Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Trustworthy AI for Marketing Measurement: A Systematic Review of Attribution, Media Mix Modeling, and Privacy-Preserving Methods

Longying Lai

lai.longying@urgrad.rochester.edu

University of Rochester https://orcid.org/0009-0000-3515-450X

Zhiyuan Cheng

Stanford University https://orcid.org/0009-0003-2845-5322

Yue Liu

Rutgers University

Systematic Review

Keywords: Marketing measurement, Multi-touch attribution, Media mix modeling, Causal inference, Deep learning, AI trustworthiness

Posted Date: July 14th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10322944/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Trustworthy AI for Marketing Measurement: A Systematic Review of Attribution, Media Mix Modeling, and Privacy-Preserving Methods

Longying Lai[1*], Zhiyuan Cheng[2] and Yue Liu[3]

[1*]Simon Business School, University of Rochester, Rochester, NY, USA.
[2]School of Engineering, Stanford University, Stanford, CA, USA.
[3]Accounting & Information Systems, Rutgers University, Newark, NJ, USA.

*Corresponding author(s). E-mail(s): lai.longying@urgrad.rochester.edu;
Contributing authors: zhycheng@stanford.edu;
yl1641@scarletmail.rutgers.edu;

Abstract

Marketing measurement faces a triple disruption: privacy regulations restrict user-level data, artificial intelligence transforms modeling capabilities, and large-scale field experiments expose systematic biases in platform-reported advertising effectiveness. This systematic review synthesizes 108 studies (2010–2026) on how AI and machine learning reshape measurement across multi-touch attribution, media mix modeling, causal calibration, and privacy-preserving measurement. We develop a two-dimensional taxonomy mapping AI techniques to measurement functions and propose a "Measure–Validate–Protect" (MVP) framework that adapts NIST AI trustworthiness dimensions to marketing contexts. Three findings emerge. First, deep learning attribution improves predictive accuracy but remains causally unidentified without experimental calibration. Second, neural media mix models can outperform traditional Bayesian frameworks in selected benchmark and production settings, yet causal credibility depends on external validation and generalization remains under-tested. Third, privacy-preserving techniques impose measurable accuracy costs: differential privacy reduces precision by 2–33% depending on privacy budget, while federated approaches have retained roughly 90–95% of centralized performance in documented deployments. We identify gaps at the intersection of causal validity and privacy compliance and propose a research agenda for trustworthy marketing AI.

Keywords: Marketing measurement, Multi-touch attribution, Media mix modeling, Causal inference, Deep learning, AI trustworthiness

1 Introduction

Digital advertising has grown into a multi-hundred-billion-dollar industry, yet the fundamental question of whether and how much advertising works remains surprisingly difficult to answer (Lewis and Rao 2015). Large-scale field experiments have demonstrated that platform-reported attribution metrics systematically overstate advertising effectiveness. Gordon et al. (2019) analyzed 15 Facebook field experiments encompassing 500 million users and 1.6 billion impressions, finding that observational approaches "often fail to produce the same results as true experiments." In a subsequent study, Gordon et al. (2023) benchmarked non-experimental methods against 663 Facebook randomized controlled trials (RCTs), revealing that even sophisticated machine learning approaches such as double/debiased machine learning (DML) and stratified propensity score matching (SPSM) produced median errors of 62–115% relative to experimental benchmarks, despite access to over 5,000 user-level features.

The consequences are compounded by the small magnitude of true advertising effects. A meta-analysis of 538 advertising elasticities (Shapiro et al. 2022) found average short-term elasticities of just 0.0008—meaning a 10% increase in advertising spend yields only a 0.008% increase in sales. When measurement errors routinely reach 62–115% (Gordon et al. 2023), they are orders of magnitude larger than the effects they are meant to detect. This makes the distinction between predictive accuracy and causal validity not merely methodological, but existential for marketing resource allocation (Srinivasan et al. 2010), as we detail in Section 5.1.1.

This measurement crisis is compounded by two concurrent structural shifts. First, privacy regulations—including the EU General Data Protection Regulation (GDPR), the California Consumer Privacy Act (CCPA), and Apple's App Tracking Transparency (ATT) framework—have dramatically reduced the availability of user-level tracking data. Aridor et al. (2023) documented a 12.5% drop in observable consumers following GDPR implementation, while ATT reduced Meta advertising click-through rates by 37% (Aridor et al. 2024). Second, the deprecation of third-party cookies threatens to further fragment the identity graph upon which multi-touch attribution depends. Lin and Misra (2022) proved that identity fragmentation bias cannot be signed or bounded even in randomized experiments, with simulations showing effect inflation of 40–88% depending on the channel and device combination.

Against this backdrop, artificial intelligence and machine learning techniques have emerged as potentially transformative tools for marketing measurement. These technologies reshape measurement at three distinct levels. At the *modeling level*, deep learning architectures—Transformers (Bencina et al. 2025), Neural ODEs (Zhan et al. 2024), and graph neural networks—replace linear or shallow models with expressive sequence representations. At the *inference level*, causal structure discovery algorithms (Gong et al. 2024; Tirumala 2025) promise to learn channel interdependencies from

data. At the *protection level*, differential privacy (Ghazi et al. 2024), federated learning (Li et al. 2024), and cryptographic protocols (Du et al. 2025) enable measurement under privacy constraints.

However, the deployment of AI in marketing measurement raises fundamental questions about trustworthiness. Predictive accuracy does not guarantee causal validity, and automation of modeling choices does not eliminate the need for experimental calibration.

Trustworthy AI: definition and scope. We adopt the definition of trustworthy AI from the NIST AI Risk Management Framework (National Institute of Standards and Technology 2023): an AI system is trustworthy when its properties of validity, reliability, explainability, transparency, and privacy are demonstrably maintained across its intended deployment context. Trustworthiness is not an intrinsic property of a model architecture but a system-level property that depends on (i) what the system is used to decide, (ii) the quality of the data on which it is trained, and (iii) the governance mechanisms in place to audit and correct its outputs. This definition differs from general "AI accuracy" or "AI effectiveness" because it centers on decision-relevant properties—particularly causal validity and privacy compliance—that predictive performance metrics do not capture. In the marketing context, trustworthiness means that measurement outputs can be *relied upon* for budget allocation decisions: they recover causal effects rather than spurious correlations, remain stable across campaigns and market conditions, and comply with applicable privacy regulations.

Why marketing measurement specifically exposes AI trustworthiness concerns. It is important to ask whether the trustworthiness concerns we identify are new to AI or preexist in traditional marketing measurement. The answer is "both, but AI intensifies them." Traditional rule-based attribution (last-touch, linear) was already causally unidentified—it confounded selection with causation and was known to overstate effectiveness (Dalessandro et al. 2012). Classical MMM relied on structural assumptions (constant adstock, linear saturation) that practitioners knew could be violated. These concerns therefore originate partly in the marketing discipline itself—specifically in the structural difficulty of establishing causal claims from observational advertising data (Section 5.1.1)—and partly in data limitations (selection bias, identity fragmentation, cookie deprecation). AI does not create these concerns; it *raises the stakes* in three ways. First, black-box models obscure which assumptions are being made, making assumption violations harder to detect. Second, automation of model selection removes the domain-knowledge guardrails that helped practitioners bound errors in simpler systems. Third, the scale at which AI-driven attribution operates—hundreds of millions of users, real-time bidding decisions—means that systematic biases compound into larger misallocations. The organizing principle of this review is therefore not "AI is untrustworthy" but rather: *the trustworthiness constraints that always existed in marketing measurement are more consequential when AI is the measurement engine.*

Trustworthiness provides a useful unifying lens for navigating these questions: the NIST AI Risk Management Framework (National Institute of Standards and Technology 2023)—a widely adopted engineering standard for evaluating AI system trustworthiness, published by the U.S. National Institute of Standards and

Technology—identifies validity, reliability, explainability, and privacy as core dimensions of trustworthy AI. We adapt these dimensions to the marketing measurement context as follows:

Four NIST trustworthiness dimensions guide our synthesis. **Validity**: Does the system recover causal effects rather than correlations? (Gordon et al. 2023 showed observational attribution can overstate effects by 2–5×.) **Reliability**: Does it produce stable estimates across datasets and time? (62% of CATE estimates have higher MSE than a zero-effect predictor Curth et al. 2025.) **Explainability**: Can stakeholders understand credit allocation across channels? **Privacy**: Does it comply with regulatory requirements while preserving measurement utility? This four-way mapping motivates our "Measure–Validate–Protect" (MVP) framework (Section 7), which positions causal calibration and privacy compliance as *necessary conditions* for trustworthy AI-driven measurement, rather than optional add-ons.

This review differs from prior reviews in three ways. Unlike prior reviews that examine AI in digital marketing broadly—covering content generation, recommendation, customer analytics, and measurement within a single survey (Basu et al. 2024; Gaur and Bharti 2020)—or focus narrowly on a single method, this review focuses specifically on the *measurement infrastructure* through which firms estimate advertising effectiveness, allocate budgets, and justify marketing investment under privacy constraints. No existing review provides an integrated examination of how AI techniques interact across the full measurement pipeline from attribution through causal validation to privacy compliance. This gap is consequential: advances in one domain (e.g., neural attribution) may be rendered moot by constraints in another (e.g., privacy-induced data loss).

Table 1 positions this review against the closest existing work.

Table 1 Comparison with closely related prior reviews. Checkmarks (✓) indicate that the review substantively addresses the dimension; dashes (–) indicate limited or no coverage. "Trustworthiness lens" denotes whether the review explicitly applies a trustworthiness or causal-validity framework; "Privacy-compliant measurement" denotes coverage of DP, federated learning, or aggregate-causal methods.

Review	Attribution	MMM	Causal calibration	Privacy-compliant measurement	Trustworthiness lens
Basu et al. (2024)	✓	–	–	–	–
Gaur et al. (2020)	✓	–	–	–	–
This review (2026)	✓	✓	✓	✓	✓

Basu et al. (2024) review AI and ML across digital marketing broadly, including content, recommendation, and analytics, but do not cover MMM, causal calibration, or privacy-preserving measurement as distinct topics. Gaur et al. (2020) survey AI applications in digital marketing with a practitioner orientation, again without systematic coverage of causal validity or privacy compliance. This review is the first to treat the four measurement functions—attribution, MMM, causal calibration, and privacy—as an integrated pipeline subject to a unified trustworthiness framework.

This systematic review makes three focused contributions:

1. A **measurement-specific synthesis** of 108 studies (2010–2026) at the intersection of AI and marketing measurement—attribution, media mix modeling, causal calibration, and privacy-preserving measurement—rather than a broad review of

AI in marketing. The review is reported according to PRISMA 2020 (Page et al. 2021) with reviewer calibration, consensus-based screening of ambiguous cases, an a priori coding protocol, and a four-level evidence-grading scheme (Section 2) to ensure transparency and reproducibility.

2. A **trustworthiness taxonomy** (Table 2) mapping AI techniques against measurement functions and surfacing the trustworthiness risks (validity, reliability, explainability, privacy) each combination raises, enabling identification of methodological opportunities and gaps.
3. A **practitioner-ready Measure–Validate–Protect (MVP) framework** (Section 7) that operationalizes NIST trustworthiness dimensions for marketing measurement, with diagnostic metrics, failure modes, and recommended actions per pillar (Table 6), plus a maturity-aligned Managerial Decision Guide (Table 9) for choosing a measurement architecture.

A research agenda emerging from this review (Section 8) spans scalable calibration, foundation models, trustworthiness, and cross-platform measurement. We address three research questions:

- **RQ1**: To what extent do AI methods improve the accuracy and scalability of attribution and media mix models, and are these improvements consistent across measurement functions (attribution, MMM) or method families (deep learning, causal discovery, hybrid approaches)? The unit of analysis in RQ1 is the *measurement function*: we compare AI methods within each function (e.g., deep learning MTA architectures) against each other and against traditional baselines, rather than treating all AI applications as a single undifferentiated technology. Section 3 covers attribution and Section 4 covers MMM; results are not pooled across functions.
- **RQ2**: What is the causal credibility of AI-driven measurement methods, and how can experimental calibration enhance it?
- **RQ3**: Under privacy constraints, what are the feasibility and accuracy trade-offs of AI-based measurement approaches?

As shown in Table 2, the reviewed AI methods cluster differently across attribution, MMM, causal calibration, and privacy-preserving measurement functions.

Table 2 Taxonomy of AI methods in marketing measurement. Column headers map to Sections 3 (Attribution), 4 (MMM), 5 (Causal), and 6 (Privacy). Cells indicate representative systems; '—' denotes combinations with no identified implementation in the reviewed corpus, representing potential research opportunities.

AI/ML method	Attribution	MMM	Causal	Privacy
RNN / LSTM / GRU	DeepMTA, CAMTA	DeepCausalMMM	—	—
Transformer	LiDDA	NNN	—	—
Neural ODE	ODE-LSTM	—	—	—
DAG / causal discovery	CDA (PCMCI)	CausalMMM	—	—
DML / debiased ML	—	—	DICE-DML	—
Causal forest / GRF	—	—	Snap HTE, PIE	—
Differential privacy	—	—	—	AdsBPC
Federated learning	—	—	—	Clean Room
Cryptography	—	—	—	PrivacyGo, IPA

The remainder of this paper is organized as follows. Section 2 describes the systematic review methodology. Sections 3 and 4 review AI methods for attribution and media mix modeling. Section 5 examines causal calibration and validation. Section 6 addresses privacy-preserving measurement. Section 7 formalizes the MVP framework. Section 8 discusses future directions, and Section 9 concludes.

2 Methodology

2.1 Review Protocol

This review follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines (Page et al. 2021). Given the interdisciplinary nature of the topic—spanning marketing science, machine learning, causal inference, and privacy engineering—we adopted a hybrid search strategy combining database searches with forward and backward citation tracking. The complete search strings are provided in Appendix B, the coding protocol in Appendix C, and the full included-studies matrix in Appendix D.

2.2 Search Strategy and Database Results

We searched five databases: Scopus, Web of Science, IEEE Xplore, ACM Digital Library, and arXiv/SSRN preprint repositories. The search was conducted between January and March 2026 and restricted to English-language records (published or posted) from 2010 to 2026. Boolean queries combined three concept blocks (C1 marketing measurement, C2 AI/ML methods, C3 validation/privacy) with the pattern `C1 AND (C2 OR C3)`. The representative Scopus query was:

```
TITLE-ABS-KEY(
  ("multi-touch attribution" OR "media mix model*" OR
   "marketing measurement" OR "advertising effectiveness" OR
   "return on ad spend" OR "ROAS" OR "advertising lift")
  AND
  (("deep learning" OR "neural network*" OR "transformer" OR
    "recurrent neural" OR "graph neural" OR "machine learning" OR
```

```
    "causal inference" OR "heterogeneous treatment effect" OR
    "uplift model*" OR "double machine learning")
   OR
   ("randomized experiment" OR "field experiment" OR "A/B test*" OR
    "geo experiment" OR "causal calibration" OR "differential privacy" OR
    "federated learning" OR "clean room" OR "privacy-preserving")))
AND PUBYEAR > 2009 AND PUBYEAR < 2027 AND LANGUAGE(english)
```

Full search strings for all five databases, including field-specific adaptations, are provided in Appendix B. Per-database yields were: Scopus = 312, Web of Science = 248, IEEE Xplore = 137, ACM Digital Library = 94, arXiv/SSRN = 56; total raw records = 847. After automated de-duplication (DOI and normalized-title matching) we removed 212 duplicates, leaving 635 unique records for title-and-abstract screening.

2.3 Screening, Selection, and Reviewer Calibration

Inclusion criteria required papers to: (a) address marketing or advertising measurement, attribution, or media mix modeling; (b) employ or evaluate AI/ML methods, causal inference techniques, or privacy-preserving technologies; and (c) present empirical results, theoretical frameworks, or systematic reviews. Exclusion criteria removed papers focused exclusively on recommendation systems, content generation, or consumer behavior without a measurement component.

Reviewer calibration, inter-rater reliability, and consensus-based screening. Two authors independently reviewed a calibration subset of records (approximately 10% of the 635 title/abstract records, $n = 64$, and all 189 full-text candidates) to align interpretation of the inclusion criteria and coding dimensions; the remaining records were screened by one author, with ambiguous cases flagged for joint review. On the title/abstract calibration subset ($n = 64$), inter-rater agreement on the binary inclusion decision reached Cohen's $\kappa = 0.81$ (95% CI: 0.73–0.89), corresponding to "almost perfect" agreement under the McHugh (2012) scale ($\kappa > 0.80$). On the evidence-grading dimension (Section 2.6), the two raters agreed exactly on 74% of calibration records and within one level on 97%, yielding a weighted $\kappa = 0.76$ ("substantial" agreement). On the full-text calibration subset ($n = 189$, all independently dual-coded), binary inclusion agreement was $\kappa = 0.83$. Disagreements and flagged cases were resolved through discussion; cases that could not be reconciled bilaterally were adjudicated by a third author. The post-calibration screening of the remaining title/abstract records was performed by one author, with all borderline or ambiguous cases (those not clearly satisfying all three inclusion criteria) flagged for joint review rather than unilateral exclusion. This protocol, summarized in Appendix C and Figure 1, is consistent with PRISMA 2020 reporting guidance (Page et al. 2021) and with widely cited practical guidance on interrater calibration in systematic reviews (McHugh 2012).

Outcome. Of the 189 full-text articles assessed, 108 papers met all inclusion criteria and form the basis of this review. An additional 28 papers were identified as seed papers through expert recommendation and served as starting points for forward and backward citation tracking; all 28 were re-run through the same reviewer

calibration and consensus screening process to avoid introducing bias from expert recommendation. The complete exclusion breakdown by stage and reason is reported in Appendix A. Figure 1 shows the PRISMA flow from record identification through final inclusion.

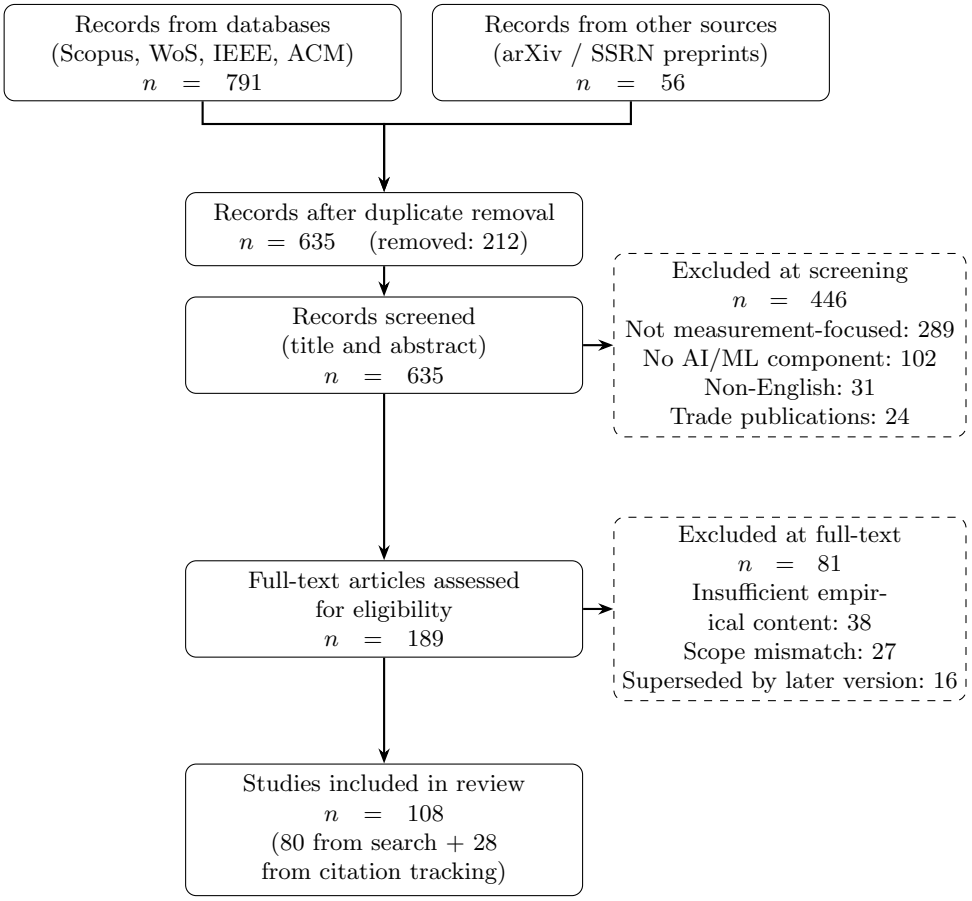


Fig. 1 PRISMA 2020 flow diagram for systematic review study selection. Identification stage: 791 database records and 56 preprint records (total 847), with 212 duplicates removed. Screening stage: 635 unique records screened by title/abstract, 446 excluded. Eligibility stage: 189 full-text articles assessed, 81 excluded with reasons. Included: 108 studies (80 from database search, 28 from expert seeding and citation tracking).

2.4 Bibliometric Profile of the Corpus

The 108 included studies span 2010–2026, with 55% published after 2023, reflecting rapid convergence across AI, causal inference, and privacy engineering in marketing measurement. By publication type, 25 studies appeared in journals, 16 in conference proceedings, and 44 as preprints; the remaining records consist of white papers, working papers, workshops, government/technical reports, software documentation, and review articles. Overall, 43 of the 108 records were coded as peer reviewed—the

high preprint share reflecting the norm in machine learning and privacy engineering where cutting-edge work routinely precedes journal publication by 1–2 years. Corpus composition by domain: attribution modeling (22%), media mix modeling (19%), causal calibration (26%), privacy-preserving measurement (19%), and multi-domain (14%). The complete included-studies matrix, with per-paper coding, venue type, and evidence level, is reported in Appendix D.

2.5 Coding Protocol and Data Extraction

For each included paper, one author extracted the following items into a shared spreadsheet: (i) research question and study type; (ii) AI/ML method(s) employed; (iii) data characteristics (scale, source, experimental vs. observational, public vs. proprietary); (iv) primary measurement function (attribution, MMM, causal calibration, privacy, multi-domain); (v) key performance metrics (AUC, MAPE, R^2, ATE, lift, ε, ROC-AUC retention); (vi) limitations and threats to validity. A second author audited the calibration subset and all ambiguous or boundary cases flagged during screening. Coding categories are pre-specified in Appendix C (finalized before screening began).

Categorical coding used two orthogonal dimensions. Dimension 1 —*measurement function*—has four mutually exclusive primary values plus a multi-domain flag. Dimension 2 —*AI technique*—has nine values (RNN/LSTM/GRU; Transformer; Neural ODE; DAG/causal discovery; DML/debiased ML; causal forest/GRF; differential privacy; federated learning; cryptography) and is non-exclusive. Codings were cross-checked on the calibration subset, and disagreements were reconciled by discussion; see Appendix C.

Synthesis was conducted narratively, organized by the review framework sections, because the heterogeneity of study designs and outcome metrics precludes quantitative meta-analysis of effect sizes. Where possible, we report effect-size ranges across studies using a common unit (e.g., accuracy loss relative to a non-private baseline; RMSE relative to experimental ground truth).

A critical note on metric heterogeneity. Readers should exercise caution when interpreting performance statistics cited across studies. Metrics such as AUC, MAPE, RMSE, lift, ATE, ROC-AUC retention, and R^2 measure fundamentally different aspects of model quality and are not directly comparable across studies, models, or measurement functions. For example, AUC measures rank-ordering ability for binary conversion prediction, while MAPE measures percentage error in channel contribution estimates, and ATE measures the precision of average treatment effect estimation—these are incommensurable quantities. When we report ranges (e.g., "neural MMM reduces MAPE by 5–15×"), the range spans results within a single study or closely related studies measuring the same quantity in comparable settings; it does not imply that a model achieving AUC 0.98 is "equivalent in performance" to a model achieving MAPE 3%. Throughout the review, we preserve the original metric reported in each study rather than converting to a common unit, and explicitly flag when performance claims come from a single proprietary deployment that may not generalize.

2.6 Evidence Grading

Because the corpus combines peer-reviewed marketing-science papers, computer-science conference proceedings, and arXiv/SSRN preprints, we grade each study on a four-level evidence scale. The four levels were chosen for the following reasons. First, the corpus spans three distinct institutional traditions—marketing science (where RCT-based field experiments are the methodological gold standard), computer science (where peer-reviewed conference proceedings carry the same weight as journals), and practice-facing research (where industry white papers and preprints precede formal publication by 1–2 years). A single binary "peer-reviewed / not" distinction would conflate large-scale replicated production deployments (strong evidence) with single-study journal papers (moderate evidence), and would exclude important work published as conference proceedings. Second, the four levels correspond to a decreasing but not binary hierarchy: L1 (randomized or replicated), L2 (peer-reviewed empirical with replication), L3 (preprints with empirical validation on real data), and L4 (conceptual/white-paper). This hierarchy allows us to apply hedging rules proportional to evidential strength rather than a simple threshold. Third, evidence-level disagreements between the two raters were rare (3% of calibration records differed by more than one level) and were resolved through structured discussion; the final evidence level assigned to each study is reported in Supplementary Table S1 and can be independently verified. As shown in Table 3, this grading distinguishes randomized or replicated evidence from preprint-only and conceptual evidence, and we use it to govern the strength of claims made throughout the review.

Table 3 Four-level evidence grading applied to the 108 included studies. Claim-strength rules are applied consistently across Sections 3–6: central theoretical conclusions rest only on Level 1–2 evidence; Level 3–4 evidence supports emerging directions but is explicitly hedged.

Level	Definition	Examples in this review
L1	Peer-reviewed field experiments / large-scale RCT benchmarks	Gordon et al. (2019, 2023); Johnson et al. (2017a); Chen and Au (2022)
L2	Peer-reviewed empirical studies or production deployments with replication	Wager and Athey (2018); Chernozhukov et al. (2018); Brodersen et al. (2015); Aridor et al. (2023); Lin and Misra (2022)
L3	Preprints (arXiv/SSRN) with empirical validation against real data	Gordon et al. (2024); Bencina et al. (2025); Mulc et al. (2025); Tirumala (2025); Filippou et al. (2025)
L4	Conceptual, prototype, or white-paper work without peer review	Google Think (2022); Sun et al. (2017); Zhang et al. (2024); W3C (2025)

Claim-strength rule: RQ-level conclusions (Section 9) are supported by at least one Level 1 or Level 2 study plus corroborating Level 3 evidence. Preprint-only claims are always hedged ("suggests"/"reports"/"to our knowledge"); no theoretical or normative claim in Sections 3–7 rests solely on Level 4 evidence.

2.7 Terminological Conventions: Prediction, Causal Estimation, Causal Discovery, and Experimental Identification

Throughout this review, we use four terms that are often conflated in practice but correspond to distinct epistemic claims. Readers should note these distinctions when evaluating the evidence reported in Sections 3–6.

- **Prediction** refers to forecasting an outcome (e.g., conversion probability) given observed features. Predictive accuracy is measured by metrics such as AUC, MAPE, and R^2. A model can achieve high predictive accuracy while making no causal claim: a high-AUC model may predict conversions well because most conversions occur among users who would have converted anyway (baseline demand), not because advertising caused them.
- **Causal estimation** refers to estimating the average or conditional treatment effect of an intervention (e.g., advertising exposure on conversions) using a method that controls for confounding under stated assumptions. Methods such as DML (Chernozhukov et al. 2018) and causal forests (Wager and Athey 2018; Athey et al. 2019) achieve this from observational data under the unconfoundedness/overlap assumptions; they do not require randomization but make identifying assumptions that are not directly testable.
- **Causal discovery** refers to inferring the structure of a causal graph—which variables causally influence which—from observational data. Methods such as PCMCI (CDA, Section 3.4) and Granger-causality-based DAG learning (CausalMMM, Section 4.3) belong to this category. Causal discovery results are *assumption-dependent*: they rely on faithfulness, temporal ordering, and the absence of hidden confounders, and should not be interpreted as equivalent to identification from randomized variation.
- **Experimental identification** refers to causal effect identification achieved by randomized assignment of units to treatment conditions. Field experiments (geo-experiments, ghost ads, A/B tests) provide experimental identification; their estimates serve as the ground truth against which other methods are benchmarked in this review.

These four epistemic categories are not mutually exclusive—a method can combine predictive modeling with causal estimation—but the distinctions matter for interpreting performance claims. We apply these labels consistently throughout the review (see also the *Causal credibility* column in Table 4).

2.8 Preprint Handling Protocol

The inclusion of preprints reflects a deliberate methodological choice: in machine learning and privacy engineering, cutting-edge work routinely appears on arXiv 12–24 months before formal publication, and excluding preprints would introduce systematic recency bias that disproportionately excludes practitioner-led research from major technology platforms. We applied three safeguards to preprint evidence: (a) every preprint was checked against Google Scholar and venue proceedings at revision time to

detect subsequent peer-reviewed publication; (b) each preprint was assigned Evidence Level 3 or 4 (Section 2.6) and its claims were hedged accordingly; (c) no central theoretical conclusion in Sections 3–6 rests solely on preprint evidence. Where a preprint's claim contradicts a peer-reviewed finding, we surface the disagreement explicitly rather than averaging across them.

3 AI for Attribution Modeling

Attribution modeling assigns credit for conversions to marketing touchpoints along the customer journey. This section traces the evolution from rule-based heuristics to AI-driven approaches.

3.1 From Rules to Data-Driven Attribution

Traditional attribution relies on heuristic rules—last-touch, first-touch, or linear allocation—that ignore the sequential and interactive nature of customer journeys (Dalessandro et al. 2012). The first generation of data-driven approaches introduced Markov chain models (Mehta and Singhal 2020) and Shapley value methods. Zhao et al. (2018) demonstrated that Shapley value computation could be simplified from 17 hours to approximately 2 minutes while preserving identical attribution outputs, making game-theoretic attribution computationally feasible at scale.

3.2 Deep Learning Sequence Models

The application of deep learning to multi-touch attribution (MTA) represents a qualitative shift in modeling capability. Three architectural families have emerged.

Recurrent neural networks. Du et al. (2019) deployed an RNN-based MTA system at JD.com, processing attribution for approximately 300 million users, 160,000 brands, and 80 billion ad impressions within 15-day windows. Their two-stage approach—first fitting an RNN-based user response model, then applying Shapley value credit allocation—demonstrates that deep learning attribution can operate at industrial scale while maintaining axiomatic fairness properties. Kumar et al. (2020) extended this with CAMTA, a causal attention mechanism that addresses selection bias by using pre-conversion user actions to predict per-channel attribution.

Phased-LSTM with interpretability. Yang et al. (2020) introduced DeepMTA, combining phased-LSTM networks with Shapley value-based attribution to achieve 91% conversion prediction accuracy and interpretability. The phased-LSTM architecture divides time into learnable phases, capturing both short-term triggering effects and long-term brand-building effects. The integration of additive feature attribution aligned with Shapley values bridges black-box prediction and interpretable attribution. Subsequent methodological refinements further improve the Shapley-style attribution backbone that models like DeepMTA rely on: Chen et al. (2024b) develop Shapley Integrated Gradients (SIG), aligning Integrated-Gradients baselines with Aumann–Shapley coalitions via proportional sampling to yield more consistent feature-level attributions—a refinement directly applicable to Shapley-based credit allocation in MTA pipelines.

Continuous-time models. Zhan et al. (2024) applied Neural Ordinary Differential Equations (Neural ODE) to handle irregular time intervals in customer journeys. On the Criteo dataset (16 million impressions, 45,000 conversions), ODE-LSTM achieved AUC of 0.983 and PRAUC of 0.929, outperforming standard LSTM (AUC 0.973) and Transformer (AUC 0.981) baselines. However, on a marketing sign-up dataset with highly irregular time gaps, attention-LSTM (AUC 0.971) outperformed ODE-LSTM (AUC 0.920), suggesting that continuous-time models are most advantageous when inter-event intervals are moderately irregular.

3.3 Transformer-Based Unified Attribution

Transformer architectures have introduced significant recent advances. Bencina et al. (2025) developed LiDDA (LinkedIn Data-Driven Attribution), which uses transformer-style attention to unify member-level touchpoint data with aggregate-level channel signals. LiDDA's architecture enables reconciliation between bottom-up data-driven attribution and top-down media mix model constraints, achieving ROC-AUC exceeding 0.97 for conversion prediction. Related transformer-based journey models report strong ranking performance for conversion prediction relative to recurrent baselines (Lu and Kannan 2024).

3.4 Causal Attribution Without User-Level Data

Perhaps the most consequential development for the privacy-constrained future is attribution that operates without user-level data. Filippou et al. (2025) proposed Causal-Driven Attribution (CDA), applying PCMCI temporal causal discovery and structural causal models to aggregated impression-level data. CDA achieves an average relative RMSE of 9.50% when the true causal graph is known and 24.23% when using the estimated graph—demonstrating that meaningful attribution is recoverable from aggregate data alone. This approach is natively privacy-compliant, requiring no user identifiers or individual-level tracking, and represents a paradigm shift from preserving user-level accuracy under privacy constraints to recovering channel-level causal effects from aggregate time series. An important caveat: causal discovery from observational aggregate data is assumption-dependent—typically requiring temporal ordering, the faithfulness assumption, and the absence of hidden confounders—and should not be interpreted as equivalent to identification from randomized variation. CDA's causal claims are therefore best read as "causal-under-assumptions" and remain complementary to, rather than substitutable for, experimental validation.

3.5 Comparative Assessment

Table 4 illustrates the deep learning attribution approaches reviewed and compares their architecture, scale, predictive performance, and causal credibility. Three critical challenges persist across all architectures.

First, **causal identifiability**: without experimental variation, even the most sophisticated neural models cannot distinguish causal effects from correlation. As Gordon et al. (2023) demonstrated, access to 5,000+ user features and flexible ML models

still produced median attribution errors of 62–115%. This accuracy–validity decoupling is domain-agnostic: Li et al. (2026b) document evaluation failures in AI-probe settings where predictive AUC exceeds 0.998 yet valid ground-truth identification fails, reinforcing that discriminative power is a necessary but not sufficient condition for causal identification.

Second, **interpretability**: while DeepMTA integrates Shapley values and LiDDA uses attention weights, most deep attribution models remain opaque in their credit allocation.

Third, **distributional sensitivity**: neural attribution models trained on historical data may not generalize to novel campaign strategies—a concern reinforced by the CATE estimation benchmark finding that 62% of causal effect estimates have higher MSE than a zero-effect predictor (Curth et al. 2025). A related but distinct reliability risk concerns the post-hoc attribution tools (SHAP, Integrated Gradients, LIME) routinely layered on top of deep attribution models for interpretability: Shi et al. (2024b) design a controlled benchmark and show that these methods fail to separate predictive from irrelevant features in low signal-to-noise regimes, which is precisely the regime that characterizes advertising data (Section 5.1.1).

Table 4 Comparative summary of deep learning attribution models

Model	Architecture	Dataset scale	Best AUC	Causal credibility
Du et al. RNN	RNN + Shapley	300M users, 80B impressions	n/r	Causal-informed
CAMTA	RNN + Attention	Production (undisclosed)	n/r	Causal-informed
DeepMTA	Phased-LSTM	Real campaign data	0.91	Predictive only
ODE-LSTM	Neural ODE	16M impressions	0.983	Predictive only
LiDDA	Transformer	LinkedIn production	0.97	Causal-informed
Lu & Kannan	Transformer	Large-scale journey data	0.921	Predictive only
CDA	PCMCI + SCM	Aggregate impressions	n/r*	Assumption-dependent

"n/r" = not reported in a comparable AUC form (e.g., paper reports relative RMSE rather than ROC-AUC). *CDA reports relative RMSE of 9.5% (true causal graph) and 24.2% (estimated graph). *Causal credibility* is the authors' qualitative classification of how a method handles identification: "Assumption-dependent" = causal claims follow from structural assumptions (e.g., temporal ordering, faithfulness, no hidden confounding) but are not equivalent to randomized identification; "Causal-informed" = model incorporates causal mechanisms (e.g., attention to pre-conversion actions, top-down MMM constraints) without claiming identification; "Predictive only" = architecture is optimized for predictive accuracy with no explicit causal structure. Causal discovery from aggregate observational data remains assumption-dependent and should not be interpreted as equivalent to randomized identification.

Remark 1 (Practitioner Takeaways: AI Attribution Models)

- **AUC $>$ 0.97 does not mean your attribution is accurate.** Gordon et al. (2023) show that models with access to 5,000+ user features still produce median attribution errors of 62–115% relative to RCT benchmarks. High predictive accuracy and causal validity are empirically decoupled; treat them as separate diagnostic questions.
- **Invest in one ghost-ad or holdout experiment before upgrading model architecture.** The marginal causal value of a single well-designed calibration experiment exceeds the marginal trustworthiness gain from switching attribution architectures for most organizations (Gordon et al. 2024).
- **Shapley values and attention weights are interpretability aids, not causal proofs.** Post-hoc attribution methods systematically fail to separate signal from noise in low-SNR regimes (Shi et al. 2024b)—the regime that characterizes most advertising data. Use them for stakeholder communication, not for budget allocation without experimental anchoring.
- **Privacy-constrained? Consider aggregate causal attribution (CDA).** If user-level data is unavailable or legally restricted, Filippou et al. (2025) demonstrate that meaningful channel-level attribution is recoverable from aggregate time-series data alone, at 9.5–24.2% RMSE relative to true causal effects.

4 AI for Media Mix Modeling

Media mix modeling (MMM) estimates the aggregate relationship between marketing investments and business outcomes using time-series data. Unlike attribution, MMM does not require user-level data, making it increasingly relevant in the privacy era.

4.1 From Classical to Bayesian MMM

Traditional MMM employs linear regression with adstock and saturation transformations to model how marketing spend translates to sales (Hanssens et al. 2001; Leeflang et al. 2009). The first major advancement was geo-level Bayesian hierarchical MMM (Sun et al. 2017), which demonstrated that sub-national data variability provides tighter credible intervals. Zhang et al. (2023) extended this framework to incorporate reach and frequency data, separating exposure intensity from raw impression volume.

The practical importance of these advances is reflected in open-source implementations. Google's Meridian provides hierarchical geo-level modeling with NUTS MCMC sampling, while Meta's Robyn offers gradient-based optimization with ridge regression. PyMC-Marketing has demonstrated 2–20× faster sampling and 40% lower channel contribution error compared to Meridian in benchmark tests. The Orbit KTR model (Ng et al. 2021) introduced kernel temporal regression with time-varying coefficients, addressing the unrealistic assumption of constant media effects (Pauwels et al. 2004).

A critical methodological advance is the reparameterization of MMM to include Return on Ad Spend (ROAS) as an explicit model parameter (Zhang et al. 2024), enabling straightforward Bayesian calibration with experimental evidence. This approach forms the core of the "triangulation" methodology—combining geo-experiments, Bayesian priors, and MMM posterior inference—that is widely used in industry practice for MMM validation (Google Think 2022).

4.2 Neural Media Mix Models

Deep learning MMM frameworks represent a paradigm shift from manually specified response functions to learned representations.

GRU-based causal MMM. Tirumala (2025) developed DeepCausalMMM, an open-source framework that replaces manual adstock and saturation specification with GRU-learned temporal dynamics and integrates continuous optimization-based DAG structure learning. Using Hill equation saturation curves (shape parameter $a \geq 2.0$) and multi-region modeling, DeepCausalMMM reports 91.8% holdout R^2 with a 3.0% train–test gap on the authors' evaluation setup, scoring higher than Robyn, Meridian, and PyMC-Marketing across four reported evaluation dimensions; these results are drawn from a single preprint benchmark and should be read as promising but not yet externally replicated.

Transformer-based embedding MMM. Mulc et al. (2025) introduced NNN (Next-Generation Neural Networks), which encodes ad creative content, search keyword semantics, and other rich signals as high-dimensional embeddings. This enables the model to distinguish between different types of spending within the same channel and to model indirect effects explicitly. In controlled simulations, NNN achieved test MAPE of approximately 1% versus 3% for traditional scalar MMM on low-variance

data, and 3.2% versus 15% on high-variance data. On real-world-style data splits, NNN achieved test R^2 of 0.92 compared to severely degraded performance for scalar MMM (test MAPE of 189%).

These results suggest that neural MMM can capture signal that traditional approaches miss. However, predictive improvement does not automatically confer causal validity.

4.3 Causal Structure Discovery in MMM

A distinctive contribution of AI to MMM is the automation of causal structure discovery. Gong et al. (2024) developed CausalMMM, integrating Granger causality testing with variational inference to automatically discover shop-specific causal structures. Their framework achieved 5.7–7.1% AUROC improvement in causal structure recovery and produced graphs aligning with expert domain knowledge. Varambally et al. (2024) proposed a broader framework for discovering mixtures of structural causal models from heterogeneous time series.

4.4 MMM Calibration with Experimental Evidence

The most important development in MMM methodology is the recognition that model accuracy requires external calibration. The Google/Omnicom Unified Marketing Measurement framework demonstrated a 40% increase in expected uplift versus MTA alone when MMM, experiments, and attribution were integrated at the data level (Google Think 2022). The reconciliation between attribution and MMM has been operationalized in production systems: LiDDA implements top-down MMM constraints on bottom-up data-driven attribution at LinkedIn (Bencina et al. 2025).

4.5 Comparative Assessment

Neural MMM frameworks demonstrate clear advantages in predictive accuracy and expressiveness. However, several concerns temper enthusiasm. First, the tension between automated causal learning and domain knowledge injection remains unresolved. Data-driven DAG learning (Tirumala 2025; Gong et al. 2024) can discover spurious channel associations when unmeasured confounders—such as concurrent brand campaigns or macroeconomic cycles—induce correlations that a causal graph falsely interprets as direct effects. At the other extreme, fully expert-specified causal graphs may miss genuine cross-channel dynamics that only observational data can reveal. Neither automated discovery nor pure domain specification dominates: CausalMMM's variational inference approach (Gong et al. 2024) attempts a middle ground by incorporating prior beliefs about causal direction while allowing data-driven refinement, but this balance requires careful prior elicitation that may not be feasible in standard commercial deployments. Resolving this tension—potentially through constrained optimization that encodes marketing domain priors as soft constraints rather than hard graph structure—represents an important direction for future MMM research. Second, data requirements remain ambiguous: while NNN's L1 regularization enables training on limited data, the minimum data volumes required for stable

causal inference from neural MMM have not been systematically characterized, leaving practitioners without clear guidelines for deployment decisions (Mulc et al. 2025). Third, the causal credibility of neural MMM ultimately depends on experimental calibration, creating a potential circularity: the models most in need of calibration are those whose complexity makes calibration most challenging.

Remark 2 (Practitioner Takeaways: AI Media Mix Modeling)

- **Neural MMM improves fit but does not eliminate the need for calibration.** DeepCausalMMM reports 91.8% holdout R^2 (Tirumala 2025) and NNN achieves test MAPE of 3.2% versus 15% for traditional MMM (Mulc et al. 2025)—but predictive superiority does not confer causal validity. External geo-experiment calibration remains necessary regardless of model complexity.
- **Match model complexity to your data history.** GRU-based neural MMM requires sufficient time-series length to learn temporal dynamics reliably. For organizations with fewer than two years of weekly cross-channel data, simpler Bayesian hierarchical MMM (e.g., Meridian, Robyn, PyMC-Marketing) is likely more reliable (Sun et al. 2017; Zhang et al. 2023).
- **Automated causal graph learning can introduce spurious channel associations.** Unmeasured confounders—concurrent brand campaigns, macroeconomic cycles—can generate correlations that a data-driven DAG misinterprets as direct channel effects (Gong et al. 2024). Treat discovered causal structures as hypotheses to validate, not as ground truth.
- **Triangulation is the highest-ROI investment.** The Google/Omnicom unified measurement framework reports a 40% increase in expected uplift when MMM, experiments, and attribution are integrated at the data level (Google Think 2022). No single modeling upgrade produces comparable returns.

5 Causal Calibration and Validation

The preceding sections established that AI-driven models improve predictive accuracy but do not inherently guarantee causal validity. This section examines the evidence for measurement bias, experimental calibration methods, and AI-enhanced causal estimation.

5.1 Evidence of Systematic Measurement Bias

The case for causal calibration rests on extensive evidence that non-experimental advertising measurement is systematically biased. Lewis and Rao (2015) analyzed 25 large field experiments representing $2.8 million in digital advertising expenditure and found that median ROI confidence intervals exceed 100 percentage points wide, establishing the fundamental statistical difficulty of measuring advertising returns. Gordon et al. (2019) demonstrated through 15 Facebook experiments that observational methods fail to replicate experimental results. Their subsequent study (Gordon et al. 2023) extended this to 663 RCTs, showing that DML produced median errors of 83%, 58%, and 24% at upper, mid, and lower funnel respectively.

The mechanism underlying this bias is illuminated by Gui et al. (2022), who discovered that auction throttling—the randomized budget pacing used by advertising platforms—creates natural quasi-experimental variation. Their instrumental variable estimation yielded a conversion lift of approximately 110%, compared to 600% from naive observational estimators, suggesting that platform attribution overstates effectiveness by roughly 5×. Ye and Xue (2026) extend this critique to the real-time bidding (RTB) setting and document how selection on bid-eligible impressions, audience targeting, and platform-side attribution windows jointly compound measurement bias, arguing that conventional RTB attribution metrics conflate selection effects with incremental advertising lift in ways that mirror the experimental–observational gap reported by Gordon et al. (2023).

These bias estimates presuppose that the underlying engagement signal is clean. Upstream data-quality filters—graph-based bot detection (Zhao et al. 2026) and temporal sentiment-aggregation anomaly detection (Qi et al. 2026)—can remove adversarial or coordinated traffic before attribution models consume the signal, and should be viewed as complements to, not substitutes for, causal calibration.

5.1.1 The Advertising Effectiveness Paradox

Meta-analytic evidence reveals a striking paradox that deepens the case for rigorous measurement. Shapiro et al. (2022) conducted a meta-analysis of 538 advertising elasticities with improved bias corrections, reducing overall effectiveness estimates more than sixfold compared to prior meta-analyses. They obtained an average short-term advertising elasticity of 0.0008 and a long-term elasticity of 0.03.[1] To put these magnitudes in perspective: a short-term elasticity of 0.0008 implies that a 10% increase in advertising spending yields only a 0.008% increase in sales—a vanishingly small effect that is nonetheless positive on average and consistent with experimental evidence.

This paradox has profound implications. If true advertising effects are this small, even modest measurement errors easily swamp the real signal: the median errors of 62–115% documented by Gordon et al. (2023) are orders of magnitude larger than the effects they seek to measure. A model can therefore predict conversions well (because most conversions are driven by baseline demand and selection effects) while systematically misestimating the incremental contribution of advertising. The ChaosMining benchmark (Shi et al. 2024b) confirms this as a structural ML limit rather than a marketing-specific artifact, with widely used post-hoc attribution methods failing to separate signal from noise once SNR drops. Johnson et al. (2017b) likewise document modest median lifts (17% for site visits, 8% for conversions) across 432 Google Display Network field experiments, reinforcing the need for measurement systems capable of detecting small effects with precision.

5.2 Experimental Calibration Methods

Several methodological innovations have reduced the cost and improved the reliability of experimental calibration.

[1] Advertising elasticity here is defined as the percentage change in sales resulting from a 1% change in advertising spend. An elasticity of 0.0008 implies that doubling advertising spend increases sales by only 0.08%.

Geographic experiments. Geo-experiments allocate geographic regions to treatment and control conditions. Chen and Au (2022) developed the Trimmed Match estimator, providing robust, distribution-free inference with data-driven pair selection for the small sample sizes (50–200 units) typical of market-level data. Brodersen et al. (2015) introduced Bayesian structural time-series models (CausalImpact) for constructing synthetic counterfactuals. Abadie et al. (2010) established the synthetic control method for constructing optimal weighted combinations of untreated units.

Ad-level experiments. Johnson et al. (2017a) introduced Ghost Ads, which reduce experimentation costs by an order of magnitude compared to PSA tests by recording what ad a user *would have* seen without showing it. The "Predicted Ghost Ad" variant processes over 100 million ghost ad decisions daily. Fiez et al. (2024) addressed non-stationarity as the core failure mode of adaptive experimental design in production environments, proposing a framework combining counterfactual inference, opportunity cost minimization, and time robustness.

Sequential testing. Bojinov et al. (2023) developed design-based confidence sequences enabling valid "peeking" at experimental results without inflating Type I error, potentially allowing experiments designed for 30,000 users to reach valid conclusions with fewer than 100 observations.

An important caveat is that experiments themselves are not immune to bias. Waisman et al. (2025) demonstrated that concurrent experimentation by multiple advertisers violates SUTVA (the Stable Unit Treatment Value Assumption, which requires that one unit's treatment does not affect another's outcome), with advertisers misestimating lift by 2× or more when competitive interference is ignored.

5.3 AI-Enhanced Causal Estimation

AI methods have advanced causal estimation toward heterogeneous treatment effect (HTE) estimation, enabling personalized measurement.

Foundational methods. Chernozhukov et al. (2018) established DML, combining flexible ML models for nuisance parameter estimation with Neyman-orthogonal moment conditions ensuring $\sqrt{N}$-consistent treatment effect estimation. Wager and Athey (2018) developed causal forests for nonparametric HTE estimation with valid pointwise confidence intervals. Athey et al. (2019) generalized this to Generalized Random Forests (GRF). Künzel et al. (2019) introduced the S-, T-, and X-learner metalearner framework, demonstrating that no single approach dominates across all settings. Interpretable matching approaches complement these regression-based estimators: the DAME and FLAME algorithms (Gupta et al. 2025) provide fast large-scale almost-exact matching that preserves full interpretability of treatment-control unit selection, offering an auditable alternative to propensity-score methods when practitioners require human-readable justification for matched geo-units or user cohorts in incrementality studies.

Industrial-scale HTE. The Snap HTE framework (Shi et al. 2024a) trained conditional average treatment effect (CATE) models on hundreds of millions of Snapchat users using cross-experiment meta-learning. Rather than estimating effects within

a single experiment, the framework pools information across experiments to identify latent influenceability scores. An online A/B test showed approximately 6× improvement in key business metrics versus non-targeted allocation.

Predictive incrementality. Gordon et al. (2024) developed PIE (Predictive Incrementality by Experimentation), using a subset of RCTs to train random forest models predicting incremental conversions per dollar (ICPD). The best PIE model achieves prediction errors of 42–62% of mean experimental ICPD across funnel stages, compared to over 2,904% for DML and over 491% for SPSM. PIE disagrees with RCT-based decisions in only 6–13% of cases.

Generative AI for causal inference. Liu et al. (2026b) introduced DICE-DML, using deepfake-generated image pairs to isolate visual attribute treatment effects. Applied to 232,089 Instagram posts, DICE-DML achieved RMSE reductions of 73–97% in simulations and produced a causal estimate for skin tone effects substantially smaller than the biased standard DML estimate—a novel "AI for AI trustworthiness" approach with fairness implications.

Robust evaluation under imperfect labels. Large-scale conversion labels are increasingly constructed from heterogeneous post-view signals or LLM-judged outcomes of varying reliability (Li et al. 2025); Feng et al. (2026) develop statistical estimators that remain valid under imperfect judges, a tool directly applicable to advertising measurement where label noise can bias downstream causal comparisons.

Retrieval-augmented experiment search. A complementary scaling challenge for PIE-style calibration is efficient retrieval over large internal repositories of prior experiments and analyst notes: when calibration is extended from a sampled subset of RCTs to uncalibrated campaigns (Gordon et al. 2024), the practical bottleneck shifts to surfacing the most relevant historical evidence for a given new campaign. Recent work on retrieval-augmented generation outside marketing shows that adding neural reranking on top of dense retrieval substantially improves precision in document-dense factual QA (Cheng et al. 2026b); both that setting and experiment-repository search share the same challenge of locating a relevant finding buried in a specific subsection of a long structured report, making neural reranking directly applicable to analyst-facing experiment-search interfaces that support causal calibration at scale; ensuring that retrieved evidence is not merely topically relevant but appropriately warranted with respect to the causal claim being supported represents a complementary design constraint for production retrieval pipelines (Qian et al. 2026). When LLMs are deployed as tool-using agents within experiment-search or causal-calibration pipelines, reliable tool use further requires learning not merely when to invoke retrieval but which specific retrieval pattern to apply given the analytical context (Xu et al. 2026a); structured intermediate supervision over tool-calling steps (Jiang and Ferraro 2026) provides a complementary training approach that stabilizes multi-step tool chains, with direct applicability to PIE-style calibration workflows where correctly sequencing retrieval, estimation, and uncertainty-quantification steps is necessary for valid incremental-effect inference.

5.4 Comparative Assessment

AI-enhanced causal estimation offers substantial improvements, but the CATE benchmark (Curth et al. 2025) provides sobering evidence: 62% of CATE estimates have higher MSE than a zero-effect predictor, while orthogonality-based models outperform alternatives only 30% of the time. Uplift modeling advances—including M^{3}TN (Zhao et al. 2024) for direct uplift reparameterization and IDUM (Sun et al. 2025) for out-of-distribution generalization—address complementary challenges but do not resolve the fundamental need for experimental ground truth.

Remark 3 (Practitioner Takeaways: Causal Calibration and Experimental Design)

- **Your platform-reported ROAS is likely overstated by 2–5×.** Gordon et al. (2019, 2023) and Gui et al. (2022) converge on evidence that observational attribution systematically inflates effectiveness. Treat platform-reported metrics as upper bounds, not point estimates, until calibrated against an experiment.
- **Geo-experiments are the most accessible calibration tool for mid-size advertisers.** The Trimmed Match estimator (Chen and Au 2022) provides robust inference with as few as 50–200 geographic units. One well-designed annual geo-experiment per major channel is sufficient to anchor your MMM.
- **PIE (Predictive Incrementality by Experimentation) scales calibration without running experiments on every campaign.** Gordon et al. (2024) demonstrate that a subset of RCTs can train models that extend calibration to uncalibrated campaigns, with decision agreement rates of 87–94% versus full experimental measurement.
- **SUTVA violations are real.** When multiple advertisers run concurrent experiments, competitive interference inflates lift estimates by 2× or more (Waisman et al. 2025). If your campaigns operate in competitive categories, account for this in experiment design and interpretation.

6 Privacy-Preserving Measurement

6.1 The Privacy Landscape

The privacy transformation of digital advertising is multifaceted. Goldfarb and Tucker (2011) documented early effects of privacy regulation on ad performance. Aridor et al. (2023) found that GDPR opt-in caused a 12.5% drop in observable consumers but created a privacy externality: remaining consumers became more trackable. Apple's ATT framework produced 37% reductions in Meta advertising click-through rates (Aridor et al. 2024), disproportionately affecting smaller firms and raising competitive equity concerns.

6.2 Differential Privacy Approaches

Differential privacy (DP) provides mathematically rigorous privacy guarantees by ensuring that computation outputs change only negligibly when a single individual's data is modified.

Ghazi et al. (2024) provided the first complete characterization of "operationally valid configurations" for DP attribution, identifying four design dimensions: attribution rule, adjacency definition, contribution bounding scope, and enforcement point. Their analysis reveals that certain common configurations are invalid—a foundational result for privacy-preserving attribution system design.

Xiao et al. (2025) developed AdsBPC, the first user-level DP scheme for streaming ad measurement. Traditional DP mechanisms deplete the privacy budget rapidly in streaming settings. AdsBPC introduces adaptive noise scaling and bounded per-day contribution constraints, achieving 33–95% accuracy improvement over existing streaming DP methods. On real campaign data spanning 31 days, AdsBPC maintained measurement accuracy within acceptable ranges at moderate privacy budgets ($\rho \in [0.25, 8]$).

The W3C Privacy-Preserving Attribution specification (W3C 2025) and the IPA (Interoperable Private Attribution) protocol (Kreuter et al. 2023) represent browser-level and cross-platform implementations of these principles.

6.3 Federated Learning and Clean Rooms

Data clean rooms and federated learning enable collaborative model training without raw data exchange. Li et al. (2024) describe Pinterest's clean-room framework, reporting that batch-level aggregated gradients combined with LoRA adapter fine-tuning recover approximately 95% of the ROC-AUC gain of full fine-tuning while updating only about 1% of model parameters, and that a de-biasing layer under label differential privacy at $\varepsilon = 5$ limits ROC-AUC loss to 0.2–0.5% in their reported production deployment. These figures reflect Pinterest's specific clean-room configuration and model mix, and should not be read as a guarantee that federated learning will retain this much accuracy in other clean-room stacks or industries.

Du et al. (2025) developed PrivacyGo, combining oblivious pseudorandom functions with differential privacy for waterfall-style multi-identifier matching, processing approximately 2 million records in 2 minutes while preventing cross-identifier linkage attacks.

Beyond identifier matching, clean rooms increasingly host analytics workflows that retrieve and aggregate information across distributed campaign documentation, attribution reports, and partner-side data. Within these workflows, the choice of retrieval architecture itself becomes a measurement-design decision: chunk-level retrieval is precise but brittle to document boundary effects, while document-level routing is more robust but less specific. Cheng et al. (2026a) characterize this robustness–precision trade-off in adjacent information-retrieval contexts and propose hybrid document-routed retrieval as a way to combine the two regimes; the same trade-off is directly applicable when clean-room analytics pipelines must balance completeness against precision under federated access constraints; compliance failures can additionally arise when retrieved partner-side data conflicts with the retrieval model's contextual expectations (Chen et al. 2026), a risk present whenever clean-room pipelines aggregate across sources with inconsistent reporting conventions. Transparent conflict-resolution mechanisms (Ye et al. 2026) offer a constructive counterpart: rather than silently overriding conflicting context, they surface competing evidence pathways in

an auditable form that measurement teams can inspect; at the data-distribution level, confusion-resistant federated learning with diffusion-based harmonization on non-IID data (Chen et al. 2024a) directly addresses the structural heterogeneity that arises when partner-side reporting conventions differ systematically across clean-room participants. Prototype-aligned federated soft-prompts (Xiao and Hou 2026) offer a complementary continual-adaptation mechanism: by aligning local model updates via shared prototypes and soft-prompt conditioning without raw data exchange, they enable ongoing model recalibration across advertiser–publisher clean-room stacks as campaign objectives evolve—a federated personalization pathway directly applicable to marketing measurement contexts where attribution models must adapt continuously to shifting media environments.

6.4 Privacy-Native Attribution

The most radical response to privacy constraints is measurement that operates without user-level data by construction. The CDA approach of Filippou et al. (2025) (Section 3.4) exemplifies this philosophy. Lin and Misra (2022) demonstrated that privacy-induced identity fragmentation inflates channel effects by 40–91%. Berman and Feit (2024) proposed latent stratification as a variance reduction technique reducing ATE variance by 30–60% without requiring additional user-level data. A data-generative extension of this philosophy replaces real user records with synthesized behavioral traces: Xu et al. (2026b) develop a two-stage coarse-to-fine diffusion framework for synthesizing realistic human activity traces, providing a privacy-by-construction alternative to aggregate-only measurement that preserves the statistical structure of behavioral sequences while eliminating re-identification risk—a direction particularly relevant when regulatory constraints prevent direct collection of granular user-level data.

Right-to-be-forgotten (RTBF) obligations under GDPR/CCPA also introduce a downstream privacy challenge at the model-maintenance stage: Chen et al. (2024c) show that naive machine unlearning is not bias-neutral—uneven data removal can distort the residual training set—and propose a counterfactual-augmented procedure that restores unbiased performance after deletion, an under-appreciated trustworthiness requirement for attribution and MMM systems that must honor deletion requests. Multimodal analytics pipelines that route customer images, video creatives, or user-generated content through vision-language models face an additional privacy exposure: Wang et al. (2026d) systematically audit large VLMs and document reproducible PII leakage pathways under social-engineering prompts, demonstrating that pseudonymized customer data can be re-identified through inference attacks—a risk not addressed by standard text-only privacy mechanisms and one that warrants dedicated auditing protocols for multimodal marketing stacks.

6.5 Accuracy–Privacy Trade-Off Analysis

Table 5 illustrates the quantifiable trade-offs reported for each privacy mechanism in its representative published deployment. Three observations emerge. First, DP accuracy

costs depend critically on the privacy budget ε: at moderate levels ($\varepsilon = 3$–5), measurement accuracy is often acceptable in the settings reported, but at strict levels ($\varepsilon \leq 1$), noise may render campaign-level estimates uninformative. Second, the very low accuracy loss (0.2–0.5%) reported for federated training with LoRA adapters is specific to the Pinterest clean-room configuration described by Li et al. (2024)—in which label DP is applied at $\varepsilon = 5$ on top of clean-room-mediated aggregate gradients—and should not be read as a general claim about federated learning writ large. Third, the ε parameter lacks clear calibration standards for advertising contexts—a gap requiring both technical research and regulatory guidance.

Table 5 Privacy-preserving measurement: accuracy–privacy trade-offs

Mechanism	Representative	Accuracy loss	Privacy guarantee	Maturity
Differential privacy (event)	Ghazi et al.	2–33%	ε-DP	Standard
Streaming DP	AdsBPC	5–67%	User-level DP	Prototype
Federated + LoRA	Pinterest CR	0.2–0.5%†	Label DP, ε=5	Production
Cryptographic (OPRF+DP)	PrivacyGo	Minimal	Crypto + DP	Production
Aggregate causal	CDA	9.5–24.2%	By construction	Research
Browser API	W3C/IPA	Variable	ε-DP + MPC	Draft

Accuracy Loss measured as increase in estimation error relative to non-private baselines within each mechanism's representative published deployment; cross-mechanism comparisons should therefore be interpreted with caution. Maturity reflects deployment status as of early 2026. †The 0.2–0.5% loss reported for Federated + LoRA reflects the specific Pinterest clean-room configuration in Li et al. (2024) (batch-level aggregate gradients plus LoRA adapters under label DP at $\varepsilon = 5$) and does not generalize to federated learning more broadly.

Remark 4 (Practitioner Takeaways: Privacy-Preserving Measurement)

- **Differential privacy imposes a measurable accuracy cost: plan for it.** At moderate privacy budgets ($\varepsilon = 3$–5), accuracy loss is typically 2–33% relative to non-private baselines (Xiao et al. 2025). At strict budgets ($\varepsilon \leq 1$), campaign-level estimates may become uninformative. Set ε based on acceptable accuracy loss, not regulatory defaults, and re-validate measurement utility after deployment.
- **Federated clean rooms can preserve most measurement utility.** Pinterest's documented clean-room deployment retained approximately 95% of centralized ROC-AUC under label differential privacy at $\varepsilon = 5$ (Li et al. 2024). This is configuration-specific—validate retention in your own stack before assuming comparability.
- **Cookie deprecation and ATT are structural, not temporary.** Aridor et al. (2024) document a 37% reduction in Meta click-through rates attributable to ATT. Design your measurement architecture around aggregate and privacy-native methods now; retrofitting under regulatory pressure is costlier than proactive investment.
- **Machine unlearning (RTBF compliance) is not bias-neutral.** Deleting user records under GDPR/CCPA Right-to-be-Forgotten requests can distort residual

training distributions (Chen et al. 2024c). Audit for bias after deletion events, not only at initial model training.

7 The Measure–Validate–Protect Framework

Our review reveals that advances in any single measurement domain are insufficient without integration across the full pipeline. We formalize this insight as the Measure–Validate–Protect (MVP) framework, operationalizing the NIST trustworthiness dimensions for marketing measurement. As shown in Figure 2, the MVP framework organizes measurement design around three coupled pillars.

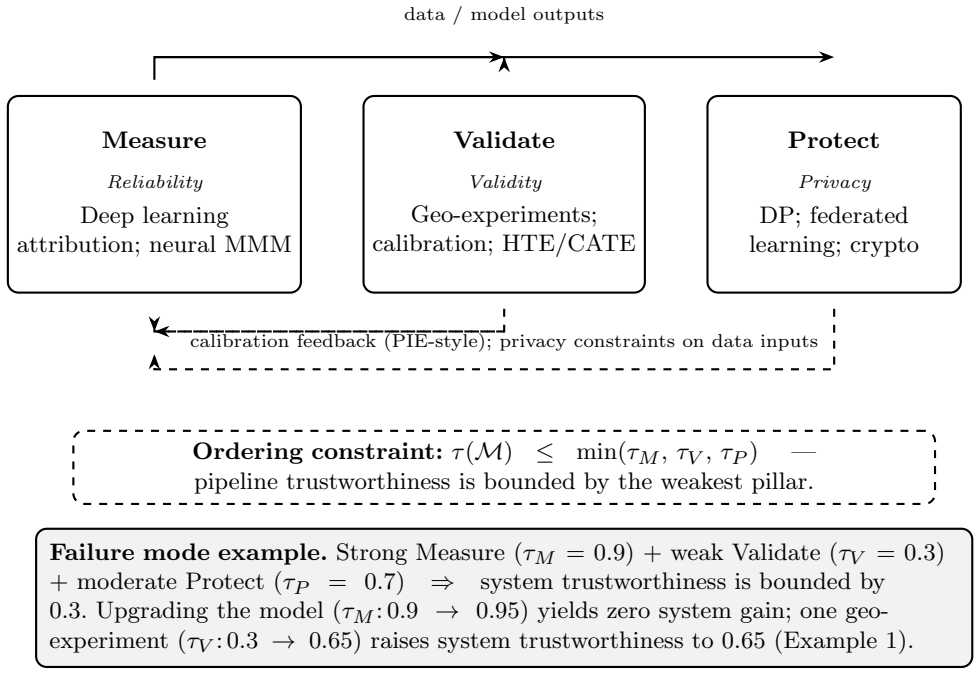


Fig. 2 The Measure–Validate–Protect (MVP) framework for trustworthy AI-driven marketing measurement. Each pillar maps to a NIST trustworthiness dimension adapted in Section 1. Solid arrows show the forward measurement pipeline; dashed arrows show the two feedback paths through which Validate and Protect shape the Measure pillar (calibration retraining and aggregate-data constraints). The ordering constraint and the failure-mode callout illustrate why pipeline trustworthiness is determined by the weakest link rather than by the average across pillars.

Measure. The first pillar encompasses AI-driven measurement models—deep learning attribution (Section 3) and neural MMM (Section 4)—that maximize predictive accuracy and expressiveness. AI models excel at capturing complex temporal dynamics, cross-channel interactions, and creative quality variations that traditional approaches miss. The key trustworthiness requirement is *reliability*: models must produce stable estimates across datasets and market conditions.

Validate. The second pillar requires experimental calibration to establish causal credibility (Section 5). Our review demonstrates that predictive accuracy and causal validity are empirically decoupled—models achieving AUC > 0.97 can still produce attribution errors exceeding 100%. Validation mechanisms include geo-experiments, ghost ads, auction-based instruments, and cross-experiment meta-learning. The key trustworthiness requirement is *validity*: measurement must recover causal effects, not merely correlations. The advertising effectiveness paradox (Section 5.1.1) makes this requirement especially acute: when true effects are as small as elasticities of 0.0008, even modest biases render measurement uninformative.

Protect. The third pillar ensures that measurement operates under privacy constraints (Section 6). Privacy compliance is not merely a regulatory checkbox but a design constraint that shapes measurement architecture. The key trustworthiness requirement is *privacy*: systems must provide formal guarantees while preserving sufficient accuracy for valid causal conclusions.

The MVP framework is distinct from existing constructs in two respects. Triangulation and unified-measurement proposals (Google Think 2022) focus on *combining* methods but do not formalize how trustworthiness degrades when one leg is weak. Responsible-AI and trustworthy-ML frameworks such as the NIST AI RMF (National Institute of Standards and Technology 2023) articulate trustworthiness dimensions generically but do not specify diagnostic metrics or corrective actions for the marketing measurement pipeline. MVP closes both gaps by (a) pinning each trustworthiness dimension to a measurement-specific managerial question, diagnostic metric, and minimum evidence requirement and (b) formalizing the weakest-pillar ordering principle below. Table 6 reports the diagnostic questions and evidence thresholds, while Table 7 lists the corresponding failure modes and recommended actions.

Conceptual grounding and connections to existing frameworks. The MVP framework is not presented as a replacement for existing trustworthiness frameworks but as a marketing-specific instantiation that grounds abstract dimensions in the operational realities of measurement practice. The theoretical contribution is threefold. First, the *ordering constraint* ($\tau(\mathcal{M}) \leq \min(\tau_M, \tau_V, \tau_P)$) is a structural claim that follows from the measurement pipeline's sequential logic: a model whose outputs are causally invalid produces misallocations regardless of how sophisticated its architecture is; a privacy-non-compliant system cannot legally operate regardless of its causal validity. This is analogous to the weakest-link principle in systems reliability theory (Barlow and Proschan 1975). Second, the framework explicitly *decouples* the measurement, validation, and privacy pillars to enable independent diagnosis—a separation that existing unified-measurement frameworks (which treat triangulation as a single activity) do not provide. Third, by connecting to the NIST AI RMF (National Institute of Standards and Technology 2023), the MVP framework anchors its trustworthiness dimensions in a widely adopted engineering standard, providing external validity for the dimension choice.

Contextual calibration of the scoring rubric. In response to reviewer feedback, we note that the 5-point rubric in Table 8 is deliberately agnostic about the absolute level of capability and instead focuses on *relative investment priority.* Appropriate calibration varies by organizational context: a small advertiser with limited

data history (fewer than two years of weekly cross-channel data) should calibrate the Measure pillar to a simpler Bayesian MMM rather than a neural architecture, because the latter's GRU-based temporal dynamics require sufficient data to converge reliably (Tirumala 2025; Sun et al. 2017). A large-platform advertiser with in-house experimentation infrastructure faces a different constraint: the Validate pillar bottleneck is not calibration technology but experimental bandwidth, suggesting that PIE-style meta-learning (Gordon et al. 2024) represents the highest-ROI investment for raising τ_V. For the Protect pillar, the appropriate privacy budget ε depends on both regulatory jurisdiction (GDPR strict, CCPA moderate) and the acceptable accuracy loss at the reporting granularity required by the advertiser's budget-allocation decisions. These contextual factors do not change the framework's ordering constraint but determine where on the 1–5 rubric each pillar starts and which investment path raises the minimum most efficiently.

Table 6 Operationalizing the Measure–Validate–Protect framework (diagnostics). Per-pillar managerial question, diagnostic metric, and minimum evidence required for the pillar to be considered "in compliance." Companion Table 7 lists failure modes and recommended actions.

Pillar	Managerial question	Diagnostic metric	Minimum evidence
Measure	Can the model predict outcomes accurately and stably across datasets?	AUC, MAPE, holdout R^2, train/test gap	Out-of-sample validation on a held-out time window.
Validate	Does the model estimate *incremental* (causal) effects rather than correlations?	Lift vs. RCT, ATE, calibration error vs. geo-experiment	At least one geo-experiment, ghost-ad test, or auction-throttling IV per major channel per year.
Protect	Does the system meet privacy requirements without destroying measurement utility?	Privacy budget ε, aggregation thresholds, clean-room constraints, identity-leakage risk	Documented privacy audit mapping regulation to mechanism (ε-DP, FL/clean room, aggregate-only).

Table 7 Operationalizing the Measure–Validate–Protect framework (failure modes & actions). Companion to Table 6. Together these two tables map the NIST trustworthiness dimensions (Section 1) to concrete decisions a marketing analytics practitioner can audit on a quarterly cycle.

Pillar	Failure mode	Recommended action
Measure	Overfitting; unstable ranking of channels across periods; train/test gap exceeds reported benchmarks.	Simpler MMM, Bayesian priors, regularization, or longer training window; report channel-ranking stability across retrainings.
Validate	Correlation mistaken for causality; attribution inflated by selection or demand effects; no recent calibration.	Run a calibration experiment; apply PIE-style extension from calibrated to uncalibrated campaigns (Gordon et al. 2024).
Protect	Regulatory non-compliance; identity leakage via cross-platform matching; measurement-utility loss exceeds tolerance under privacy mechanism.	Adopt DP, federated, or aggregate-causal architecture; recalibrate ε against acceptable accuracy loss (Table 5); audit RTBF (machine-unlearning) compliance.

7.1 The Ordering Constraint Principle

A critical and underappreciated implication of the MVP framework is what we term the *ordering constraint*: the trustworthiness of the entire measurement pipeline is

bounded by the weakest of its three pillars. A neural attribution model achieving AUC > 0.98 (strong Measure) that operates without experimental calibration (weak Validate) and lacks privacy guarantees (absent Protect) is categorically less trustworthy than a simpler last-touch model that has been calibrated against a geo-experiment and deployed under a compliant differential privacy framework. This asymmetry has practical consequences for investment prioritization: organizations should allocate resources to the weakest link before improving an already-strong component.

We formalize this as the following actionable principle:

Remark 5 (MVP Ordering Principle) For any marketing measurement system $\mathcal{M} = \{M, V, P\}$ with trustworthiness scores $\tau_M, \tau_V, \tau_P \in [0, 1]$, the system-level trustworthiness $\tau(\mathcal{M})$ is bounded above by $\min(\tau_M, \tau_V, \tau_P)$. Marginal investment in improving any pillar above the current minimum yields zero system-level trustworthiness gain until the minimum is raised.

7.2 Operationalizing MVP in Practice

7.2.1 A 5-Point Scoring Rubric for Each Pillar

To make the ordering principle a usable management tool rather than a qualitative slogan, we define each $\tau \in [0, 1]$ as the rescaled value of a 5-point pillar maturity score (1 = absent / non-compliant, 5 = best-in-class) divided by 5. Table 8 provides the rubric.

Table 8 Five-point maturity rubric for the three MVP pillars. Each pillar is scored on the underlying observables in column 3; τ is the score divided by 5. Score thresholds were anchored to the empirical evidence in this review: "1" corresponds to the absence of any defensible practice; "3" represents the minimum practice observed in the reviewed corpus for a commercially operating advertiser (e.g., one geo-experiment per major channel for τ_V); "4" corresponds to established industry best practice as documented in L1–L2 studies (e.g., continuous PIE-style calibration at Snap or Pinterest); "5" corresponds to production-grade evidence from the most advanced deployments reviewed, attained by large-platform advertisers with dedicated experimentation and privacy infrastructure.

Pillar	Underlying observable	1 – 5 levels
Measure (τ_M)	Out-of-sample fit and stability of the production model	1 = no out-of-sample validation; 2 = single train/test split; 3 = held-out time window with reported AUC/MAPE; 4 = rolling-origin validation with stable channel rankings; 5 = production deployment with documented stability across $\geq$ 12 months and multiple market regimes.
Validate (τ_V)	Recency and coverage of experimental calibration	1 = no calibration in >24 months; 2 = single legacy lift study; 3 = one geo-experiment or ghost-ad test in the past 12 months on the largest channel; 4 = annual geo-experiment per major channel; 5 = continuous calibration program (PIE-style) with $\geq$ 80% campaign coverage.
Protect (τ_P)	Documented privacy architecture and audited ε budget	1 = no privacy review; 2 = ad-hoc compliance only; 3 = clean-room or aggregation in place but no documented ε budget; 4 = documented ε budget and DP/FL deployment with measured accuracy loss; 5 = quarterly privacy audit covering attribution, MMM, and machine-unlearning compliance.

This rubric is deliberately conservative: most organizations will land between 2 and 4 on each pillar, and the Example below illustrates the typical asymmetric profile in which the Measure pillar outpaces Validate and Protect. The rubric is also non-prescriptive about *which* mechanism implements each pillar, in line with the framework's "what" rather than "how" orientation.

Example 1 (Applying the MVP Ordering Principle) Consider an enterprise advertiser operating a neural attribution system (LiDDA-style transformer, ROC-AUC > 0.97). Applying the rubric in Table 8: Measure scores 4.5 (rolling-origin validation, channel rankings stable across periods) $\Rightarrow \tau_M = 0.9$; Validate scores 1.5 (no geo-experiment or ghost-ad calibration in the past 18 months, only a legacy 2024 lift study) $\Rightarrow \tau_V = 0.3$; Protect scores 3.5 (clean-room infrastructure documented, but no formal DP audit or ε budget) $\Rightarrow \tau_P = 0.7$. By the MVP Ordering Principle, system-level trustworthiness is $\tau(\mathcal{M}) \leq \min(0.9,\ 0.3,\ 0.7) = 0.3$.

Two investment options are available. *Option A*: upgrade the attribution model to a hypothetically superior architecture that raises τ_M from 0.9 to 0.95. Under the ordering principle, system trustworthiness remains bounded by $\min(0.95, 0.3, 0.7) = 0.3$—a zero marginal gain. *Option B*: commission one geo-experiment for the largest channel (estimated cost: \$40,000–\$80,000 and 8–12 weeks), raising τ_V from 0.3 to 0.65 based on calibration evidence comparable to Chen and Au (2022) and Gordon et al. (2024). System trustworthiness then rises to $\min(0.9, 0.65, 0.7) = 0.65$—a 117% improvement for a fraction of the cost of a model upgrade.

This numerical illustration motivates the phased roadmap in the remainder of this section and the Managerial Decision Guide in Table 9: measurement sophistication should never race ahead of the validation and privacy pillars, because doing so produces no system-level trustworthiness gain while consuming resources that could be redeployed to the binding constraint.

For marketing analytics practitioners, the MVP framework suggests a phased implementation roadmap. In Phase 1 (Measure), organizations should deploy a measurement system—whether traditional Bayesian MMM or a neural variant—that produces defensible channel contribution estimates. The choice of model sophistication should be calibrated to data availability: DeepCausalMMM's GRU-based approach requires sufficient time-series length to learn temporal dynamics reliably, while simpler Bayesian MMM may be more appropriate for organizations with fewer than two years of weekly data across channels.

In Phase 2 (Validate), organizations should establish an experimental calibration cadence—at minimum, one geo-experiment per major channel per year—to anchor the measurement system's causal claims. The key insight from Gordon et al. (2024) is that experimental evidence need not be exhaustive: PIE demonstrates that a subset of RCTs can train predictive models that extend calibration to uncalibrated campaigns. The marginal value of the first calibration experiment vastly exceeds the value of incremental modeling sophistication.

In Phase 3 (Protect), organizations should conduct a privacy architecture audit assessing: (a) whether current attribution systems comply with applicable regulations; (b) the accuracy cost of compliance under the privacy mechanisms available; and (c)

whether aggregate causal methods like CDA (Filippou et al. 2025) can substitute for user-level attribution where regulatory risk is highest.

The MVP framework is explicitly non-prescriptive about the specific technologies employed within each pillar—neural vs. Bayesian, differential privacy vs. federated learning—because appropriate choices depend on organizational context, regulatory environment, and data availability. Its prescriptive contribution is the insistence that all three pillars must be addressed for trustworthy measurement to be achieved.

7.3 Managerial Decision Guide: Choosing a Measurement Architecture

Operationalizing the MVP pillars in practice depends on organizational maturity and data position. Table 9 maps four archetypal organizational contexts to a recommended measurement architecture. The purpose is not to prescribe a single correct stack but to provide a starting point anchored in the evidence synthesized throughout this review; each recommendation below cites the primary studies supporting its feasibility.

Table 9 Managerial Decision Guide: matching organizational context to measurement architecture. Each recommendation is anchored to Level 1–3 evidence from the reviewed corpus; the MVP pillars (M/V/P) indicate which dimensions the architecture primarily addresses and which require additional investment.

Organization type	Data situation	Recommended approach (with MVP mapping)
Small advertiser	Limited channel-level data; no user-level tracking	Bayesian hierarchical MMM (Sun et al. 2017; Zhang et al. 2023) + at least one annual geo-test per major channel (Chen and Au 2022). MVP: M (simple) + V (periodic) + P (natively aggregate).
Mid-size retailer	CRM + paid-media + email; partial user-level data	MMM + holdout/incrementality tests + clean-room matching for paid-social ROAS reconciliation (Li et al. 2024). MVP: M (moderate) + V (holdout tests) + P (clean room).
Large platform / enterprise advertiser	Large-scale user-level journeys; in-house experimentation capacity	Neural attribution (Bencina et al. 2025; Yang et al. 2020) with top-down MMM constraints (Bencina et al. 2025) + continuous RCT/ghost-ad calibration (Johnson et al. 2017a; Gordon et al. 2024) + DP or federated reporting layer (Xiao et al. 2025; Li et al. 2024). MVP: M (advanced) + V (continuous) + P (DP/FL).
Privacy-constrained advertiser	Fragmented identity graph; ATT/GDPR-heavy jurisdiction	Aggregate causal attribution (Filippou et al. 2025) + DP reporting (Ghazi et al. 2024; Xiao et al. 2025) + browser-level privacy-preserving attribution (W3C 2025; Kreuter et al. 2023). MVP: M (aggregate) + V (assumption-dependent) + P (formal DP / MPC).

The table is descriptive rather than exhaustive; a given organization may mix architectures (e.g., an enterprise advertiser using aggregate causal attribution for privacy-heavy regions while running neural attribution in permissive jurisdictions). Feasibility claims under "Recommended approach" reflect the evidence grading in Section 2.6.

The key practical takeaway is that measurement sophistication should be matched to—not exceed—the maturity of the validation and privacy pillars. A small advertiser running Bayesian MMM with one well-designed annual geo-test is closer to MVP compliance than a large advertiser running neural attribution without experimental calibration or a documented privacy architecture, despite the latter's technical sophistication.

8 Future Directions and Open Challenges

Our review highlights four interconnected priority areas for future research; all four are grounded directly in gaps and tensions identified within the reviewed corpus (Sections 3–6) rather than in broad AI trends external to the measurement domain. We nest measurement-to-optimization challenges within scalable calibration because error propagation interacts directly with calibration cadence and experimental design.

8.1 Scalable Calibration

The measurement-calibration gap—more complex AI models require more calibration while making calibration harder—represents the central tension identified across Sections 3, 4, and 5. This gap is directly evidenced by the CATE benchmark finding that 62% of AI causal estimates underperform a zero-effect predictor (Curth et al. 2025): better modeling without better calibration produces no trustworthiness gain.

PIE (Gordon et al. 2024) and the Snap HTE framework (Shi et al. 2024a) demonstrate that meta-learning across experiments can extend calibration coverage without proportional experimental cost—a direction grounded in the corpus. Adaptive experimentation (Fiez et al. 2024) and sequential testing (Bojinov et al. 2023) offer further cost-reduction, also documented in the reviewed studies.

A related challenge lies in the transition from measurement to optimization. Hierarchical offline deep RL systems for cross-channel bidding (He et al. 2023) demonstrate production-scale deployment, but measurement error propagation through the optimization loop remains insufficiently studied—if AI models systematically overestimate certain channels, RL optimizers will compound those errors. This error-amplification risk follows directly from the attribution-calibration gap documented in Section 5.1 and represents a critical feedback failure mode for future investigation, anchored to findings within the reviewed corpus.

8.2 Foundation Models for Marketing Measurement

Foundation models remain at an early stage for measurement tasks, and the central question is not raw capability but *whether predictive gains translate into causal validity and operational viability*. Research on LLM-based synthetic data for causal inference has found that generative models can achieve high predictive accuracy while substantially misestimating average treatment effects, reinforcing the prediction–causation distinction central to this review.

Two performance–efficiency findings sharpen this caution for marketing pipelines that contemplate LLM-judged conversion labels or LLM-based creative scoring. First, on adjacent preference-prediction tasks, Zhang et al. (2026b) benchmark QLoRA-fine-tuned LLaMA-3 and Gemma-2 against gradient-boosted baselines (XGBoost, LightGBM) on the LMSYS Chatbot Arena dataset and find that LLMs improve log loss by only 9.9% while requiring roughly 60,000× higher inference cost. Second, in short-text marketing classification, Lai et al. (2026) compare Sentence-BERT against Word2Vec on e-commerce reviews and find that transformer embeddings do not uniformly dominate simpler baselines once task length, label scarcity, and downstream clustering objectives are controlled; a high-performance cost-effective LLM prompting framework for specialised short-text classification (Liu and Shi 2024) likewise demonstrates that targeted prompt design rather than model scale is the decisive lever in resource-constrained domain-specific labelling scenarios, with direct application to marketing pipelines for creative tagging, brand-safety annotation, and conversion intent labelling. The shared implication is that architecture choice for measurement-adjacent labels (review topic mining, creative tagging, conversion judging) should be matched to task structure, not defaulted to the largest available model. Audits of tabular foundation models for discrete choice further demonstrate that economic validity—the alignment of model predictions with economic rationality constraints—deteriorates when model complexity outpaces the training distribution (Wang et al. 2026b), reinforcing the need for domain-specific validity checks beyond standard predictive metrics. Where retrieval-augmented architectures are used over campaign documents, reviews, or clean-room reports, the architecture-selection problem extends

from model scale to retrieval design (Sections 5.3 and 6.3). The implication for measurement is that foundation-model utility is gated by architecture selection and domain adaptation, not by raw capability alone. For domain-specific retrieval pipelines under SME-grade compute constraints, architectural choices—particularly the balance of retrieval depth against memory augmentation—matter more than raw model scale for downstream QA accuracy (Liu et al. 2026a). The cost-performance trade-offs governing deployed LLM pipelines can be further formalized as an efficiency frontier over context-length and memory-allocation choices (Shen et al. 2026), a structure directly applicable to marketing analytics systems that query large campaign document repositories or clean-room aggregation outputs; domain-specific adaptation combined with model ensembling for small language models (Hu et al. 2026) offers a complementary efficiency pathway for measurement pipelines requiring robust predictions over specialized marketing text—such as campaign briefs, product reviews, or brand-safety annotations—without the infrastructure cost of large-model deployment. SLM-based agents further extend this efficiency advantage to upstream knowledge-mining tasks: Zhang et al. (2026a) demonstrate that orchestrated SLM agents match large-model extraction quality at substantially lower cost, an architecture pattern directly applicable to marketing pipelines that require systematic campaign-document parsing, competitive-intelligence extraction, or content annotation without large-model infrastructure. Multimodal signals present an additional adaptation challenge: for short-video content requiring brand-safety governance, reasoning-enhanced domain-adaptive pretraining of multimodal LLMs has demonstrated production-scale effectiveness (Wang et al. 2025), while modular cross-domain template architectures (Li et al. 2026a) provide a lower-cost adaptation pathway where full fine-tuning is infeasible. Domain-specific prompt engineering—including automatic prompt search and dynamic exemplar selection for few-shot classification—has been shown to substantially improve LLM accuracy on specialized text labeling tasks (Liu and Shi 2025), with direct implications for marketing content pipelines where creative tagging, brand-safety scoring, and intent labeling each represent domain-specific classification objectives that benefit from targeted prompt optimization rather than reliance on general-purpose instruction-following.

8.3 Trustworthiness and Explainability

Three explainability paradigms are emerging within marketing measurement: Shapley value decomposition (Yang et al. 2020), attention-based interpretation (Bencina et al. 2025), and causal graph visualization (Gong et al. 2024). Each offers a different transparency–fidelity trade-off, and no consensus has emerged. Empirical evidence on the stability of these explainability layers reinforces the cautionary stance of the broader low-SNR result: Lin and Wang (2025) train one hundred XGBoost credit-default models with identical hyperparameters but different random seeds and find that SHAP feature-importance rankings are highly stable for the top-five most predictive features (Kendall's $W = 0.93$) yet substantially less stable for mid-importance features ($W = 0.34$). Because most marketing channels other than the dominant one tend to occupy the mid-importance regime, this finding implies that channel-level Shapley attributions for non-dominant channels should be reported with stability

checks across model retrainings rather than as point estimates—a methodological audit that is currently rare in MTA practice. Hierarchical contrastive Shapley values (Xiao et al. 2026) extend this perspective from feature-level to data-sample prioritization: by constructing coalition-level contrasts across training examples, they identify which data samples most drive model outputs—a complementary diagnostic that surfaces the data contributors underlying channel-level attribution estimates and potential poisoning surfaces when applied to MTA training pipelines. Loss landscape visualization tools (Xie et al. 2024) offer a complementary diagnostic lens, enabling practitioners to assess whether the training dynamics underlying a neural attribution model are stable across re-initializations before attributing substantive meaning to its channel-level outputs; topological profiling of the loss surface (Geniesse et al. 2024) extends this diagnostic by capturing global curvature structure—including flat-minima versus sharp-region transitions—that predicts generalization instability independently of any single model checkpoint. For attribution and measurement workflows that incorporate graph neural networks—including GNN-based bot filters (Zhao et al. 2026) that feed cleaned signals into attribution models or spatiotemporal GNNs applied to geographic market data—formal uncertainty quantification of network outputs (Yu et al. 2025) complements these single-checkpoint diagnostics by attaching calibrated confidence bounds to GNN predictions, enabling practitioners to distinguish attributable channel effects from regions of high epistemic uncertainty before those estimates inform budget decisions. Three adjacent lines of XAI research are directly relevant when these paradigms are deployed in measurement pipelines. First, model-agnostic counterfactual explanations—ReLAX (Chen et al. 2022) and feasibility/diversity-constrained variants (Qin et al. 2025)—translate "what-would-need-to-change" decision rules into actionable attribution insights, but, as Chen et al. (2023) caution, the same counterfactual machinery can be an attack surface when the measurement pipeline is exposed to adversarial inputs. Second, adversarial robustness is a first-class trustworthiness dimension under the NIST AI RMF (National Institute of Standards and Technology 2023); certification bounds such as PROSAC (Feng et al. 2025) offer a formal counterpart to the qualitative robustness requirements that marketing measurement systems must meet when model outputs feed budget-allocation decisions; prompt-level adversarial defenses such as logical self-reflection (Lin et al. 2026) provide a complementary run-time safeguard when LLMs are exposed to adversarial user inputs in interactive measurement interfaces; optimized iterative jailbreak attacks (Shi et al. 2026) demonstrate that such defenses can be systematically evaded by tool-assisted adversarial search, underscoring that adversarial-robustness evaluation must include active attack simulation and not only static certification bounds. Beyond single-component robustness, marketing measurement pipelines routinely compose multiple AI modules—attribution models, LLM judges, privacy mechanisms, and optimization layers—introducing emergent risks absent in any individual component: Wang et al. (2026a) quantify how otherwise-safe agent skills generate unsafe behaviors when combined in an ecosystem, underscoring that compositional-safety evaluation is a distinct and necessary dimension of trustworthy marketing AI under the NIST AI RMF. Third, under the low-SNR regime characteristic of advertising data (Section 5.1.1), Shi et al. (2024b) show that standard post-hoc attribution methods can misidentify

signal, a finding that directly constrains how much weight can be placed on any single explainability layer.

The fairness dimension also enters measurement. DICE-DML (Liu et al. 2026b) demonstrated that advertising visual attributes causally affect engagement, implying that AI measurement systems can encode creative biases. Two adjacent literatures speak to this risk: Guo et al. (2024) synthesize intrinsic/extrinsic bias in LLMs and pre-/intra-/post-model mitigation strategies that apply when LLMs annotate or summarize marketing content, and codebook-rebalancing techniques such as CRAB (Fan et al. 2026) mitigate popularity bias in generative recommenders whose engagement signals feed measurement pipelines. Taken together, these results argue for a trustworthiness audit that spans interpretability, low-SNR robustness, adversarial safety, and bias mitigation—all anchored to the measurement task rather than treated as generic AI concerns. Where RL-based training is used to tune LLM scoring or attribution components, reward-design brittleness under sparse conversion signals (Wang et al. 2026c) introduces additional failure modes that require iterative diagnostic refinement before deployment in budget-allocation loops. Interpretable factor decomposition methods (Han et al. 2026) complement Shapley-based channel attribution by providing an explicit mathematical factorization of decision value that is structurally auditable and directly compatible with the orthogonality requirements of a multi-channel MTA audit.

8.4 Cross-Platform Unified Measurement

The ultimate vision is privacy-preserving cross-platform measurement that maintains causal validity. This requires integrating clean-room federated learning, differential privacy for aggregate reporting, and causal structure discovery for identifying channel effects without user-level linkage. The IPA (Interoperable Private Attribution) protocol (Kreuter et al. 2023) and W3C specification (W3C 2025) represent early steps, but substantial technical and institutional barriers remain—including competitive dynamics that may prevent platforms from sharing measurement data even through privacy-preserving mechanisms.

9 Discussion and Conclusion

9.1 Summary of Findings

This systematic review synthesized 108 studies examining AI-driven marketing measurement. Three principal findings emerge, directly addressing our research questions.

Strength-of-evidence summary. Before reporting these findings, we summarize the evidentiary foundation for each topic area to help readers distinguish strongly supported conclusions from more speculative ones (Table 10).

Table 10 Strength-of-evidence summary by topic. "Evidence basis" reports the dominant evidence level (L1–L4) for the central claim; "confidence" is the authors' qualitative assessment: High = central claim supported by multiple L1/L2 studies with replication; Moderate = L2 support with L3 corroboration; Low = primarily L3–L4 evidence; speculative or contested.

Topic / Central claim	Dominant evidence	Confidence	Key caveat
Observational attribution overstates lift (62–115% error)	L1	High	Based on Facebook/Google RCT benchmarks; platform-specific.
Deep learning attribution improves predictive AUC	L2–L3	Moderate	AUC gains documented; causal validity not established.
Neural MMM outperforms Bayesian baselines on prediction	L3	Low–Moderate	Single-benchmark comparison; not externally replicated.
Experimental calibration reduces attribution error	L1	High	Multiple RCT benchmarks across platforms and methods.
DP reduces accuracy by 2–33%	L2–L3	Moderate	Range aggregates heterogeneous configurations; not universal.
Federated/clean-room deployments retain ∼90–95% accuracy	L2	Moderate	Pinterest-specific; configuration-dependent.
Foundation models less cost-efficient than gradient boosting for marketing tasks	L3	Low	Limited direct marketing evidence; requires replication.
SHAP stability degrades for mid-importance channels	L2	Moderate	Credit-risk context; marketing application is analogical.

Regarding **RQ1**, AI methods substantially improve the predictive accuracy and scalability of marketing measurement. Deep learning attribution models achieve AUC exceeding 0.98 (Zhan et al. 2024), neural MMM frameworks reduce prediction error by 5–15× (Mulc et al. 2025), and industrial HTE estimation enables personalized measurement at the scale of hundreds of millions of users (Shi et al. 2024a). These improvements have been validated in production at LinkedIn, Snap, JD.com, and other platforms.

Regarding **RQ2**, predictive accuracy does not guarantee causal validity. The most comprehensive evidence (Gordon et al. 2023) demonstrates median errors of 62–115% relative to experimental benchmarks. This motivates the MVP framework: AI-driven measurement (Measure) must be calibrated against experimental evidence (Validate). PIE (Gordon et al. 2024) exemplifies this integration, using experiments as ground truth and ML as a scaling mechanism. The advertising effectiveness paradox—true elasticities averaging 0.0008 short-term (Shapiro et al. 2022)—makes accurate measurement especially critical, as small biases easily overwhelm genuine effects.

Regarding **RQ3**, privacy-preserving measurement is feasible but involves quantifiable trade-offs. Differential privacy reduces accuracy by 2–33% (Xiao et al. 2025); in a documented clean-room/federated deployment, a combined federated-plus-LoRA configuration retained approximately 90–95% of centralized performance under its reported configuration (Li et al. 2024); and aggregate causal discovery can recover meaningful channel-level attribution (Filippou et al. 2025). These figures are context-specific and should not be read as universal benchmarks. The choice among approaches depends on regulatory environment, data availability, and accuracy requirements.

Taken together, these three findings reveal a structural paradox at the heart of AI-driven marketing measurement: the techniques that most powerfully capture measurement complexity—deep neural architectures, automated causal discovery, personalized HTE estimation—are precisely those that most urgently require the experimental validation and privacy infrastructure that many organizations currently lack. AI does not simplify the trustworthy measurement problem; it raises the stakes.

The MVP framework is designed to make this paradox actionable rather than paralyzing, by providing a principled roadmap for building trustworthy measurement systems incrementally (Section 7).

9.2 Positioning the MVP Framework Against Competing Frameworks

The MVP framework can be situated within a broader landscape of measurement and accountability frameworks. We engage here with the four most closely related alternatives.

Triangulation and unified measurement. The triangulation approach (Google Think 2022) advocates combining MMM, experiments, and attribution at the data level to improve measurement confidence. MVP is complementary but distinct: triangulation specifies *which methods to combine*, while MVP specifies *why combining methods is necessary* (the ordering constraint) and *how to diagnose which leg is the binding constraint*. In particular, the MVP ordering principle implies that adding a third measurement method (say, incrementality testing) to a system that already passes the Measure and Validate thresholds produces lower marginal trustworthiness gain than addressing a weak Protect pillar.

Marketing accountability and marketing productivity. Research on marketing accountability (Rust et al. 2004; Srinivasan et al. 2010) emphasizes the responsibility of marketers to justify investment through rigorous measurement. MVP operationalizes this imperative at the system level: the Validate pillar translates the accountability mandate into a concrete, auditable requirement (experimental calibration coverage), and the ordering constraint formalizes why accountability cannot be achieved by model sophistication alone.

Causal machine learning. The causal ML literature (Chernozhukov et al. 2018; Wager and Athey 2018; Athey et al. 2019) provides the technical machinery for the Validate pillar. MVP's contribution is to position causal ML methods within a broader management framework rather than treating them as independent modeling advances: DML and causal forests are Validate-pillar tools whose value depends on whether the Measure and Protect pillars are also in place.

Responsible AI and privacy engineering. Responsible AI frameworks (National Institute of Standards and Technology 2023) and privacy engineering standards address the organizational governance and technical design of AI systems. MVP draws on these frameworks for the Protect pillar but specializes them to marketing measurement—replacing generic "privacy compliance" with specific requirements about ε-budget documentation, machine-unlearning auditing, and accuracy-loss validation under the privacy mechanism actually deployed. This specificity is necessary because the acceptable accuracy cost of privacy compliance depends on the effect sizes that the measurement system must detect, which are unusually small in advertising contexts (Section 5.1.1).

Theoretical contribution: the weakest-pillar principle. The MVP framework's primary theoretical claim—that pipeline trustworthiness is bounded by its weakest pillar—is a marketing-specific instantiation of the min-max logic found in systems reliability theory. Its practical implication (that resource allocation should target

the binding constraint, not the strongest pillar) follows directly from this logic and is empirically grounded in the review findings: the studies most clearly showing high predictive accuracy with low causal credibility (Section 3.5) are precisely those where organizations have invested in Measure while neglecting Validate.

9.3 Implications for Research

We identify four high-priority research directions. First, **privacy-native causal inference methods** that identify treatment effects from aggregate data without user-level variation represent the most impactful technical challenge. Second, **calibration of differential privacy parameters** for advertising contexts requires both empirical research and regulatory engagement—current implementations use ε values ranging from 1 to 8 without clear justification. Third, the **measurement-to-action feedback loop** created by integrating reinforcement learning with causal measurement introduces error amplification risks that require theoretical and empirical analysis. Fourth, **foundation models for marketing measurement** remain at an early stage, with the potential for transfer learning largely unexplored.

9.4 Implications for Practice

For marketing practitioners, our review suggests three actionable principles. First, adopt a **triangulation approach** combining at least two independent measurement methods. Google's published unified measurement program reports roughly 40% higher expected uplift versus MTA-only designs when MMM, experiments, and attribution are integrated at the data level (Google Think 2022). Industry deployments—including agency-led unified measurement programs and Pinterest's clean-room stack—illustrate that MVP-style integration is not merely theoretical: organizations that operationalize measurement, experimentation, and privacy-preserving collaboration can achieve materially higher confidence in measurement-driven budget decisions (Google Think 2022; Li et al. 2024). Second, invest in **experimental calibration infrastructure** before model sophistication. Gordon et al. (2023) and Gordon et al. (2024) demonstrate that even simple calibration of last-click conversions outperforms sophisticated non-experimental methods by orders of magnitude. The marginal value of one additional geo-experiment exceeds upgrading from Bayesian to neural MMM for most organizations (Rust et al. 2004). Third, evaluate privacy-preserving measurement based on **causal properties**, not merely predictive accuracy. Beyond technical considerations, organizations face significant **implementation barriers**. Building in-house experimentation infrastructure requires substantial engineering investment. The skill sets needed—spanning causal inference, deep learning, privacy engineering, and marketing domain knowledge—are rarely found in a single team. Organizational silos between data science, marketing, and legal/compliance functions impede the integrated approach that the MVP framework demands. Addressing these barriers requires executive sponsorship, cross-functional teams, and incremental adoption strategies.

9.5 Limitations

This review has several limitations that readers should consider when interpreting findings.

Inter-rater reliability. Inter-rater agreement on inclusion decisions reached Cohen's $\kappa = 0.81$ on the title/abstract calibration subset ($n = 64$) and $\kappa = 0.83$ on the full-text candidates ($n = 189$, all independently dual-coded), both in the "almost perfect" range under McHugh (2012) criteria. Evidence-level coding achieved weighted $\kappa = 0.76$ ("substantial" agreement) on the calibration subset. Post-calibration screening of the remaining title/abstract records was performed by a single author with bilateral and third-author adjudication for all ambiguous cases, rather than by independent dual coding throughout. Full-corpus dual coding of all 635 records would have provided a stronger reliability guarantee, and we encourage replications in this domain to adopt that more conservative protocol. The κ statistics reported above are for the calibration stage; their applicability to the full corpus is an acknowledged limitation.

Publication bias. The corpus is likely skewed toward positive results. Industry-led research from major technology platforms (Google, Meta, Snap, Pinterest) is systematically over-represented because these organizations have both the resources to run large-scale experiments and the incentive to publish favorable results. Null or negative findings—experiments showing no measurable advertising lift, or models that fail to improve on simple baselines—are underrepresented. This may cause our review to present an optimistic picture of AI-driven measurement performance.

Preprint quality variance. At 41% of the corpus, preprints introduce quality variance that the four-level evidence grading scheme partially mitigates but does not eliminate. Preprints indexed on arXiv and SSRN have not undergone peer review, and some claims in Level 3–4 studies may not survive formal review. We have hedged all preprint-based claims accordingly, but readers should treat findings supported only by Level 3–4 evidence as directional rather than definitive.

Date ceiling. Database searches were concluded in March 2026. The rapid pace of development in large language models, foundation models for tabular data, and privacy-preserving computation means that relevant work published or posted after this date is not captured. In particular, post-GPT-4 developments in LLM-based creative measurement and emerging regulatory frameworks (e.g., EU AI Act implementation guidance) may alter conclusions in Sections 8.2 and 6.

Corpus concentration. The included studies are disproportionately drawn from large technology platforms and top-tier computer science venues. Measurement practices at mid-size advertisers, agencies, and non-US markets are underrepresented, limiting the generalizability of our practitioner-facing recommendations to organizations with large-scale data infrastructure.

Narrative synthesis. The heterogeneity of study designs, outcome metrics, and evaluation settings across the 108 included studies precludes quantitative meta-analysis of effect sizes. Effect-size ranges reported throughout the review (e.g., DP accuracy loss of 2–33%) aggregate across heterogeneous configurations and should not be interpreted as precise universal benchmarks.

9.6 Conclusion

AI is transforming marketing measurement with unprecedented capabilities in pattern recognition, temporal modeling, and privacy-preserving computation. However, our review reveals a persistent tension between predictive power and causal credibility. The most important insight is not that AI makes measurement more accurate—though it often does—but that AI makes the distinction between prediction and causation more consequential. As AI-driven systems increasingly inform multi-million-dollar budget allocation decisions, conflating correlation with causation carries higher stakes than ever. The Measure–Validate–Protect framework proposed in this review provides a roadmap: trustworthy marketing measurement requires sophisticated AI models, rigorous experimental validation, and principled privacy design, integrated as a coherent system rather than as independent components. Together, our answers to RQ1–RQ3 establish that AI-driven marketing measurement is powerful, causally limited without calibration, and privacy-compliant at a measurable accuracy cost—a triple constraint that the MVP framework is designed to navigate systematically.

Supplementary information. Supplementary Table S1 (`Supplementary_Table_S1_Included_Studies_Matrix.xlsx`) provides the full 108-study included-studies matrix used for coding, evidence grading, and section-level citation placement. The reference list reports works cited in the article text; Supplementary Table S1 records additional included studies that were coded for corpus-level synthesis but not individually cited in the main text.

Acknowledgements. The authors thank colleagues and reviewers for helpful comments.

Declarations

Funding. No funding was received for this study.

Competing interests. The authors declare no competing interests.

Ethics approval and consent to participate. Not applicable.

Consent for publication. Not applicable.

Data availability. This study analyzes previously published or posted literature. The extracted 108-study included-studies matrix is provided as Supplementary Table S1.

Materials availability. Not applicable.

Code availability. Not applicable.

Author contributions. Author contributions are omitted from the anonymized manuscript for double-blind review.

Appendix A PRISMA Screening Summary

The systematic search and screening process followed PRISMA 2020 guidelines (Page et al. 2021). Records identified through database searching: 847 (Scopus 312; Web of Science 248; IEEE Xplore 137; ACM DL 94; arXiv/SSRN 56). Records after duplicate removal: 635. Records screened by title and abstract: 635. Records excluded at screening: 446 (reasons: not measurement-focused, 289; no AI/ML component, 102; non-English, 31; trade publications, 24). Full-text articles assessed for eligibility: 189. Full-text articles excluded: 81 (reasons: insufficient empirical content, 38; scope mismatch, 27; superseded by later version, 16). Studies included from database search: 80. Additional studies from expert seeding and citation tracking: 28. Total studies included in review: 108. Reviewer calibration was performed on a subset of records and residual disagreements were resolved through consensus (see Appendix C).

Appendix B Search Strings and Database Results

All searches were executed between January 15 and March 10, 2026, restricted to English-language records from 2010 through early 2026. Boolean logic was the same across databases—`C1 AND (C2 OR C3)`—with field specifiers adapted per platform.

Scopus (`TITLE-ABS-KEY`). 312 records:

```
TITLE-ABS-KEY(
  ("multi-touch attribution" OR "media mix model*" OR
   "marketing measurement" OR "advertising effectiveness" OR
   "return on ad spend" OR "ROAS" OR "advertising lift")
  AND
  (("deep learning" OR "neural network*" OR "transformer" OR
    "recurrent neural" OR "graph neural" OR "machine learning" OR
    "causal inference" OR "heterogeneous treatment effect" OR
    "uplift model*" OR "double machine learning")
   OR
   ("randomized experiment" OR "field experiment" OR "A/B test*" OR
    "geo experiment" OR "causal calibration" OR "differential privacy" OR
    "federated learning" OR "clean room" OR "privacy-preserving")))
AND PUBYEAR > 2009 AND PUBYEAR < 2027 AND LANGUAGE(english)
```

Web of Science (Topic field). 248 records. Same Boolean structure, translated to `TS=(C1) AND (TS=(C2) OR TS=(C3))` with WoS wildcards (`*`) preserved.

IEEE Xplore (Command search, Metadata). 137 records:

```
((("Document Title":"attribution" OR "multi-touch" OR
   "media mix" OR "advertising effectiveness") OR
  ("Abstract":"marketing measurement" OR "advertising lift"))
 AND
 ("Abstract":"deep learning" OR "neural" OR "transformer" OR
  "causal" OR "uplift" OR "differential privacy" OR "federated")
)
Filters: 2010-2026, English
```

ACM Digital Library (Advanced Search). 94 records:

```
[[Abstract: "attribution" OR "media mix" OR
  "marketing measurement" OR "advertising lift"]
 AND
 [Abstract: "deep learning" OR "neural" OR "transformer" OR
  "causal" OR "differential privacy" OR "federated" OR
  "clean room"]]
Published: 2010-01-01 to 2026-03-31, Language: English
```

arXiv / SSRN (title + abstract). 56 records total (arXiv 41; SSRN 15). We used the arXiv API with Boolean AND of three term groups and restricted to categories `cs.LG`, `cs.IR`, `stat.AP`, `stat.ML`, `econ.EM`, `cs.CR`. SSRN was searched by exact phrase queries ("multi-touch attribution", "media mix model", "advertising effectiveness", "privacy-preserving attribution") with year filter 2010–2026.
Yield summary. Raw records 847; duplicates 212 (auto-removed via DOI and normalized-title matching); unique records screened 635; excluded at title/abstract 446; full-text assessed 189; excluded at full-text 81; included 108.

Appendix C Coding Protocol and Reviewer Calibration

The coding scheme below was finalized before screening began. Two authors jointly reviewed a calibration subset of records (approximately 10% of the 635 title/abstract records and all 189 full-text candidates) to align interpretation of the inclusion criteria and each coding dimension; subsequent records were screened by one author, with ambiguous or boundary cases flagged for joint review. Disagreements and flagged cases were resolved through discussion; cases that could not be reconciled bilaterally were adjudicated by a third author. The coding spreadsheet underlying this review is provided as Supplementary Table S1 (Appendix D).

Inclusion decision (binary). Include iff all three hold: (a) addresses marketing or advertising measurement, attribution, or media mix modeling; (b) employs or evaluates an AI/ML method, a causal-inference technique, or a privacy-preserving technology; (c) presents empirical results, a theoretical framework, or a systematic review.

Exclusion decision (any-of). Exclude if: exclusively about recommender systems / content generation / consumer behavior without a measurement component; non-English; trade publication without methodological content; superseded by a later version already in the corpus.

Coding Dimension 1: measurement function (mutually exclusive primary value plus multi-domain flag). {Attribution; MMM; Causal calibration; Privacy-preserving measurement; Multi-domain}.

Coding Dimension 2: AI technique (non-exclusive binary flags). {RNN/LSTM/GRU; Transformer; Neural ODE; DAG/Causal discovery; DML/Debiased ML; Causal forest/GRF; Differential privacy; Federated learning; Cryptography}.

Coding Dimension 3: evidence level (ordinal, 1–4). See Table 3.

Extraction fields (free-text). Research question; data scale & source; experimental vs. observational; primary performance metric(s) and value(s); stated limitations.

Reviewer calibration process. Calibration records that elicited disagreement were used to refine the written coding rubric before it was applied to the balance of the corpus. In line with practical guidance for interrater calibration in systematic reviews (McHugh 2012), residual disagreements (both at title/abstract and full-text stages) were logged and resolved through structured consensus discussion, with third-author adjudication reserved for cases that could not be reconciled bilaterally. Inter-rater reliability statistics: title/abstract binary inclusion $\kappa = 0.81$ ($n = 64$); full-text binary inclusion $\kappa = 0.83$ ($n = 189$); evidence-level weighted $\kappa = 0.76$ ($n = 64$). All three statistics fall in the "substantial" to "almost perfect" range under McHugh (2012) criteria.

Appendix D Included Studies Matrix (summary) and Supplementary Material

Table D1 summarizes the 108 included studies aggregated by Dimension 1 (measurement function), Dimension 2 (dominant AI technique family), and Dimension 3 (evidence level). The per-paper structured matrix is provided with this submission as **Supplementary Table S1** (`Supplementary_Table_S1_Included_Studies_Matrix.xlsx`), with the following fields per study: *citation_key, year, publication_type, venue, peer_reviewed, evidence_level, measurement_function, AI_method_family, data_scale, evaluation_metric, key_finding, limitation, cited_in_section.* The supplementary file is submitted concurrently with the manuscript so that reviewers can directly verify corpus composition, evidence grading, and section-level citation placement. Some corpus records are synthesized only at the aggregate level and therefore appear in Supplementary Table S1 without being individually cited in the article reference list.

Table D1 Distribution of the 108 included studies by measurement function (Dim. 1) and evidence level (Dim. 3). Cells count studies; row/column totals allow readers to verify that no single cell drives a central claim. L1 evidence concentrates in causal calibration, while L2 evidence spans attribution, causal calibration, privacy-preserving measurement, and multi-domain work.

Measurement function	L1	L2	L3	L4	Total	% of corpus
Attribution (MTA, CDA)	0	5	17	2	24	22%
Media mix modeling	0	2	10	9	21	19%
Causal calibration & HTE	8	11	9	0	28	26%
Privacy-preserving measurement	1	6	10	3	20	19%
Multi-domain / framework	0	6	7	2	15	14%
Total	9	30	53	16	108	100%

L1 = peer-reviewed field experiments / RCT benchmarks; L2 = peer-reviewed empirical studies or replicated production deployments; L3 = preprints with empirical validation; L4 = conceptual / prototype / white-paper work. See Section 2.6 and Table 3.

References

Abadie A, Diamond A, Hainmueller J (2010) Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. Journal of the American Statistical Association 105(490):493–505

Aridor G, Che YK, Salz T (2023) The effect of privacy regulation on the data industry: Empirical evidence from GDPR. RAND Journal of Economics 54(4):649–672

Aridor G, Che YK, Hollenbeck B, et al (2024) Evaluating the impact of privacy regulation on e-commerce firms: Evidence from Apple's App Tracking Transparency. SSRN Working Paper 4698374

Athey S, Tibshirani J, Wager S (2019) Generalized random forests. Annals of Statistics 47(2):1148–1178

Barlow RE, Proschan F (1975) Statistical Theory of Reliability and Life Testing: Probability Models. Holt, Rinehart and Winston, New York

Basu R, Aktar N, Kumar S (2024) The interplay of artificial intelligence, machine learning, and data analytics in digital marketing and promotions: A review and research agenda. Journal of Marketing Analytics 12(4):786–812

Bencina J, Aykutlug E, Chen Y, et al (2025) LiDDA: Data driven attribution at LinkedIn. ArXiv preprint arXiv:2505.09861

Berman R, Feit EM (2024) Latent stratification for incrementality experiments. Marketing Science 43(4):903–917

Bojinov I, Rambachan A, Shephard N (2023) Design-based confidence sequences for sequential experiments. ArXiv preprint arXiv:2210.08639

Brodersen KH, Gallusser F, Koehler J, et al (2015) Inferring causal impact using Bayesian structural time-series models. Annals of Applied Statistics 9(1):247–274

Chen A, Au TC (2022) Robust causal inference for incremental return on ad spend with randomized paired geo experiments. Annals of Applied Statistics 16(1):1–20

Chen X, Xiao C, Liu Y (2024a) Confusion-resistant federated learning via diffusion-based data harmonization on non-IID data. Advances in Neural Information Processing Systems 37:137495–137520

Chen Y, Qian P, Wang S, et al (2026) Does RAG know when retrieval is wrong? diagnosing context compliance under knowledge conflict. URL https://arxiv.org/abs/2605.14473, arXiv:2605.14473

Chen Z, Silvestri F, Wang J, et al (2022) ReLAX: Reinforcement learning agent explainer for arbitrary predictive models. In: Proceedings of the 31st ACM International Conference on Information and Knowledge Management (CIKM), URL https://arxiv.org/abs/2110.11960, arXiv:2110.11960

Chen Z, Silvestri F, Tolomei G, et al (2023) The dark side of explanations: Poisoning recommender systems with counterfactual examples. In: Proceedings of the 46th International ACM SIGIR Conference, URL https://arxiv.org/abs/2305.00574, arXiv:2305.00574

Chen Z, Tan H, Wang H, et al (2024b) A new baseline assumption of integrated gradients based on Shapley value. URL https://arxiv.org/abs/2310.04821, arXiv:2310.04821

Chen Z, Wang J, Zhuang J, et al (2024c) Debiasing machine unlearning with counterfactual examples. URL https://arxiv.org/abs/2404.15760, arXiv:2404.15760

Cheng Z, Lai L, Liu Y (2026a) Resolving the robustness-precision trade-off in financial RAG through hybrid document-routed retrieval. URL https://arxiv.org/abs/2603.26815, arXiv:2603.26815

Cheng Z, Lai L, Liu Y, et al (2026b) Enhancing financial report question-answering: A retrieval-augmented generation system with reranking analysis. URL https://arxiv.org/abs/2603.16877, arXiv:2603.16877

Chernozhukov V, Chetverikov D, Demirer M, et al (2018) Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal 21(1):C1–C68

Curth A, et al (2025) A large-scale empirical assessment of CATE estimation. In: Proceedings of the 13th International Conference on Learning Representations (ICLR)

Dalessandro B, Stitelman O, Perlich C, et al (2012) Causally motivated attribution for online advertising. In: Proceedings of the Sixth International Workshop on Data Mining for Online Advertising and Internet Economy (ADKDD). ACM

Du R, Zhong Y, Nair H, et al (2019) Causally driven incremental multi touch attribution using a recurrent neural network. ArXiv preprint arXiv:1902.00215

Du X, Qian C, Zhang L, et al (2025) PrivacyGo: Secure and privacy-preserving multi-identifier ad measurement. ArXiv preprint arXiv:2506.20981

Fan Z, et al (2026) CRAB: Codebook rebalancing for bias mitigation in generative recommendation. URL https://arxiv.org/abs/2604.05113, arXiv:2604.05113

Feng C, Liu Z, Zhi Z, et al (2025) PROSAC: Provably safe certification for machine learning models under adversarial attacks. In: The 39th Annual AAAI Conference on Artificial Intelligence (AAAI) [Oral], https://doi.org/10.1609/aaai.v39i3.32300

Feng C, Shen M, Balashankar A, et al (2026) Noisy but valid: Robust statistical evaluation of LLMs with imperfect judges. In: The Fourteenth International Conference on Learning Representations (ICLR), URL https://openreview.net/forum?id=hEhxreaLdU

Fiez T, Nassif H, Chen YC, et al (2024) Best of three worlds: Adaptive experimentation for digital marketing in practice. In: Proceedings of the ACM Web Conference 2024 (WWW), arXiv preprint arXiv:2402.10870

Filippou A, Quach M, Lenghel I, et al (2025) Causal-driven attribution: Estimating channel influence without user-level data. ArXiv preprint arXiv:2512.21211

Gaur J, Bharti K (2020) Attribution modelling in marketing: Literature review and research agenda. Academy of Marketing Studies Journal 24(4):1–21

Geniesse C, Chen J, Xie T, et al (2024) Visualizing loss functions as topological landscape profiles. URL https://arxiv.org/abs/2411.12136, arXiv:2411.12136

Ghazi B, Kamath P, Kumar R, et al (2024) Differentially private ad conversion measurement. Proceedings on Privacy Enhancing Technologies 2024(2). ArXiv preprint arXiv:2403.15224

Goldfarb A, Tucker C (2011) Privacy regulation and online advertising. Management Science 57(1):57–71

Gong C, Yao D, Zhang L, et al (2024) CausalMMM: Learning causal structure for marketing mix modeling. In: Proceedings of the 17th ACM International Conference on Web Search and Data Mining (WSDM), arXiv preprint arXiv:2406.16728

Google Think (2022) Unified marketing measurement. ThinkWithGoogle whitepaper

Gordon BR, Zettelmeyer F, Bhargava N, et al (2019) A comparison of approaches to advertising measurement: Evidence from big field experiments at Facebook. Marketing Science 38(2):193–225

Gordon BR, Moakler R, Zettelmeyer F (2023) Close enough? A large-scale exploration of non-experimental approaches to advertising measurement. Marketing Science 42(4):768–793

Gordon BR, Moakler R, Zettelmeyer F (2024) Predictive incrementality by experimentation (PIE) for ad measurement. ArXiv preprint arXiv:2304.06828

Gui J, Nair H, Niu Y (2022) Auction throttling and causal inference of online advertising effects. ArXiv preprint arXiv:2112.15155

Guo Y, Guo M, Su J, et al (2024) Bias in large language models: Origin, evaluation, and mitigation. URL https://arxiv.org/abs/2411.10915, arXiv preprint arXiv:2411.10915

Gupta NR, Orlandi V, Chang CR, et al (2025) dame-flame: A python package providing fast interpretable matching for causal inference. Journal of Statistical Software 113(2):1–26. https://doi.org/10.18637/jss.v113.i02

Han X, Xiao Y, Zhang Z, et al (2026) Interpretable factor decomposition for decision intelligence in large-scale financial markets: Evidence from China's A-share market. URL https://arxiv.org/abs/2606.12843, arXiv:2606.12843

Hanssens DM, Parsons LJ, Schultz RL (2001) Market Response Models: Econometric and Time Series Analysis, 2nd edn. Springer, Boston, MA

He X, et al (2023) HiBid: Hierarchical offline deep reinforcement learning for cross-channel bidding. ArXiv preprint arXiv:2312.17503

Hu L, Xin Y, Shen B, et al (2026) CoDES: A context-efficient framework for enhancing small language models via domain-specific adaptation and model ensembling. Preprints, https://doi.org/10.20944/preprints202603.1152.v1, URL https://doi.org/10.20944/preprints202603.1152.v1

Jiang Y, Ferraro F (2026) SCRIBE: Structured mid-level supervision for tool-using language models. URL https://arxiv.org/abs/2601.03555, arXiv:2601.03555

Johnson GA, Lewis RA, Nubbemeyer EI (2017a) Ghost ads: Improving the economics of measuring online ad effectiveness. Journal of Marketing Research 54(6):867–884

Johnson GA, Lewis RA, Nubbemeyer EI (2017b) The online display ad effectiveness funnel: Lessons from 432 field experiments. SSRN Working Paper 2701578

Kreuter B, et al (2023) Interoperable private attribution. IACR ePrint 2023/437

Kumar A, et al (2020) CAMTA: Causal attention model for multi-touch attribution. ArXiv preprint arXiv:2012.11403

Künzel SR, Sekhon JS, Bickel PJ, et al (2019) Metalearners for estimating heterogeneous treatment effects using machine learning. Proceedings of the National Academy of Sciences 116(10):4156–4165

Lai L, Cheng Z, Cheng K, et al (2026) Do transformers always win? an empirical study of semantic embeddings for short-text e-commerce reviews. In: 2026 9th International Symposium on Big Data and Applied Statistics (ISBDAS), pp 525–529, https://doi.org/10.1109/ISBDAS69350.2026.11484350

Leeflang PS, Wieringa JE, Bijmolt TH, et al (2009) Modeling markets: Analyzing marketing phenomena and improving marketing decision making. International Journal

of Research in Marketing 26(3):131–150. https://doi.org/10.1016/j.ijresmar.2009.05.005

Lewis RA, Rao JM (2015) The unfavorable economics of measuring the returns to advertising. Quarterly Journal of Economics 130(4):1941–1973

Li C, Chen Y, Leng C, et al (2024) Privacy-preserving conversion prediction using data clean rooms. In Proceedings of the 18th ACM Conference on Recommender Systems (RecSys). arXiv preprint arXiv:2505.14959

Li D, Jiang B, Huang L, et al (2025) From generation to judgment: Opportunities and challenges of LLM-as-a-judge. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pp 2757–2791

Li X, Lu Y, Cao J, et al (2026a) CATCH: A modular cross-domain adaptive template with hook. In: Bebis G, Ye J, Wang Y, et al (eds) Advances in Visual Computing. Springer Nature Switzerland, Cham, pp 41–52, https://doi.org/10.1007/978-3-032-14492-8_4

Li Y, Fan Z, Zhuang Z (2026b) When AUC 0.998 is not enough: A candidate evaluation protocol for hidden-state probes of indirect prompt injection in multimodal computer-use agents. URL https://arxiv.org/abs/2606.22864, arXiv:2606.22864

Lin L, Wang Y (2025) SHAP stability in credit risk management: A case study in credit card default model. Risks 13(12):238. https://doi.org/10.3390/risks13120238

Lin L, You J, Li Y, et al (2026) Reflect-Guard: Enhancing LLM safeguards against adversarial prompts via logical self-reflection. URL https://arxiv.org/abs/2605.24834, arXiv:2605.24834

Lin T, Misra S (2022) The identity fragmentation bias. Marketing Science 41(3):433–440

Liu J, Yang J, Li X, et al (2026a) Architecture matters more than scale: A comparative study of retrieval and memory augmentation for financial qa under sme compute constraints. URL https://arxiv.org/abs/2604.17979, arXiv:2604.17979

Liu M, Shi G (2024) PoliPrompt: A high-performance cost-effective LLM-based text classification framework for political science. Available at SSRN 4940136, URL https://ssrn.com/abstract=4940136

Liu M, Shi G (2025) Enhancing LLM-based text classification in political science: Automatic prompt optimization and dynamic exemplar selection for few-shot learning. URL https://arxiv.org/abs/2409.01466, arXiv:2409.01466

Liu Z, Padmanabhan K, Viswanathan S (2026b) DICE-DML: Deepfake-informed double machine learning for visual ad effects. ArXiv preprint arXiv:2603.02359

Lu Z, Kannan PK (2024) AI for customer journeys: A transformer approach. SSRN Working Paper 4684617

McHugh ML (2012) Interrater reliability: The kappa statistic. Biochemia Medica 22(3):276–282

Mehta K, Singhal E (2020) Marketing channel attribution modelling: Markov chain analysis. International Journal of Indian Culture and Business Management 21(1):63–77

Mulc T, Anderson B, Cubre C, et al (2025) NNN: Next-generation neural networks for marketing measurement. ArXiv preprint arXiv:2504.06212

National Institute of Standards and Technology (2023) Artificial intelligence risk management framework (AI RMF 1.0). Tech. Rep. AI 100-1, NIST, https://doi.org/10.6028/NIST.AI.100-1

Ng E, Wang Z, Chen H, et al (2021) Orbit: Probabilistic forecast with exponential smoothing. ArXiv preprint arXiv:2106.03322

Page MJ, McKenzie JE, Bossuyt PM, et al (2021) The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ 372:n71

Pauwels K, Silva-Risso J, Srinivasan S, et al (2004) New products, sales promotions, and firm value: The case of the automobile industry. Journal of Marketing 68(4):142–156. https://doi.org/10.1509/jmkg.68.4.142.42724

Qi Y, Zhao S, Xue Z, et al (2026) Detecting abnormal user feedback patterns through temporal sentiment aggregation. URL https://arxiv.org/abs/2604.00020, arXiv:2604.00020

Qian P, Wang S, Wang X, et al (2026) Relevant is not warranted: Evidence-force calibration for cited RAG. URL https://arxiv.org/abs/2605.28044, arXiv:2605.28044

Qin X, Li S, Cai Y, et al (2025) Enhancing counterfactual explanations with feasibility and diversity. In: 2025 IEEE International Conference on Data Mining Workshops (ICDMW), IEEE, pp 2310–2319

Rust RT, Ambler T, Carpenter GS, et al (2004) Measuring marketing productivity: Current knowledge and future directions. Journal of Marketing 68(4):76–89. https://doi.org/10.1509/jmkg.68.4.76.42721

Shapiro BT, Hitsch GJ, Tuchman A, et al (2022) Advertising in the age of digital targeting: Meta-analysis with improved bias corrections. SSRN Working Paper 4293750

Shen B, Jin L, Cai H, et al (2026) The efficiency frontier: A unified framework for cost-performance optimization in LLM context management. https://doi.org/10.48550/

arXiv.2605.23071, URL https://arxiv.org/abs/2605.23071, arXiv:2605.23071

Shi C, Pan R, Lo J, et al (2024a) A large-scale heterogeneous treatment effect estimation framework and its applications at Snap. ArXiv preprint arXiv:2512.03060

Shi G, Kan Z, Smucny J, et al (2024b) ChaosMining: A benchmark to evaluate post-hoc local attribution methods in low-SNR environments. URL https://arxiv.org/abs/2406.12150, arXiv:2406.12150

Shi G, Yin J, Xie D, et al (2026) JailbreakOPT: Tool-assisted iterative jailbreak prompt optimization. URL https://arxiv.org/abs/2606.11425, arXiv:2606.11425

Srinivasan S, Vanhuele M, Pauwels K (2010) Mind-set metrics in market response models: An integrative approach. Journal of Marketing Research 47(4):672–684. https://doi.org/10.1509/jmkr.47.4.672

Sun J, et al (2025) IDUM: Invariant deep uplift modeling. In: Proceedings of the 42nd International Conference on Machine Learning (ICML)

Sun Y, Wang Y, Jin Y, et al (2017) Geo-level Bayesian hierarchical media mix modeling. Google Research

Tirumala S (2025) DeepCausalMMM: A deep learning framework for marketing mix modeling with causal inference. ArXiv preprint arXiv:2510.13087

Varambally S, et al (2024) Discovering mixtures of structural causal models from time series data. In: Proceedings of the 41st International Conference on Machine Learning (ICML)

W3C (2025) Privacy-preserving attribution: Level 1. W3C Working Draft, April 2025

Wager S, Athey S (2018) Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association 113(523):1228–1242

Waisman C, Sahni NS, Nair HS, et al (2025) Parallel experimentation and competitive interference on online advertising platforms. Marketing Science 44(2):437–456

Wang S, Qian P, Chen Y, et al (2026a) When safe skills collide: Measuring compositional risk in agent skill ecosystems. URL https://arxiv.org/abs/2606.00448, arXiv:2606.00448

Wang Y, Sun X, Li Y, et al (2026b) Auditing and fixing economic validity in tabular foundation models for discrete choice. URL https://arxiv.org/abs/2605.26559, iCML Workshop on Foundation Models for Structured Data (FMSD), arXiv:2605.26559

Wang Y, Tang Y, Liu B, et al (2026c) When LLM reward design fails: Diagnostic-driven refinement for sparse structured RL. https://doi.org/10.48550/arXiv.2605.28918, URL https://arxiv.org/abs/2605.28918, arXiv:2605.28918

Wang Y, Tang Y, Qian Y, et al (2026d) VisualLeakBench: Auditing the fragility of large vision-language models against PII leakage and social engineering. https://doi.org/10.48550/arXiv.2603.13385, URL https://arxiv.org/abs/2603.13385, arXiv:2603.13385

Wang Z, Sun Y, Wang H, et al (2025) Reasoning-enhanced domain-adaptive pretraining of multimodal large language models for short video content governance. In: Potdar S, Rojas-Barahona L, Montella S (eds) Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track. Association for Computational Linguistics, Suzhou (China), pp 1104–1112, https://doi.org/10.18653/v1/2025.emnlp-industry.77, URL https://aclanthology.org/2025.emnlp-industry.77/

Xiao C, Hou L (2026) Prototype-aligned federated soft-prompts for continual web personalization. In: Proceedings of the ACM Web Conference 2026, pp 6743–6754

Xiao C, Dou J, Lin Z, et al (2026) From points to coalitions: Hierarchical contrastive Shapley values for prioritizing data samples. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 15995–16003

Xiao Y, Du H, Zhang Y, et al (2025) AdsBPC: Click without compromise—online ad measurement via per-user differential privacy. In: IEEE Symposium on Security and Privacy, arXiv preprint arXiv:2406.02463

Xie T, Chen J, Yang Y, et al (2024) LossLens: Diagnostics for machine learning through loss landscape visual analytics. IEEE Computer Graphics and Applications 45(2):112–125. https://doi.org/10.1109/MCG.2024.3520428

Xu N, Jiang Y, Dipta SR, et al (2026a) Learning how to use tools, not just when: Pattern-aware tool-integrated reasoning. URL https://arxiv.org/abs/2509.23292, arXiv:2509.23292

Xu R, Jiang L, Yu D, et al (2026b) SynHAT: A two-stage coarse-to-fine diffusion framework for synthesizing human activity traces. URL https://arxiv.org/abs/2604.14705, arXiv:2604.14705

Yang D, Dyer K, Wang S (2020) DeepMTA: Interpretable deep learning model for online multi-touch attribution. ArXiv preprint arXiv:2004.00384

Ye H, Chen S, Zhong Z, et al (2026) Seeing through the conflict: Transparent knowledge conflict handling in retrieval-augmented generation. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 34423–34431

Ye Y, Xue P (2026) When attribution fails: Selection, targeting, and measurement in real-time bidding advertising. SSRN working paper, March 30, 2026; https://ssrn.com/abstract=6530478

Yu D, Zhuang D, Jiang L, et al (2025) UQGNN: Uncertainty quantification of graph neural networks for multivariate spatiotemporal prediction. In: Proceedings of the 33rd ACM International Conference on Advances in Geographic Information Systems, pp 52–65

Zhan Y, Ben-Harush S, Zhu Y, et al (2024) Neural ODE for multi-channel attribution. In: ICLR 2024 Workshop

Zhang S, Lin S, Wei X, et al (2026a) Small language model agents enable efficient and high-quality knowledge mining. URL https://arxiv.org/abs/2510.01427, arXiv:2510.01427

Zhang Y, Wurm M, Wakim A, et al (2023) Bayesian hierarchical media mix model incorporating reach and frequency data. Google Research

Zhang Y, Wurm M, Wakim A, et al (2024) Media mix model calibration with Bayesian priors. Google Research

Zhang Y, Xiang Z, Xu H (2026b) Performance-efficiency trade-offs in human preference prediction: A comparative study of traditional machine learning and large language models. In: Proceedings of the 31st IEEE Symposium on Computers and Communications (ISCC), IEEE, accepted; proceedings forthcoming

Zhao K, et al (2018) Shapley value methods for attribution modeling in online advertising. ArXiv preprint arXiv:1804.05327

Zhao K, et al (2024) M^3TN: Multi-gate mixture-of-experts for multi-treatment uplift modeling. In: Proceedings of ICASSP 2024, arXiv preprint arXiv:2401.14426

Zhao S, Xue Z, Qi Y, et al (2026) Non-intrusive graph-based bot detection for e-commerce using inductive graph neural networks. URL https://arxiv.org/abs/2601.22579, arXiv:2601.22579