CrowdWave Accuracy Framework

Calibrated AI predictions with documented, predictable accuracy

February 2026

Calibration reduces prediction error by 79%, enabling reliable AI-simulated research

79%	1.9	20+	5M+
Error reduction vs. naive LLM	Mean absolute error (points)	Validated domains	Human survey responses

Core insight: Raw LLMs are 25% less accurate than expert forecasters. Calibration against human survey data closes this gap for established topics.

Source: CrowdWave validation (27 test cases); ForecastBench 2025 (Forecasting Research Institute)

The accuracy problem: Raw LLMs and synthetic surveys fail at rates unacceptable for business decisions

LLM Forecasting Accuracy

(ForecastBench, Oct 2025)

System	Brier Score	Gap
Superforecasters	0.081	—
GPT-4.5	0.101	+25%
GPT-4	0.131	+62%
Median public	0.150+	+85%

Synthetic Survey Correlation

(Dig Insights, 2025)

Task	Correlation
Known events	0.85 ✓
Future events	0.50 ⚠️
New concepts	**0.30** ✗

The paradox

Synthetic data works for what you already know — and fails at what you actually need to predict.

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

Accuracy Spectrum: Three zones determine appropriate use and required validation

HIGH ACCURACY

±2-3 pts

Question types:

Trust scales

Awareness (Y/N)

Party ID

Bipartisan rankings

Action: Use for decisions

MEDIUM ACCURACY

±4-5 pts

Question types:

Satisfaction (1-5)

NPS / Recommend

Concern levels

Tech comfort

Action: Use for direction

LOW ACCURACY

±8-15 pts

Question types:

Purchase intent

Price sensitivity

Polarized politics

Novel behaviors

Action: Validate first

Source: CrowdWave accuracy testing (27 test cases, 6 domains)

High-accuracy zone: Stable attitudes with abundant benchmark data

Why these work

Stable

Low volatility over time

Abundant data

Multiple benchmark sources

Low emotion

Factual, not affective

Training aligned

LLM data matches reality

Validated performance

Question	Calibrated	Actual	Error
Trust in scientists	77%	77%	0
% Independent	44%	45%	1 pt
Smartphone 50+	89%	90%	1 pt
Employee engaged	32%	31%	1 pt

Source: Gallup (N=13,000+); Pew Research (N=5,000+); AARP 2025 (N=3,838)

Medium-accuracy zone: Industry-specific calibration required for NPS and satisfaction

NPS variance by industry (LLMs assume 35-40 for all — actual range is 30-65)

65
Manufacturing

61
Healthcare

55
Retail

46
Fintech

42
Education

30
Software

B2B vs. B2C splits reveal further variance

Industry	B2B	B2C	Gap

Low-accuracy zone: Intent requires conversion factors; polarized topics require segmentation

Intent-to-Action Gap

"Very likely"
30%
"Likely"
15%
"Might consider"
5%

Rule

Multiply top-2-box by ×0.30

Partisan Segmentation Required

Topic	R	D	Gap
Immigration	75%	25%	50
Climate	25%	70%	45
Gun violence	35%	70%	35

Warning

Never predict a single number for polarized topics. The "average" represents no one.

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

Eight documented bias patterns with systematic corrections

Under-prediction biases

Senior tech adoption	**×1.30-1.65**
Status quo preference	**+15-20 pts**
Emotional intensity	**×1.20-1.30**
Cruise/travel satisfaction	**+15 pts**

Over-prediction biases

AI concern (general)	**×0.90**
Intent-to-action	**×0.30-0.55**
Life satisfaction	**-3 to -5 pts**

Structural biases

Partisan averaging	**Segment**
Open-end polish	**20% low-quality**

Validation sources

AARP 2025 (N=3,838)
Pew/YouGov 2025
Gallup 2025 (N=13K+)
Behavioral meta-analysis

Source: Validated calibrations documented in CALIBRATION_MEMORY.md and BIAS_COUNTERMEASURES.md

Validation results: Calibration brings 100% of predictions within 5 points of actual

Before/After Comparison

Metric	Naive	Calibrated
Mean error	9.1 pts	**1.9 pts**
Within 2 pts	7%	**81%**
Within 5 pts	30%	**100%**

79% error reduction
Statistically significant at $p < 0.0001$

Sample Predictions

Prediction	Naive	Cal.	Actual
50+ smartphone	72%	89%	90%
Independent	35%	44%	45%
AI concern	58%	50%	48%
Mfg NPS	40	64	65
Cruise sat.	78%	91%	90%
Engaged	38%	32%	31%

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

Executive audiences require role-specific calibration

C-suite concern multipliers by role

Concern	CEO	CFO	CHRO	CMO	Insight
Cyber	1.30	1.40	**1.60**	0.90	CHROs most concerned
AI	0.90	1.05	**1.40**	1.10	CEOs least concerned
Transformation	1.50	1.15	**1.70**	1.40	CHROs leading change
Uncertainty	1.35	**1.50**	1.50	1.25	CFOs feel it most

Key finding

CHROs are 75% more concerned about AI than CEOs. Generic "executive" predictions miss these role-based variations.

Application

Always segment by role when surveying C-suite. Use multipliers above to adjust baseline predictions.

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

Domain coverage: 20+ validated categories with documented accuracy

Fully Validated (±2-4 pts)	Partial Validation (±4-6 pts)	Gaps / Low Confidence
Trust in institutions	Executive attitudes	Purchase intent
Political identity	Healthcare decisions	Price sensitivity
Technology adoption	Workplace engagement	Novel behaviors
NPS by industry	Financial attitudes	Rural/urban splits
Consumer concerns	Media consumption	Emerging tech
Travel/hospitality		

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

Use case matrix: Match application to accuracy zone

By Confidence Level

HIGH CONFIDENCE

Concept testing
Audience sizing
Trend validation
Priority ranking
Benchmark comparison

MEDIUM CONFIDENCE

Hypothesis generation
Early-stage screening
Directional guidance

VALIDATE FIRST

New product concepts
Pricing research
High-stakes decisions

Not Recommended Without Validation

Purchase conversion

Use A/B testing instead

Polarized topics

Must segment by party

Novel behaviors

No training data

Legal/regulatory

Requires human evidence

Source: CrowdWave accuracy framework; industry best practices

Competitive differentiation: Documented accuracy vs. unvalidated claims

Capability	Raw LLM	Competitors	CrowdWave
Documented accuracy	✗	"95%" ⚠️	27 tests ✓
Human validation	✗	Unclear	5M+ ✓
Bias corrections	✗	None	8 patterns ✓
Domain calibrations	✗	Generic	20+ ✓
Confidence scoring	✗	✗	Per-question ✓
Known limitations	✗	✗	Documented ✓

Competitor claims

"95% accuracy" — testimonials only, no methodology published, no test cases documented

27 test cases, 79% error reduction, full methodology transparency, known limitations documented

Source: Competitive analysis; COMPETITIVE_BENCHMARKS.md

Methodology: 10-phase production workflow with ensemble estimation

1
Config

2
Priors

3
Behavior

4
Survey

5
Ensemble

6
Verify

7
Calibrate

8

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

Performance metrics

79% Error reduction	1.9 MAE (points)
20+ Domains	5M+ Human responses

Accuracy by zone

- **±2-3 pts** — Trust, awareness → Decisions
- **±4-5 pts** — Satisfaction, NPS → Direction
- **±8-15 pts** — Intent, polarized → Validate

Differentiation: Documented accuracy. Known limits. Transparent methodology.

Source: CrowdWave Accuracy Framework, February 2026

Appendix

Appendix A: Demographic calibration multipliers

Segment	Emotional	Digital	Price
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 framework

Tier 1

Fed, Pew, Gallup, AARP

Probability sample
N > 1,000
Published methodology

Tier 2

McKinsey, Deloitte, JD Power

Large N
Established methodology
Industry standard

Tier 3

YouGov, Harris, Morning Consult

Online panels
Useful for trends
Directional guidance

Minimum sample requirements: Topline: 400 | Subgroups: 800-1,000 | Rare: 2,500+

Source: AAPOR standards; VALIDATION_METHODOLOGY.md

CrowdWave

Documented accuracy. Known limits. Transparent methodology.

February 2026