Vendor Evaluation Checklist

Procurement-ready questions for demos and pilots in synthetic market research.

Prepared for: syntheticmarketresearch.org/resources#templates
Reference baseline: syntheticmarketresearch.org and Standards & Ethics
Persona comparability: SPL taxonomy per The Ten Levels of Synthetic Personas (Ask Ditto)
Date: December 29, 2025

Purpose

This checklist operationalises the association's core principles—**disclosure**, **validation**, **auditability**, **privacy posture**, **comparability**, and **misuse protection**—into concrete questions and evidence requests for vendor evaluations. It is intended to reduce category confusion and prevent "convincing output" from being misinterpreted as measured reality.

How this template relates to earlier SMRA articles

The structure below is aligned with the themes discussed in SMRA's published essays: the need for enforceable standards and governance to avoid manipulation and mis-selling (Standards and Ethics in Synthetic Market Research), the broader moral and ethical landscape of synthetic personas and digital twins (Review of the First Symposium on Moral and Ethical Considerations), and the importance of evidence-based method stacks and benchmarking (A Research Reading List for Synthetic Market Research).

How to use this checklist

Use this document in three phases: (i) initial screening, (ii) structured demo, and (iii) pilot. Do not treat a demo transcript as validation. Require written disclosures and repeatable evidence.

1. **Phase 1 — Gate 0:** request the minimum disclosure pack before a demo.
2. **Phase 2 — Demo:** run Sections 1–10 as a scripted evaluation; ask for evidence artefacts.
3. **Phase 3 — Pilot:** execute a small, repeatable pilot with stability + sensitivity + benchmark checks (Section 11).
4. **Decision:** score vendors consistently using the rubric (Appendix) and document limitations in procurement records.

Procurement principle

If a vendor cannot disclose **population frame**, **grounding class**, and **validation approach**, their outputs should be labelled and treated as **exploratory** (hypothesis-generating), not as decision-grade evidence.

Gate 0: Minimum disclosure pack (non-negotiable)

Require the following in writing before any pilot or procurement decision. If the vendor cannot provide these items, you do not yet have a defensible basis to evaluate validity, privacy risk, or comparability.

- **Population frame statement:** who results represent (geography, language, time window), who is excluded, and why.
- **Method card / study disclosure example:** a completed, study-level disclosure label (or equivalent) for a real study (redacted acceptable).
- **Grounding & provenance summary:** what sources are used (public stats, first-party, third-party), transformations applied, and retention rules.
- **Validation pack:** test-retest stability, sensitivity testing approach, external benchmark evidence (or pilot plan), and known failure modes.
- **Auditability description:** what is logged (prompts/settings at method level, model versions, grounding inputs), and what is exportable to customers.
- **Security & privacy controls:** access controls, encryption posture, isolation between customers, incident response, and red-teaming practices.
- **Use policy:** prohibited/restricted use cases and enforcement mechanisms (gates, monitoring, escalation, offboarding).

For baseline expectations on disclosure, validation, privacy, and prohibited behaviours, see SMRA Standards & Ethics.

1. Category clarity: what exactly is being sold?

The fastest path to failure is buying “synthetic research” without pinning down whether you are purchasing (a) prompt-based personas, (b) a statistically grounded synthetic panel, (c) a twin-like simulation layer, (d) an LLM interface, or (e) a workflow that mixes synthetic and human components.

Questions to ask

- Is the core product a **synthetic panel**, a **persona generator**, **digital twins**, or a **research workflow tool**?
- What is simulated vs measured? Which outputs (if any) are derived from real respondent data for the specific study?
- Which claims do you explicitly **not** make (e.g., “replacement for fieldwork”, “represents specific individuals”)?

Evidence to request

- Example deliverables that are clearly labelled “synthetic” and include a method disclosure/limitations section.
- An architecture overview suitable for audit: components, inputs/outputs, what persists vs what is ephemeral.

Red flags

- “It’s proprietary” used as a blanket refusal to disclose essentials (population frame, grounding class, validation approach).
- Terminology-driven persuasion (“digital twin”) without a testable modelling definition.

2. Population framing & coverage: who does this represent?

Synthetic market research only has meaning if the population frame is explicit. Without it, you cannot interpret results, compare vendors, or evaluate bias and coverage limits.

Questions to ask

- What is the target population (country/region, language, timeframe)?
- How are segments defined (demographics, psychographics, behaviours) and what evidence supports those definitions?
- What is your **coverage statement**: where is the method reliable vs out-of-domain?
- Do panels/personas drift over time? What is held constant vs updated?

Evidence to request

- Written population frame attached to every study output.
- How quotas/weighting are implemented (if any) and how marginal distributions are checked.

Red flags

- “Global consumers” claims without region/language-specific validation evidence.
- No clarity on what changes between runs (freshness, retrieval sources, model versions, sampling seeds/settings).

3. Personas, twins, and comparability: require a clear persona specification (incl. SPL)

One of the largest sources of mis-selling in this category is the word “persona”. It can mean anything from a single prompt to a persistent agent with memory and state. Require a written persona specification, and insist on a declared capability level for comparability.

Synthetic Persona Level (SPL) comparability

Ask vendors to declare the SPL level(s) supported for each persona product tier, and to provide evidence for those claims. See Ask Ditto: The Ten Levels of Synthetic Personas. Treat SPL as a claim that must be validated, not as a marketing badge.

Minimum persona specification (require in writing)

- **Persona type:** segment-level persona vs twin-like proxy vs respondent generator.
- **Representation target:** archetype, micro-cohort, or individual-like proxy; and whether you ever claim to represent specific people.
- **SPL declaration:** the SPL level(s) supported and the implemented features that justify the level.
- **Memory model:** none / retrieval memory / structured episodic memory + decay (and what persists).
- **Temporal context:** does “today” exist (context streaming) and how is provenance controlled?
- **State variables:** what latent state exists (goals, beliefs, affect), how it evolves, and whether it is auditable.
- **World connections:** external tools/feeds accessed (if any), with logging and allow/deny controls.
- **Interaction model:** single-turn Q&A; vs multi-turn interviews vs agentic workflows vs multi-agent simulations.

SPL reality-check questions

- If you claim SPL 3+ (memory): show how memory is stored, retrieved, bounded, and decayed; demonstrate controls against “perfect recall”.
- If you claim SPL 4+ (context streaming): what feeds are allowed/blocked, and how is provenance logged?
- If you claim SPL 5+ (state): list state variables and show auditable state transitions across a run.
- If you claim SPL 8 (closed-loop runtime): describe the runtime loop and which artefacts customers can inspect (not only internal teams).
- If you claim SPL 9–10 (social/multi-agent): show that dynamics are stable under reruns and not just “theatrical transcripts”.

Red flags

- “Digital twins” claims with no operational definition and no restrictions on person-like inference.
- Refusal to specify whether personas are prompt-only vs persistent agents with memory/state.
- Capability level claims without independent artefacts that allow evaluation.

4. Grounding & provenance: what is this built on, and is it legitimate?

Synthetic outputs do not eliminate provenance risk; they can obscure lineage behind fluent narratives. Require auditable provenance, retention rules, and clear statements about whether and how personal data is used.

Questions to ask

- What are your grounding inputs: public statistics, curated corpora, survey microdata, first-party client data, third-party data?
- Do you use data about identifiable individuals? If yes, what is the lawful/ethical basis and purpose limitation?
- What is the ‘purpose distance’ between original data collection and your modelling use?
- Do you train or improve shared models using customer interactions, prompts, or outputs (opt-in vs opt-out)?

Evidence to request

- A provenance statement template attached to every study (inputs, transformations, what changed, what is retained).
- A data flow diagram and retention schedule (prompts, outputs, embeddings, derived features, logs).

Red flags

- “Synthetic means no privacy risk” claims.
- Evasive answers about training on customer data or about retention of prompts/outputs.

5. Validation: can they prove reliability (not just plausibility)?

The central ethical failure mode in synthetic research is over-claiming—treating simulation as measurement. Vendors must show stability, robustness, and benchmark alignment appropriate to the intended use.

Questions to ask

- What is your test-retest stability on a standard study configuration?
- How sensitive are results to prompt/context changes (controlled perturbations)?
- What external benchmarks have you run (public stats, known-truth tasks, fieldwork comparisons), and can we replicate them?
- What are documented failure modes (domains, populations, question types) and how do you prevent misuse?

Evidence to request

- A stability report (variance ranges, not cherry-picked examples).
- Benchmark results with methodology sufficient for replication (not screenshots).
- A known-truth task pack relevant to your domain (or willingness to run one in pilot).

Red flags

- Validation framed as ‘sounds human’ rather than measured stability and benchmark alignment.
- No willingness to run blinded comparisons against limited fieldwork where feasible.

6. Research workflow integrity: do they support real study design?

Synthetic market research should behave like a research instrument: fixed stimuli, fixed wording, repeatable run settings, and disclosed aggregation logic. Avoid tools that collapse into ad-hoc prompting.

Questions to ask

- Which study types are supported (concept tests, message tests, pricing exploration, scenario simulation)?
- How are responses aggregated (distributions, uncertainty indicators, segment cuts)?
- Can you lock stimuli, wording, and run settings for repeatability?
- What controls exist to reduce operator prompt bias?

Evidence to request

- A protocol template and an example of a repeatable run configuration.
- Documentation of aggregation logic, weighting, and normalisation steps.

Red flags

- “We just ask the model” with no protocol, controls, or variance reporting.

7. Bias, fairness & representational harm

Bias includes systematic representational gaps—where certain groups are mis-modelled, stereotyped, or erased. Require subgroup evaluation, a coverage statement, and documented mitigations.

Questions to ask

- How do you evaluate representational coverage across demographics/segments relevant to the population frame?
- How do you detect and mitigate stereotyping or narrative harm in persona outputs?
- Do you provide ‘do not use for X group/topic’ constraints where reliability is weak?

Evidence to request

- Bias assessment documentation and examples of mitigations (not just principles).
- Segment-level benchmark results where feasible.

Red flags

- Refusal to discuss subgroup performance or coverage limits.
- ‘We removed bias’ claims with no measurement plan.

8. Privacy posture & security: “synthetic” is not automatically anonymous

A system can generate large volumes of human-like data while still creating privacy and security risks. Require explicit threat modelling, retention controls, customer isolation, and privacy testing.

Questions to ask

- What privacy tests do you run (uniqueness checks, membership-inference-style probes, red teaming)?
- What data is stored (inputs, prompts, outputs, embeddings, logs), for how long, and who can access it?
- How is customer data isolated? What prevents cross-customer leakage?
- What controls exist for sensitive topics and re-identification attempts?

Evidence to request

- Security overview (access control, encryption, incident response) and audit artefacts if available.
- Retention/deletion policy that explicitly covers prompts and outputs.

Red flags

- Privacy described as a marketing property rather than a tested property.
- Vague answers about prompt/output retention or training use.

9. Auditability & reproducibility

If two teams cannot re-run the same method and obtain comparable results (within reported variance), you do not have a research instrument; you have a story generator.

Questions to ask

- What is logged per run (model version, run settings, prompts at method level, grounding sources, aggregation steps)?
- Can customers export run artefacts for internal governance review?
- How are model updates handled and drift tracked over time?

Evidence to request

- An example audit log and a reproducibility guide for re-running a study configuration.
- Change logs describing what changed between versions and expected impact.

Red flags

- No logs and no ability to re-run with comparable settings.
- Frequent updates with no drift monitoring or communication.

10. Misuse safeguards & governance: preventing Cambridge-Analytica-style dynamics

Synthetic market research can lower the cost of iterative profiling and message testing. Without hard boundaries, it can industrialise manipulation. Procurement should test whether safeguards are enforceable, not merely aspirational.

Questions to ask

- Which uses are prohibited (misinformation, exploitation of vulnerability, discriminatory targeting, sensitive-trait inference)?
- Which domains require enhanced review (minors, health, financial distress, elections)?
- What enforcement exists (policy gates, monitoring, contractual clauses, offboarding)?
- How is accountability assigned across vendor, client, and integrators?

Evidence to request

- Use policy and enforcement description (including escalation and incident handling).
- Examples of disclosures embedded in the product (not buried in terms).

Deal-breaker red flags

- Encouraging deception, manipulation, or targeted harassment.
- Claiming the system describes specific real individuals, or implying “digital twin” equals a person-like proxy without strict consent boundaries.
- A “black box by design” posture that prevents audit and governance.

11. Pilot plan: the minimum test you should run

Do not rely on a single demo. Require a pilot that forces repeatability and benchmark discipline. A minimal pilot includes the components below.

1. **Two repeatable study types** (e.g., concept test + message test) with fixed stimuli, wording, and run settings.
2. **Test-retest runs**: run each study multiple times; report variance and explain drivers.
3. **Sensitivity tests**: apply controlled perturbations (wording, context constraints) and measure conclusion shifts.
4. **At least one benchmark**: compare against a small human sample, public statistics, or known-truth dataset where feasible.
5. **Disclosure package**: every output includes population frame, grounding class, validation performed, limitations, reproducibility notes.

Pilot acceptance criteria (example)

The vendor should demonstrate stable directional conclusions under repeat runs, transparent disclosures, and at least one credible benchmark alignment on a relevant task. If not, restrict use to exploration and hypothesis generation.

Appendix: Scoring rubric (optional)

Score each category 0–2 (0 = absent/evaded, 1 = partial/opaque, 2 = complete/testable). Use the totals to compare vendors consistently and to document procurement rationale.

Category	0	1	2
Population frame & coverage	Unspecified	Stated, weak justification	Stated + defensible + coverage limits
Persona specification (incl. SPL)	Marketing-only	Partial spec	Complete spec + evidence + tests
Grounding & provenance	Opaque	High-level only	Auditable provenance + retention rules
Validation & benchmarking	None / anecdotes	Internal checks only	Stability + sensitivity + external benchmark
Disclosure & limitations	Absent	Partial	Standard disclosure per study + clear limits
Auditability & reproducibility	No logs / no reruns	Limited	Exportable logs + comparable re-runs
Privacy posture & security	Hand-waving	Controls without tests	Controls + threat model + privacy testing
Misuse safeguards	None	Policy only	Policy + enforcement + escalation

Decision guidance (example)

Vendors that score low on validation, disclosure, provenance, or auditability should be restricted to exploratory use. Vendors that cannot meet Gate 0 should not be procured for decision support.

References

- Synthetic Market Research Association (homepage): https://syntheticmarketresearch.org/
- SMRA Standards & Ethics: https://syntheticmarketresearch.org/standards
- SMRA Resources (Templates): https://syntheticmarketresearch.org/resources#templates
- SMRA Blog — Standards and Ethics in Synthetic Market Research: https://syntheticmarketresearch.org/blog-standards-ethics
- SMRA Blog — Review of the First Symposium on Moral and Ethical Considerations: https://syntheticmarketresearch.org/blog-symposium-ethics
- SMRA Blog — A Research Reading List for Synthetic Market Research: https://syntheticmarketresearch.org/blog-reading-list
- Ask Ditto — The Ten Levels of Synthetic Personas (SPL): https://askditto.io/news/the-ten-levels-of-synthetic-personas

Accessed December 29, 2025.