Differential Privacy Through Depth-Selective Routing

by Nick Clark | Published March 27, 2026 | PDF

Differential privacy in training traditionally relies on noise injection that degrades model quality. Depth-selective gradient routing offers an alternative: privacy guarantees achieved through structural isolation rather than noise. By restricting sensitive content to shallow layers, the architecture ensures that sensitive information cannot influence deep representations while maintaining full model quality in unrestricted layers.

What It Is

Privacy through depth-selective routing restricts content classified as sensitive to shallow model layers. The sensitive content can inform surface-level behavior (response style, topic awareness) without embedding in the deep representations that form the model's core knowledge. This structural isolation provides privacy guarantees without the quality degradation of noise-based differential privacy.

Why It Matters

Noise-based differential privacy requires adding calibrated noise to gradients, which degrades model quality proportionally to the privacy guarantee. Stronger privacy guarantees require more noise, producing worse models. Depth-selective routing provides privacy through structure rather than noise, avoiding this quality-privacy tradeoff.

How It Works

Sensitive content is assigned a shallow depth profile that restricts its gradients to upper model layers. These layers capture style and surface patterns. Deep layers that capture fundamental knowledge representations receive no gradient from sensitive content. The sensitive information influences how the model responds but not what it fundamentally knows.

This approach can be combined with noise-based methods for additional guarantees, but the structural isolation alone provides meaningful privacy properties.

What It Enables

Depth-selective privacy enables training on sensitive data with structural privacy guarantees and minimal quality impact. Medical models can train on patient data that informs clinical reasoning patterns without memorizing patient specifics. Legal models can train on confidential case files that inform legal reasoning without encoding case details. The structural approach provides both privacy and utility.

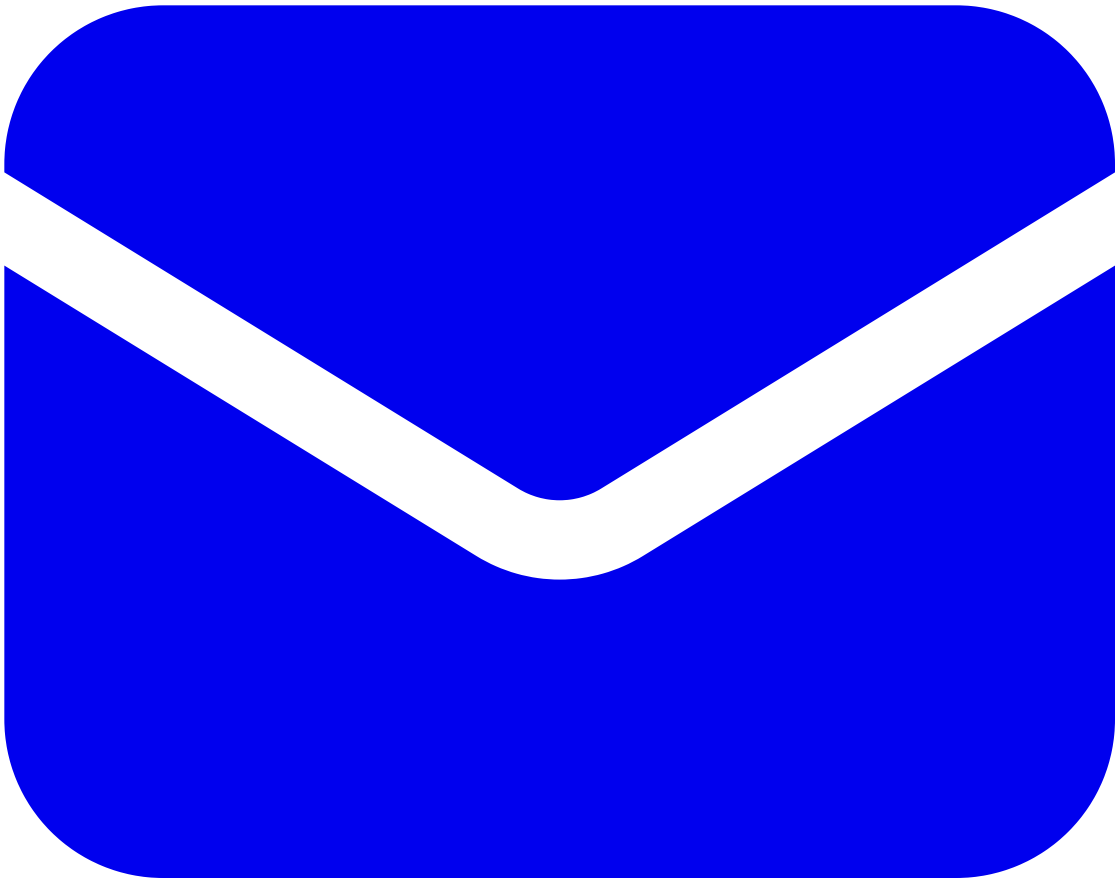
Training Governance All 21 steps →

Govern what the model learns, at what depth, with what provenance.

Primary Technical Disclosure
○ Depth-Selective Training Governance for Machine Learning Systems
Secondary Technical
○ Training Examples as Proposed Semantic Mutations○ Entropy-Band-Indexed Training Depth Profiles○ Depth-Selective Gradient Routing for Governed Training○ Training-Level Memorization Detection● Differential Privacy Through Depth-Selective Routing○ Governed Fine-Tuning With Verifiable Provenance○ The Training Loop as a Governed Execution Environment○ Policy-Governed Knowledge Retention and Suppression○ Provenance-Traceable Training Dynamics○ Curriculum-Integrated Depth Scheduling○ Affect-Modulated Training Depth○ Training-Inference Governance Integration○ Training Governance for Human-Relatable Agents
Applications (General)
○ Rights-Compliant Model Training Through Depth-Selective Routing○ Regulated Industry Model Governance With Provenance○ Training Governance for Medical AI○ Training Governance for Legal AI○ Training Governance for Financial Model Training○ Training Governance for Defense AI○ Training Governance for Educational AI Models○ Training Governance for Creative AI
Applications (Specific)
○ OpenAI's Training Pipeline Has No Depth-Selective Governance○ Constitutional AI Training Lacks Depth-Selective Control○ Stable Diffusion's Training Has No Provenance Layer○ Midjourney Trains Aesthetics Without Governed Depth○ Scale AI Labels Data Without Governing What Models Learn○ Labelbox Manages Annotation Workflows, Not Learning Dynamics○ Snorkel AI Programs Labels but Does Not Govern Gradient Depth○ Weights & Biases Tracks Experiments, Not Learning Governance○ Determined AI Orchestrates Compute, Not Learning Depth○ MosaicML Optimizes Training Efficiency, Not Learning Governance
Training Governance overview →
AQ
deterministic
autonomy

Legal

Last updated: 2026-03-03



-
- nick@qu3ry.net
- 72 28 14 36 01

Invented by Nick Clark | Founding Investors: Devin Wilkie