mdt.

Ex Datis, Claritas

# Me estáis usando los datos sintéticos MAL

Generar encuestas sintéticas no es innovación: es la forma más sofisticada de engañarte a ti mismo.

Manuel Delgado Tenorio · 25 de febrero de 2026

Publicado originalmente en www.manueldelgado.com/blog/me-estais-usando-los-datos-sinteticos-mal



**Llevo años, muchos años, hablando con cierta frecuencia sobre datos sintéticos (por ejemplo, aquí, aquí o aquí). Hace tiempo, predije que veríamos una explosión en su uso y, efectivamente, así ha sido. Siempre defenderé su valor y su utilidad, pero ha llegado el momento de decir esto: os estáis pasando.**

De verdad que no entiendo cómo hemos pasado tan rápido de "*vamos a generar un dataset sintético para analizar y compartir datos evitando riesgos de privacidad*" a "*en lugar de hacer encuestas a seres humanos, se las vamos a hacer a chatbots y nos vamos a creer los resultados*".

Cada vez más a menudo, pasan por mi feed ejemplos de uso de datos sintéticos en investigación de mercados y, en cada ocasión, me echo las manos a la cabeza por lo que leo. El caso más reciente con el que me he topado es este relato sobre cómo Colgate se ahorró pasta por un tubo (*pun intended*) preguntándole a un LLM si compraría su producto en lugar de, hmm, preguntarle a las personas que potencialmente podrían comprar su producto.

Llegado a este punto, me surge una duda genuina: **¿realmente necesito explicar por qué es mala idea dejar de preguntarle a tu mercado y contentarte con lo que te dice una máquina?**

¿Acaso necesitaba el marketing más herramientas para cavar su propia tumba?

## El bucle que no lleva a ningún sitio

Vamos a pensarlo un momento. Un modelo de lenguaje genera respuestas sintéticas a partir de patrones estadísticos que ha aprendido de su corpus de entrenamiento. Si esas respuestas reproducen un patrón conocido (pongamos, que la gente prefiere Colgate a una marca blanca), lo único que has descubierto es... algo que el modelo ya sabía.

Has creado un bucle: entrenas con datos del mundo, generas datos sintéticos, los analizas y, sorpresa, encuentras lo que ya estaba en los datos de entrenamiento. **No has aprendido nada nuevo.**

Y si los resultados *no* reproducen un patrón conocido, tienes un problema aún mayor: no tienes forma de saber si has descubierto algo genuino o si el modelo simplemente ha alucinado.

En ambos casos, el resultado es inútil. Y en ambos casos, **estás completamente desconectado de lo que pasa en la mente real de tus consumidores reales.**

## No es estadística: es numerología

Hay algo que suena muy convincente cuando un proveedor te presenta un estudio con datos sintéticos: hay tablas, hay porcentajes, hay intervalos de confianza. Parece ciencia. Pero no lo es, y la razón es tan fundamental que da un poco de vergüenza tener que explicarla: los conceptos estadísticos como "muestra representativa", "margen de error" o "tamaño muestral efectivo" **solo tienen sentido cuando tus datos proceden de una muestra aleatoria de una población definible.**

Las respuestas de un LLM no son una muestra aleatoria de nada. No existe una "población" de la que el modelo esté extrayendo respuestas. Puedes calcular

medias y desviaciones típicas, pero no puedes hacer inferencias sobre ningún grupo humano a partir de esos números. **Es como calcular la media de los números de teléfono de tus contactos:** el cálculo es correcto, pero el resultado no significa absolutamente nada.

Chris Chapman, un investigador con décadas de experiencia en Google y antes en Microsoft, lo expresó de forma demoledora en un webinar reciente: cuando seis modelos de OpenAI intentaron replicar datos de aprobación presidencial en EEUU (probablemente, el dato más encuestado del planeta), el error medio fue de 22 puntos.**Veintidós.** Y los modelos ni siquiera se ponían de acuerdo entre ellos. Si no funciona con la pregunta más estudiada del mundo, ¿cómo funcionará con un dentífrico?

## Las encuestas no miden: escuchan

Quizá, el problema más profundo es que **usar datos sintéticos para sustituir encuestas revela una incomprensión fundamental de lo que es una encuesta.**

Una encuesta no es un termómetro que mide una "temperatura de opinión" que existe ahí fuera, objetiva e inmutable. Las respuestas de una encuesta son actos de comunicación: las personas construyen sus respuestas en tiempo real, influidas por el contexto, por cómo se formula la pregunta, por su estado de ánimo, por su motivación para participar. Eso no es un *bug* del sistema, es la *feature*. Es precisamente esa riqueza, esa complejidad humana, lo que hace que la investigación de mercado tenga valor.

Un LLM no tiene motivación. No tiene contexto vital. No construye respuestas a partir de experiencias personales. Y, para rematar, los modelos entrenados con RLHF (que son prácticamente todos los que se usan comercialmente, hoy en día) tienen un sesgo de aquiescencia positiva: están entrenados para ser agradables, para decirte lo que quieres oír. Que es, exactamente, **lo contrario de lo que necesitas** cuando le preguntas a tu mercado si tu producto le parece bueno. "Sí, tu idea es fantástica" es lo último que quieres escuchar de una herramienta de investigación, y es lo primero que te va a decir un chatbot.

## El marketing no necesita más atajos

El marketing lleva décadas arrastrando una crisis de credibilidad. Demasiadas decisiones basadas en intuiciones disfrazadas de datos. Demasiadas *vanity metrics*. Demasiado miedo (¿o incapacidad?) para hablar de resultados económicos, de impacto en el negocio. Sustituir a los consumidores reales por simulaciones de consumidores no es innovación: es la última vuelta de tuerca de esa misma dinámica. **Es encontrar una forma aún más eficiente de engañarte a ti mismo.**

Si quieres saber qué piensan tus clientes, pregúntales. A ellos. A los de verdad. Los datos sintéticos tienen usos legítimos y poderosos (privacidad, testing, aumento de datasets pequeños), pero la investigación de mercado no es uno de ellos. No porque la tecnología sea mala, sino **porque la pregunta que intentas responder solo puede responderla un ser humano.**



## Sobre el autor

### Manuel Delgado Tenorio

Manuel Delgado Tenorio es asesor independiente de estrategia de datos e inteligencia artificial, profesor e investigador. Cofundó y vendió una *startup* de inteligencia artificial, fue socio de la práctica de Data & Analytics para servicios financieros en EY y responsable de Data Analytics para el sur de Europa en AWS. Es profesor asociado en IE University, donde imparte la primera asignatura de grado sobre Customer Lifetime Value en España, investiga sobre el CLV en su doctorado en la Universidad Complutense de Madrid y es autor del libro "Customer Lifetime Value (CLV): No todos tus clientes valen lo mismo".

**manueldelgado.com** · **LinkedIn** · **X**

Recibe cada artículo en tu correo con la newsletter Ex Datis, Claritas:
manueldelgado.com/newsletter/suscribete