CrowdWave Accuracy Framework

February 2026 | Chief Accuracy Officer

Calibration reduces prediction error by 79%, making AI-simulated surveys reliable for directional research decisions

79%

Error reduction
vs. naive LLM

1.9

Mean absolute error
(points)

20+

Validated
domains

5M+

Human survey
responses

> **Key finding:** Raw LLMs are 25% less accurate than expert forecasters. Calibration against human survey data closes this gap, enabling reliable predictions for established topics.

Source: CrowdWave validation testing (27 test cases, 6 domains); ForecastBench 2025

Raw LLM predictions fail at rates unacceptable for business decisions; synthetic surveys show only 0.30 correlation for new concepts

The accuracy problem is well-documented in peer-reviewed research

System	Brier Score	Gap to Expert
Superforecasters (human experts)	0.081	—
GPT-4.5 (best available LLM)	0.101	25% worse
GPT-4	0.131	62% worse
Median public forecaster	0.150+	85% worse

Synthetic survey validation (Dig Insights, 2025)

Prediction task	Correlation	Business utility
Backcasting (known events)	0.85	Acceptable
Forecasting (future events)	0.50	Limited
New product concepts	0.30	Unacceptable

Source: Forecasting Research Institute (ForecastBench, Oct 2025); Dig Insights synthetic validation study (N=500, 30 movies)

Accuracy varies predictably by question type, enabling appropriate use case selection and confidence calibration

Accuracy spectrum based on validated testing

Accuracy zone	Error range	Question types	Recommendation
High	±2-3 points	Trust scales, awareness, party ID, bipartisan rankings	Use for decisions
Medium	±4-5 points	Satisfaction, NPS, concern levels, technology comfort	Use for direction
Low	±8-15 points	Purchase intent, price sensitivity, polarized topics	Validate first

> **Implication:** Match question type to use case. High-accuracy questions support decisions; low-accuracy questions require human validation before action.

Source: CrowdWave accuracy testing (27 test cases); ACCURACY_BY_QUESTION_TYPE.md

High-accuracy zone (±2-3 points): Trust, awareness, and demographic questions can be used with confidence

Validated performance on high-accuracy question types

Question type	Validated MAE	Example	Calibrated vs. Actual
Trust/confidence scales	2 pts	"How much do you trust scientists?"	77% vs. 77%
Awareness (Yes/No)	<3 pts	"Have you heard of X?"	—
Political party ID	1 pt	% Independent	44% vs. 45%
Bipartisan issue rankings	2-3 pts	Top concerns (economy, healthcare)	—
Basic demographics	<2 pts	Device ownership, behavior frequency	—

Why these work:

- Stable attitudes with abundant benchmark data
- Low emotional volatility
- Minimal recency bias
- Training data aligns with current reality

Source: Gallup (N=13,000+), Pew Research (N=5,000+), validated calibrations

Medium-accuracy zone (±4-5 points): Satisfaction and NPS require industry-specific calibration

Calibrated NPS baselines by industry (Survicate 2025, N=5.4M responses)

Industry	Median NPS	B2B	B2C	LLM default error
Manufacturing	65	66	62	-25 pts
Healthcare	61	38	70	-20 pts
Retail/Ecommerce	55	55	54	-15 pts
Fintech	46	—	—	-10 pts
Software	30	29	47	+5 pts

Key insight: LLMs assume NPS of 35-40 for all industries. Actual variance is 30+ points. Industry-specific calibration is required for accuracy.

Source: Survicate NPS Benchmark 2025 (599 companies, 5.4M responses); Retently 2025

Low-accuracy zone (±8-15 points): Intent questions require conversion factors; polarized topics require segmentation

Intent-to-action gap (validated)

Stated response	Actual conversion	Correction factor
"Very likely" to purchase	25-35%	×0.30
"Likely" to purchase	10-20%	×0.15
"Might consider"	3-8%	×0.05

Partisan segmentation required (Pew Research, Feb 2025)

Topic	Overall	Republican	Democrat	Gap
Illegal immigration	48%	75%	25%	50 pts
Climate change	45%	25%	70%	45 pts
Gun violence	52%	35%	70%	35 pts

> **Rule:** Never predict a single number for polarized topics. The "average" doesn't represent anyone.

Source: Academic meta-analysis (intent-action gap); Pew Research Feb 2025 (N=5,086)

Eight documented LLM bias patterns enable systematic correction

Bias patterns with validated correction factors

Bias pattern	Direction	Correction	Validation source
Senior tech adoption	Under-predicts	×1.30-1.65	AARP 2025 (N=3,838)
AI concern (general)	Over-predicts	×0.90	Pew/YouGov 2025
Status quo preference	Under-predicts	+15-20 pts	Behavioral research
Intent-to-action	Over-predicts	×0.30-0.55	Meta-analysis
Emotional intensity	Under-predicts	×1.20-1.30	Pet owner study (N=173)
Life satisfaction (uncertainty)	Over-predicts	-3 to -5 pts	Gallup 2025
Partisan averaging	Incorrect	Segment	Pew 2025
Open-end quality	Over-polished	20% low-quality	Industry benchmark

Source: Validated calibrations documented in CALIBRATION_MEMORY.md

Calibration reduced mean absolute error from 9.1 points to 1.9 points across 27 test cases

Validation testing results

Metric	Naive LLM	Calibrated	Improvement
Mean absolute error	9.1 pts	1.9 pts	79%
Predictions within 2 pts	7%	81%	+74 pts
Predictions within 5 pts	30%	100%	+70 pts

Example predictions vs. actuals

Prediction	Naive	Calibrated	Actual	Error
Adults 50+ smartphone ownership	72%	89%	90%	1 pt
Political independents	35%	44%	45%	1 pt
AI "very concerned"	58%	50%	48%	2 pts
Manufacturing NPS	40	64	65	1 pt

Source: CrowdWave validation testing (27 test cases, 6 domains); ACCURACY_TESTS.md

Executive audiences require role-specific calibration; CHROs are 75% more concerned about AI than CEOs

C-suite concern calibration by role (Conference Board 2026, N=1,732)

Concern	CEO	CFO	CHRO	CMO
Cyberattacks	×1.30	×1.40	×1.60	×0.90
AI disruption	×0.90	×1.05	×1.40	×1.10
Business transformation	×1.50	×1.15	×1.70	×1.40
Economic uncertainty	×1.35	×1.50	×1.50	×1.25

Key insight: Generic "executive" predictions miss role-based variations. CHROs show 40% higher AI concern than CEOs; CMOs show 40% lower cyber concern than CHROs.

Source: Conference Board Global C-Suite Survey 2026 (N=1,732 executives)

Calibration data covers 20+ domains with 100+ multipliers from 15 authoritative sources

Domain coverage matrix

Domain	Status	Accuracy	Primary sources
Trust in institutions	✅ Validated	±2 pts	Edelman, Pew, Gallup
Political identity	✅ Validated	±1 pt	Gallup (N=13K+)
Technology adoption	✅ Validated	±3 pts	AARP, Pew
NPS by industry	✅ Validated	±4 pts	Survicate (N=5.4M)
Executive attitudes	✅ Validated	±4 pts	Conference Board
Consumer concerns	✅ Validated	±3 pts	Pew, McKinsey
Travel/hospitality	✅ Validated	±3 pts	CLIA, JD Power
Healthcare attitudes	⚠️ Partial	±5 pts	KFF, Gallup
Purchase intent	⚠️ Partial	±10 pts	Apply intent gap
Price sensitivity	❌ Gap	Unknown	Needs validation

Source: CALIBRATION_MEMORY.md, CALIBRATION_EXPANSION.md (combined ~50KB documentation)

Recommended applications align with accuracy zones; high-stakes decisions require validation

Use case guidance by confidence level

Confidence	Applications	Accuracy basis
High	Concept testing, audience sizing, trend validation, priority ranking, benchmark comparison	±2-5 pts, 20+ validated domains
Medium	Hypothesis generation, early-stage screening, directional guidance	±5-8 pts, calibration applied
Low	New product concepts, pricing research, emerging categories	Validate with human sample

Not recommended without human validation:

- Exact purchase conversion (use A/B testing)
- Polarized political topics (segment by party)
- Novel behaviors with no training data
- Regulatory or legal evidence requirements

Source: CrowdWave accuracy framework; industry best practices

Competitive differentiation: documented accuracy, known limits, transparent methodology

Comparison to alternatives

Capability	Raw LLM	Competitors	CrowdWave
Documented accuracy	None	"95%" (unvalidated)	27 test cases
Human validation data	None	Unclear	5M+ responses
Bias corrections	None	None documented	8 patterns
Domain calibrations	None	Generic	20+ domains
Confidence scoring	None	None	Per-prediction
Known limitations	None	None documented	Full transparency

Differentiation: Other vendors claim magic. We document our methodology, show our work, and tell you when NOT to trust the output.

Source: Competitive analysis; COMPETITIVE_BENCHMARKS.md

Implementation requires 10-phase methodology with ensemble estimation and continuous calibration

Production workflow (simplified)

Phase	Function	Accuracy impact
2	Anchor on prior benchmark data	Critical
5	Ensemble simulation (3 independent runs)	Reduces variance 40%
6	Verification against live data	Correction opportunity
7	Confidence calibration	Quality signal
9	QA checklist	Error prevention

Ensemble approach

Run	Strategy	Weight
Run 1	Conservative (anchor on priors)	40%
Run 2	Signal-forward (assume effects)	35%
Run 3	Heterogeneity (model variance)	25%

Source: MASTER_SIMULATION_SYSTEM.md (10-phase methodology, 38KB)

Summary: Calibrated predictions deliver 79% error reduction with transparent accuracy by question type

Key metrics

Measure	Performance
Error reduction vs. naive LLM	79%
Mean absolute error (calibrated)	1.9 points
Predictions within 5 points	100%
Validated domains	20+
Human data foundation	5M+ responses
Documented bias patterns	8

Accuracy spectrum

Zone	Error	Use case
High	±2-3 pts	Trust, awareness, demographics → Decisions
Medium	±4-5 pts	Satisfaction, NPS, concern → Direction
Low	±8-15 pts	Intent, price, polarized → Validate

Source: CrowdWave Accuracy Framework, February 2026

Appendix

Appendix A: Demographic calibration multipliers

Segment	Emotional intensity	Digital adoption	Price sensitivity
Women 60+	×1.30	×1.35	×0.85
Women 18-59	×1.10	×1.00	×1.00
Adults 50-69	—	×1.30	—
Adults 70-79	—	×1.40	—
Adults 80+	—	×1.50	—
High-income ($150K+)	—	+0.3	×0.60
Parents (child context)	+0.6	—	×0.80

Source: AARP Tech Trends 2025 (N=3,838); validated calibration studies

Appendix B: Source quality tiers

Tier	Sources	Quality criteria
Tier 1	Federal Reserve, Pew Research, Gallup, AARP	Probability sample, N>1,000, published methodology, peer review
Tier 2	McKinsey, Deloitte, Conference Board, JD Power	Large N, established methodology, industry standard
Tier 3	YouGov, Harris Poll, Morning Consult	Online panels, useful for trends, directional guidance

Minimum sample size requirements

Analysis type	Minimum N
Topline estimates	400
Subgroup analysis	800-1,000
Rare populations	2,500+

Source: AAPOR standards; VALIDATION_METHODOLOGY.md

Appendix C: Documentation inventory

Document	Size	Purpose
MASTER_SIMULATION_SYSTEM.md	38.5KB	Complete 10-phase methodology
CALIBRATION_MEMORY.md	25.2KB	Master calibration reference
CALIBRATION_EXPANSION.md	21.2KB	Extended domain coverage
VALIDATION_METHODOLOGY.md	22.1KB	Accuracy tracking framework
ACCURACY_TESTS.md	19KB	27 documented test cases
BIAS_COUNTERMEASURES.md	13.9KB	8 patterns with corrections
COMPETITIVE_BENCHMARKS.md	11.2KB	Market positioning

Total system documentation: ~150KB

Source: CrowdWave documentation repository

CrowdWave

February 2026