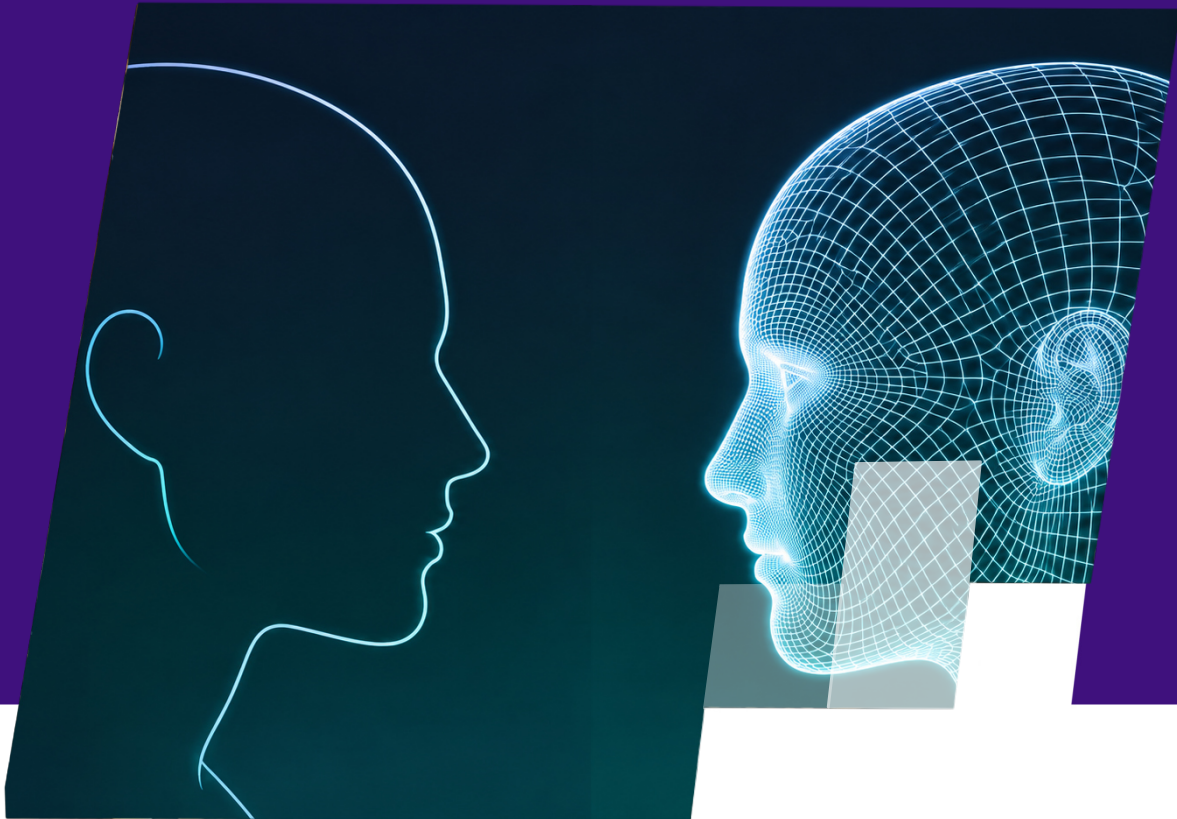




The Synthetic Shortcut

When AI Can Speed Up Research, and When Real Human Choice Still Has to Lead

A market research perspective on synthetic respondents, digital twins, pricing, and new product decisions.

September 2026

Executive Summary

- AI can meaningfully accelerate research when used for the right tasks, including hypothesis generation, survey design, concept pressure-testing, and synthesis.
- The core limitation of AI-generated synthetic responses is that they use historical patterns to predict future behavior. They are strongest when the question resembles something the market has already seen.
- For new products, new price points, and unfamiliar combinations of features, synthetic responses can produce confident answers without a reliable behavioral anchor.
- Pricing is where this risk matters most. Real customers construct value in context, shaped by trade-offs, comparisons, and irrational behaviors that current models do not reliably reproduce.
- The strongest approach is not AI or human research. It is a disciplined division of labor: use AI to accelerate the process, but anchor high-stakes pricing, elasticity, willingness-to-pay, and new product decisions in real human choice data.

Should We Just Use AI for This?

It comes up in more and more research and insights meetings, and the pressure behind it is real. A new generation of AI tools promises survey results in minutes instead of weeks, at a fraction of the cost, and leadership has noticed. Some of that promise is genuine. For the right tasks, these tools are fast, cheap, and good enough.

This paper is not an argument against them. **It is an argument for knowing where the line sits.** Used well, AI is a powerful accelerant for parts of the research process. Used for decisions it cannot support, it produces confident answers that can quietly mislead.

What follows maps what is actually being sold, explains the single assumption that limits all of it, shows what real human research still uniquely provides, and makes the case that the strongest approach is not AI or humans. It is a **deliberate division of labor** between them. The aim is simple: to know when AI can help, and when real human response must lead.

What's Being Sold, and the Assumption Underneath All of It

What's on the market

Walk into any market research function today and you will hear a growing vocabulary of AI-generated alternatives to fieldwork with real humans. The labels include **synthetic data, synthetic respondents, virtual audiences, AI personas, and digital twins**. Behind each label sits a commercial product actively being sold to research and strategy teams.

At the simpler end of the spectrum, some teams bypass purpose-built tools altogether and prompt general-purpose large language models such as ChatGPT, Gemini, or Claude directly, asking them to answer survey questions or respond to product concepts as a specified type of consumer. No training, customization, or primary data is required. At the more sophisticated end, purpose-built platforms ground simulated responses in proprietary panel, survey, or behavioral data. For example, Qualtrics offers synthetic research capabilities through Edge Audiences, and Toluna's HarmonAIze Personas offers synthetic personas based on its human panel.

Digital twins sit further along this same continuum. Under a strict definition, a digital twin is **a model of a specific known individual**, built from that person's own prior survey responses and behavioral data. That distinction matters: a synthetic persona may represent a type of consumer, while a digital twin attempts to replicate how a particular respondent would answer. This concept is quickly taking hold in the market research world. Ipsos announced a formal partnership with Stanford University's Politics and Social Change Lab in August 2025 specifically to build and validate digital twin panels for market and public opinion research, grounding them in Ipsos's proprietary KnowledgePanel. Beyond major research firms, dedicated platforms have already commercialized digital twins for market research. Panoplai boasts clients including HubSpot, Diageo, and Alphabet, while Soulmates.ai claims 93 percent fidelity to real human responses. At the frontier, agent-based platforms such as Aaru and Artificial Societies run simulations of entire populations.

The Synthetic Response Continuum

From general-purpose prompting to fully simulated populations



The assumption underneath all of it

The spectrum of sophistication runs from a single prompt to a general-purpose chatbot to a multi-million-dollar population simulation infrastructure. What does not change across that spectrum is **the underlying logic**.

Every one of these approaches, at any level of sophistication, does the same thing. It uses patterns from historical data to generate predictions about future behavior. A model is trained, fine-tuned, or prompted on what consumers have previously said, chosen, or done. It uses those learned patterns to simulate how a consumer might respond to a new question or a new concept. **The architecture varies. The mechanism does not.**

There is genuine positive performance evidence in certain conditions. A recent Qualtrics-conducted replication study published in early 2026 found that a fine-tuned synthetic model deviated from human responses by an average of 0.07 standard deviations, against 0.87 to 0.88 standard deviations for general-purpose models, a small effect by conventional benchmarks (McLean, Qualtrics / Greenbook, February 2026). A separate 2025 study found that synthetic respondents achieved 90 percent of human test-retest reliability for purchase intent tasks across 57 surveys covering established personal care products, suggesting outputs are stable enough to provide directional signals on known product concepts (Maier et al., 2025).

The condition underlying both findings, however, is one that is often less visible in commercial positioning. The results hold for well-established categories with abundant behavioral history, where the model has relevant signal to draw from. Strip that condition away and the core limitation of every historically trained system becomes visible. The Qualtrics study makes this explicit. In the same paper that produced the 0.07 standard deviation finding, the author notes that the model works by pattern recognition from training data, that emerging topics outside the training corpus are areas where the model has fewer patterns to learn from, and that the study played to the model's strengths, validating

one set of use cases rather than all of them (McLean, Qualtrics / Greenbook, February 2026). A model is strongest when the new task closely resembles patterns it has encountered. For an existing brand in a known category, it has signal. For a new product the market has not yet bought, a new price point the category has not experienced, or a genuinely novel combination of attributes, there is **nothing to anchor to**.

The model does not pause. It produces a confident-looking output drawn from the nearest familiar thing. The evidence for this failure is direct. Practitioners who tested synthetic respondents on new ideas not yet widely present online found the models could not produce accurate responses, because the training data did not contain them (Dig Insights, 2025). Even Ipsos's own digital twin pilot, separate from the Stanford initiative mentioned above, a test using matched human respondents and AI twins, found that synthetic respondents could reproduce broad themes and generate creatively rich ideas, but showed weak alignment with humans when asked to evaluate new product concepts, including a mismatch in preferred concept choice. The digital twins were found to be more rational and objective, while the human counterparts they were trained to mimic were more multidimensional in their responses (Ipsos, Conversations with AI Part 5.0, 2024).

There is not yet evidence that current systems reliably solve this; it follows from how these systems work. It applies at every point on the sophistication spectrum, from the team prompting ChatGPT with a survey question to the enterprise running a population simulation. All of them are **using the past to predict the future**. That works well enough when the future closely resembles the past. It breaks precisely when the decision concerns something the market has never seen, never priced, and never chosen. Those are exactly the decisions this paper addresses.

What Real People Give You That a Model Can't

The first thing that currently only real people can do is react to something genuinely new. A model works from what already exists. A person can look at a product the market has never seen, consider it in context, and tell you what they think. That alone is reason enough to keep humans in the loop for new launches, new concepts, and unfamiliar combinations of features.

But novelty is only part of the issue. The bigger question is what a buying decision actually is. People do not evaluate products in isolation, and they do not walk around with fixed answers waiting to be retrieved. They build value in the moment, shaped by what they are comparing against, how the choice is framed, what sits next to it, and the trade-offs they are being asked to make.

That is the part AI-generated responses have not yet closed. The obvious objection is that AI models are trained on human behavior, so surely they pick up our quirks along with everything else. They do not, at least not in a way you can trust. When researchers put today's models through classic tests that expose human bias, the models were irrational too, but not in human ways. They also contradicted themselves,

giving different answers to the same question from one run to the next (Macmillan-Scott and Musolesi, 2024).

A large recent test makes the point directly. A Columbia Business School team built digital twins, the most personalized version of this technology, training a model on each person's answers to more than 500 questions, then checked whether the twins actually matched the people they were built from. They barely did. The twins tracked their real humans only weakly, with a correlation of roughly 0.2 where 1.0 would be a perfect match, and did only marginally better in practice than a generic chatbot. They were also consistently too rational, ironing out the impulsive, contradictory texture of real choice. Their verdict is worth borrowing: today's digital twins are less a mirror than a funhouse mirror, and anyone leaning on them risks decisions built on a distortion rather than the real thing (Peng et al., 2025).

The point is not that synthetic respondents are useless. They can help pressure-test ideas, generate hypotheses, and extend existing research. But they cannot yet replace observing how real people respond to real choices. For lower-risk exploration, that may be enough. For pricing, it is not.

Why You Can't Let a Model Set Your Price

Pricing is where a confident wrong answer costs the most. A message test that comes back a little too positive may send you down a weaker creative route. A price set on a mistaken elasticity bleeds margin on every unit sold, and you may not find out until it is already in market. That asymmetry is why pricing deserves harder scrutiny than almost any other research question before you hand it to a model.

The reason is simple: price is not evaluated in isolation. A \$9.99 option does not feel cheap or expensive in the abstract. It feels cheap, expensive, safe, indulgent, or sensible depending on what sits next to it, what it is being compared against, and what trade-off the customer is making in that moment. That is why good pricing research does not simply ask people to name a price. It observes the choices they make between alternatives and infers value from behavior. A model has no behavior to observe. It has only the prices, claims, and patterns it has already absorbed.

This points to a real risk, but also to where the value actually comes from. Given the right data, a model is useful. It reads patterns, flags signals, and speeds up analysis. Without that data, it does not stop and say so. It produces a plausible number anyway. This is not a bug to be patched but a property of how the systems work. OpenAI's own researchers put it plainly: models hallucinate because standard training and evaluation procedures reward guessing over admitting uncertainty, and errors of this kind arise from ordinary statistical pressure rather than from some exotic failure (Kalai et al., 2025). A model that has never seen your category will still answer a question about it, whether or not that answer is grounded in reliable evidence. The value and the risk come from the same place, which is why the model has to be kept working from real evidence rather than filling the gap on its own. And even when the data is there, a deeper problem remains: the model can reproduce the numbers without reproducing the behavior behind them.

The Columbia digital-twin study shows this directly. It tested comparative choice, the kind of trade-off at the heart of pricing. The attraction effect is simple: add a third option that is clearly worse than one of the original choices, and the better original choice starts to look more attractive. The decoy is not there to be chosen. It is there to change the comparison. In the Columbia study, people made product choices described by price and quality features across categories like printers, TVs, and telephone plans. Humans showed the attraction effect. Their digital twins did not (Peng et al., 2025). So the failure is not abstract. In a product-choice setting with price built into the stimulus, the models failed to reproduce the comparison-driven behavior that makes pricing hard in the first place.

The willingness-to-pay evidence lands in the same place. In a Harvard and Microsoft study, researchers directly tested whether LLMs could estimate what consumers would pay for product features using conjoint-style tasks. Their answer was not that models are useless. It was that they are not ready to replace human responses. LLM-derived willingness-to-pay estimates were sometimes comparable to human studies, but often inaccurate, and sometimes wrong-signed. In other words, the model sometimes treated a feature as something consumers would pay more for when human respondents treated it as something that reduced value, or the reverse. That is not a rounding error. That is the model pointing the pricing decision in the wrong direction.

Fine-tuning with prior human survey data improved alignment, including for new product features, but only within the same product category and population. The gains did not generalize reliably to new product categories or customer segments. The key point is that the model became useful only after being corrected by relevant human data, which underscores that these tools are most defensible for extending human research, not replacing the human responses that make pricing estimates valid in the first place (Brand, Israeli and Ngwe, 2023).

The conclusion is simple: a model's read on price is shaped by how it was built and what it was fed, not by how customers actually weigh price against value in a live decision. That does not make it worthless. It makes it a first draft. It can help generate hypotheses, pressure-test scenarios, and speed up analysis. But it cannot be trusted to set elasticity, willingness to pay, or a final price without human choice data anchoring the estimate.

The Right Role for AI in Research

The mistake is to treat this as one question with one answer, as if a study were a single indivisible act that either belongs to a model or belongs to a person. It is not. A study is a sequence. You frame a question, design an instrument, decide who to ask, gather what they say, and make sense of it. AI is not uniformly good or bad across that sequence. It is very good at some stages, ordinary at others, and currently unsuitable for a few. The work is knowing which is which.

One of the fullest accounts of that division comes from a recent Journal of Marketing study that tested LLMs as collaborators across the research workflow rather than as a wholesale replacement for it. Its

finding cuts both ways. Humans and models have complementary skills, and a **human-model hybrid outperformed both the human-only and the model-only** version of the same work (Arora, Chakraborty and Nishimura, 2025). The lesson is not that AI is weak. It is that AI is strong in a different place than people are, and the return comes from the pairing, not from either half run alone.

That strength clusters at the two ends of a study. At the front, in design and framing, a model is a fast and tireless thinking partner. It can draft screener logic, surface hypotheses you had not considered, pressure-test an attribute list before you field it, and play back how a concept might read to a skeptic before a single respondent sees it. None of this asks the model to be right about the market. It asks it to be generative, and generativity is what it has. At the back of a study, the same holds. In qualitative analysis and synthesis, the model matches human performance on surfacing themes and drafting summaries, turning a week of manual work into an afternoon of review (Arora, Chakraborty and Nishimura, 2025). Similar evidence has emerged in quantitative analysis: ChatGPT has been shown to support tasks such as selecting analysis procedures, generating statistical syntax, and presenting results, although the work still requires expert review to check suitability, functionality, and accuracy (Prandner, Wetzelhütter and Hese, 2025). And in between, wherever you already hold real human data in a category, a model can extend it. This is the same supplement role the Harvard team reached from the pricing side earlier, and it holds only under their condition, that relevant human data is already in hand (Brand, Israeli and Ngwe, 2023).

Everything above is either upstream of the decision, helping you ask a sharper question, or downstream of it, helping you digest an answer you already collected from real people. The decision sits in the middle. Establishing what a customer will actually pay, what a feature is worth, and how demand moves with price cannot yet be delegated, because it requires observing how real people construct value under real uncertainty. That is not a support task. It is the decision, and **the decision is where human data has to lead.**

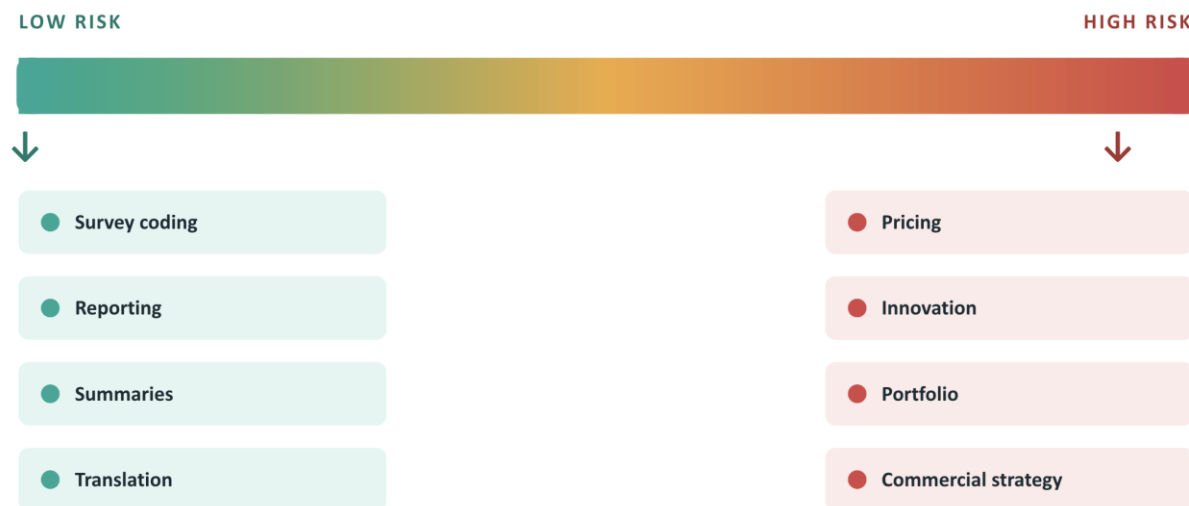
There is a cost question sitting underneath all of this, and it deserves a straight answer. Human research is slower and more expensive than generating synthetic responses and that gap is exactly why the pressure to substitute is real. The framework here is not a case for spending on humans everywhere. It is a case for spending on them where the cost of a wrong answer is high and the decision turns on something the model cannot see. A concept screen for an established line, an early-stage hypothesis, a first pass at synthesis: these tolerate a fast, cheap, approximate answer, and AI can support them. A price, an elasticity, a willingness-to-pay estimate, or the reception of something genuinely new: these are decisions where a confident error compounds across every unit sold, and that is where the more expensive method pays for itself.

The rule is not humans or AI. It is to match the cost of the method to the cost of being wrong.

So the division of labor is not a truce between two rival methods. It is a sequence with a single rule. For any stage of a study, ask one question: does the answer require watching real people weigh real trade-offs, or can it be assembled from what is already known? If it can be assembled, the model can provide significant valuable input. If it requires a real human decision, human data leads and the model supports. Framing, generation, extension, and synthesis fall on the first side. For high-stakes decisions, pricing, willingness to pay, elasticity, and genuinely new product evaluation should remain **anchored in human choice data**. The strongest research function is not the one that resists AI, and it is not the one that hands everything to it. It is the one that knows, study by study and stage by stage, exactly where that line falls, and puts the right method on each side of it.

Where AI Fits: A Risk Spectrum

Matching AI respondent use cases to the level of business risk involved



The right side isn't off-limits — it's where human judgment and oversight matter most.

EPIC's Own Work at the Frontier

We do not observe this boundary from the outside. EPIC actively builds and stress-tests synthetic-respondent methods for choice-based conjoint, so that our guidance rests on hands-on evidence rather than vendor claims. Our most recent experiment involved a toothpaste-focused conjoint benchmarked against a quota-matched human sample. We generated synthetic respondents in the same demographic proportions as the real study, then tested a progression of approaches that varied how we queried the model: directly prompting it for a choice; a two-stage design in which the model gives an opinion first and a second step turns that into a selection; and an embedding-based method, adapted from semantic-similarity work now emerging in the literature (Maier et al., 2025), that asks the model about each feature and matches those preferences to the products on offer.

The results track this paper's thesis closely. The embedding-based method reproduced human preferences substantially more accurately than simply prompting the model for a choice. It also did something direct prompting could not: instead of collapsing every subgroup onto a single averaged consumer, it produced real variation between subgroups, the way an actual population does.

But key limits remained. The differences it produced between subgroups did not reliably match those seen among the real groups, sometimes they pointed in the wrong direction entirely. And closing that gap was not something the model did on its own. It required calibrating the embedding method's prompts with what we already knew about human behavior. In other words, synthetic respondents can be designed to reproduce known subgroup patterns; a general-purpose model will not reliably find those patterns by itself. The method was therefore strongest exactly where this paper says synthetic data is strongest, recovering known, average preferences in a familiar category, and weakest where human choice still has to lead: measuring how specific segments genuinely differ, and how price sensitivity changes between them.

We treat this as the point, not a setback. Building these methods ourselves allows us to tell clients, with specifics rather than caution in the abstract, where a synthetic shortcut may be useful and where it introduces risk. If a genuine breakthrough arrives, it will come from teams rigorously testing these methods against human evidence. EPIC intends to be among the first to introduce it responsibly, while continuing to anchor high-stakes decisions in real human choice data until the evidence supports doing otherwise.

How EPIC Can Help Teams Use AI in Research with Confidence

This is not a question of whether AI belongs in research. It does. The real question is where it can accelerate the work, and where the decision still needs to be anchored in real human choice data.

EPIC helps insight, pricing, product, and commercial teams make that distinction with confidence. Using choice-based conjoint, pricing simulation, segmentation, and cost-adjusted revenue modeling, EPIC can:

- Measure willingness to pay, price elasticity, feature value, and trade-offs using real customer responses.
- Test new products, new concepts, price changes, pack changes, and unfamiliar feature combinations before they reach market.
- Model how customers respond across alternative scenarios, including “none of these”, trade-up, cannibalization, and portfolio impact.
- Segment customers by preference, value perception, and price sensitivity, not demographics alone.
- Translate the evidence into a clear recommendation for pricing, packaging, product strategy, and market validation.

The result is not just a preference readout. It is a commercial decision framework that shows what customers will choose, what they will pay for, where margin risk sits, and which pricing, product, or pack architecture is most likely to convert demand into profitable growth.

References

- Arora, N., Chakraborty, I. and Nishimura, Y. (2025). AI-human hybrids for marketing research: leveraging large language models (LLMs) as collaborators. *Journal of Marketing*, 89(2), 43-70.
- Brand, J., Israeli, A. and Ngwe, D. (2023, revised 2026). Using LLMs for Market Research. Harvard Business School Marketing Unit Working Paper No. 23-062.
- Dig Insights (2025). The truth about synthetic data: can AI replace survey respondents? diginsights.com/resources/truth-about-synthetic-data
- Ipsos (2024). Conversations with AI, Part 5.0. ipsos.com
- Ipsos (2025). Ipsos partners with Stanford University to pioneer the future of market research with synthetic data. [Ipsos.com](https://ipsos.com)
- Kalai, A. T., Nachum, O., Vempala, S. S., & Zhang, E. (2025). Why language models hallucinate (arXiv:2509.04664). arXiv. <https://arxiv.org/abs/2509.04664>
- Macmillan-Scott, O. and Musolesi, M. (2024). (Ir)rationality and cognitive biases in large language models. *Royal Society Open Science*, 11(6), 240255. (Also available as arXiv:2402.09193.)
- Maier, B.F. et al. (2025). LLMs reproduce human purchase intent via semantic similarity elicitation of Likert ratings. arXiv:2510.08338. arxiv.org/abs/2510.08338
- McLean, D. (2026). Testing synthetic data against academic benchmarks: a replication study. Qualtrics / Greenbook, February 2026. greenbook.org
- Peng, T., Gui, G., Brucks, M., Merlau, D.J., Fan, G.J. et al. (2025, revised 2026). Digital Twins as Funhouse Mirrors: Five Key Distortions. arXiv:2509.19088; Columbia Business School Research Paper No. 5518418. [Arxiv.org/abs/2509.19088](https://arxiv.org/abs/2509.19088)
- Prandner, D., Wetzelhütter, D. and Hese, S. (2025). ChatGPT as a data analyst: an exploratory study on AI-supported quantitative data analysis in empirical research. *Frontiers in Education*, 9, 1417900. <https://doi.org/10.3389/feduc.2024.1417900>



Predicting Value. Powering Decisions.

EPIC Insights provides expert-led product and pricing conjoint software that develops actionable retail, wholesale and food service insights in just **7 business days**. By combining a cutting-edge research

platform with deep domain expertise, EPIC empowers global leaders like **Coca-Cola, L'Oréal, Kraft Heinz, PepsiCo, Decathlon, Philips and Harley Davidson** to make precise, data-driven commercial decisions at scale.

To meet with one of our experts visit

EPICINSIGHTS.IO