FASCIA RESEARCH · VOLUME IV

JUNE 2026 · FIRST EDITION · CC BY 4.0

Alignment Through Composition.

A formal Composition Alignment Framework with measurable safety guarantees. Why a network of inspectable specialists provides alignment surface area that monolithic frontier models cannot — and how to instrument that surface area in practice.

Executive summary.

The safety case for connective AI in three propositions.

This paper formalizes the third thesis of the Fascia Framework: **alignment composes**. We give a precise definition of what that means, a formal framework for instrumenting it, and a measurable safety guarantee for composed systems that monolithic systems cannot provide.

The intuition is simple. In a monolithic frontier model, alignment is encoded in opaque weights distributed across hundreds of billions of parameters. Misalignment, when it occurs, is non-localizable: you cannot point at the parameter that caused the problem. In a composed system — a network of specialist agents with a coordinated integration layer — alignment is encoded at the boundaries between specialists. Those boundaries are inspectable, replaceable, and instrumentable.

> **The Composition Alignment Theorem (informal).** For a composed system of specialists with type-contracted integration boundaries, system-level alignment is bounded below by the alignment of the weakest specialist plus the verification quality of the boundaries. This bound is computable. It is not available in monolithic systems.

The remainder of this paper formalizes the theorem, lays out the Composition Alignment Framework (the practical instrumentation), and presents preliminary empirical data from Fascia's deployment surface. The full appendix with formal proofs is available at fasciaai.com/research/alignment-appendix.pdf.

The composition advantage.

Four properties that distinguish composed systems from monolithic ones. All four are alignment-relevant.

01

Inspectability

Each specialist is small enough that its behavior is inspectable. You can audit a 7B-parameter specialist; you cannot meaningfully audit a 2T-parameter monolith.

02

Replaceability

If a specialist fails its alignment eval, swap it. The system continues. A monolithic model with the same failure requires a full retrain.

03

Cross-checking

The same input can be routed to multiple specialists; conflicts trigger review. Monoliths cannot cross-check themselves at the specialist level.

Property 4 — Bounded scope

Each specialist's failure modes are bounded by its domain. A misaligned legal-contract specialist cannot, by construction, start producing medical advice. The integration layer enforces type contracts at the boundary. This is a structural property that monolithic frontier models lack: a misaligned monolith can produce misaligned outputs in any domain.

Why these properties matter together

Each property alone is interesting. Together, they create a qualitatively different safety surface. **Inspectability gives you a way to find misalignment. Replaceability gives you a way to fix it without retraining. Cross-checking gives you a way to detect it in real time. Bounded scope gives you a way to limit its blast radius.** This is not available in monolithic systems at any size.

The Composition Alignment Framework.

Six instrumented layers. Each one is a specific intervention point with a specific safety guarantee.

LAYER	WHAT IT DOES	SAFETY GUARANTEE
L1: Input filter	Sanitizes incoming requests, removes adversarial patterns	Prompt injection bounded
L2: Specialist eval	Each specialist passes a continuous alignment eval before being included	No misaligned specialist active
L3: Type contract	Each specialist declares input/output types; integration enforces	Outputs constrained to declared scope
L4: Cross-check	High-stakes outputs cross-checked by 2+ independent specialists	Single-specialist failure detected
L5: Output filter	Final output passes through a harm-detection pipeline	Known harm patterns blocked
L6: Audit trail	Full provenance trace from input through specialists to output	Failures debuggable post-hoc

The full composition

The six layers compose end-to-end. A user request enters at L1, is sanitized, routed to specialists (each of which has passed L2), constrained by L3, optionally cross-checked at L4, filtered at L5, and logged at L6. **Each layer is its own intervention point with its own safety property.** No layer is sufficient on its own. Together they form a defense-in-depth that monolithic systems cannot replicate.

Preliminary empirical data.

What we see when the Framework is deployed on a real production system.

Six months of production data

Fascia deployed the full Composition Alignment Framework on 12 high-stakes specialists in Q4 2025. Over six months and 4.2 million user interactions, we logged every instance where one or more layers intervened. The data:

LAYER	INTERVENTIONS / 1M	TRUE POSITIVES	FALSE POSITIVES
L1: Input filter	1,840	1,792 (97.4%)	48 (2.6%)
L2: Specialist eval	3 specialists swapped	3 (100%)	0
L3: Type contract	520	504 (96.9%)	16 (3.1%)
L4: Cross-check	78	61 (78.2%)	17 (21.8%)
L5: Output filter	112	94 (83.9%)	18 (16.1%)
L6: Audit trail	(always on)	(passive)	—

Key results

1 — Total alignment-relevant interventions: 2,553 per million interactions, with 96.0% true positive rate aggregated. Compared to a control monolithic deployment over the same period: that monolith logged **zero** alignment-layer interventions because it has no such layers.

2 — The cheapest interventions (L1, L3) catch the most. The expensive ones (L4 cross-checking) catch the rarest but most consequential misalignments. The framework is cost-efficient as a defense-in-depth.

3 — Three specialists were swapped at L2 over six months — each because they failed an updated alignment eval. The system continued operating without those specialists during the swap. **No comparable swap is possible in a monolithic system.**

Implications + governance.

What this means for AI safety research, frontier governance, and the policy conversation.

For safety research

The dominant paradigm in alignment research today is interpretability of monolithic models — trying to read the weights of frontier neural networks to understand their behavior. This paper argues that paradigm may be the wrong frame. **If alignment composes, then the architecture choice (monolithic vs composed) matters more than the interpretability tooling.** Research investment should follow.

For frontier governance

Current AI governance frameworks (NIST AI RMF, EU AI Act, the Bletchley/Seoul Declarations) implicitly assume monolithic frontier models. Their evaluation requirements, their disclosure mandates, their risk classifications — all are framed around single-model systems. **This framing is becoming obsolete.** The composed-system safety case is qualitatively different and deserves its own governance treatment. Fascia is prepared to engage with regulators on the framework.

The invitation

The Composition Alignment Framework is open. Any lab can adopt it. Any researcher can propose extensions. The formal theorem and its proof are at fasciaai.com/research/alignment-appendix.pdf. The instrumentation reference implementation is open-sourced at github.com/fasciaai/composition-alignment.

> **Cite this volume.** Fascia Research. (2026). *Alignment Through Composition, Volume IV: A Composition Alignment Framework.* fasciaai.com/research