EduSphere technology stack: a complete 2025 open-source architecture guide

The optimal stack for a knowledge graph-based, multi-tenant educational platform in 2025 centers on PostgreSQL as the unified data backbone (with Apache AGE for graph queries and pgvector for semantic search), **Drizzle ORM with row-level security for multi-tenancy**, **NATS JetStream for lightweight event-driven messaging**, and **Vercel AI SDK for LLM abstraction** between Ollama (dev) and cloud APIs (production). This architecture prioritizes true open-source licensing (Apache 2.0/MIT), Docker-first local development, and a clear upgrade path to Kubernetes production. Every recommendation below has been validated against 2024–2025 release cycles, GitHub activity, and license compliance.

The platform spans 12 technical domains — from monorepo tooling to AI agent sandboxing — and the choices interlock. A PostgreSQL-centric data layer eliminates operational complexity by consolidating graph, vector, and relational data in one engine. NATS replaces heavier Kafka-like brokers with a **20 MB binary** that starts in milliseconds. GraphQL Federation via Yoga + Hive Gateway provides a fully MIT-licensed alternative to Apollo's restrictive ELv2. Apollo GraphQL Apollo GraphQL The frontend pairs React + Vite (for SPA performance) with Expo SDK 54 (for offline-capable mobile), sharing ~70–80% of code.

---

1. Nx and pnpm dominate monorepo tooling

For a TypeScript monorepo containing multiple microservices, a React web app, a React Native mobile app, and shared packages, **Nx (~27.4K GitHub stars, MIT, v21.2+)** is the strongest choice. Nx It outperforms Turborepo by **5–7x on large repos** due to selective cache file restoration, Nx offers first-class plugins for React, Expo, Nest.js, and Node.js, and includes code generators for scaffolding new services consistently. Its `nx affected` command enables efficient CI by running only tasks impacted by changes. Nx Turborepo (v2.8, ~26K stars, MIT) remains excellent for simpler setups and teams already in the Vercel ecosystem, Turborepo Wisp CMS but lacks Nx's plugin depth and distributed CI capabilities. Rush (v5.154, ~6.2K stars) targets strict enterprise versioning but has the smallest community. npm Trends

pnpm workspaces (v9.x, ~30K stars, MIT) is the consensus package manager for new monorepos. It delivers **60–80% disk savings** via content-addressable storage, runs 3–5x faster than npm, Glen Thomas and prevents phantom dependencies by design. Medium Its `workspace:` protocol cleanly references cross-package dependencies. pnpm Yarn 4 Berry's Plug'n'Play mode, while innovative, still causes ecosystem compatibility issues. Turborepo

The recommended folder structure separates `apps/` (deployable frontends), `services/` (backend microservices), `packages/` (shared libraries like types, UI components, utils), `tools/` (build configs), and `infrastructure/` (Docker, Helm, Terraform). This pattern enables each service to have its own Dockerfile while sharing TypeScript types and utilities through workspace packages. Medium

2. PostgreSQL unifies graph, vector, and relational data

The graph database landscape splits between expressiveness and operational simplicity. **Neo4j Community Edition (v2025.12, ~13K stars)** offers the most expressive query language (Cypher) and the best ecosystem for

knowledge graphs — pattern-matching relationships like `(Concept)-[:PREREQUISITE_OF]->(Concept)` is natural and intuitive. However, its **AGPLv3 + Commons Clause license is NOT truly open-source** by OSI definition, DB News and CE lacks clustering, RBAC, and multi-database support. Wikipedia

For a 100% open-source system, **Apache AGE (Apache 2.0, ~3K stars)** is the strategic choice. This PostgreSQL extension adds openCypher graph query support alongside standard SQL, enabling graph traversal and relational queries in the same transaction. GitHub The critical advantage: **AGE coexists with pgvector in the same PostgreSQL instance**, delivering graph + vector + relational capabilities without additional infrastructure. SurrealDB (~28.8K stars) offers multi-model convenience but uses BSL 1.1 licensing GitHub Caperaven and has less expressive graph syntax than Cypher. Markaicode NebulaGraph (~11.7K stars, Apache 2.0) targets massive-scale distributed graphs GitHub — overkill for most educational knowledge graphs.

No property graph database natively supports RDF/OWL ontologies. For RDF/SPARQL compliance, specialized triple stores (Apache Jena, Blazegraph, Virtuoso) would be needed as an additional layer. For EduSphere's semantic relationships (CONTRADICTS, MENTIONS, RELATES_TO, PREREQUISITE_OF), the property graph model with labeled edges is more natural and performant.

For vector search, **pgvector (PostgreSQL License, ~4K stars)** keeps everything in one database. With pgvectorscale, it achieves **471 QPS at 99% recall on 50M vectors** Firecrawl — sufficient for educational content. The alternative is **Qdrant (Apache 2.0, ~9K stars, Rust-based)**, which offers superior filtering and a dedicated REST/gRPC API, AIMultiple but requires a separate service. For embedding generation, use **Ollama with `nomic-embed-text`** Ollama **(768 dims)** in development and **`all-MiniLM-L6-v2` via sentence-transformers** in production.

The state-of-the-art RAG pattern for education is **HybridRAG**: parallel vector retrieval (semantic similarity) and graph retrieval (structured relationships), with results fused before LLM generation. Memgraph Research shows this achieves **96% factual faithfulness** (NVIDIA/BlackRock) NetApp Community and **91.4% retrieval accuracy** in course-oriented QA (KA-RAG framework). MDPI

3. Yjs and Hocuspocus power real-time collaboration

Yjs (v13.6, ~20.5K stars, MIT, 1.9M weekly npm downloads) dominates the CRDT landscape GitHub with **3–4x the market adoption of Automerge** npm Trends and consistently superior performance: processing 260K edits in ~0.5s versus Automerge's ~4.7s, with lower memory usage and smaller encoded documents. Its ecosystem spans every major editor (ProseMirror, Tiptap, CodeMirror 6, Monaco, Lexical) Yjs Velt and includes network providers for WebSocket, WebRTC, and Cloudflare Durable Objects. GitHub Automerge (v3, ~5.7K stars) offers a more intuitive JSON-like API and better cross-platform native support (Swift, Rust), GitHub but its niche is structured data rather than text-heavy collaboration.

Hocuspocus (v2.x, MIT) provides production-ready server infrastructure over raw `y-websocket`. Tiptap It adds lifecycle hooks (`onAuthenticate`, `onChange`, `onStoreDocument`), built-in JWT authentication, Tiptap debounced persistence, Redis scaling for multi-server deployment, and a webhook extension for notifying external services on document changes. tiptap For high-scale scenarios (100K+ concurrent connections), **y/hub** by Yjs creator Kevin Jahns streams through Redis without holding Y.Doc in memory, but it carries an AGPL license.

The proven **offline-first pattern** uses dual providers: `y-indexeddb` persists the document locally for instant load, while `HocuspocusProvider` syncs over WebSocket when online. GitHub CRDTs eliminate traditional

merge conflicts by design GitHub — concurrent text insertions are ordered deterministically, and map values use last-writer-wins semantics. For **PostgreSQL persistence**, store CRDT binary updates as BYTEA rows, append incrementally, and compact periodically by merging all updates into a single snapshot (the y-postgresql library automates this with a configurable flushSize). To make CRDT data queryable in SQL, use Hocuspocus's onChange hook to extract JSON and denormalize into regular relational columns.

4. faster-whisper leads the transcription pipeline

faster-whisper (~14K stars, MIT) is the production recommendation for Docker-based transcription. Google Using the CTranslate2 engine, it runs **~4x faster than original Whisper with identical accuracy** Towards AI and uses **50–70% less VRAM**. On CPU with int8 quantization, it benchmarks at 14s versus whisper.cpp's 46s for the same audio. GitHub whisper.cpp (~38K stars, MIT) excels on edge/embedded devices Google with its zero-dependency C++ implementation and minimal memory footprint, but performs poorly on GPU. Alibaba Original Whisper (~75K stars) serves as the accuracy reference Google but requires heavy PyTorch containers.

The video processing pipeline chains **FFmpeg (v6/7)** for transcoding, segment extraction, thumbnail generation, and HLS/DASH packaging, OTTVerse with **Bento4** for fragmented MP4 (CMAF) creation. Medium Output streams upload to MinIO as S3-compatible storage, and **hls.js (16.3K stars)** handles browser-side adaptive playback. npm Trends

For the video player, **Video.js (v8.23, ~39.4K stars,** npm Trends **Apache 2.0)** has the strongest plugin ecosystem SaaSHub for annotation — videojs-overlay provides time-based HTML overlays, GitHub videojs-markers adds timeline markers, and dedicated annotation plugins support moment/range comments. GitHub Video.js v10 is in development with Mux and the Plyr creator. Mux For canvas-based annotation drawing over video frames, **Konva.js (v10, ~14K stars, MIT) with react-konva** delivers NPM Compare **~2.5x better rendering performance** than Fabric.js, first-class React integration via declarative components, and a layer system NPM Compare for separating annotation types. Synchronize annotations using requestAnimationFrame polling video.currentTime for ~60fps accuracy.

5. Vercel AI SDK abstracts the LLM layer cleanly

The AI agent framework decision hinges on TypeScript support. **CrewAI and AutoGen are Python-only**, disqualifying them for this TypeScript-native platform. The recommended three-layer architecture:

- **Vercel AI SDK (v6, ~15K stars, Apache 2.0, 2.8M weekly npm downloads)** for LLM provider abstraction GitHub — its unified API lets you swap between ollama('llama3.1') in development and openai('gpt-4o') in production with zero code changes Medium
- **LangGraph.js (~2.3K stars, MIT, 529K weekly npm downloads)** for graph-based agent workflow orchestration — defining pedagogical flows as state machines (assess → quiz → explain → debate → reassess) Medium
- **LlamaIndex.TS (MIT)** for RAG and knowledge graph integration — best-in-class data connectors and indexing strategies

LiteLLM (~34.5K stars, MIT) serves as an optional proxy layer providing cost tracking, load balancing, fallback chains, and rate limiting across LLM providers. It adds **8ms P95 latency** at 1K RPS. GitHub For local

development, **Ollama (~120K stars)** supports 100+ models — **Llama 3.1 8B** (4–6 GB VRAM) for general tutoring and **Phi-4 14B** (8–10 GB) for reasoning-heavy explanations.

For user-buildable agents, store agent definitions as **JSON configurations validated with Zod schemas**, offering templates (Quizzer, Explainer, Debater) that users customize via a UI with dropdowns for personality, difficulty, and topic scope. **FlowiseAI (~35K stars, Apache 2.0)** provides an open-source visual agent builder (DronaHQ) for advanced users. Agent sandboxing for multi-tenant safety uses **gVisor (Apache 2.0, by Google)** — a user-space kernel providing medium-strength isolation with only 10–20% performance overhead, seamlessly integrated with Docker and Kubernetes. (SoftwareSeni) (SoftwareSeni)

For agent communication, adopt **MCP (Model Context Protocol, 97M monthly SDK downloads)** for tool integrations and the **OpenAI function calling format** as the internal standard. MCP was donated to the Linux Foundation in December 2025 (Medium) and has become the dominant agent-to-tool protocol.

6. Row-level security with Drizzle ORM scales multi-tenancy

For a SaaS educational platform with potentially thousands of tenants, **PostgreSQL row-level security (RLS) with a single shared schema** is the clear winner over schema-per-tenant or database-per-tenant approaches. RLS adds invisible WHERE clauses at the query-optimizer level, enforcing tenant isolation at the database layer (Logto) with minimal performance overhead when using simple policies indexed on (tenant_id). The safe pattern with connection pooling: always use (SET LOCAL app.current_tenant = '...') inside transactions, which auto-resets on transaction end — no leakage risk even with PgBouncer in transaction mode.

Drizzle ORM (v1.0 beta, ~32.7K stars, Apache 2.0, 545% download growth in 2024) is the recommended ORM because it provides **native RLS support** (TheDataGuy) via (pgPolicy()) and (enableRLS()) directly in schema definitions. Its SQL-first, schema-as-code approach generates optimized single-statement queries (up to **14x lower latency** than N+1-prone ORMs), ships at ~7.4 KB min+gzipped with zero dependencies, and has excellent serverless/edge performance. (Nihar Daily) (TheDataGuy) Prisma (v6.x, ~44K stars) is a strong alternative for teams preferring higher-level abstraction, (DesignRevision) but it lacks native RLS schema support — requiring raw SQL in migration files and client extensions for SET commands.

Keycloak (v26.5, Apache 2.0, ~25K stars, CNCF incubating) handles multi-tenant authentication (ZITADEL) with its native **Organizations feature** (fully supported since v26). (BootLabs TechBlog) Use a single realm with one organization per tenant — this scales to thousands of organizations while enabling cross-org user sharing, centralized SSO, and per-org role-based access control. (Phasetwo) Custom protocol mappers inject (tenant_id) into JWT tokens. (Notebook) **Zitadel (~10K stars, Apache 2.0 → AGPL v3)** is a compelling Go-based alternative with built-in multi-tenancy, (Htdocs) but its recent license change to AGPL v3 is a consideration.

For tenant isolation in messaging, start with the **pool model** (shared NATS subjects with tenant ID in message headers) for simplicity, evolving to per-tenant subjects with ACLs as isolation requirements grow.

7. NATS JetStream is the lightest production-ready broker

NATS with JetStream (v2.11, Apache 2.0, ~17.1K stars, CNCF project) is the optimal message broker for this platform. (GitHub) (Go Packages) Its server binary is **~20 MB**, starts in milliseconds, and runs with **128–256 MB RAM** for local development — dramatically lighter than Kafka's 1.5–3 GB JVM footprint. Subject wildcarding ((content.>), (quiz.*)) naturally maps to educational event hierarchies like (content.created),

annotation.added, quiz.completed, and agent.message. It includes a built-in KV store (for ephemeral state like collaboration cursors) and WebSocket support for real-time features.

Apache Kafka (v4.1, Apache 2.0, ~31.8K stars) finally eliminated ZooKeeper dependency in v4.0 (March 2025) GitHub and remains unmatched at extreme throughput. **Redpanda (v24.3, BSL 1.1, ~10K stars)** Redpanda offers the best Kafka-compatible developer experience with its single C++ binary, built-in schema registry, and web console — but it is **not open-source** (BSL restricts offering it as a streaming service). GitHub For an educational platform at moderate event volume, NATS JetStream provides the right balance of simplicity, performance, and true open-source licensing. Migration to Kafka is possible later if throughput demands exceed NATS capabilities.

For the API layer, **GraphQL Yoga + Hive Gateway (MIT, ~8.3K stars)** provides Apollo Federation v2 compatibility without Apollo's ELv2 license restrictions. It's the **fastest JS-based gateway** (nearly 2x faster than competitors) and supports all runtimes (Node, Deno, Bun, Cloudflare Workers). Each microservice exposes a federated GraphQL subgraph, composed by Hive Gateway at the edge. **Hive** serves as the open-source schema registry for breaking change detection. For internal service-to-service communication, use **NATS for async events** and **gRPC for synchronous calls** where needed.

8. React + Vite with shadcn/ui for the web, Expo for mobile

React + Vite (v6, 100K+ stars) beats Next.js for this educational SPA platform — it delivers near-instant dev server startup, millisecond HMR, Medium smaller bundles without framework overhead, and full control over architecture. Strapi Use Next.js only if the platform has significant SSR/SEO requirements for public-facing pages. CodeParrot The state management pairing of **TanStack Query (v5.90, ~46.8K stars, 11.7M weekly npm downloads)** for server/API state and **Zustand (v5.0, ~54.9K stars, 11.3M weekly downloads)** npm Trends for client-only UI state is the industry-standard combination — simple, performant, and TypeScript-native.

shadcn/ui (~85–94K stars, MIT) is the recommended UI library. Built on Radix primitives + Tailwind CSS, it provides copy-paste components with full code ownership, Makers Den maximum customizability, and strong AI-assisted development support. SW Habitation +2 **Mantine (v7.10, ~28K stars, MIT)** is the faster-time-to-market alternative with 100+ components SW Habitation Pmbanugo including built-in forms, notifications, date pickers, Makers Den and charts. Both are excellent — shadcn/ui for maximum control, Mantine for maximum feature coverage.

Expo SDK 54 (React Native 0.81, ~40K stars, MIT) is the mobile framework Expo — officially recommended by the React Native team React Native and used by Coinbase, Discord, and NFL Network. Bitcot Its offline capabilities center on **expo-sqlite** (full SQLite with WAL mode) Medium paired with TanStack Query for offline-first data patterns. Universal apps share ~70–80% of code across iOS, Android, and Web.

9. Docker Compose with Traefik for local dev, Kubernetes for production

Traefik (v3.6, ~61.5K stars, MIT) is the recommended reverse proxy/API gateway. GitHub Its automatic service discovery via Docker labels eliminates manual nginx.conf maintenance as services scale, Traefik and it provides built-in Let's Encrypt SSL, Januschung BestCloudPlatform a monitoring dashboard, and native Kubernetes Ingress Controller support. NGINX's raw throughput advantage matters only at extreme traffic volumes irrelevant to an educational platform. Docker Hub Notably, the NGINX Ingress Controller is retiring in March 2026, making Traefik the more strategic Kubernetes choice. Traefik

MinIO (RELEASE.2025-10, ~49K stars, AGPLv3) provides S3-compatible object storage for local development. The same AWS SDK code works against MinIO locally and AWS S3 in production — just swap the endpoint and credentials. Arifszn Use the quay.io/minio/minio Docker image with console access on port 9001. GitHub

For **Docker Compose**, split compose files by concern (infrastructure, services, dev tools) and use profiles (docker compose --profile backend up) for selective startup. Network isolation separates public and internal networks, with named volumes for all persistent data. The Dev World

For **CI/CD**, GitHub Actions with Nx's affected commands provides the most sophisticated monorepo pipeline — running only builds and tests impacted by changes, Graphite with pnpm and .nx cache directories persisted across runs. This reduces build times by **60–80%**. Business Compass LLC Production deploys to **Kubernetes via Helm charts** using a shared template chart pattern — one generic microservice chart with per-service values.yaml overrides. GitHub

Consolidated technology decisions

Domain	Primary Choice	License	Alternative
Monorepo	**Nx v21 + pnpm v9**	MIT	Turborepo v2.8
Graph Database	**Apache AGE (PostgreSQL)**	Apache 2.0	Neo4j CE (AGPL+CC)
Vector Search	**pgvector + pgvectorscale**	PostgreSQL	Qdrant (Apache 2.0)
CRDT/Collaboration	**Yjs v13 + Hocuspocus v2**	MIT	Automerge v3
Transcription	**faster-whisper**	MIT	whisper.cpp
Video Player	**Video.js v8**	Apache 2.0	Custom w/ hls.js
Annotation Canvas	**Konva.js v10 (react-konva)**	MIT	Fabric.js v7
AI Framework	**Vercel AI SDK v6 + LangGraph.js**	Apache 2.0/MIT	Mastra
RAG	**LlamaIndex.TS**	MIT	LangChain.js
LLM Abstraction	**Vercel AI SDK providers**	Apache 2.0	LiteLLM proxy
ORM	**Drizzle ORM v1**	Apache 2.0	Prisma v6
Auth	**Keycloak v26 (Organizations)**	Apache 2.0	Zitadel
Message Broker	**NATS JetStream v2.11**	Apache 2.0	Redpanda (BSL)
API Gateway	**GraphQL Yoga + Hive Gateway**	MIT	Apollo Router (ELv2)
Frontend	**React + Vite v6**	MIT	Next.js 15

Domain	Primary Choice	License	Alternative
State Management	**TanStack Query v5 + Zustand v5**	MIT	Jotai
UI Components	**shadcn/ui**	MIT	Mantine v7
Mobile	**Expo SDK 54**	MIT	Bare React Native
Reverse Proxy	**Traefik v3.6**	MIT	NGINX
Object Storage	**MinIO**	AGPLv3	—
Multi-tenancy	**PostgreSQL RLS**	—	Schema-per-tenant

Conclusion

The architecture's central insight is **PostgreSQL maximalism**: by consolidating graph queries (Apache AGE), vector search (pgvector), CRDT persistence, relational data, and row-level security into a single PostgreSQL instance, EduSphere avoids the operational tax of managing 3–4 separate database systems. This doesn't prevent future separation — pgvector can be replaced by Qdrant, AGE by Neo4j — but it provides the simplest viable starting point that still handles sophisticated knowledge graph traversal and semantic search.

The **hybrid LLM strategy** via Vercel AI SDK's provider system is production-proven: a single function call routes to Ollama locally or OpenAI/Anthropic in production with zero code changes, eliminating the common pitfall of dev/prod configuration drift. NATS JetStream's **subject wildcarding** (`education.content.>`) maps naturally to educational event taxonomies and eliminates Kafka's heavyweight JVM overhead for teams that don't need trillion-event-per-day throughput.

Two forward-looking risks deserve monitoring. First, the AI agent framework landscape is evolving at breakneck speed — MCP reached 97M monthly SDK downloads within a year of launch, `Medium` and new orchestration frameworks emerge monthly. Anchoring on Vercel AI SDK's stable abstraction layer (20M+ monthly downloads, backed by Vercel/Next.js team) provides insurance against churn in lower layers. Second, Apache AGE's community (~3K stars) is smaller than Neo4j's, meaning fewer tutorials and integrations — mitigate this by designing the graph access layer behind an interface that could swap to Neo4j if AGE's development stalls.