BytesAndBrains: A Shared Runtime for Machine Learning Outside the Data Center

Matthew Love
Bytes and Brains LLC
matthewlove@bytesandbrains.ai

June 2026

Abstract

Machine learning increasingly runs outside the data center, and the strategies the field has produced for that setting (federated learning [1], gossip learning [2], split learning [3], peer-to-peer inference) share little infrastructure: each ships with its own reference code, its own RPC stack, and its own training loop. BytesAndBrains is a runtime layer these strategies can share. It records a workload into an intermediate representation serialized as ONNX [4], partitions the recording across nodes, and manages the structured byte layer between them. The runtime is sans-IO [5]: a state machine driven by the host program, with every distributed behavior testable in-process. The extension surface is composed of *Components*, typed plug-ins that satisfy *Roles*, runtime contracts the engine dispatches against. This paper describes the design, states its boundaries, and positions BytesAndBrains relative to Flower [6], FedML [7], NVIDIA FLARE [8], Ray [9], Substra [10], libp2p [11], TensorFlow Federated [12], Hivemind [13], and Spark [14].

1 Introduction

Machine learning runs in two operating envelopes. One is the data center: homogeneous clusters, fast interconnect, a single operator, one trust domain. The other is everything outside the data center: phones, sensors, hospital networks, on-premises fleets, and peer-to-peer overlays. The strategies the field has produced for the second envelope (federated learning [1], gossip learning [2], split learning [3], federated-split hybrids [15], peer-to-peer inference [16]) all address a single underlying question: how to run machine learning where the data already lives.

These strategies do not compose. Each ships with its own reference implementation: a one-off RPC stack, a paradigm-specific training loop, and a coordinator process or peer-sampling implementation hard-wired to one transport. A research group attempting to compare three strategies on the same dataset writes three networking stacks. The cost is borne in every empirical study and every reproducibility effort.

BytesAndBrains is a runtime layer these strategies can share. It takes a description of a computation, partitions it across nodes, and manages the bytes between them. Federated, gossip, split, and peer-to-peer learning become bindings on the same surface. The remainder of this paper describes the design (Sections 3 through 7), walks a small example (Section 8), names BytesAndBrains' boundaries (Section 9), positions it relative to existing systems (Section 10), and concludes (Section 11).

2 Background and Motivation

We are not aware of an open system that occupies the layer this paper describes; Section 10 examines the nearest candidates. The layers around it are well populated (Figure 1). PyTorch [17], TensorFlow [18], and JAX [19] provide tensor compute and model-authoring DSLs. Ray [9] and Spark [14] provide cluster scheduling and actor-based execution. libp2p [11] provides peer-to-peer transports, multiplexing, peer identity, and discovery primitives. Between these layers sits a gap: a runtime layer that authors computation, partitions it across nodes that may be reachable only through networks it does not own, executes each partition without assuming what kind of process hosts it, and ships bytes between participants without committing to any particular transport.

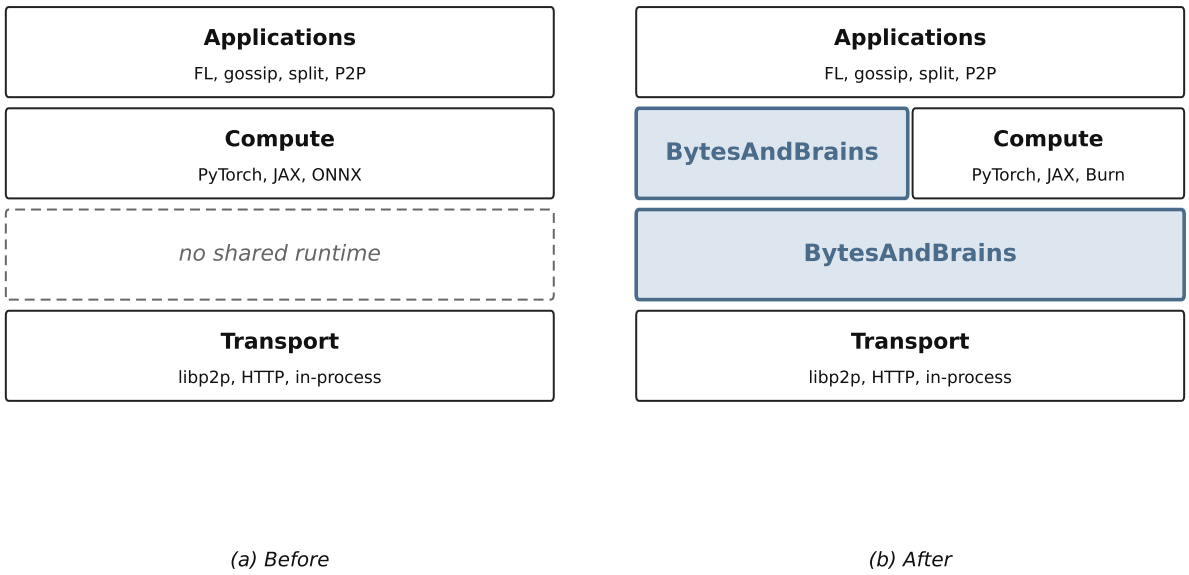

Figure 1: The layering view. (a) PyTorch, JAX, and ONNX provide tensor compute; libp2p, HTTP, and in-process channels provide transport. Between them sits a gap: no shared runtime layer. (b) BytesAndBrains is both the authoring DSL (*Module*, *Graph*, *Roles*) and the runtime layer beneath it (IR, compiler, engine, wire envelope); tensor frameworks plug in as Backend Components.

The strategies the field has produced for the second envelope are technically valuable. Federated averaging [1] established the coordinator-driven round model and is the de facto starting point for cross-device learning. Gossip learning [2] generalized it to a fully decentralized exchange with age-weighted merging. Split learning [3] partitions the model itself between participants. Push-sum aggregation [20] established asynchronous averaging primitives still used in production gossip protocols. Their reference implementations are standalone artifacts glued to bespoke RPC stacks and training loops; framework-level reuse across them requires substantial reimplementation.

A runtime layer that consolidates these strategies must satisfy specific constraints. It is a framework rather than a library, because partitioning is a global concern that user code cannot make in isolation. It uses a portable intermediate representation, so the same computation survives crossing language and process boundaries. It is sans-IO [5], so the host program retains exclusive control over its runtime and sockets. It owns the byte layer but stops there; existing transports are well-developed, and it does not invent its own. It is composable at the role level, so the building blocks of the field's strategies (compute backends, models, indices, aggregators, peer-sampling overlays, custom protocols) plug in as distinct, swappable components.

3 System Overview

BytesAndBrains performs three operations.

First, BytesAndBrains records the user's computation into a portable intermediate representation. A workload is authored as a *Module*, a typed handle whose recording method walks the description once and produces a graph in the IR. From that point forward, the IR is the program; the host language no longer participates in subsequent processing.

Second, the compiler turns the recording into per-node programs. Placement is expressed by the author at module granularity: each node class in a deployment (a sensor, a trainer, an aggregator) is authored as a Module, and Modules compose — a parent embeds child Modules and calls their entries. Each top-level Module is a compile target selected by name, and a node compiles and installs the target it will run. Cross-node edges are typed wire operations in the recording; the compiler's wire passes hoist them to the node

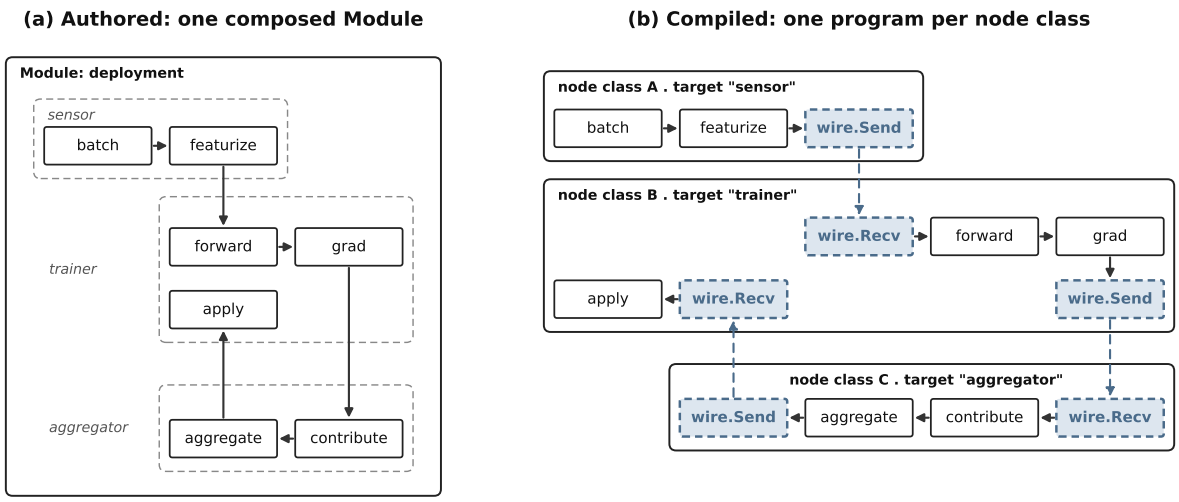

Figure 2: Module composition and partitioning. (a) The authored computation: one composed Module whose children are the three node classes of a deployment — a sensor that featurizes local data, a trainer that computes an update, and an aggregator that merges updates and returns parameters. Cross-module edges are ordinary dataflow; the recording carries no wire operations and no placement. (b) After compilation: any child Module is selectable as a compile target, each node class compiles to its own per-node program, and the compiler lowers every cross-target edge to typed `wire.Send` / `wire.Recv` operations (dashed, slate-blue), deriving correlation and receive routes from the graph topology. The update–parameters exchange between trainer and aggregator becomes a request–response pair. One recording yields three node programs that exchange typed values through one structured envelope.

boundary, recognize request–response round trips from the graph topology, synthesize the matching receive roots on the peer, and stamp the routing table (Figure 2). The author writes the send once; the correlation, parking, and reply plumbing is compiler-derived.

Third, BytesAndBrains owns the byte layer between nodes. Every byte sent or received rides as a structured envelope, whose encoding, routing, and correlation the framework owns. The host owns the sockets that move the envelope. A transport adapter mediates the boundary.

The discipline that holds the design together is sans-IO. The library crate contains no asynchronous runtime, no shared sockets, and no thread-pool dependency. The runtime is a state machine. The host calls a poll function on it; the state machine drains pending inputs, walks the program to a fixed point, and returns outbound envelopes. The host ships those envelopes and feeds inbound bytes back. All other runtime decisions (event-loop selection, transport choice, reconnection policy, logging) belong to the host. BytesAndBrains is a coordination and authoring layer; Component composition handles backend, transport, security, and deployment-shape concerns.

4 Intermediate Representation

A framework that partitions computation across nodes requires a representation it can analyze, rewrite, and serialize independently of the host language. BytesAndBrains uses one IR in two forms: an in-memory form the compiler and engine work on, and ONNX [4] as its serialization.

The working form is *Program*: an interned, immutable dataflow graph. Operations carry a type, a domain, and decoded attributes; values are typed sites, each either a graph input or the output of a producing operation. A Program is recursive — it holds hash-consed child programs under a lexical symbol table, and a `Call` operation names a child scope — which is how a recording expresses composition beyond a single flat graph. Program identity is structural: a name-independent denotation hash identifies the same computation across processes and doubles as its on-wire identity.

The persistent form is ONNX. ONNX defines a typed, directed acyclic graph format and an opset of named operations, and serves as the de facto cross-framework format for trained models, with tooling and

runtime support across PyTorch [17], TensorFlow [18], JAX [19], and ONNX Runtime. A Program serializes to an ONNX `ModelProto` (each program node becomes a `FunctionProto`; the child tree flattens into the model's function list) and reconstructs from one. The round trip is lossless: reconstruction preserves structural equality and the denotation hash, as an invariant of the design rather than a convention. The proto barrier is crossed exactly twice, once when a recording is persisted or shipped and once when a node installs it, and never at runtime: compiler passes are functions over Program, and the engine consults no proto. Three properties make this arrangement appropriate.

Portability. The persistent IR is a serializable protocol buffer. A recording produced by a host in one language can be read by a host in another. The authoring surface is in Rust, but the IR does not constrain that choice; alternative front-ends in Python or other languages remain admissible.

Community vocabulary. ONNX op definitions are agreed by a community of toolmakers and framework authors who care about how forward and backward passes are described, what a typed tensor means, and how control-flow nodes behave. A backend author who already understands ONNX implements one for BytesAndBrains without learning a new IR. Models authored elsewhere export to ONNX format; Backend Components consume ONNX nodes directly or translate to their preferred runtime (a CPU backend executes ONNX nodes directly; an ExecuTorch backend translates and caches; an ONNX Runtime backend passes through).

Layered extension. BytesAndBrains defines additional opsets on top of ONNX's base (runtime intrinsics, inter-node value transfer, role-specific verbs) while serializing to a single graph proto. The in-memory form adds structure the flat proto lacks (recursion, entry graphs, interning) but introduces no second serialization format: everything that leaves the process is ONNX, so ONNX-aware tooling applies directly to persisted recordings, and models authored elsewhere embed by importing their ONNX export into a recording.

Typed contract for cross-node values. The IR is typed end-to-end: every Component slot, every cross-node edge, every wire envelope carries a typed contract that both ends must agree on. The framework surfaces an error when an envelope arrives with a type the receiver does not recognize; it does not provide a mechanism for renegotiating types at runtime. Two nodes that exchange typed values agree on the type system at deployment time, and cross-version compatibility is a deployment concern, not an IR concern.

The cost of this choice is concrete. ONNX's opset is the community's opset; decisions baked into it cannot be relitigated locally. Versioning must coordinate with the broader ecosystem. These costs are accepted in exchange for portability and ecosystem reach.

5 Runtime Model

A node is a runtime container hosting an execution engine. The engine owns the per-node state required to run a partitioned program: the dispatch table mapping operations to bound Components, the slot table holding intermediate values, the frontier of operations ready to fire, the ingress queue for external events, the tracking for long-running asynchronous operations, and a small set of framework primitives (a scheduler, an event bus, a request tracker, and an outbound envelope queue).

The container is a state machine. The host calls a poll function on it. The function drains pending inputs, walks the program's directed acyclic graph to a fixed point, and returns a vector of outbound envelopes. External threads interact with the engine exclusively through the ingress queue, which constitutes the single thread-safe boundary. Within the poll cycle, every operation has exclusive access to engine state.

Three design properties follow.

Single-threaded dispatch, concurrent execution. The engine dispatches operations sequentially within a poll cycle. Each op borrows mutable references to shared framework primitives, and the single-threaded dispatch loop is what keeps the engine state machine straightforward to reason about. Op *execution* is not constrained to that loop. A Contract method that returns a deferred completion releases the engine to drain the rest of the frontier while the op runs to completion on whatever runtime the host supplies (a thread pool,

an async executor, a GPU stream, a remote service). At this layer, BytesAndBrains is a networked scheduler: the engine orchestrates which op fires when, and the Component implementation chooses how to execute. Components are free to be fully threaded if their runtime supports it.

Sans-IO discipline. The library crate contains no asynchronous runtime dependency, no shared sockets, and no spawned tasks. Two consequences follow. Every system built on it is testable in-process: a harness instantiates multiple nodes in one process, connects them by in-memory channels, and exercises the full distributed behavior without opening sockets. The library also ports to any host runtime that can shuttle bytes and call poll, decoupling framework logic from the host's runtime choice.

Snapshot as a first-class concept. Components that maintain state implement two methods: serialize current state to bytes and reconstruct an instance from bytes. The compiler assigns each stateful instance a stable snapshot key at build time. A node-wide snapshot walks the dispatch table, captures each Component's state, and emits a blob; a restore rebuilds the same shape. Snapshot captures runtime state for replay-based debugging of distributed failures, or for migration of a Node between hosts. A captured snapshot plus the recorded ingress stream reproduces the failing trajectory deterministically, and the same blob hands a Node off when a host is decommissioned.

Iteration without in-graph loops. A single execution of a program is a strict dataflow DAG; the IR has no loop operation. Rounds — the unit of iteration in federated and gossip learning — are repeated executions of the same installed program, each in a fresh execution frame. Three triggers create executions: the host invokes a target by name, the clock component's interval timer spawns an execution of an enrolled program each period, and lifecycle phases fire enrolled programs when the host signals a phase transition. The engine recycles per-program frame and readiness allocations, so steady-state rounds allocate nothing. State that persists across rounds (the model being refined, the aggregation buffer) lives in stateful Components, not in the graph.

Memory model. BytesAndBrains distinguishes external byte payloads (transport-owned, ephemeral) from internal byte ownership (framework-owned, durable). External boundaries entering the engine (network adapter ingress, application event push, async completion) take payloads as borrowed slices (`&[u8]`) and copy them into framework-owned storage inside the boundary call. Many transport stacks hand the framework a pointer into their own memory (libp2p reads, QUIC completion buffers, kernel ring buffers); the framework must not keep a reference past the call and must not assume the transport's allocator owns the buffer. The copy is where ownership transfers from transport to framework. Inside the framework, byte ownership flows normally: a `Vec<u8>` moves between framework-owned slots, and an `Arc<...>` handle threads backend-managed tensor buffers across slots without deep copy. The single mandatory copy is at the boundary.

Error semantics at the boundary. External boundaries use fallible allocation and a per-node byte budget. On allocation failure or budget exceedance, the engine emits a typed bus event naming the failure and the surface it occurred on (wire or application), drops the offending bytes, and continues processing other envelopes and events. The engine never panics, never aborts, and never stalls at the boundary. Inside the engine, normal Rust patterns apply: components are designed to play by the runtime contract, so a per-component allocation failure inside a Contract body is a process abort, not a framework-handled event. The fallibility line lives at the engine boundary itself. Per-source caps (per-event, per-invoke, per-completion) reject oversize single payloads with typed errors; the cumulative byte budget guards sustained load. The two layers compose: the per-source cap protects against single adversarial inputs, and the budget protects against ensemble pressure.

Observability. The runtime is instrumented through three surfaces. Tracing spans bound each poll cycle and its sub-phases (ingress drain, frontier drain, async completions, outbound flush). Each operation invocation opens a structured span carrying the op's name, kind, domain, and dispatch identity, so traces filter and group by op kind or by specific dispatch instance. The event bus carries node-, infrastructure-, and

application-level events as the in-process observability surface; the host subscribes to drive metrics, logs, and external tracing without coupling to a specific telemetry backend.

Multiple targets per Node. A Node is a runtime container, not a Module. A single Node can install multiple Module targets, and those targets share the named Component bindings the host registered at construction. Authors compose distinct behaviors as separate Modules (a client target that trains and uploads, a server target that aggregates and rebroadcasts); the host installs each target onto the Node it wishes to run that behavior. Bindings are looked up by name, so two targets referencing the same Component name resolve to the same instance. The canonical example is federated learning, where one peer may take on both the client and the server role against a shared model: the Node installs a `client` Module and a `server` Module, both targets bind the same Backend, the same model parameter store, and the same PeerSelector, and the engine dispatches against the shared instances without coordination between the targets. The same shape covers gossip-and-coordinate hybrids, validator-and-participant peers, and any deployment where one runtime container plays multiple roles in the larger computation.

6 The Byte Layer

When a partitioned program needs to send a value to another node, the framework wraps the value in a structured envelope and asks the host to ship it. When the host receives an envelope, it hands it back to the framework, which routes it to the destination. The envelope is the contract between framework and host. Above it, BytesAndBrains owns dispatch and encoding. Below it, the host owns sockets, encryption, identity, and the network stack.

Two semantically distinct messaging planes share the envelope format. The *data plane* carries values across the cross-node edges the compiler inserted: encoding, request and response correlation, batching of multiple edges to a common peer in a single cycle, and routing to the correct destination slot. The user does not see this plane; partitioning produces it. The *control plane* carries protocol traffic for Components that speak to peers: gossip protocols, DHT-style overlays, consensus algorithms. Inbound envelopes for these protocols route to the appropriate Component; the Component owns encoding and decoding of its own messages. Both planes use the same envelope format. The envelope's protocol identifier distinguishes them.

The transport-adapter contract is small. Outbound: encode the envelope as framed bytes and ship through whatever channel the host has chosen. Inbound: decode framed bytes as an envelope and pass to the framework. The adapter consumes whatever peer identity, multiplexing, encryption, and discovery the underlying transport supplies; the framework does not assume any of them.

The choice of transport is a deployment concern. Peer-to-peer overlays, request-response HTTP, and in-process channels (for tests, simulators, or embedded multi-node deployments inside one host process) all satisfy the same contract. BytesAndBrains owns the envelope and terminates at the adapter boundary.

Identity and replay. The envelope carries a transparent sender identifier. BytesAndBrains does not authenticate that identifier; production deployments use transports with built-in identity (libp2p peer ids [11], mutual TLS, or equivalent). The envelope carries request and response correlation but no anti-replay nonce; the same transport layer supplies replay protection. Side-channel mitigation (constant-time codec implementations, padding, traffic shaping) lives in the Component that handles the sensitive data.

Backend-mediated tensor receive. A wire receive whose destination slot binds a Backend Component routes through that Backend rather than through a generic decoder. The engine caps the inbound fill's byte length against per-envelope caps, charges the length against the node's ingress byte budget, and hands the framework-owned bytes to the Backend by value. The Backend chooses the materialization strategy: zero-copy adoption when alignment permits, a buffer drawn from its own pool, or a fresh allocation. The resulting backend-native tensor is wrapped in an internal handle that carries the byte charge so the slot-table writer can release it on overwrite or eviction. This is the framework-to-Component handoff, not an external boundary; the borrowed-slice rule that governs transport ingress does not apply, because the Backend lives inside the framework ecosystem and plays by the runtime contract. The Backend interface (Section 7) defines

the single contract method that participates in this hand-off and supplies a default that delegates to a shared wire decoder, so backends without tensor pooling work without modification.

6.1 Network reality

Production networks impose constraints BytesAndBrains names but does not solve in isolation. The framework provides composition seams; Components and protocols implement the policies.

NAT traversal. Peer-to-peer overlays solve NAT through hole-punching, relays, and rendezvous coordination [21, 11]. BytesAndBrains is transport-agnostic; libp2p adapters consume these mechanisms directly. Lifecycle phases and interval timers let protocols express NAT-friendly behaviors (client-first-speak, server-driven heartbeat) as ordinary graph programs rather than transport-specific code.

Intermittent connectivity. Asynchronous federated rounds and disconnection-tolerant gossip are addressed by communication-efficient FL protocols [22], low-communication training across loosely-coupled clusters with H-step inner loops and outer-sync aggregation [23, 24], and production federated systems [25]. Hop-local back-off rides as a typed wire op and per-site liveness is tracked through the φ-accrual failure detector [26]. The PeerSelector role provides the seam where disconnect-aware cohort selection (drop, stash, requeue) is expressed; the Aggregator role determines what a partial round means.

Asymmetric bandwidth. Communication-efficient federated learning explicitly addresses uplink-dominated cost asymmetry through gradient compression, sketching, and structured updates [22, 27], with theoretical foundations in privacy-aware estimation under communication constraints [28], algorithmic underpinnings in local-update decentralized SGD [29], and control-variate-corrected aggregation [30]. The Codec role provides the encode-decode seam where compression schemes compose with the wire layer; the Aggregator role consumes the decompressed payloads. BytesAndBrains ships no codec; it defines where one composes.

Async timeouts and backpressure. Tail-latency mitigation through per-call deadlines is established practice [31]. Dapper-style per-envelope deadlines [32] are derived at compile time from graph-walk chain depth and refined at runtime through Karn-Partridge RTT smoothing [33, 34]. Backpressure under congestion is well-studied in event-driven systems [35]: hop-local back-off rides as a typed wire op, receiver-side detection composes with φ-accrual liveness, and the PeerSelector role provides the seam where exponential-backoff and circuit-breaker policies are expressed.

7 Extension Surface

Pluggability distinguishes BytesAndBrains from a one-paradigm library. The extension surface is built from two abstractions: Components and Roles.

A **Component** is a typed plug-in. The only structural requirement is a lifecycle contract: a stable type identifier, a serialization method, and a reconstruction method. The engine dispatches operations to Components. Components are otherwise free to assume whatever internal shape their logic requires.

A **Role** is a runtime contract defining a category of behavior. Role traits name the recurring categories: a compute backend executes tensor operations; a data loader produces batches; an aggregator combines parameter shards from multiple peers; an index looks up values by key or similarity; a peer selector draws peers from a view of the network [36]; a codec transforms values at the wire boundary; a wire extension participates in envelope handling. Custom protocols need no dedicated Role: a protocol is an ordinary Component that declares its own opset and speaks it to peers.

Each Role corresponds to a trait. A Component implements the Role traits it intends to satisfy. A single-Role Component is a single-purpose plug-in. A multi-Role Component exposes several surfaces from one piece of state (Figure 3). A Kademlia-style DHT [21] naturally exposes three surfaces from one piece of state: it is an index (values stored in the overlay are retrievable by key), a peer-sampling source (peers are drawn from its routing table), and a protocol (it speaks its own message types to peers). Recent variants such as the learned-CDF distributed hash table [37] keep the same structure while replacing the placement

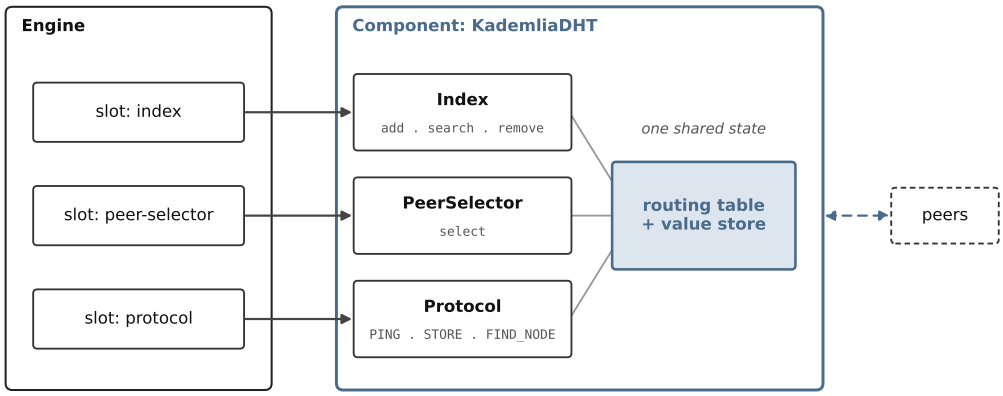

Figure 3: Multi-role Component composition. A Kademlia-style distributed hash table holds one piece of state (a routing table and value store) and exposes three surfaces over it: `Index` (`add`, `search`, `remove`), `PeerSelector` (`select`), and a protocol opset it speaks to peers through the wire envelope (`PING`, `STORE`, `FIND_NODE`). The compiler binds the same component instance at all three engine slots; the engine dispatches each call to the matching method set. The framework carries no opinion on how many Roles a Component implements.

function with a trained model, and decentralized inference overlays [16] compose an indexing Role with a protocol opset over the same envelope. The engine dispatches by Role without requiring knowledge of how many Roles any given Component carries.

The surface intentionally lacks a central registry, a plug-in protocol, or any compile-time ritual beyond trait implementation; polymorphism comes from the trait system alone. Library authors ship Components in standalone crates. Users bind those Components to Roles at node construction. The framework remains small, and most of the system's behavior ships as Components.

Backend interface. The compute backend Role is the surface for tensor compute. A Backend Component declares an associated `Tensor` type (the backend's native tensor handle, which is typically an `Arc`-shared handle around a backend-managed buffer for cheap intra-node clones), implements either a per-operation method surface (one method per mandatory tensor primitive) or a whole-graph execute method, and supplies one contract method that participates in the wire path: it consumes framework-owned wire bytes by value and returns a backend-native tensor. A default delegates to a shared wire decoder, so backends without tensor pooling continue to work. This is the framework's single hook for routing cross-node tensor values through the Backend's allocator. Existing tensor runtimes (a CPU backend, an ExecuTorch wrapper, a Burn integration, an ONNX Runtime adapter) implement the Backend Role by satisfying its trait contract; dispatch lands against the contract surface without coupling to any particular runtime.

Federated-learning composition seams. Production federated learning addresses gradient poisoning, membership inference, model inversion, and differential privacy as distinct threats [25, 27, 38, 39, 40]. The framework ships none of these defenses; it provides composition seams: compiler-time bindings that attach a concrete Component at a named slot. The aggregator binding is where Byzantine-robust reductions compose [38]. The codec binding is where encrypt-aggregate-decrypt and DP noise compose. The peer-selector binding is where attesting cohort gating composes. Hierarchical and multi-tier federated topologies — D2D inside clusters, intermediate aggregation between tiers, server reconciliation [41] — express through the same seams: the compiler-stage hook is where tier-aware partitioning lands, and the Aggregator and PeerSelector roles host the per-tier behavior. Federated personalization at the objective level [42] composes through per-Module objective bindings rather than runtime-level personalization shims; clustered federated learning [43] expresses through lifecycle-phase cohort assignment as observable graph state. Deployments that need these defenses or composition shapes choose Component implementations that supply them; the framework is silent on policy.

Layered responsibility. BytesAndBrains names each concern, owns one layer of the stack, and exposes a composition seam for the rest.

Concern	Owner	Composition seam
Tensor compute	Backend Component	Compiler backend binding
Network transport	Host + adapter	Wire envelope
Identity / authentication	Transport adapter	Below the envelope
Gradient-poisoning defense	Aggregator Component	Compiler aggregator binding
Differential privacy	Codec pair	Compiler codec binding
Peer attestation	PeerSelector	Compiler peer-selector binding
NAT traversal	Transport + protocol Components	libp2p adapter + lifecycle phases
Adaptive deadlines	Framework runtime	Per-hop deadline budgets + RTT tracking

8 Example: Two Modules, Two Node Programs

The shape the preceding sections describe is compact in practice. The listing below is schematic — concrete names will evolve with the implementation — but the shape is the design commitment: a worker and a coordinator each authored as a Module, each compiled as its own target, with the backend resolved as a compile-time binding and the host loop as the sans-IO boundary from Section 5.

```
// Two behaviors, each authored as a Module over shared Components.
let worker = worker_module();           // computes locally, sends its update
let coordinator = coordinator_module(); // receives, aggregates, replies

// Each node compiles the Module it will run. The backend is a binding.
let program = Compiler::new()
    .bind_backend::<CpuBackend>()
    .compile(worker)?;

let mut node = Node::new(config);
node.install(program)?;

// The sans-IO host loop: poll, ship envelopes, feed inbound bytes back.
let mut events = Vec::new();
node.poll(now, &mut events);
for event in events.drain(..) {
    if let Event::SendEnvelope { peer, bytes } = event {
        transport.send(peer, bytes);   // host-owned IO
    }
}
node.deliver_inbound(peer, inbound_bytes)?;
```

Three things are worth noting. First, the coordinator is the same few lines compiled from the other Module with its own bindings; the two Modules are authored together, their cross-node edges reference each other, and the compiler derives the matching send and receive plumbing on each side. Second, changing strategy means changing Modules and bindings, not infrastructure: the wire operations, correlation, envelope encoding, and poll loop are identical across strategies. Third, because the host owns IO, a test harness runs both nodes in one process and hands envelopes across directly — the full two-node exchange, including encoding and routing, executes without a socket.

9 Boundaries and Limitations

Placement is manual. Partitioning is expressed by the author at module granularity; there is no cost model and no automatic partitioner. A split-learning cut is authored as a module boundary with explicit wire operations, not discovered by the compiler. Automatic operator-level placement is future work.

Failure recovery is policy, not mechanism. The framework detects and surfaces failures — φ-accrual liveness, typed send- and dial-failure events, per-envelope deadlines — but does not decide what they mean. The PeerSelector decides whether a silent peer stays in the cohort; the Aggregator decides what a partial round produces; a cross-node edge whose peer never replies fails by deadline, and the framework does not reschedule the partition elsewhere. Dynamic membership likewise lives in Components (the peer ledger and selector), not in the engine.

Programs are installed, not accepted. The envelope format can carry serialized programs, but execution requires installation, and installation is a host decision. The framework does not validate, sandbox, or attest a program received from the network. An installed graph can only invoke opsets that locally installed Components expose, under the node's byte and operation budgets, which bounds what a foreign graph can do — but deciding whose programs to install is provenance policy, and it belongs to the host.

No empirical evaluation yet. This is a position paper. The overheads the design accepts (the boundary copy, sequential poll dispatch, proto decode at install) are unquantified, and there is no head-to-head study against Flower or Hivemind on a shared workload — the comparison the consolidation argument ultimately owes. This is the most important missing artifact and the nearest-term future work.

10 Related Work

Backends as enablers. TFLite, ExecuTorch, Core ML, ONNX Runtime, and Burn are inference and training backends that integrate as Backend Components. They are enablers of the runtime layer, not competitors with it; the integration shape is the same for a minimal CPU backend as for accelerator and platform-specific runtimes.

Ray and PyTorch. Ray [9] provides actor-based distributed execution with a global object store, designed for homogeneous clusters with low-latency interconnect. PyTorch [17] provides the model-authoring DSL. Their combination is the canonical reference for distributed training inside a data center, with strong support from a single operator running a homogeneous fleet. The combination is unsuitable for heterogeneous, WAN-scale deployments: actors carry runtime weight, the object store assumes shared infrastructure, and placement is built around scheduling decisions rather than overlay topology. BytesAndBrains targets these gaps.

DeepSpeed, Megatron-LM, and PyTorch FSDP. These three systems address a layer of the stack distinct from the one this paper targets. DeepSpeed's ZeRO [44] partitions optimizer states, gradients, and parameters across data-parallel workers to enable training models with parameter counts that exceed single-device memory. Megatron-LM [45] specializes in tensor and pipeline parallelism for transformer-style architectures. PyTorch FSDP [46] shards model parameters across data-parallel workers and gathers them on demand. All three assume a single trust domain, synchronous collective operations, and a low-latency interconnect (NVLink, InfiniBand, or equivalent). They are appropriate when the goal is to train a single large model as quickly as possible inside a homogeneous cluster. BytesAndBrains does not compete with them; it targets workloads where the assumptions of homogeneous hardware, low-latency interconnect, and single-trust-domain operation do not hold.

Flower. Flower [6] is a federated learning framework supporting custom strategies, with PyTorch and TensorFlow integration and a substantial set of supported aggregation algorithms. Flower selects one paradigm (federated learning with a coordinator) and ships strategies within that paradigm. Implementing gossip learning on top of Flower runs against the grain of its API; implementing peer-to-peer inference falls outside Flower's scope. For workloads bounded by the federated learning paradigm, Flower offers a more direct path than BytesAndBrains; for workloads composing across paradigms, the two systems sit at different layers.

TensorFlow Federated. TensorFlow Federated [12] is Google's production federated learning stack. It exposes a high-level Federated Learning API and a Federated Core for expressing federated computations. Like Flower, it is single-paradigm: a coordinator-driven federated computation model with TensorFlow as the bound tensor backend. BytesAndBrains sits at a layer that accepts TFF-style federated rounds as one composition among others.

FedML, NVIDIA FLARE, and OpenFL. These systems are the nearest neighbors to the claim this paper makes, because each supports more than one strategy under one roof. FedML [7] covers federated, split, and decentralized training in a single Python library with MPI-, MQTT-, and gRPC-based communication backends. NVIDIA FLARE [8] generalizes the coordinator round into pluggable controller and executor workflows and is deployed in production federated settings. OpenFL [47] provides task-runner and interactive APIs over a defined federation topology. All three are orchestration frameworks: they schedule paradigm-shaped workflows over an RPC layer they own, with the training step expressed as framework callbacks in Python. None records the computation into a portable IR, none partitions a single program across nodes (split learning in FedML is a distinct workflow, not a compiler output), and none embeds as a sans-IO state machine in a host that keeps its own sockets. They demonstrate demand for multi-strategy consolidation at the workflow layer; BytesAndBrains addresses the same demand one layer down, at the program representation and the byte layer.

Hivemind. Hivemind [13] is a peer-to-peer distributed training library for PyTorch, with decentralized averaging, distributed mixture-of-experts, and DHT-coordinated training. It is paradigm-specific (PyTorch-bound, training-focused) and supplies its own RPC and DHT stack. Moshpit SGD [48] establishes the convergence guarantees for Hivemind-style averaging under churn, and Petals [49] extends the paradigm into collaborative LLM inference. BytesAndBrains sits at a layer that hosts these systems' coordination shapes as compositions: PeerSelector and Aggregator roles plus a custom protocol Component express the same shape over the wire envelope, with the backend choice unbound.

libp2p. libp2p [11] provides transports, multiplexing, DHT-style discovery, and a peer-identity model. It does not provide an authoring DSL, an IR, a compiler, or a dispatch model. Building distributed machine learning on bare libp2p is a substantial engineering effort. BytesAndBrains treats libp2p as one valid transport adapter beneath the envelope, not as a competing layer.

Spark. Spark [14] solves an adjacent problem: bulk-synchronous data-parallel computation on a cluster, with a high-level functional API and a runtime that handles partitioning and shuffles. It is not designed for streaming workloads, heterogeneous fleets, or stateful protocols with control-plane traffic (gossip exchanges, peer sampling, leader election). The planning layers differ: BytesAndBrains plans against a typed graph with explicit cross-node edges, Spark against a dataset abstraction with bulk-synchronous stages. The two systems answer different shapes of question.

Substra. Substra [10] is a federated learning platform targeting regulated industries. It provides orchestration, asset management, audit trails, and coordinator-driven workflows. It is a platform with workflow primitives, not a runtime with authoring primitives. BytesAndBrains and Substra occupy different layers and could coexist: BytesAndBrains expressing the per-node computation Substra orchestrates.

In summary, BytesAndBrains targets a layer we have not found addressed by existing open systems. It is sans-IO rather than actor- or service-based, records computation into a portable IR rather than expressing it as framework callbacks, and manages the byte layer between nodes rather than the scheduling layer above them. It treats federated, gossip, split, and peer-to-peer strategies as placement-and-binding problems on a single runtime layer, rather than as distinct paradigms with their own coordinators and runtimes.

11 Conclusions

This paper has presented BytesAndBrains, a shared runtime for machine learning outside the data center. It records a user's computation into an intermediate representation serialized as ONNX [4], partitions the

recording across nodes through the compiler, executes each partition as a state machine, and ships bytes between participants through a single structured envelope. The runtime is sans-IO [5], single-threaded in dispatch within each node yet free in execution, and decoupled from any particular transport. Above it, learning strategies compose as Components satisfying Role contracts. Below it, any transport that can ship bytes can carry the envelopes.

The design's usefulness scales with the Components available for it: each additional Backend, Codec, Aggregator, PeerSelector, and Protocol widens the set of strategies expressible without new infrastructure. Building out that set, alongside the empirical evaluation Section 9 identifies as open, is the immediate future work. Development is public at https://github.com/Bytes-Brains/bytesandbrains.

References

[1] H. Brendan McMahan et al. "Communication-Efficient Learning of Deep Networks from Decentralized Data". In: *Proceedings of Artificial Intelligence and Statistics (AISTATS)*. 2017. URL: https://arxiv.org/abs/1602.05629.

[2] István Hegedűs, Gábor Danner, and Márk Jelasity. "Gossip Learning as a Decentralized Alternative to Federated Learning". In: *Distributed Applications and Interoperable Systems (DAIS)*. 2019. URL: https://arxiv.org/abs/1906.05882.

[3] Praneeth Vepakomma et al. *Split learning for health: Distributed deep learning without sharing raw patient data*. 2018. URL: https://arxiv.org/abs/1812.00564.

[4] ONNX. *Open Neural Network Exchange specification*. URL: https://onnx.ai/onnx/.

[5] Sans-IO. *Network protocols, sans I/O*. URL: https://sans-io.readthedocs.io/.

[6] Daniel J. Beutel et al. *Flower: A Friendly Federated Learning Framework*. 2020. URL: https://arxiv.org/abs/2007.14390.

[7] Chaoyang He et al. *FedML: A Research Library and Benchmark for Federated Machine Learning*. 2020. URL: https://arxiv.org/abs/2007.13518.

[8] Holger R. Roth et al. *NVIDIA FLARE: Federated Learning from Simulation to Real-World*. 2022. URL: https://arxiv.org/abs/2210.13291.

[9] Philipp Moritz et al. "Ray: A Distributed Framework for Emerging AI Applications". In: *USENIX Symposium on Operating Systems Design and Implementation (OSDI)*. 2018. URL: https://arxiv.org/abs/1712.05889.

[10] Mathieu N. Galtier and Camille Marini. *Substra: a framework for privacy-preserving, traceable and collaborative Machine Learning*. 2019. URL: https://arxiv.org/abs/1910.11567.

[11] libp2p. *A modular network stack*. URL: https://libp2p.io/.

[12] TensorFlow Federated. *An open-source framework for machine learning and other computations on decentralized data*. URL: https://www.tensorflow.org/federated.

[13] Hivemind. *Decentralized deep learning in PyTorch*. URL: https://github.com/learning-at-home/hivemind.

[14] Matei Zaharia et al. "Apache Spark: A Unified Engine for Big Data Processing". In: *Communications of the ACM* 59.11 (2016), pp. 56–65. DOI: 10.1145/2934664. URL: https://doi.org/10.1145/2934664.

[15] Chandra Thapa et al. "SplitFed: When Federated Learning Meets Split Learning". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2022. URL: https://arxiv.org/abs/2004.12088.

[16] Fei Fang et al. "PlanetServe: A Decentralized, Scalable, and Privacy-Preserving Overlay for Democratizing Large Language Model Serving". In: *Proceedings of the USENIX Symposium on Networked Systems Design and Implementation (NSDI)*. 2026. URL: https://arxiv.org/abs/2504.20101.

[17] Adam Paszke et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2019. URL: https://arxiv.org/abs/1912.01703.

[18] Martín Abadi et al. "TensorFlow: A System for Large-Scale Machine Learning". In: *USENIX Symposium on Operating Systems Design and Implementation (OSDI)*. 2016. URL: https://arxiv.org/abs/1605.08695.

[19] James Bradbury et al. *JAX: composable transformations of Python+NumPy programs*. 2018. URL: https://github.com/google/jax.

[20] David Kempe, Alin Dobra, and Johannes Gehrke. "Gossip-based computation of aggregate information". In: *Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*. 2003, pp. 482–491. DOI: 10.1109/SFCS.2003.1238221. URL: https://doi.org/10.1109/SFCS.2003.1238221.

[21] Petar Maymounkov and David Mazières. "Kademlia: A Peer-to-Peer Information System Based on the XOR Metric". In: *International Workshop on Peer-to-Peer Systems (IPTPS)*. 2002, pp. 53–65. DOI: 10.1007/3-540-45748-8_5. URL: https://doi.org/10.1007/3-540-45748-8_5.

[22] Jakub Konečný et al. *Federated Learning: Strategies for Improving Communication Efficiency*. 2017. URL: https://arxiv.org/abs/1610.05492.

[23] Arthur Douillard et al. *DiLoCo: Distributed Low-Communication Training of Language Models*. arXiv preprint arXiv:2311.08105. 2023. URL: https://arxiv.org/abs/2311.08105.

[24] Ji Qi et al. *DiLoCoX: A Low-Communication Large-Scale Training Framework for Decentralized Cluster*. arXiv preprint arXiv:2506.21263. 2025. URL: https://arxiv.org/abs/2506.21263.

[25] Keith Bonawitz et al. "Towards Federated Learning at Scale: System Design". In: *Proceedings of Machine Learning and Systems (SysML)*. 2019. URL: https://arxiv.org/abs/1902.01046.

[26] Naohiro Hayashibara et al. "The φ Accrual Failure Detector". In: *Proceedings of the 23rd IEEE International Symposium on Reliable Distributed Systems (SRDS)*. 2004. URL: https://dl.acm.org/doi/10.1109/RELDIS.2004.1353004.

[27] Peter Kairouz et al. "Advances and Open Problems in Federated Learning". In: *Foundations and Trends in Machine Learning* 14.1–2 (2021). URL: https://arxiv.org/abs/1912.04977.

[28] John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. "Privacy Aware Learning". In: *Journal of the ACM* 61.6 (2014), 38:1–38:57. DOI: 10.1145/2666468. URL: https://arxiv.org/abs/1210.2085.

[29] Sebastian U. Stich. "Local SGD Converges Fast and Communicates Little". In: *International Conference on Learning Representations (ICLR)*. 2019. URL: https://arxiv.org/abs/1805.09767.

[30] Sai Praneeth Karimireddy et al. "SCAFFOLD: Stochastic Controlled Averaging for Federated Learning". In: *Proceedings of the 37th International Conference on Machine Learning (ICML)*. 2020. URL: https://arxiv.org/abs/1910.06378.

[31] Jeffrey Dean and Luiz André Barroso. "The Tail at Scale". In: *Communications of the ACM* 56.2 (2013), pp. 74–80. URL: https://research.google/pubs/pub40801/.

[32] Benjamin H. Sigelman et al. *Dapper, a Large-Scale Distributed Systems Tracing Infrastructure*. Tech. rep. Google Research, 2010. URL: https://research.google/pubs/pub36356/.

[33] Phil Karn and Craig Partridge. "Improving Round-Trip Time Estimates in Reliable Transport Protocols". In: *Proceedings of ACM SIGCOMM*. 1987. DOI: 10.1145/55482.55484. URL: https://dl.acm.org/doi/10.1145/55482.55484.

[34] Vern Paxson et al. *Computing TCP's Retransmission Timer*. 2011. URL: https://datatracker.ietf.org/doc/html/rfc6298.

[35] Matt Welsh, David Culler, and Eric Brewer. "SEDA: An Architecture for Well-Conditioned, Scalable Internet Services". In: *Proceedings of the 18th ACM Symposium on Operating Systems Principles (SOSP)*. 2001. URL: https://www.sosp.org/2001/papers/welsh.pdf.

[36] Márk Jelasity, Alberto Montresor, and Ozalp Babaoglu. "T-Man: Gossip-based fast overlay topology construction". In: *Computer Networks* 53.13 (2009), pp. 2321–2339. DOI: 10.1016/j.comnet.2009.03.013. URL: https://doi.org/10.1016/j.comnet.2009.03.013.

[37] Shengze Wang et al. “A Distributed Learned Hash Table”. In: *IEEE/ACM Transactions on Networking* (2025). URL: https://arxiv.org/abs/2508.14239.

[38] Dong Yin et al. “Byzantine-Robust Distributed Learning: Towards Optimal Statistical Rates”. In: *Proceedings of the 35th International Conference on Machine Learning (ICML)*. 2018. URL: https://arxiv.org/abs/1803.01498.

[39] Reza Shokri et al. “Membership Inference Attacks Against Machine Learning Models”. In: *Proceedings of the IEEE Symposium on Security and Privacy (S&P)*. 2017. URL: https://arxiv.org/abs/1610.05820.

[40] Martín Abadi et al. “Deep Learning with Differential Privacy”. In: *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS)*. 2016. URL: https://arxiv.org/abs/1607.00133.

[41] Seyyedali Hosseinalipour et al. “Multi-Stage Hybrid Federated Learning over Large-Scale D2D-Enabled Fog Networks”. In: *IEEE/ACM Transactions on Networking* (2022). DOI: 10.1109/TNET.2022.3143495. URL: https://arxiv.org/abs/2007.09511.

[42] Virginia Smith et al. “Federated Multi-Task Learning”. In: *Advances in Neural Information Processing Systems 30 (NIPS 2017)*. 2017. URL: https://arxiv.org/abs/1705.10467.

[43] Christopher Briggs, Zhong Fan, and Peter Andras. “Federated Learning with Hierarchical Clustering of Local Updates to Improve Training on Non-IID Data”. In: *Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN)*. 2020. URL: https://arxiv.org/abs/2004.11791.

[44] Samyam Rajbhandari et al. “ZeRO: Memory Optimizations Toward Training Trillion Parameter Models”. In: *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC)*. 2020. URL: https://arxiv.org/abs/1910.02054.

[45] Mohammad Shoeybi et al. *Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism*. 2019. URL: https://arxiv.org/abs/1909.08053.

[46] Yanli Zhao et al. “PyTorch FSDP: Experiences on Scaling Fully Sharded Data Parallel”. In: *Proceedings of the VLDB Endowment* 16.12 (2023), pp. 3848–3860. URL: https://arxiv.org/abs/2304.11277.

[47] G. Anthony Reina et al. *OpenFL: An open-source framework for Federated Learning*. 2021. URL: https://arxiv.org/abs/2105.06413.

[48] Max Ryabinin et al. “Moshpit SGD: Communication-Efficient Decentralized Training on Heterogeneous Unreliable Devices”. In: *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*. 2021. URL: https://arxiv.org/abs/2103.03239.

[49] Alexander Borzunov et al. “Petals: Collaborative Inference and Fine-tuning of Large Models”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*. 2023, pp. 558–568. DOI: 10.18653/v1/2023.acl-demo.54. URL: https://arxiv.org/abs/2209.01188.