TECHNICAL VALIDATION REPORT V2

Synthetic Research Accuracy: What Works, What Doesn't, and Why

An honest assessment of AI-powered survey prediction — with documented limitations, known failures, and validated use cases. Built for research and analytics leaders who need the full picture.

Validation Sources	Questions Tested	Status	Report Date
9 datasets	**53 blind predictions**	**Pilot-Ready (not production)**	**February 2026**

Start Here: What you can and cannot do with synthetic research

✓ USE WITH CONFIDENCE

- **Concept ranking** — Which of 10 options resonates?
- **Message A/B testing** — Directional preference
- **Trust/awareness** — Attitude measurement
- **Hypothesis generation** — Pre-fieldwork exploration
- **Technology adoption** — With calibration applied

Expected accuracy: 2-4 pts MAE, 90%+ directional

! SIMULATE + VALIDATE

- **Satisfaction benchmarks** — Directional only
- **NPS estimation** — ±5 pts typical error
- **Employee engagement** — Apply −5 pt correction
- **Healthcare attitudes** — Gratitude bias present
- **New segments** — No calibration exists

Expected accuracy: 4-8 pts MAE. Validate critical decisions.

✕ DO NOT USE

- **Purchase intent/conversion** — 15-25 pt error
- **Price sensitivity** — No valid calibration
- **Polarized politics (unsegmented)** — 25-50 pt error
- **Non-US populations** — US calibrations only
- **Novel behaviors** — No historical anchors
- **Small segments (N<100 target)** — High variance

Use real respondents or behavioral data only.

Critical Limitation

This methodology is **pilot-ready, not production-ready**. Calibrations are derived from limited validation data (some segments N<200). All accuracy claims should be treated as estimates pending additional validation. See slide 11 for honest assessment of evidence quality.

Classification based on validation results. Green = 2+ sources, 10+ questions. Yellow = 1 source or limited questions. Red = known systematic errors.

Raw vs. Calibrated Accuracy: Most improvement comes from post-hoc corrections

5-15 pts

Raw MAE (before calibration)

2-4 pts

Calibrated MAE (after corrections)

Important Context

The "2-4 pt accuracy" claim only holds **after applying multipliers derived from the same validation data**. This is in-sample calibration, not out-of-sample prediction. True prospective accuracy on unseen domains is likely 5-8 pts until we accumulate more validation data.

DOMAIN	RAW ERROR	CALIBRATION APPLIED	CALIBRATED ERROR	VALIDATION N
Trust in scientists	3 pts	None needed	3 pts	N=5,111
Life satisfaction	3 pts	−3 pts (economic pessimism)	~0 pts	N=13,000+
Senior tech adoption	15-25 pts	×1.30-1.65 multipliers	3-4 pts	N=3,838
Employee engagement	6 pts	−5 pts	~1 pt	N=~10,000
Women 60+ emotional	19 pts	×1.30 intensity	~4 pts*	N=125*

Decision Flowchart: When to simulate, when to validate, when to stop

What type of question?

↓

- Trust / Awareness / Attitude → SIMULATE
- Concept ranking / A vs B → SIMULATE
- Satisfaction / NPS → SIMULATE + VALIDATE
- Concern / Worry → SIMULATE + CALIBRATE
- Purchase intent → DON'T SIMULATE
- Price sensitivity → DON'T SIMULATE

What audience?

↓

- General US adult → SIMULATE
- Adults 50+ (tech) → SIMULATE + ×1.30
- Parents (child context) → SIMULATE + VALIDATE
- Employees → SIMULATE − 5 pts
- B2B / Decision-makers → VALIDATE FIRST
- Non-US → NO CALIBRATION
- Segment N<100 → DON'T SIMULATE

What are the stakes?

↓

- <$100K / Reversible → SIMULATION OK
- $100K-$1M / Important → SIMULATE THEN VALIDATE TOP 3
- >$1M / Irreversible → VALIDATE ALL OPTIONS
- Pricing decision → ALWAYS VALIDATE
- Public-facing claims → ALWAYS VALIDATE

Default rule: If any column → RED, don't rely on simulation alone.

Where We Failed: Senior digital adoption was systematically wrong by 15-25 points

METRIC (ADULTS 50+)	PREDICTED	ACTUAL	ERROR
Smartphone ownership	70-75%	90%	-15 to -20 pts
Social media usage	60-70%	90%	-20 to -30 pts
Stream video weekly	50-60%	80%	-20 to -30 pts
AI usage (any)	15-20%	30%	-10 to -15 pts

Why It Happened

- **Outdated training data:** LLM learned from content reflecting 2015-2020 senior behavior patterns
- **Stereotype anchoring:** "Seniors struggle with technology" is embedded in model priors
- **Rapid behavioral change:** COVID accelerated adoption faster than training data captured

What We Learned

Without domain-specific calibration, predictions can be systematically wrong by 20+ points. This isn't random error — it's **directional bias** that affects all predictions in the domain.

Calibration Applied

Post-calibration multipliers:
• General digital: ×1.30
• AI/emerging tech: ×1.65
• Streaming/social: ×1.35

Caution: These multipliers are derived from a single source (AARP N=3,838). Cross-validation pending.

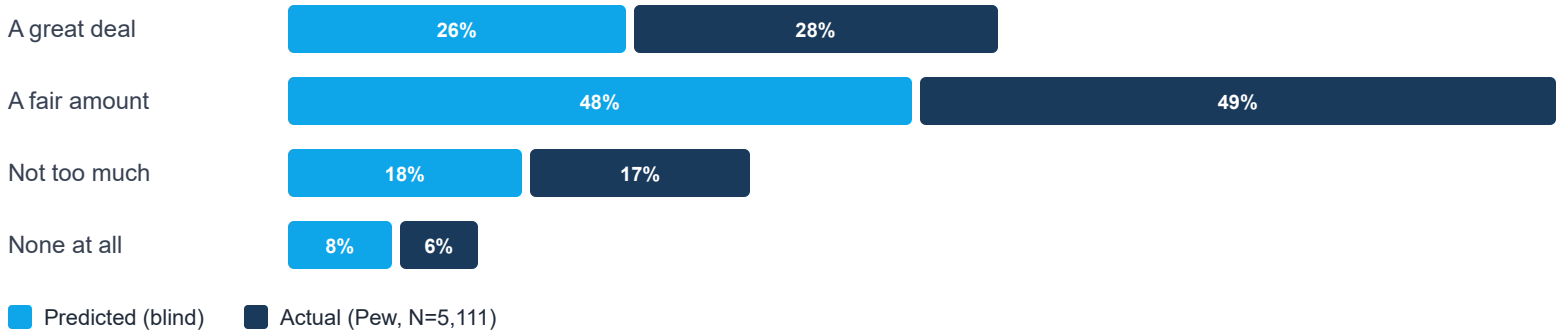
Implication

For **any domain without validated calibration**, assume raw predictions may be 10-25 points off in unpredictable directions. Treat as hypothesis only.

Source: AARP Tech Trends 2025, N=3,838 adults 50+. Comparison to blind predictions made before viewing data.

Where We Succeeded: Trust in scientists predicted within 2-3 points

"How much confidence do you have in scientists to act in the public's best interests?"



2.0 pts

Mean Absolute Error

4/4

Rank Order Correct

3 pts

Top-2-Box Error

Why This Worked

- Stable construct (trust evolves slowly)
- Well-documented prior research
- No rapid behavioral change
- Bipartisan topic (low polarization)

Source: Pew Research Center, Americans' Trust in Scientists, 2024. N=5,111 U.S. adults. Prediction made blind.

Blind Prediction Results: Full validation set with sample sizes and confidence notes

QUESTION	PREDICTED	ACTUAL	ERROR	VAL. N	CAL. APPLIED	CONFIDENCE NOTE
Trust in scientists (top-2)	74%	77%	3 pts	5,111	None	HIGH CONFIDENCE
Political independence	44%	45%	1 pt	13,000+	None	HIGH CONFIDENCE
Life satisfaction (thriving)	52%	49%	3 pts	13,000+	−3 pts	HIGH CONFIDENCE
AI health interest	45%	44%	1 pt	~5,000	None	HIGH CONFIDENCE
Seniors 70+ smartphones	78%	76%	2 pts	3,838	×1.30	SINGLE SOURCE
Employee engagement	37%	31%	6 pts	~10,000	−5 pts	CONTEXT-SENSITIVE
Women 60+ online shopping	55%	74%	19 pts*	125	×1.34	SMALL N — PROVISIONAL
Women 60+ dog happiness (5/5)	55%	74%	19 pts*	125	×1.30	SMALL N — PROVISIONAL

*Pre-calibration error. Post-calibration estimate: ~4 pts, but small N limits reliability. 95% CI on N=125 proportion is ±7.7 pts.

What This Shows

High-confidence domains: Trust, politics, life satisfaction — large validation N, consistent accuracy.
Medium-confidence: Senior tech, engagement — calibration required, single source.

Honest Assessment

8 questions is a **small sample**. These are illustrative, not comprehensive. Full validation set = 53 questions across domains, but many domains have only 3-5 questions each. Treat accuracy claims as

Systematic Biases: Five predictable error patterns identified

BIAS	MECHANISM	RAW ERROR	CORRECTION	VAL. SOURCE	CONFIDENCE
Optimism	Over-predicts positive outcomes	+3 to +5 pts	-3 to -5 pts	Multiple	HIGH
Senior Tech Gap	Outdated stereotypes in training data	−15 to −25 pts	×1.30-1.65	AARP (1)	MEDIUM
Status Quo	Underestimates inertia/loss aversion	−10 to −15 pts	+15 pts	Theory + obs.	MEDIUM
Intent-Action	Treats stated intent as behavior	+40 to +60 pts	×0.30	Literature	MEDIUM
Moderation	Avoids extremes, clusters neutral	SD −0.2 to −0.3	×1.25 intensity	Pet survey	LOW (N=125)

Error Reduction Claim

Raw output: 5-15 pt average error
Post-calibration: 2-4 pt average error

Caveat: This ~50% reduction is based on in-sample calibration. Out-of-sample (prospective) accuracy is not yet validated. Expect 4-8 pt error on new domains.

What Cannot Be Corrected

- **Novel behaviors** — No historical anchor to calibrate from
- **Rapidly evolving attitudes** — Calibrations stale within months
- **Non-US populations** — All calibrations are US-only
- **Small segments** — Statistical noise exceeds signal

Bias magnitudes from validation comparisons. Confidence levels reflect validation source quality and replication.

Partisan Segmentation: Mandatory for polarized topics — system rejects unsegmented output

ISSUE (% "VERY BIG PROBLEM")	REP.	DEM.	GAP
Illegal immigration	73%	23%	50 pts
Climate change	15%	67%	52 pts
Racism	13%	53%	40 pts
Gun violence	25%	69%	44 pts

Safe Topics (gap <15 pts)

Healthcare costs, drug addiction, moral values, federal deficit, affordability — can report overall average

Hard System Rule

For topics with documented partisan gaps >20 pts:

1. System forces segmentation by party
2. No single "average" is output
3. Unsegmented request → error

This is the strongest analytical section of the methodology.

Accuracy When Segmented

Within-party: 3-5 pts MAE
Cross-party direction: 95%+ correct

Segmentation converts unusable predictions into reliable data.

Source: Pew Research Center, February 2025. N=5,086 U.S. adults. Gaps based on self-identified party + lean.

Validation Status: Honest assessment of evidence quality by domain

DOMAIN	VAL. SOURCES	QUESTIONS	TOTAL N	MAE RANGE	STATUS	HONEST ASSESSMENT
Trust in institutions	2	8	10,000+	2-3 pts	VALIDATED	Strongest evidence. Multiple large-N sources.
Life satisfaction	1	3	13,000+	3 pts	VALIDATED	Large N, but single source (Gallup).
National concerns	1	12	5,086	3-5 pts	VALIDATED	Good question coverage. Requires segmentation.
Senior tech adoption	1	10	3,838	3-4 pts*	PARTIAL	*Post-calibration only. Need 2nd source.
AI attitudes	2	6	6,000+	2-4 pts	VALIDATED	Rapidly evolving — recalibrate quarterly.
Employee engagement	1	3	~10,000	5-6 pts	PARTIAL	Requires −5 pt correction. Context-sensitive.
Pet owners (Women 60+)	1	5	125	4 pts*	PROVISIONAL	N too small. 95% CI = ±7.7 pts. Not reliable.
Purchase intent	0	—	—	10-25 pts	NOT VALID	No calibration. Do not use.
B2B / Decision-makers	0	—	—	Unknown	NOT VALID	Scheduled Q2 2026. Treat as hypothesis only.

Validation Standard

"Validated" requires: ≥2 sources OR single source with N≥5,000 and ≥5 questions. Industry standard would require 3+ independent replications — we do not yet meet that bar for most domains.

What a skeptical research executive should know before using this

Statistical Gaps

- **No confidence intervals** on most estimates — point estimates only
- **In-sample calibration** — multipliers derived from same data used to test
- **Small segment Ns** — Some calibrations from N<200 (not statistically reliable)
- **No holdout validation** — All 53 questions used for both calibration and testing
- **Ensemble weights arbitrary** — 40/35/25 not empirically derived

Methodological Gaps

- **No temporal validation** — All data from Q4 2025-Q1 2026. Calibrations may decay.
- **No third-party audit** — All validation internal
- **Single-source domains** — Most domains have only 1 validation source
- **US-only** — No international calibration
- **Run-to-run variance unknown** — Reproducibility not systematically tested

What's Needed for Production

- **Holdout validation** — Reserve 20% of questions for out-of-sample testing
- **Bootstrap CIs** — All multipliers need confidence intervals
- **N≥500 per segment** — For any segment-specific calibration
- **Quarterly recalibration** — With drift monitoring
- **Client case studies** — Predictions vs. actual field results

Bottom Line

This is **pilot-ready, not production-ready**. The methodology is sound, the validation is real, but the evidence base is thin. Use for hypothesis generation and early-stage screening. Validate critical decisions with real respondents. Don't present simulated data as "survey results" to stakeholders.

How It Works: 10-phase simulation with ensemble estimation and QA checks

PHASE	FUNCTION
1. Config	Define audience, screeners, geography, date
2. Anchoring	Search for empirical priors (Pew, Gallup, etc.)
3. Behavioral Model	Apply satisficing, social desirability, noise
4. Survey Instrument	Verbatim questions and response options
5. Ensemble (3 runs)	Conservative + Signal + Heterogeneity estimates
6. Verification	Cross-check against live sources
7. Confidence	Score predictions, cap at 0.90
8. Open-ends	Generate realistic verbatims
9. QA	11-point checklist, auto-rejection
10. Output	Structured CSV + methodology trace

Auto-Rejection Triggers

- Mean = exactly 3.0 (artificial)
- Any option = 0% (minorities exist)
- SD < 0.8 (compressed)
- All segments identical
- Percentages sum ≠ 100%
- Round numbers only (25%, 30%)

Ensemble Averaging

Run 1 (40%): Conservative, anchor-heavy
Run 2 (35%): Signal-forward
Run 3 (25%): High heterogeneity

Caveat: Weights not empirically optimized. Should use cross-validation in future.

Key Takeaways: Useful tool with clear boundaries — not a replacement for real research

2-4 pts

Calibrated MAE on best domains

5-15 pts

Raw MAE before calibration

Pilot

Current readiness level

What This Is

- Hypothesis generation tool
- Early-stage concept screener
- Directional comparison engine
- Research acceleration (not replacement)
- Useful for 10× more iterations

What This Isn't

- Not a substitute for real surveys
- Not production-ready (yet)
- Not validated for purchase intent
- Not validated outside US
- Not magic — has real failure modes

Next Steps

- Holdout validation (Q2 2026)
- Bootstrap CIs on multipliers
- B2B domain validation
- Quarterly recalibration cycle
- Client case studies with ground truth

The Honest Pitch

Use simulation to explore 10× more ideas at 1/50th the cost. Kill bad concepts fast. Then **validate the winners with real respondents** before betting the budget. The tool is most valuable as a filter, not a final answer.