В работе представлена архитектура многослойной нейросети с интегрированным особым защитным слоем, реализующим концепцию “иммунитета” искусственной нейронной сети. Ранее предлагался подход с использованием двух независимых моделей, предлагаемая архитектура предполагает встраивание аналитического защитного слоя непосредственно в структуру нейросети. Данный слой функционирует как специализированный модуль валидации, осуществляющий многоступенчатый контроль входных данных (детекция промт-инъекций, несанкционированных команд) и выходных ответов (анализ утечек конфиденциальной информации, признаков атак инверсии модели и установления принадлежности). Особенностью изучения архитектуры является определения возможности обучения защитного слоя совместно с основной сетью при сохранении его изолированности от остальных компонентов. Приводятся результаты моделирования, демонстрирующие снижение успешности атак. Архитектура обеспечивает гибкость настройки и совместимость с популярными фреймворками машинного обучения.