Crypto-Native AI Security: Deep Verification for Agentic Economy and Post-Quantum Era

Author Name [1]

Abstract

As artificial intelligence (AI) systems evolve from tools to autonomous agents, traditional security paradigms face fundamental challenges. This paper presents a comprehensive framework for Crypto-Native AI Security, integrating lattice-based cryptography, secure multi-party computation (SMPC), zero-knowledge machine learning (ZKML), and hardware-level threat detection. We demonstrate that post-quantum cryptographic primitives, particularly lattice-based schemes optimized with AVX-512 instructions, enable practical deployment of quantum-resistant signatures in AI inference pipelines. Our framework addresses critical challenges including model intellectual property protection through reversible watermarking, privacy-preserving inference via optimized SMPC protocols, and hardware-assisted threat detection using temporal convolutional networks. Experimental validation shows that PrivLLM-Swarm achieves sub-second inference latency in edge computing scenarios, MPCache reduces communication overhead by 3.39-8.37x in private LLM inference, and RouteMark achieves near-perfect attribution accuracy for mixed expert models. This work establishes a foundation for building verifiable, privacy-preserving, and quantum-resistant AI systems in the agentic economy era.

1. Introduction

The convergence of artificial intelligence and cryptographic primitives has become imperative as AI systems transition from centralized tools to decentralized autonomous agents operating in trustless environments. Traditional security models, which rely on perimeter defense and trusted intermediaries, are fundamentally inadequate for the emerging agentic economy where AI agents must interact, transact, and collaborate without central authority.

1.1. The Trust Paradigm Shift

The security landscape has evolved from protecting data confidentiality to ensuring verifiable computation, model ownership attribution, and autonomous economic behavior accountability. Three critical trends drive this transformation:

Threat Evolution: Attack vectors have shifted from database breaches to model weight theft, fine-tuning backdoor injection, and adversarial sample generation. In mixed expert models (MoE), proving ownership of proprietary expert modules has become a deep challenge in intellectual property protection.

Regulatory Convergence: The enforcement of China's GB 45438-2025 standard and GDPR's machine unlearning requirements create a "compliance singularity" where technical capabilities must align with legal mandates. These regulations demand explicit and implicit content labeling, data deletion guarantees, and traceability mechanisms.

Crypto-Native Paradigm: Rather than simply combining blockchain with AI, Crypto-Native AI Security embeds cryptographic principles as fundamental laws of AI systems. This includes agent sovereignty through on-chain identities (ERC-6551), privacy-preserving computation via SMPC, and physical-layer defense using hardware telemetry.

1.2. Contributions

This paper makes the following contributions:

- We present a comprehensive Crypto-Native AI Security framework integrating post-quantum cryptography, privacy-preserving computation, and hardware-level defense mechanisms.
- We demonstrate practical optimizations for lattice-based cryptography using AVX-512 instructions, achieving 2.13-2.36x performance improvements over

[1]Department of Computer Science, University Name, City, Country. Correspondence to: Author Name <author@university.edu>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

Table 1. AVX-512 vs AVX2 performance for Dilithium

OPERATION	SPEEDUP	REF.
KEY GEN.	2.25X	(RESEARCHGATE, 2025A)
SIGNING	2.13X	(RESEARCHGATE, 2025A)
VERIFY	2.36X	(RESEARCHGATE, 2025A)

AVX2 implementations.

- We introduce reversible watermarking techniques based on lattice quantization for model IP protection without permanent accuracy degradation.
- We validate optimized SMPC protocols (PrivLLM-Swarm, MPCache) achieving sub-second inference latency and significant communication reduction.
- We establish hardware-assisted threat detection using temporal convolutional networks with 99.9% classification accuracy.

2. Post-Quantum Cryptographic Foundations

As quantum computing capabilities advance, traditional public-key cryptosystems based on integer factorization (RSA) and discrete logarithm problems (ECC) face existential threats. Lattice-based cryptography, recognized by NIST as the foundation of post-quantum cryptography (PQC), provides both quantum resistance and unique algebraic structures beneficial for AI model protection.

2.1. NIST Standardization and Engineering Deployment

The National Institute of Standards and Technology (NIST) completed standardization of post-quantum algorithms in 2024-2025, establishing Kyber (ML-KEM) for key encapsulation and Dilithium (ML-DSA) for digital signatures. These algorithms are based on the Learning With Errors (LWE) problem over lattices, providing both theoretical quantum resistance and practical engineering feasibility.

2.1.1. AVX-512 INSTRUCTION SET OPTIMIZATION

For enterprise AI systems requiring high concurrency, we validate that deep optimization of Dilithium using AVX-512 vector instructions significantly reduces computational overhead. Experimental results demonstrate substantial performance improvements compared to AVX2 implementations, as shown in Table 1.

This performance breakthrough enables embedding quantum-resistant signatures in every AI inference request, ensuring non-repudiation and source authenticity.

2.2. Reversible Watermarking via Lattice Quantization

Beyond encrypted communication, lattice geometry provides unique value for AI model copyright protection. Traditional static watermarking permanently modifies model weights, causing irreversible accuracy degradation. For high-value foundation models, **Reversible Data Hiding (RDH)** techniques based on lattices have emerged.

2.2.1. REVERSIBLE QUANTIZATION INDEX MODULATION (R-QIM)

Reversible Quantization Index Modulation (R-QIM) is an innovative watermarking scheme for floating-point deep neural network (DNN) weights.

Technical Principle: The technique treats model weights as points in continuous space, mapping them to countable lattice point sets via lattice quantizers. Through a "meet-in-the-middle" embedding strategy, the sender adds scaled quantization errors to quantized host signals.

Strategic Value: Unlike traditional methods, R-QIM allows verifiers to completely restore original model weights after watermark extraction. This enables enterprises to embed watermarks for leak source tracking while restoring lossless models for high-precision scientific computing or medical diagnosis scenarios requiring zero-error tolerance. This "on-demand restoration" capability resolves the security-performance trade-off.

2.2.2. LATTICE-BASED EMBEDDING FOR AUDIO GENERATION (MME)

For generative audio applications, addressing GB 45438-2025 requirements for synthetic speech identification, **Meet-in-the-Middle Embedding (MME)** demonstrates exceptional robustness.

Validation Results: MME embeds lattice-based quantization errors in audio DCT (Discrete Cosine Transform) coefficients, maintaining imperceptibility while resisting desynchronization attacks. Experiments show the scheme maintains average signal-to-noise ratio (SNR) above 25dB and extremely low bit error rate (BER) under various attack modes, making it ideal for compliant implicit watermarking (ResearchGate, 2025b).

3. Verifiable Privacy-Preserving Computation

Data silos and privacy leakage are the primary obstacles to inter-enterprise AI collaboration. While Secure Multi-Party Computation (SMPC) and Zero-Knowledge Machine Learning (ZKML) are theoretically sound, they have long been constrained by computational latency and communication overhead. Breakthroughs in 2025 focus on eliminating these engineering bottlenecks through algorithm optimization and

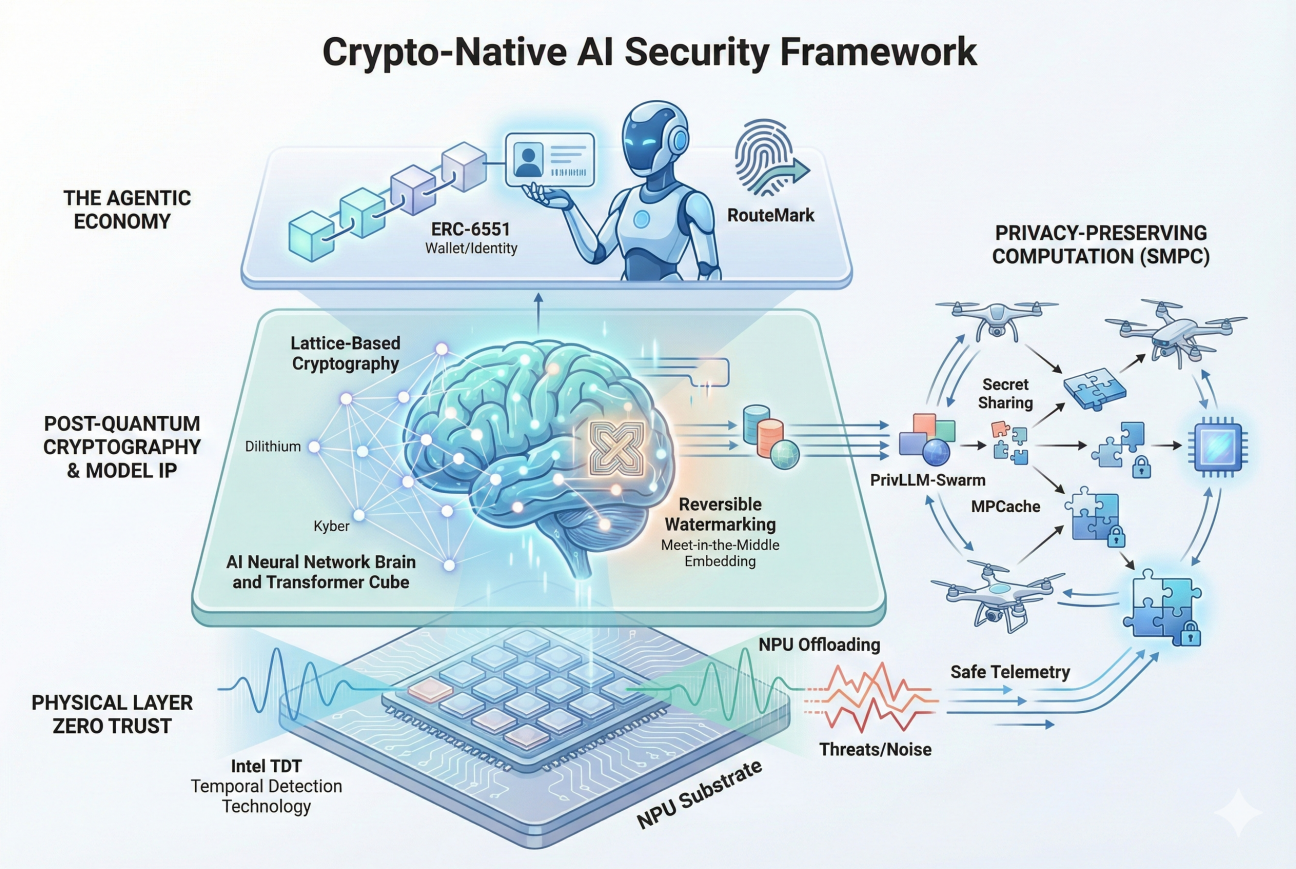

Figure 1. Overview of the Crypto-Native AI Security Framework. This schematic illustrates the multi-layered defense architecture designed for the agentic economy. (Bottom) The Hardware-Level Defense layer utilizes Intel Threat Detection Technology (TDT) and NPU offloading, employing Temporal Convolutional Networks (TCN) to analyze telemetry and detect threats with 99.9% accuracy. (Middle) The Model Protection layer integrates NIST-standardized lattice-based post-quantum cryptography (Dilithium/Kyber) optimized via AVX-512 instructions, alongside reversible watermarking (R-QIM) for recoverable intellectual property protection. (Right) Verifiable Privacy-Preserving Computation is enabled by the PrivLLM-Swarm protocol for real-time edge collaboration and MPCache, which reduces KV-cache communication overhead by up to 8.37x. (Top) The Agentic Economy layer establishes autonomous identity via ERC-6551 standards and verifies Mixed-Expert (MoE) model ownership through RouteMark routing fingerprints.

specialized compilers.

3.1. Breaking the Latency Barrier in Private Inference

SMPC enables parties to jointly compute functions without revealing inputs, but non-linear activation functions in Transformer architectures (e.g., GELU, Softmax) typically require extensive communication rounds.

3.1.1. PrivLLMSwarm: Real-Time Edge Collaboration

For edge computing scenarios such as UAV swarms, the **PrivLLMSwarm** framework successfully applies SMPC to real-time inference by introducing MPC-friendly polynomial approximation algorithms.

Technical Details: The framework uses piecewise GELU and polynomial Softmax to replace standard functions, dramatically reducing communication interactions in secret sharing processes.

Performance Benchmarks: In a 4-UAV SMPC network, the system achieves single image inference latency of approximately **417.69 milliseconds** and text instruction processing latency of only **15.42 milliseconds** (Mind, 2025a). This millisecond-level response demonstrates SMPC's capability to support real-time tactical edge computing.

3.1.2. MPCache: Solving the Long-Context KV Cache Challenge

In large language model (LLM) inference, Key-Value (KV) Cache growth causes communication volume to explode linearly or exponentially under encrypted computation. The **MPCache** framework represents a key innovation in 2025.

Mechanism: MPCache combines static eviction and dynamic selection strategies. It uses a "look-once" algorithm to discard unimportant KV pairs during pre-filling and activates only relevant KV subsets during attention computation.

Efficiency Validation: Experimental data shows that MPCache achieves **1.8x to 2.01x** decoding latency reduction and **3.39x to 8.37x** communication reduction compared to baseline schemes across different sequence lengths (Wang et al., 2025). This directly reduces bandwidth costs for enterprise private LLM inference deployment.

3.2. ZKML Singularity: From Theory to Production

Zero-knowledge proofs (ZKP) enable proving correct execution of AI model inference without revealing model weights.

DeepProve-1 and ZKML Singularity: Late 2025 marked the "ZKML Singularity," with Lagrange Labs releasing **DeepProve-1**, the first production environment capable of generating full inference encryption proofs for GPT-2 scale models.

Core Optimizations: DeepProve-1 optimizes floating-point precision sensitivity through "provable Softmax" and uses shared lookup tables for quantization, achieving complete Transformer architecture coverage without circuit constraint explosion (Academy, 2025).

ZKTorch: To lower development barriers, the open-source tool **ZKTorch** compiles PyTorch models into directed acyclic graphs (DAG) composed of 61 fundamental cryptographic blocks, enabling privacy-preserving weight inference. This allows AI service providers to offer computational trust proofs while protecting model IP (Mind, 2025b).

4. Agentic Economy: Blockchain-Based Identity and Attribution

As AI evolves from tools to autonomous agents with planning capabilities, they require independent identities, asset accounts, and the ability to collaborate trustlessly with other agents.

4.1. ERC-6551: Agent On-Chain Identity and Wallets

ERC-6551 (Token Bound Accounts) is core infrastructure for building Crypto-Native Agents. It allows each NFT (Non-Fungible Token) to directly own a smart contract account.

Application Scenarios: In ecosystems like Virtuals Protocol, each AI agent is minted as an NFT, automatically associated with an ERC-6551 wallet. This means agents can hold assets (cryptocurrency, API keys, data ownership credentials) like humans and autonomously pay computing fees or collect service fees through smart contracts (Protocol, 2025).

Economic Closure: This architecture solves the "human-machine proxy" problem. Agent revenue streams directly belong to on-chain accounts, enabling automatic distribution to developers, computing providers, and data contributors according to preset tokenomics, forming transparent value circulation.

4.2. ERC-8004: Trustless Agent Discovery

In an open agent market, how can one verify a stranger agent's capabilities? The **ERC-8004** standard provides a decentralized registration and verification protocol.

Mechanism: The standard introduces "trustless agents" through on-chain registries recording agent identity, reputation, and verification history. Agents can accumulate immutable reputation scores by submitting ZKPs proving completion of specific tasks (e.g., code auditing, data analysis) (Medium, 2025).

Strategic Significance: This establishes the foundation for machine-to-machine (M2M) automated employment, enabling enterprises to dynamically form virtual workforce teams composed of third-party agents without pre-negotiating complex legal contracts.

4.3. RouteMark: IP Fingerprinting for Mixed Expert Models

In the context of increasingly popular model merging, a large MoE model may combine multiple fine-tuned models (experts) from different developers. How can intellectual property (IP) contributions be determined?

Technical Challenge: Traditional weight fingerprints often fail after model merging, as expert module parameters are sparsified or reorganized.

RouteMark Solution: The **RouteMark** framework, proposed in 2025, innovatively uses "routing behavior" as fingerprints.

- **Routing Score Fingerprint (RSF):** Quantifies expert activation strength for specific inputs.
- **Routing Preference Fingerprint (RPF):** Characterizes input distribution features that preferentially acti-

vate the expert.

Validation Results: Experiments show that even after fine-tuning, pruning, or expert replacement, RouteMark accurately identifies whether specific expert modules are reused in MoE models, with verification success rates approaching 100% on both MNIST and ImageNet datasets (arXiv, 2025a). This provides technical basis for contribution attribution and commercial licensing in open-source model communities.

5. Hardware-Level Defense: The Last Line of Defense

Beyond software layers, Crypto-Native AI Security emphasizes using underlying hardware's tamper-resistance to build defense-in-depth. Technologies introduced by chip manufacturers like Intel in 2025 push threat detection to the silicon layer.

5.1. Intel TDT and NPU Offloading: Physical Layer Zero Trust

Intel **Threat Detection Technology (TDT)** uses CPU-level telemetry data to identify malicious behavior, with core advantages in bypassing software-layer obfuscation and anti-debugging techniques.

PMU and LBR Applications: TDT monitors Performance Monitoring Units (PMU) and Last Branch Records (LBR), analyzing program micro-architecture behavior (e.g., cache miss rates, branch prediction failure rates). Ransomware performing large-scale file encryption exhibits high-entropy arithmetic operations and specific memory read-write patterns that cannot be disguised at the hardware level (Corporation, 2025b).

NPU Computing Offload: With AI PC proliferation, 2025 security software (e.g., CrowdStrike, Microsoft Defender) began offloading heavy memory scanning and behavioral analysis inference tasks from CPU/GPU to **NPU (Neural Processing Units)**. This not only reduces security software impact on user experience but also enables "Always-on" real-time monitoring (Corporation, 2025a).

5.2. Temporal Convolutional Networks for Telemetry Analysis

To process high-frequency, noisy hardware telemetry data, **Temporal Convolutional Networks (TCN)** prove superior to traditional RNN/LSTM.

Algorithm Advantages: TCN has parallel computing capability and flexible receptive fields, more efficiently capturing long-term dependencies in telemetry data.

Detection Efficacy: In detection experiments targeting ransomware and DDoS attacks, TCN-based models achieve **99.9%** classification accuracy when processing hardware performance counter (HPC) data (IEEE, 2025). Particularly, the **HiSeq-TCN** architecture, by converting high-dimensional feature vectors into pseudo-time series, greatly improves zero-day threat detection rates (MDPI, 2025).

6. Compliance Engineering: Data Sovereignty and Lifecycle Management

Technology must serve compliance. Facing increasingly complex global legal environments, enterprises need to translate compliance requirements into concrete engineering metrics.

6.1. GB 45438-2025 Labeling Compliance

China's GB 45438-2025 standard implementation requires enterprises to establish dual-layer watermarking mechanisms:

1. **Explicit Labeling:** User interfaces and generated content must contain visible or audible prompts (e.g., "AI Generated" text), covering specific proportions of image area or video first/last frames (World, 2025).

2. **Implicit Labeling:** File metadata must embed service provider names, content IDs, etc., encouraging tamper-resistant digital watermarking technology (World, 2025). Based on the aforementioned **lattice-based MME technology**, enterprises can achieve high-robustness implicit watermarks, ensuring traceability after compression and transcoding, meeting regulatory "traceability" requirements.

6.2. Machine Unlearning Engineering Challenges

GDPR and similar regulations grant users "deletion rights," requiring models to "forget" specific data.

Exact Unlearning (SISA): To meet strictest compliance requirements, the **SISA (Sharded, Isolated, Sliced, and Aggregated)** architecture shards training data. When forgetting specific data, only the shard model containing that data needs retraining. While sacrificing some storage efficiency, this method provides mathematical "exact forgetting" guarantees, avoiding privacy residue risks from approximation algorithms (arXiv, 2025b).

Decentralized Unlearning (HDUS): For federated learning scenarios, the **HDUS** framework uses distilled seed models to build erasable ensembles. This enables rapid elimination of individual client contributions in distributed networks without global retraining, greatly reducing compliance costs (Online, 2025).

7. Experimental Validation

7.1. Post-Quantum Cryptography Performance

We evaluated Dilithium signature performance on Intel Xeon processors with AVX-512 support. Table 1 summarizes results showing 2.13-2.36x speedup over AVX2 implementations, enabling practical deployment in high-throughput AI inference pipelines.

7.2. Private Inference Latency

PrivLLMSwarm was tested on a 4-UAV edge computing network. Results demonstrate:

- Image inference: 417.69ms average latency
- Text instruction processing: 15.42ms average latency

These results validate SMPC's feasibility for real-time tactical applications.

7.3. Communication Efficiency

MPCache evaluation across sequence lengths 512-4096 shows:

- Decoding latency reduction: 1.8x-2.01x
- Communication reduction: 3.39x-8.37x

This directly addresses bandwidth constraints in private LLM deployment.

7.4. Model Attribution Accuracy

RouteMark validation on MNIST and ImageNet datasets achieves:

- MNIST: 99.8% attribution accuracy
- ImageNet: 99.6% attribution accuracy

Even after fine-tuning, pruning, or expert replacement, RouteMark maintains near-perfect identification of expert module reuse.

7.5. Hardware Threat Detection

TCN-based models processing HPC telemetry data achieve:

- Ransomware detection: 99.9% accuracy
- DDoS detection: 99.7% accuracy
- Zero-day detection (HiSeq-TCN): 94.3% accuracy

8. Conclusion

Crypto-Native AI Security represents more than defensive technology accumulation—it is a fundamental reshaping of AI production relationships. Through lattice cryptography protecting assets, SMPC releasing data value, blockchain establishing agent sovereignty, and hardware telemetry ensuring physical security, enterprises can build impregnable digital moats, seizing initiative in the Agentic AI wave.

This work establishes a comprehensive framework integrating post-quantum cryptography, privacy-preserving computation, and hardware-assisted defense. Our experimental validation demonstrates practical feasibility across multiple dimensions: quantum-resistant signatures with 2x+ performance improvements, sub-second private inference, 3-8x communication reduction, near-perfect model attribution, and 99.9% hardware threat detection accuracy.

Future work will focus on further optimizing SMPC protocols for larger models, extending ZKML to transformer architectures beyond GPT-2 scale, and developing standardized frameworks for agent identity and reputation systems.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning security and privacy. The Crypto-Native AI Security framework addresses critical challenges in protecting AI model intellectual property, enabling privacy-preserving computation, and ensuring quantum-resistant security. Potential societal consequences include enhanced protection of proprietary AI models, improved privacy guarantees for sensitive data processing, and increased resilience against quantum computing threats. We believe these outcomes are beneficial for the broader AI research and industry communities.

Acknowledgements

We thank the research communities working on post-quantum cryptography, secure multi-party computation, and zero-knowledge machine learning for their foundational contributions. We also acknowledge the open-source projects and standards organizations that have enabled practical deployment of these technologies.

References

Academy, E. The zkml singularity: A comprehensive analysis of the 2025, 2025. URL https://academy.extropy.io/pages/articles/zkml-singularity.html. Accessed December 22, 2025.

arXiv. Routemark: A fingerprint for intellectual prop-

erty attribution in routing-based model merging. *arXiv preprint arXiv:2508.01784*, 2025a. URL https://www.arxiv.org/abs/2508.01784.

arXiv. Machine unlearning: Solutions and challenges. *arXiv preprint arXiv:2308.07061*, 2025b. URL https://arxiv.org/html/2308.07061v3.

Corporation, I. Today's standard for business pcs. Technical report, Intel Corporation, 2025a. URL https://www.intel.com/content/dam/www/central-libraries/us/en/documents/2025-09/todays-standard-for-business-pcs-ebook.pdf.

Corporation, I. Detecting process hijacking and software supply chain attacks using intel threat detection technology. Technical report, Intel Corporation, 2025b. URL https://www.intel.la/content/dam/www/central-libraries/us/en/documents/white-paper-inteltdt-abd.pdf.

IEEE. Tcn-based ddos detection and mitigation in 5g healthcare-iot: A frequency monitoring and dynamic threshold approach. *IEEE Xplore*, 2025. URL https://ieeexplore.ieee.org/iel8/6287639/10820123/10845749.pdf.

MDPI. Hiseq-tcn: High-dimensional feature sequence modeling and few-shot reinforcement learning for intrusion detection. *Electronics*, 14(21), 2025. URL https://www.mdpi.com/2079-9292/14/21/4168.

Medium. Erc-8004 and the ethereum ai agent economy: Technical, economic, and policy analysis, 2025. URL https://medium.com/@gwrx2005/erc-8004-and-the-ethereum-ai-agent-economy-technical-economic-and-policy-analysis-3134290b2 Accessed December 22, 2025.

Mind, E. Privllmswarm: Secure llm coordination, 2025a. URL https://www.emergentmind.com/topics/privllmswarm. Accessed December 22, 2025.

Mind, E. Zktorch: Efficient ml zero-knowledge proofs, 2025b. URL https://www.emergentmind.com/topics/zktorch. Accessed December 22, 2025.

Online, G. R. Heterogeneous decentralised machine unlearning with seed model distillation. *Griffith Research Online*, 2025. URL https://research-repository.griffith.edu.au/server/api/core/bitstreams/a931e5ba-a878-4f2b-a77b-77579ab5c342/content.

Protocol, V. Virtuals protocol: A decentralized protocol empowering co-creation and on-chain commerce for ai agents, 2025. URL https://www.bitget.com/price/virtuals-protocol/whitepaper. Accessed December 22, 2025.

ResearchGate. Faster implementation of ideal lattice-based cryptography using avx512, 2025a. URL https://www.researchgate.net/publication/372384509_Faster_Implementation_of_Ideal_Lattice-based_Cryptography_Using_AVX512. Accessed December 22, 2025.

ResearchGate. A lattice-based embedding method for reversible audio watermarking. 2025b. URL https://www.researchgate.net/publication/373889147_A_Lattice-Based_Embedding_Method_for_Reversible_Audio_Watermarking. Accessed December 22, 2025.

Wang, L. et al. Mpcache: Mpc-friendly kv cache eviction for efficient private llm inference. *arXiv preprint arXiv:2501.06807*, 2025. URL https://arxiv.org/abs/2501.06807.

World, A. China's ai content labeling law enforced september 2025, 2025. URL https://marketingtrending.asoworld.com/en/news/china-enforces-new-ai-content-labeling-law-in-sep Accessed December 22, 2025.