=Strongly disagree 6=Not sure
Local food consumption	Thinking one year into the future do you believe that you will use/eat/buy more locally produced food than today?	1=Kyllä 2=Ei
Purchasing behavior	How often have you purchased soft drinks in the last six months?	1=Not at all 2=Less frequently 3=Once every couple of weeks 4=About once a week 5=Several times a week 6=Daily
Purchasing behavior	How often do you purchase plant-based cheeses?	1=Not at all 2=Less frequently 3=Once every couple of weeks 4=About once a week 5=Several times a week 6=Daily

Comparison between real and simulated distributions - Composite Distribution Distance (CDD)

To evaluate how closely AI-simulated survey response distributions reproduce empirically observed distributions, we use a **Composite Distribution Distance (CDD)**. The CDD provides a single, interpretable summary of distributional discrepancy while preserving sensitivity to complementary aspects of mismatch that are relevant for ordinal survey data. The metric is used throughout this white paper to assess the quality of the validation results reported herein.

Distribution representation

For each item, both the empirical (“real”) and AI-simulated responses are represented as discrete probability distributions over the response scale:

KAAVA

where:

- k denotes the number of response categories (e.g., a 5-point Likert scale),
- p_i and q_i represent the probability of response category i,
- distributions are normalized such that **KAAVA**

Component metrics

The Composite Distribution Distance combines three established discrepancy measures, each capturing a distinct aspect of distributional difference.

1. *Normalized Earth Mover's Distance (EMD)*

Earth Mover’s Distance quantifies the minimum amount of probability mass that must be shifted along the ordered response scale to transform the simulated distribution into the empirical distribution. Because response categories are ordinal, EMD directly reflects the magnitude of misplacement in scale units.

To ensure comparability across different scale lengths, EMD is normalized by the maximum possible distance on the scale:

KAAVA

This normalization bounds the metric to the interval [0,1], where 0 indicates identical distributions and 1 indicates maximal ordinal displacement.

2. *Kolmogorov-Smirnov Distance (KS)*

The Kolmogorov–Smirnov distance captures the largest absolute difference between the cumulative distribution functions (CDFs) of the empirical and simulated distributions:

KAAVA

KS is particularly sensitive to pronounced local discrepancies, identifying the single point on the response scale at which the mismatch between distributions is greatest.

3. *Jensen-Shannon Distance (JS)*

The Jensen–Shannon distance measures the overall difference in distributional shape and overlap. It is derived from the Jensen–Shannon divergence, which symmetrically compares each distribution to their average:

KAAVA

Using logarithms with base 2, the Jensen–Shannon distance is bounded in [0,1]. Unlike EMD, JS does not depend on category order; instead, it captures global differences in probability allocation across categories.

Composite Distribution Distance (CDD)

The final Composite Distribution Distance is defined as the equally weighted mean of the three component metrics:

KAAVA

The resulting value lies in the interval [0,1], with:

- **0** indicating perfect correspondence between simulated and empirical distributions, and
- larger values indicating increasing distributional discrepancy.

For interpretive convenience, results may also be expressed as a fit score:

KAAVA

where higher values indicate closer alignment with empirical data.

Rationale for the Composite Approach

Each component metric captures a distinct failure mode that may arise in distributional validation:

- **EMD** reflects ordinal misplacement along the response scale,
- **KS** identifies the most severe localized deviation,

- **JS** captures global shape differences and loss of distributional overlap.

No single metric adequately characterizes all of these aspects in isolation. By combining them, the Composite Distribution Distance provides a balanced and robust summary measure suitable for evaluating the fidelity of AI-generated survey distributions.

Agreement between real and simulated response distributions was assessed using a normalized Composite Distribution Distance, expressed as percentage agreement for interpretability. Grades were assigned using predefined thresholds to facilitate qualitative interpretation: A+ (≥95%), A (90–94.9%), B (85–89.9%), and C (<85%). These grades are intended as descriptive summaries rather than strict accept–reject criteria, reflecting the expected variation across item types. Higher agreement was typically observed for demographic and behavioral items, while attitudinal and normative items showed greater dispersion, consistent with known properties of survey response behavior.

4.2 Results

Across 19 survey items spanning attitudes, behaviors, and demographic characteristics, the synthetic population produced response distributions that closely aligned with established real-world survey benchmarks. Overall agreement between simulated and observed distributions was high, with an average CDD agreement of 88.8%. Most items achieved A or B grades, indicating strong correspondence at the distributional level (Table 3).

Importantly, deviations between real and simulated distributions were rarely extreme and tended to preserve overall shape and ordering, even where agreement was lower. This suggests that synthetic populations capture meaningful population-level patterns rather than producing arbitrary or noisy responses.

Taken together, these results indicate that synthetic populations can generate credible, population-level survey signals across a wide range of domains. While not a substitute for real-world data collection, they provide a robust foundation for early-stage insight generation, question testing, and scenario exploration, enabling organizations to enter fieldwork with clearer hypotheses and better-designed instruments.

Table 3. Comparison of real-world and synthetic survey response distributions with agreement scores

Topic	Real distribution (%)	kansa.io distribution (%)	CDD and grade
Trust in institutions & Democracy	1: 12 2: 75 3: 3 4: 10	1: 22 2: 63 3: 9 4: 6	89.2% (B)
Perceived local safety	1: 45.82: 46.33: 7.14: 0.8	1: 322: 563: 114: 1	89.3% (B)
Perceived happiness	0–4: 2.45–6: 6.37: 11.98: 369–10: 43.5	2–4: 95–6: 127: 198: 329–10: 28	83.4% (C)
Native language	No: 10.8Yes: 89.2	No: 12Yes: 88	98.7% (A+)
Investing in education	Strongly agree: 17Somewhat agree: 44.9Hard to say: 26.8Disagree: 11.4	Strongly agree: 14Somewhat agree: 49Hard to say: 25Disagree: 12	97.1% (A+)
Taxpaying	Strongly agree: 15.5Somewhat agree: 28.6Hard to say: 16.9Disagree: 39.1	Strongly agree: 5Somewhat agree: 25Hard to say: 22Disagree: 48	85.5% (B)
Outsourcing public services	Agree: 18.5Hard to say: 14.3Disagree: 67.1	Agree: 17Hard to say: 21Disagree: 62	88.6% (B)

Teachers authority	Agree: 84.5Hard to say: 10.4Disagree: 5.1	Agree: 79Hard to say: 15Disagree: 6	84.4% (C)
Left-right political self-placement	Left: 14.8Slightly left: 30Slightly right: 35.9Right: 19.3	Left: 7Slightly left: 31Slightly right: 47Right: 15	90.9% (A)
Financial coping	Cutting back: 17.8Manageable: 34.8Comfortable: 42.5	Cutting back: 18Manageable: 35Comfortable: 47	96.5% (A+)
Belonging (rural/urban- identity)	Urban: 33.4Rural: 22.6Both: 43.9	Urban: 29Rural: 19Both: 52	93.0% (A)
Worry about finances	Daily/weekly: 33Monthly: 24Rarely/never: 36	Daily/weekly: 37Monthly: 24Rarely/never: 39	81.7% (C)
Social media use: Facebook	Daily: 83Weekly: 7Rarely/never: 9	Daily: 79Weekly: 10Rarely/never: 11	87.0% (B)
AI attitudes	Agree: 35Neutral: 20Disagree: 31Don't know: 5	Agree: 16Neutral: 32Disagree: 31Don't know: 21	81.2% (C)
Fake news & Disinformation	Agree: 64Neutral: 20Disagree: 13Don't know: 2	Agree: 79Neutral: 14Disagree: 1Don't know: 6	81.1% (C)
Local food consumption	1=Kyllä 2=Ei	1=Kyllä 2=Ei	
Purchasing behavior	None/rare: 57Weekly or more: 26	None/rare: 46Weekly or more: 27	89.6% (B)
Purchasing behavior	None/rare: 91Weekly or more: 4	None/rare: 88Weekly or more: 2	93.5% (A)

Conclusions

This study demonstrates that synthetic populations grounded in large language models can reproduce real-world survey response distributions with a high degree of fidelity across a diverse set of topics. Across demographic, behavioral, and attitudinal items, simulated responses showed strong alignment with established survey benchmarks, indicating that synthetic populations are capable of capturing meaningful population-level patterns rather than producing arbitrary or generic outputs.

The results also highlight important boundaries. Agreement was highest for structurally stable variables, such as demographics and consumption behaviors, while greater variation was observed for attitudinal and normative questions—an expected pattern that mirrors the inherent sensitivity of such items in traditional survey research. This reinforces the view that synthetic surveys should not be treated as replacements for real-world data collection, but as complementary tools that are particularly valuable in the early stages of research, strategy, and product development.

Taken together, these findings suggest that synthetic populations can meaningfully reduce the cost, time, and friction associated with survey-based insight generation. By enabling rapid hypothesis testing, iterative question design, and early scenario exploration, they help organizations enter real-world fieldwork with sharper questions and clearer priorities. In this way, synthetic populations support a shift from episodic surveying toward more continuous, informed, and adaptive decision-making.

About the team behind kansa.io and call to action

We are a Helsinki-based company of psychologists and IT professionals exploring new methods for understanding populations through data and simulation. Our work bridges behavioral science and AI for more agile, evidence-based decision-making.

We’re looking for research partners and organizations interested in testing the next version of our model. Reach out if you’d like to collaborate.