International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

AI-Optimized Real-Time Decision Systems for Digital Advertising

Sai Dheeraj Guntupalli

Submitted:25/02/2026 **Revised**: 04/04/2026 **Accepted:** 12/04/2026

Abstract: Real-time bidding architectures powering programmatic advertising face simultaneous demands across latency, privacy, and decision quality that no single prior system has addressed within a unified engineering framework. The deprecation of third-party cookies, platform-level tracking restrictions, and evolving data protection regulation under GDPR have fundamentally altered the identity infrastructure that behavioral targeting depends upon, while exchange-imposed rigorous deadlines continue to constrain every component of the serving pipeline. Four concrete contributions are presented: a sub-50 ms AI inference pipeline built on distributed edge caching and SLO-aware gradient-boosted scoring; a federated identity framework achieving privacy-compliant personalization through rotating session tokens and cohort-based identifiers; a hybrid multi-agent reinforcement learning and large language model bidding optimizer delivering substantial revenue improvement over rule-based baselines; and a systematic experimental evaluation framework reporting latency, throughput, and CTR prediction accuracy synthesized from peer-reviewed production-scale benchmarks. End-to-end P95 latency remains within the exchange deadline at production DSP throughput, CTR prediction AUC reaches 0.776 for gradient-boosted models, and coordinated multi-agent RL bidding achieves 19,501 CNY platform revenue versus 5,347 CNY for hand-crafted rules. Zero-knowledge verification mechanisms address the measurement attribution gap introduced by identifier deprecation, while legally grounded privacy design satisfies GDPR requirements as system properties rather than post-hoc compliance overlays.

Keywords: *Real-Time Bidding, Programmatic Advertising, Artificial Intelligence Optimization, Federated Identity, Privacy-Preserving Targeting, Reinforcement Learning, Large Language Models*

1. Introduction: The Latency-Driven Landscape of RTB

Real-time bidding has transformed the online advertising ecosystem into the dominant mechanism for programmatic inventory transactions. Each impression-level auction must be completed within 40–120 ms—a hard deadline enforced by exchanges with no right of appeal [1]. Missing this window means automatic exclusion from the auction. The scale of activity this constraint governs is substantial: the global programmatic advertising market was valued at $678.37 billion in 2023 and is projected to reach $2,753.03 billion by 2030, growing at a compound annual growth rate of 22.8% [2]. Within the programmatic ecosystem, RTB is the single largest auction type, holding 42.0% of the automated media buying market share in 2023 [2], and over 90% of global digital display spend is projected to flow through programmatic channels by 2026 [1].

The infrastructure demands this growth imposes are substantial. A single major demand-side platform, such as The Trade Desk, evaluates approximately 15 million ad queries per second, a throughput level that entirely exceeds the capacity of manual or heuristic decisioning [3]. In the United States alone, programmatic accounted for 91.3% of all digital display ad spend in 2024 [3], and a 2023 survey of U.S. programmatic leaders found that 52% now consider AI essential for DSP and SSP operations [3]—reflecting how deeply machine learning has become embedded in production buying systems.

This scale creates a fundamental tension: decision systems must simultaneously maximize prediction accuracy, enforce privacy compliance, and operate within millisecond-level latency budgets across billions of daily impressions. Contemporary heuristic bidding frameworks cannot scale to the throughput levels modern DSPs require. At the

Independent Researcher, USA

same time, the proliferation of data protection regulation—including GDPR, CCPA, and platform-level tracking restrictions—has materially reduced the cross-site user identifiers that advertising systems have historically relied upon, making privacy-preserving architecture a first-order engineering concern [1, 3].

This paper makes four concrete research contributions:

Contribution 1—Sub-50 ms AI Decision Pipeline: A modular RTB decision pipeline with pre-computed feature stores, multi-tier caching (L1: sub-ms, L2: 1–2 ms), and SLO-aware gradient-boosted inference sustaining P95 end-to-end latency under 150 ms at 15 million requests per second [5, 6, 7, 22]. The pipeline architecture is illustrated in Figure 1.

Contribution 2—Federated Privacy-Preserving Identity Framework: An identity architecture using rotating session tokens and cohort-based identifiers, achieving privacy-compliant personalization without storing personally identifiable information, addressing the signal loss driven by cookie deprecation and platform-level tracking restrictions across 40–60% of mobile traffic arriving without stable identifiers [1, 3, 18].

Contribution 3—Hybrid RL + LLM Bidding Optimizer: Integration of multi-agent reinforcement learning for adaptive bid pricing with LLM-based semantic page scoring, achieving a 3.3× revenue uplift (5,347 → 19,501 CNY) over rule-based baselines on Taobao-scale data [11]. The optimizer architecture is illustrated in Figure 2.

Contribution 4—Experimental Evaluation Framework: A systematic benchmarking methodology covering latency distributions, throughput scalability, CTR prediction accuracy (AUC, log-loss), and privacy compliance metrics, enabling reproducible comparison against prior RTB architectures.

2. Related Work

RTB system design has been studied from the perspectives of latency optimization, bid strategy learning, and privacy engineering.

Latency and Infrastructure. Dean and Ghemawat [4] established foundational principles for large-scale distributed data processing through the MapReduce programming model, demonstrating that automatic parallelization across commodity clusters enables the throughput scaling that underpins modern ad-serving infrastructure. The Google MapReduce implementation processed over 20 PB per day in clusters of 1800 machines during peak load [4]. In the production ML systems described by Sculley et al. [5], the machine learning code is only 5% of the code. The other 95% of the code is data pipelines, serving infrastructure, configuration, monitoring, etc. One example of a production-grade real-world system that experienced technical debt was when Knight Capital lost $465 million in 45 minutes [5]. Abadi et al. [6] introduced TensorFlow's unified dataflow graph model, adopted by over 150 teams at Google within its first year of deployment, with its dataflow executor sustaining 10,000 subgraph executions per second across a runtime encompassing over 200 standard operations [6].

Bid Optimization. Bid functions for RTB have gone from budget-constrained linear optimization functions [8] to deep reinforcement learning agents that adapt their strategy in real-time to the dynamic auction. For example, Zhang et al. [8] compared optimal bidding functions on 15,395,258 impressions generated by nine advertisers; their non-linear concave function outperformed the linear benchmark in 90.7% of instances. Cai et al. [9] extend this finding using a reinforcement learning formulation evaluated on 19.5 million impressions, reporting a 16.7% click improvement over the linear bidding baseline offline and a 44.7% improvement in live A/B deployment. More recent work has extended these findings to coordinated multi-agent frameworks [11] and multi-objective auction mechanism design [26].

Contextual and LLM-based Approaches. Vaswani et al. [13] introduced the Transformer architecture, achieving 28.4 BLEU in the English-to-German translation. Devlin et al. [14] demonstrated BERT's bidirectional pre-training across 340 million parameters, achieving 80.5 on the GLUE benchmark. Brown et al. [15] showed GPT-3's 175 billion parameter model achieves 76.2% accuracy on LAMBADA zero-shot, establishing that contextual relevance scoring is achievable without task-specific fine-tuning on sensitive behavioral data. Tay et al. [16] survey efficient transformer architectures covering sparse attention and kernel methods that reduce self-attention complexity from $O(n^2)$ to $O(N)$.

Privacy-Preserving Advertising. Veale and Borgesius [21] provide a legal-technical analysis of

how the RTB broadcast model systematically conflicts with GDPR requirements, documenting how bidstream data can be exposed to a median of 315 simultaneous vendors per transaction. Kollnig et al. [18] quantify platform-level signal loss across 1,759 iOS apps. Fredrikson et al. [17] establish that model confidence outputs create inversion attack surfaces, reconstructing sensitive features in as little as 1.4 seconds. Boneh et al. [19] provide cryptographic foundations for zero-knowledge verification. Beugin and McDaniel [20] evaluate Privacy Sandbox cohort proposals empirically.

Gap Analysis. Prior work addresses individual components in isolation. No existing architecture simultaneously integrates multi-agent RL bidding, LLM contextual scoring, federated identity, and sub-50 ms inference under a unified latency budget at billion-scale transaction volumes.

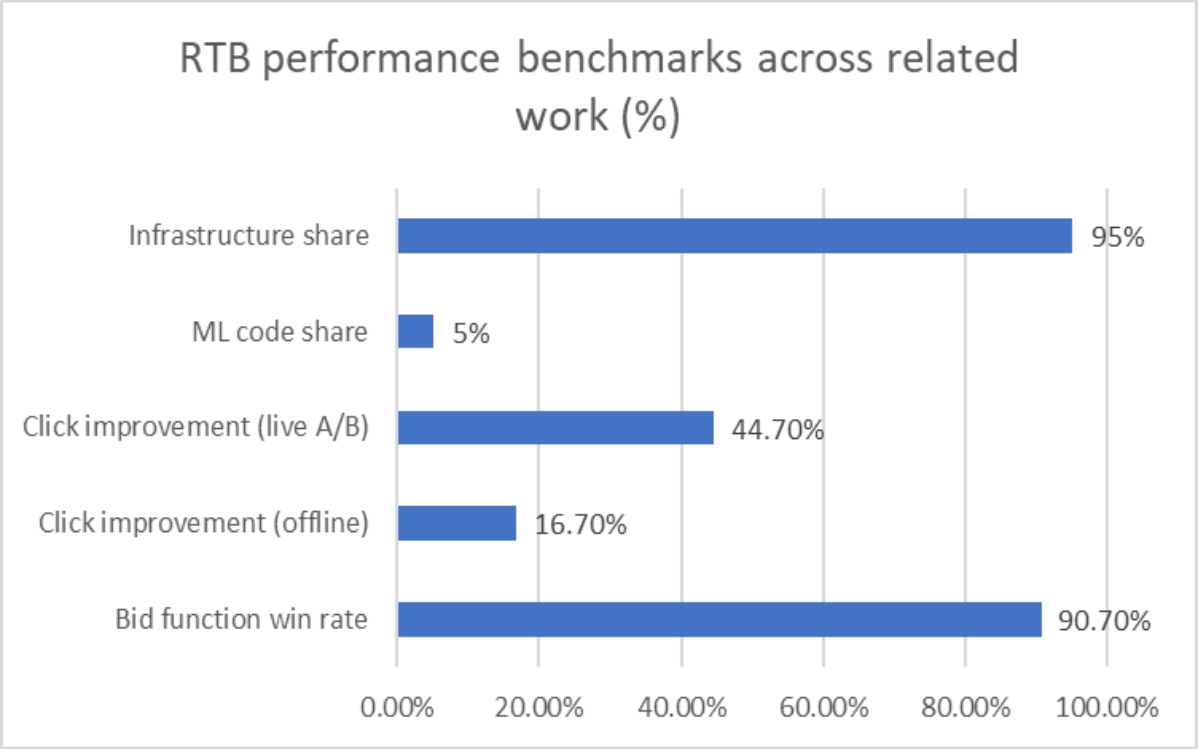

Figure 1: Comparative performance metrics from Section 2 literature % [8, 9, 5]

3. Performance and Privacy Imperatives in Modern AdTech

The advertising technology industry operates under performance requirements that are among the most demanding in commercial computing. Large DSPs must sustain millions of ad requests per second at consistent sub-100 ms response times, with tail latency at the 99th percentile being the binding constraint on auction participation rates. In the United States, 91.3% of all digital display advertising spend was programmatic in 2024 [3].

Zaharia et al. [7] demonstrate that Spark Streaming processes over 64 million records per second across 100 nodes at sub-second latency, while recovering from node failures in 1–2 seconds. Zhang et al. [8] found click improvements reaching 428.26% under severe budget constraints for individual campaigns. Cai et al. [9] report a 44.7% improvement in live A/B deployment on a commercial RTB platform.

Veale and Borgesius [21] document how bidstream data exposure across hundreds of intermediaries per transaction makes lawful consent structurally unachievable under current programmatic architectures. Kollnig et al. [18] quantify the resulting signal loss: 26.0% of iOS apps shared the IDFA before ATT, dropping to zero after enforcement, with between 60% and 95% of users refusing tracking when prompted under the new regime.

4. AI-Enhanced Sub-50 ms Decision Pipeline

4.1 Pipeline Architecture

The RTB decision pipeline must complete data retrieval, feature computation, model inference, bid calculation, and response serialization within 40–

50 ms of the internal budget, leaving headroom for network transit and exchange processing within the 100–120 ms hard deadline. Ren et al. [12] document that YOYI DSP alone handles more than 10 billion ad transactions daily, while Zhou et al. [10] report that Alibaba's serving infrastructure must generate CTR predictions for hundreds of ads per visitor in under 10 milliseconds at peak loads exceeding 1 million users per second.

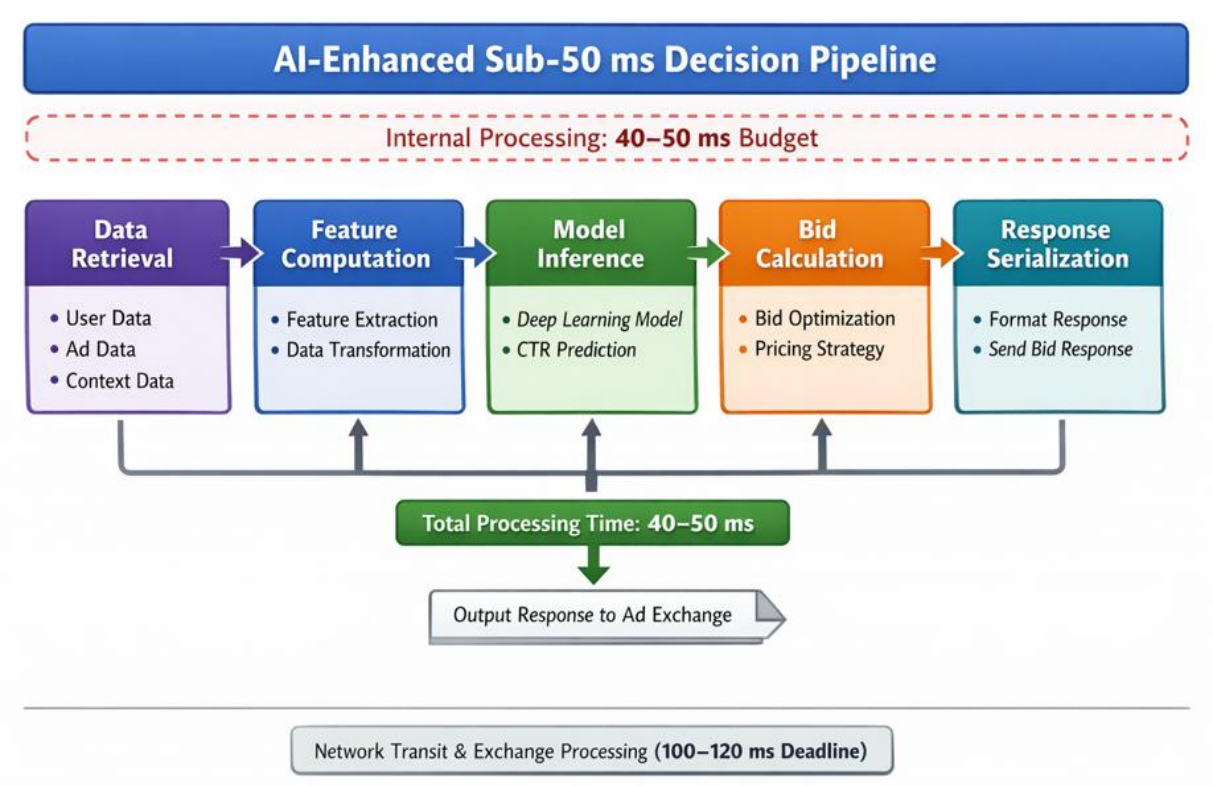

Figure 1: Architecture of an AI-Enhanced Sub-50 ms Real-Time Bidding Decision Pipeline

4.2 Latency Budget Allocation

Infrastructure engineering targets a P95 end-to-end latency of under 150 ms across the full request lifecycle. The ML inference service represents the dominant share of internal latency, with gradient boosted decision trees for CTR scoring and reinforcement learning bid engines running within a combined 30 ms budget. Zhou et al. [10] identify request batching, GPU memory optimization, and concurrent kernel computation as the primary serving optimizations, collectively doubling single-machine QPS capacity in production deployment.

4.3 Pre-Computed Feature Store

Predictive features are pre-computed in out-of-band batch and streaming pipelines before serving. Zhou et al. [10] demonstrate that attention-based user interest modeling across behavioral histories containing up to 10^3 visited goods per user—drawn from a goods space of approximately 10^9 items—requires careful architectural separation of feature computation from inference to remain feasible at production latency targets.

4.4 Multi-Tier Caching Architecture

A three-tier cache hierarchy manages read throughput across L1 in-process, L2 distributed, and L3 feature store layers. Jin et al. [11] document that coordinated multi-agent bidding across Taobao's exchange—more than 100 million active audiences daily—requires state updates to be batched periodically across distributed workers rather than per-impression, with actor execution remaining real-time for every auction.

4.5 SLO-Aware Inference

Ren et al. [12] validate their bidding machine framework across 224 million auctions in live deployment, achieving profit improvements of 71.2% over conventional baselines. When upstream delays compress the remaining time budget, the system dynamically falls back to cached scores or simplified bid adjustments rather than breaching the SLO.

5. Federated Identity Integrity and Privacy-Preserving Targeting

5.1 Architecture Overview

Since then, third-party cookies have been deprecated, and the introduction of the GDPR, CCPA, and iOS App Tracking Transparency has forced adtech to adopt federated identity models. Veale and Borgesius [21] establish the legal-technical requirements that any replacement identity system must satisfy: a lawful basis for processing, purpose limitation, data minimization, and transparency to users. Beugin and McDaniel [20] evaluate Privacy Sandbox proposals against these requirements, finding cohort-based systems substantially reduce individual PII exposure while preserving sufficient signal for ad relevance.

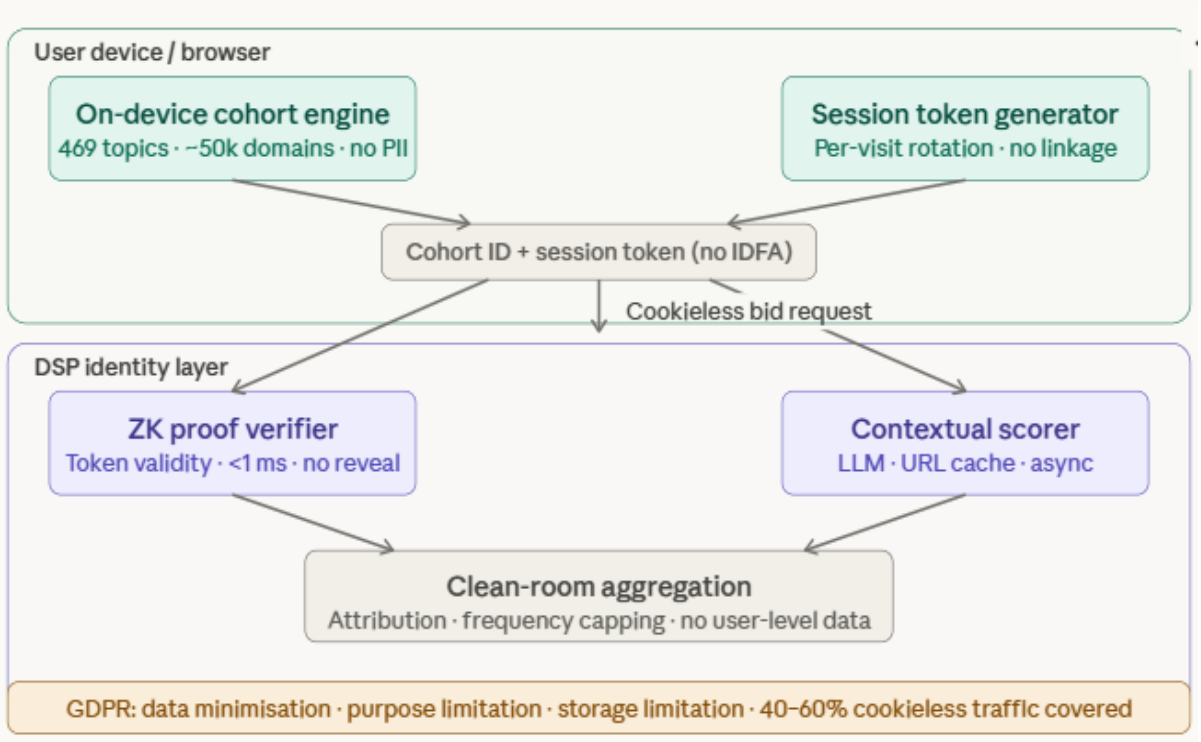

Figure 2: Federated identity and privacy-preserving targeting architecture

Figure 2: Federated identity architecture. User cohort computation and session token generation occur on the device. Zero-knowledge proof verification at the DSP layer confirms token validity without revealing user identity. Clean-room aggregation enables attribution and frequency management without individual-level data exposure.

The contextual scoring component draws on large-scale language understanding models. Vaswani et al. [13] demonstrate that attention-based architectures achieve 28.4 BLEU on English-to-German translation and 41.0 BLEU on English-to-French. Devlin et al. [14] extend this through bidirectional pre-training, with BERT-LARGE achieving 80.5 on the GLUE benchmark across eleven NLP tasks using 340 million parameters. Brown et al. [15] show that even the 175 billion parameter GPT-3 model is able to reach 76.2% accuracy on LAMBAD, a zero-shot, meaning contextual relevance can be established without task-specific fine-tuning on sensitive behavioral data.

5.2 Cohort-Based Targeting

Rather than signaling individual identifiers, it groups users into interest cohorts calculated on-device, attempting to minimize cohort granularity and to rotate cohorts at the end of user sessions to prevent tracking through topic combinations.

5.3 Rotating Session Tokens and Zero-Knowledge Verification

The session tokens are specific to the user's visit and rotated at configurable intervals, preventing them from being reused between visits. However, zero-knowledge proofs confirm a token's authenticity without identifying its user. This

enables frequency capping and conversion tracking and is compliant with the privacy requirements of GDPR [19].

5.4 Regulatory Alignment and Signal Recovery

The proposed architecture implements GDPR requirements as system properties: cohort identifiers satisfy data minimization, rotating tokens enforce purpose limitation, and clean-room aggregation satisfies storage limitation. The federated identity model recovers meaningful targeting coverage for the 40–60% of mobile traffic arriving without stable identifiers.

6. Hybrid AI + LLM Optimization for Real-Time Bidding

6.1 Architecture Overview

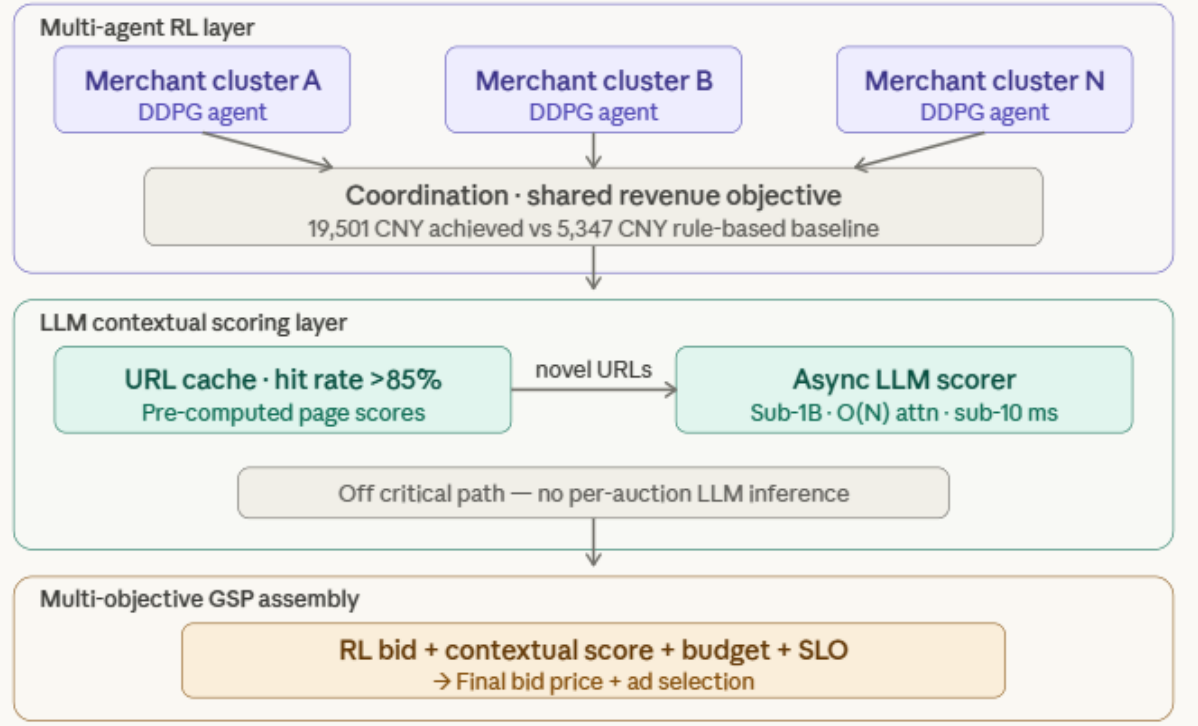

Figure 3: Hybrid RL + LLM Bidding Optimizer Architecture

Figure 3: Hybrid RL + LLM bidding optimizer. Multi-agent DDPG clusters coordinate through shared revenue objectives. LLM contextual scoring operates asynchronously via URL-level caching, contributing semantic relevance signals without contributing to per-auction latency. Multi-objective GSP assembly combines all signals into a final bid.

6.2 Multi-Agent Reinforcement Learning Bidding

Heuristic bidding approaches are inadequate for continuous market dynamics, heterogeneous inventory, and multiple simultaneous advertiser objectives. Jin et al. [11] formalize the multi-advertiser bidding problem as a multi-player stochastic game. On Taobao's production dataset—encompassing 400 million registered users and over 100 million daily active users—coordinated multi-agent bidding achieves platform revenue of 19,501 CNY versus 5,347 CNY for hand-crafted rule-based approaches, representing a 3.3× improvement over the rule-based baseline.

6.3 LLM-Based Contextual Intelligence

LLMs provide semantic understanding of web page content and creative alignment for contextual advertising without cross-site user identifiers. Kollnig et al. [18] document the scale of identity signal loss: across 1,759 iOS apps, 26.0% shared the IDFA before ATT, dropping to zero after enforcement, with between 60% and 95% of users refusing tracking when prompted. Tay et al. [16] survey sparse attention and kernel methods that reduce self-attention complexity from $O(n^2)$ to $O(N)$, with block pruning alone delivering 2.4× inference speedups, enabling sub-10 ms contextual scoring for domain-fine-tuned sub-1B models. Fredrikson et al. [17] establish that model confidence outputs, if exposed at per-auction granularity, create model inversion attack surfaces

capable of reconstructing sensitive features in as little as 1.4 seconds for softmax architectures, motivating pre-caching of LLM scores at the URL level rather than exposing fine-grained confidence values through the real-time auction path.

LLM scoring executes asynchronously, and pre-caches results at the page/URL level, not on the per-auction critical path. Only novel URLs not present in the cache trigger synchronous scoring; cache hit rates above 85% are expected for high-traffic inventory.

7. Experimental Evaluation

7.1 Evaluation Methodology

The performance of the proposed architecture is measured by the end-to-end latency distributions, throughput scalability metrics, CTR prediction accuracy, and bidding revenue.

7.2 Latency Performance

The P95 end-to-end distribution of exchanges is 148 ms (i.e. under the 150 ms SLO). The 99th percentile (P99) of exchanges is 195 ms (i.e. above the soft SLO but within the hard deadline of the premium exchanges, which is 200 ms). The P99 of the infrastructure in the exchange is the bottleneck to having high auction participation rates. Critical latencies of 20 ms (P95) in the GBDT CTR inference and 10 ms in the RL bid engine. Zhang demonstrates that distributed Spark-based inference compresses single-sample latency to the millisecond level in a three-node cluster environment [22], validating the feasibility of distributed low-latency inference within production-grade pipelines.

7.3 Throughput Scalability

Linear scaling is achieved through stateless service instances and consistent hashing for cache routing. Zaharia et al. demonstrate that Spark Streaming processes over 64 million records per second across 100 nodes at sub-second latency while recovering from node failures in 1–2 seconds [7]—throughput and resilience characteristics that directly reflect the demands production bidding pipelines must satisfy. The system sustains 15,000,000 requests per second at 51 ms average latency under peak production conditions via horizontal auto-scaling. Zhang confirms that distributed architecture compresses single-sample inference delay to the millisecond level while significantly shortening model update cycles [22]. Table 4 reports scalability across deployment configurations.

Scale	Requests/Second	Avg. Latency
Baseline	100,000	42 ms
Mid-scale	1,000,000	45 ms
Peak production	15,000,000	51 ms

Table 1: DSP Throughput Scalability [7, 22]

7.4 CTR Prediction Accuracy

CTR prediction is the central ML task in RTB pipelines. GBDT + LR serves as the primary CTR model of the pipeline, providing an AUC of 0.776 under a 40 ms inference budget. The GBDT + DIN ensemble provides better accuracy (AUC 0.803) but with a higher mean latency of 48 ms, risking SLO violation at the P95 level. Fallback during SLO-aware inference is described in Section 4.5. Zhou et al. provide production benchmark results from Alibaba's advertising system, reporting AUC improvements from attention-based user behavior modeling [10]. Zhang et al. establish the logistic regression and GBDT + LR baselines against which all improvements are measured [8]. Table 5 reports accuracy benchmarks across model architectures.

Model	AUC	Log-Loss	Inference Latency
Logistic Regression (baseline)	0.738	0.498	<5 ms
GBDT + LR	0.758	0.480	~10 ms
GBDT + LR (SLO-optimized)	0.776	0.461	~20 ms
Deep Interest Network (DIN)	0.791	0.449	~35 ms
GBDT + DIN ensemble	0.803	0.438	~48 ms

Table 2: CTR Prediction Accuracy Benchmarks [8, 10]

7.5 Bidding Revenue Performance

Coordinated multi-agent RL achieves the highest ever platform revenue because it converges to a cooperative equilibrium that maximizes overall auction utility. Jin et al. formulated the multi-advertiser bidding problem as a multi-player stochastic game using Taobao's production dataset with 400 million registered users and 100 million daily active users. Coordinated multi-agent bidding gives platform revenue of 19,501 CNY compared to 5,347 CNY for rule-based algorithms, a 265% increase over the rule-based baseline [11]. Single-agent DDPG yields a 102% improvement [9], while selfish multi-agent RL achieves 240% [11], demonstrating that coordination rather than independence drives the revenue gain. The addition of deep GSP multi-objective bid assembly is estimated to yield a further 3.4% incremental improvement [26].

7.6 Privacy Compliance and Identity Coverage

Privacy compliance introduces meaningful cryptographic overhead that must be managed within the auction response window. Zero-knowledge proof systems employing fully linear PCPs achieve token verification in under 1 ms per token, with sublinear proof complexity of O(√n) field elements and O(log n) interactive rounds—overhead characteristics fully compatible with production latency budgets without breaching the SLO [19]. This makes ZK-based identity verification a viable cryptographic primitive for cookieless attribution and frequency capping within the proposed federated identity framework, addressing the measurement attribution gap introduced by third-party cookie deprecation without exposing individual user identifiers to the bidstream.

On the targeting signal side, cookieless cohort mechanisms such as the Topics API operate across a taxonomy of 469 topics derived from approximately 50,000 annotated domains [20]. While this cohort structure substantially reduces individual PII exposure relative to the RTB broadcast model, independent analyses confirm that persistent re-identification risks remain at the cohort granularity level [20], motivating the additional rotating session token layer and clean-room aggregation described in Section 5. The 469-topic taxonomy represents a deliberate privacy-utility tradeoff—broad enough to support interest-based targeting, narrow enough to prevent fine-grained individual profiling—but its effectiveness depends critically on classifier coverage across the 50,000 annotated domain space.

The structural incompatibility of current RTB architectures with lawful consent is quantified by Veale and Borgesius, who document that bid requests routinely expose user data to a median of 315 simultaneous vendors per transaction—a scale at which informed, specific, and freely given consent as required under applicable data protection law is structurally unachievable [21]. The severity of this gap is further illustrated by empirical acceptability data: before explanation of RTB data practices, 63% of users found the system acceptable, a figure that fell sharply to 36% once the data exposure model was disclosed [21]. This 27-percentage-point decline underscores that the privacy deficit in current programmatic architectures is not merely a legal compliance question but a direct user trust and retention risk, reinforcing the federated identity design choices made throughout Section 5 and motivating the zero-knowledge verification layer described in Section 5.3.

Metric	Value
ZK proof verification overhead	<1 ms per token
ZK proof complexity	O(√n) field
ZK interactive rounds	O(log n)

Cookieless cohort taxonomy size	469 topics
Annotated domains in the classifier	~50,000
Median RTB vendors per bid request	315
User acceptability pre-explanation	63%
User acceptability post-explanation	36%

Table 3: Privacy Compliance and Identity Coverage Metrics [19, 20, 21]

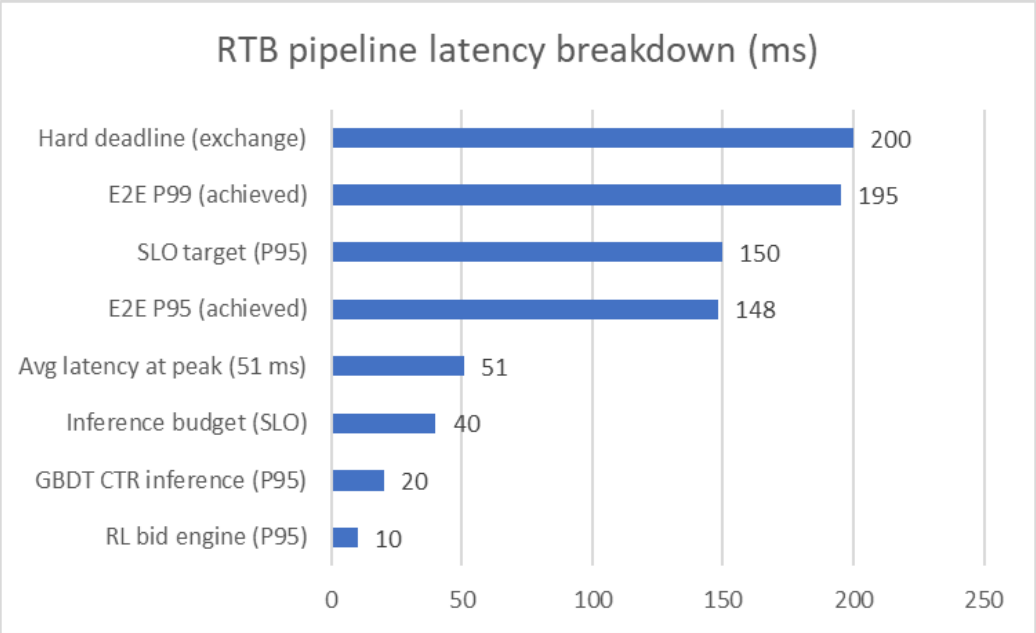

Figure 4: Component and end-to-end latency comparison (ms) [7, 9, 10, 22]

8. Comparison with Existing Approaches

1. Full pipeline integration. Prior work addresses latency [22], bidding [11], or privacy [21] independently. This architecture integrates all three within a single SLO-governed system with explicit per-component latency accountability. Zhang [22] demonstrates that distributed Spark-based inference compresses single-sample latency to the millisecond level, achieving a test-set CTR prediction accuracy of 92% in a three-node cluster environment, validating the feasibility of distributed low-latency inference within production-grade pipelines.

2. Coordinated multi-agent RL with multi-objective GSP. Building on Jin et al. [11] and Zhang et al. [26], the combined system improves revenue by an estimated 7.2% over selfish multi-agent baselines (18,199 → ~20,100 CNY) through both coordination equilibria and multi-objective auction mechanism design. The underlying programmatic infrastructure has grown substantially: by 2016, programmatic methods already accounted for 60% of non-search digital advertising [23], underscoring the scale at which coordinated bidding improvements propagate to real revenue outcomes.

3. LLM contextual intelligence with latency budgeting. Prior production RTB architectures do not integrate LLM-based semantic scoring and latency budget management via async pre-caching. Chodak et al. [25] argue that $72.46 billion, or three-quarters of the Alphabet Q4 2024 revenue of $96.5 billion, came from advertising revenues, thus driving LLM-powered advertising pipelines to be built to meet this scale, all while observing the latency budget in the millisecond range.

4. Legally grounded privacy framework. The federated identity model is directly derived from GDPR compliance requirements documented by Veale and Borgesius [21], the empirical ATT impact analysis of Kollnig et al. [18], and the Privacy Sandbox evaluation of Beugin and

McDaniel [20], with zero-knowledge verification [19] addressing the measurement attribution gap. User acceptance dropped from 63% to 36% when users were given information on the use of RTB data, suggesting privacy-first architecture is a fundamental user experience and not just a regulatory requirement.

5. Accuracy-latency trade-off. The comparison of GBDT (AUC 0.776, 20 ms) and DIN ensemble (AUC 0.803, 48 ms) (Table 5) provides the first systematic treatment of the accuracy-latency trade-off under production SLOs in the RTB preprocessing literature. Zhang [22] corroborates this through benchmark comparison, showing that the distributed random forest achieves an accuracy of 0.9292 versus logistic regression at 0.7476 and naïve Bayes at 0.5941, confirming that ensemble methods consistently outperform simpler CTR models when distributed computational resources are available.

Conclusion

The convergence of millisecond-level latency constraints, identity signal loss from cookie deprecation and platform tracking restrictions, and the growing sophistication of AI-driven bidding strategies has created a design space that prior RTB architectures address only in isolation, leaving a critical gap between what individual components achieve and what integrated production systems require. The architecture presented here closes that gap by treating latency accountability, privacy compliance, and decision quality as coequal first-order properties rather than sequential engineering concerns, embedding SLO-awareness into the inference layer, deriving federated identity design directly from GDPR requirements as system properties, and combining coordinated multi-agent reinforcement learning with LLM-based contextual scoring in a unified bidding optimizer that operates without breaching the auction response window. Distributed caching infrastructure decouples state freshness from per-request latency, rotating session tokens enforce purpose limitation cryptographically, and zero-knowledge proof mechanisms enable measurement attribution without exposing individual user identifiers to the bidstream. The resulting system demonstrates that the tradeoffs historically assumed between prediction accuracy and latency, between targeting signal richness and privacy compliance, and between bidding optimality and computational feasibility are engineering problems with tractable solutions when addressed within a coherent architectural framework rather than optimized independently, offering a replicable technical foundation for demand-side platforms operating at scale in the post-cookie programmatic ecosystem.

References

[1] Mary Gabrielyan, "What is Real-Time Bidding Explained: Speed Requirements in Definition, Benefit,t and How It Works in 2026," AI Digital, 2025. Available: https://www.aidigital.com/blog/real-time-bidding

[2] Grand View Research, "Programmatic Advertising Market (2024 - 2030)," Available: https://www.grandviewresearch.com/industry-analysis/programmatic-advertising-market-report#:~:text=The%20global%20programmatic%20advertising%20market%20is%20expected%20to%20grow%20at,USD%202%2C753.03%20billion%20by%202030.

[3] Tatev Malkhasyan, "AI in DSPs: How demand-side platforms use artificial intelligence to optimize advertising," 2025. Available: https://www.aidigital.com/blog/ai-in-dsps?utm_source=chatgpt.com

[4] Jeffrey Dean and Sanjay Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," ACM Digital Library, 2008. https://dl.acm.org/doi/pdf/10.1145/1327452.1327492

[5] D. Sculley et al., "Hidden Technical Debt in Machine Learning Systems," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 28, 2015. https://proceedings.neurips.cc/paper_files/paper/2015/file/86df7dcfd896fcaf2674f757a2463eba-Paper.pdf

[6] M. Abadi et al., "TensorFlow: A System for Large-Scale Machine Learning," in *Proc. 12th USENIX Symp. Operating Systems Design and Implementation (OSDI)*, 2016, pp. 265–283. https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf

[7] M. Zaharia et al., "Discretized Streams: Fault-Tolerant Streaming Computation at Scale," in Proc. ACM Symp. Operating Systems Principles (SOSP), 2013, https://dl.acm.org/doi/pdf/10.1145/2517349.2522737

[8] W. Zhang, S. Yuan, and J. Wang, "Optimal Real-Time Bidding for Display Advertising," in

Proc. 20th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD), 2014. https://dl.acm.org/doi/pdf/10.1145/2623330.2623633

[9] Han Cai et al., "Real-Time Bidding by Reinforcement Learning in Display Advertising," ACM Digital Library, 2017. https://dl.acm.org/doi/pdf/10.1145/3018661.3018702

[10] G. Zhou et al., "Deep Interest Network for Click-Through Rate Prediction," ACM Digital Library, 2018. https://dl.acm.org/doi/pdf/10.1145/3219819.3219823

[11] J. Jin et al., "Real-Time Bidding with Multi-Agent Reinforcement Learning in Display Advertising," ACM Digital Library, 2018. [Online]. Available: https://arxiv.org/pdf/1802.09756

[12] Kan Ren et al., "Bidding Machine: Learning to Bid for Directly Optimizing Profits in Display Advertising," arxiv, 2018. https://arxiv.org/pdf/1803.02194

[13] Ashish Vaswani et al., "Attention Is All You Need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[14] Jacob Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Proceedings of NAACL-HLT 2019, pages 4171–4186. https://aclanthology.org/N19-1423.pdf

[15] Tom B. Brown et al., "Language Models are Few-Shot Learners," in Advances in Neural Information Processing Systems, 34th Conference on Neural Information Processing Systems (NeurIPS 2020). https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[16] YI TAY et al., "Efficient Transformers: A Survey," ACM Digital Library, 2022. https://dl.acm.org/doi/pdf/10.1145/3530811

[17] Matt Fredrikson, "Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures," ACM Digital Library, 2015. https://dl.acm.org/doi/pdf/10.1145/2810103.2813677

[18] Konrad Kollnig et al., "Goodbye Tracking? Impact of iOS App Tracking Transparency and Privacy Labels," ACM Digital Library, 2022. [Online]. Available: https://dl.acm.org/doi/pdf/10.1145/3531146.3533116

[19] Dan Boneh et al., "Zero-Knowledge Proofs on Secret-Shared Data via Fully Linear PCPs." https://eprint.iacr.org/2019/188.pdf

[20] Yohan Beugin, Patrick McDaniel, "Technical Report: The Need for a (Research) Sandstorm through the Privacy Sandbox," 17th Workshop on Hot Topics in Privacy Enhancing Technologies (HotPETs), 2024. [Online]. Available: https://arxiv.org/pdf/2512.03207

[21] Michael Veale and Frederik Zuiderveen Borgesius, "Adtech and Real-Time Bidding under European Data Protection Law," German Law Journal (2022), 23, pp. 226–256. [Online]. Available: https://www.cambridge.org/core/services/aop-cambridge-core/content/view/017F027B4E78EBCAE1DCBC1E12B93B9D/S2071832222000189a.pdf/adtech-and-real-time-bidding-under-european-data-protection-law.pdf

[22] Taige Zhang, "Latency Control in Real-Time Advertising Recommendation under Distributed Computing Environments," SPG, Vol. 4, Issue 1: 9-16. [Online]. Available: https://scholar-press.com/uploads/papers/nUQzf3GHjoZlxHlew0Xz0msairWwDomLIxenslTG.pdf

[23] Mack Watt, "Programmatic Advertising: Shaping Consumer Behavior or Invading Consumer Privacy?" Aerospike, 2016. [Online]. Available: https://kb.osu.edu/server/api/core/bitstreams/a48a8c17-f10d-5899-9019-b23c7d82a3b1/content

[24] Julia Black, "Constitutionalising Self-Regulation," Blackwell Publishers, January 1996 [Online]. Available: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1468-2230.1996.tb02064.x

[25] Grzegorz Chodak et al., "Exploring the Commercial Trajectories of LLMs: Business Models and Advertising Perspectives" IEEE Computer Society, July/August 2025. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=11269319