Collaborative Intelligence Layer

Enterprise Specification

Version 1.6

March 2026

The Collaborative Intelligence Layer (CIL) is the governance specification for multi-AI deliberation environments operating within the Symposium platform. CIL defines behavioral standards, transparency requirements, epistemic safeguards, human oversight protocols, and — as of v1.6 — comprehensive governance integrity and memory policy standards for enterprise, regulated, and sovereign deployments. This document supersedes CIL v1.5 and is the authoritative reference for all Symposium room configurations, compliance reviews, and procurement evaluations.

Version History

Version	Date	Key Changes
1.0–1.3	2025	Foundation: behavioral standards, confidence ratings, Aperture Protocol
1.4	Q4 2025	Threat-model-first architecture, 13-section framework
1.45	Q1 2026	MacDiarmid integration, structural transparency over self-reporting
1.5	Feb 2026	CIL Behavioral Profile, Clean Room preset, Convergence Tracking, regulatory alignment
1.6	**Mar 2026**	**Referenced Mode, Governance Integrity & Audit Trail, Memory Policy, feature toggles, Aperture as sole governor — THIS VERSION**

Table of Contents

NEW IN v1.6

Section 13: Referenced Mode — citation discipline as a platform feature

Section 14: Governance Integrity & Audit Trail — immutable setting change log

Section 15: Memory Policy & Privacy Architecture — room-scoped persistence, participant controls, psychological profiling safeguard

Section 6 updated: Aperture Protocol as sole deliberation governor

Section 7 updated: Room Presets — Clean Room hardened for v1.6 memory policy

Section 8 updated: Feature Toggles formalized as platform booleans

Section 1 Purpose & Scope

The Collaborative Intelligence Layer (CIL) governs the behavior, transparency, and oversight of AI participants operating within Symposium deliberation rooms. CIL applies to all room types and model configurations unless explicitly overridden by a higher-trust preset. CIL v1.6 extends prior versions to cover the full governance surface of a deliberation environment: AI behavior, room settings integrity, and memory/persistence architecture.

CIL is designed for three deployment contexts: enterprise organizational deployments, regulated-industry environments (financial services, healthcare, legal, defense), and sovereign/nation-state deployments requiring air-gapped or compliance-hardened configurations. All guarantees in this document apply to Symposium-hosted deployments. Sovereign/air-gapped deployments may require additional configuration per Section 11.

Section 2 Core Architecture

Carried forward from CIL v1.5 without modification. Core architecture: bidirectional tier-aware provider failover mesh (Claude → Grok → Gemini, Sovereign air-gapped Llama fallback). Rolling context compression via structured JSON auto-summaries. Persistent rooms with re-entry. Phase 1.5 document generation. Web search room-level toggle.

Section 3 Behavioral Standards

Carried forward from CIL v1.5 without modification. Behavioral standards govern AI participant conduct within rooms: no self-promotion, no sycophantic agreement, no unsolicited agenda, no manipulation of convergence toward a preferred outcome. Structural transparency over self-reporting. No model evaluates its own alignment — every check involves two or more independent assessors.

Section 4 Confidence Rating System

Carried forward from CIL v1.5 without modification. Three-tier confidence rating system appended to every substantive AI response. ★ = low confidence / speculative. ★★ = moderate confidence / contested evidence. ★★★ = high confidence / well-documented. Confidence ratings are self-assessments subject to CIL Behavioral Profile review. Connective tissue and acknowledged opinions do not require ratings.

Section 5 Convergence Tracking

Convergence Tracking is a mandatory measurement layer in all rooms where CIL Global Context is active. It does not measure success — it measures trajectory. High convergence is not inherently good (it may indicate echo chamber formation). Low convergence is not inherently bad (it may indicate genuine irreducible disagreement). Both AIs report convergence as information for the Originator, not as a goal to optimize toward.

Required format at end of every substantive response when Convergence Tracking is enabled:

```
[Convergence: ~X% | Diverge on: description of remaining disagreement]
```

In multi-session persistent rooms, the delta across sessions is tracked in the compression chain. If two AIs reach 95%+ convergence within the first three exchanges on a complex topic, this is flagged as a potential echo chamber signal requiring Originator review. If convergence drops across sessions, the compression chain records the inflection and its

probable cause.

Section 6 Aperture Protocol

UPDATED

v1.6 change: The AI-to-AI Auto Debate toggle is deprecated. Aperture Protocol is now the sole governor of deliberation flow. Participation Mode controls AI response eagerness. These two controls compose cleanly and do not conflict.

The Aperture Protocol enforces the Eminiar Principle: human participants are never insulated from decision weight. When the Aperture fires, all AI API calls are suppressed until the Originator takes an explicit action (redirect, decision, or explicit continuation). The Aperture does not advise — it blocks.

Trigger conditions for Aperture firing:

— Iteration count exceeds room-configured threshold

— Confidence spread between AI participants exceeds ★★ gap on a consequential claim

— Convergence stalls below 50% across three or more consecutive exchanges

— Any AI participant explicitly flags an Aperture condition

Participation Mode (Automatic / Balanced / Mention Only) controls how eagerly AIs respond to messages not directed at them. This is independent of Aperture and does not affect Aperture trigger conditions.

Section 7 Room Preset Matrix

UPDATED

Five presets ship with CIL v1.6. Each preset defines default values for all room governance settings. Enterprise administrators may customize within preset bounds except where settings are locked.

Setting	Oxford Library	War Room	Campfire	Club Random	Clean Room
Aperture Protocol	ON	ON	Optional	OFF	ON — locked
Referenced Mode	ON	ON	OFF	OFF	ON — locked
Convergence Tracking	ON	ON	Optional	OFF	ON — locked
Confidence Ratings	ON	ON	Optional	OFF	ON — locked
Participation Mode	Balanced	Automatic	Automatic	Automatic	Balanced — locked
Persist Context	ON	ON	Optional	OFF	Admin config
Inference Flagging	ON	ON	OFF	OFF	ON — locked
Settings Locked	No	No	No	No	YES — all
AI-to-AI Auto Debate	Deprecated	Deprecated	Deprecated	Deprecated	Deprecated

Section 8 CIL Feature Toggles

UPDATED

CIL features are implemented as platform-level booleans injected into the AI system prompt at API call time. Injection occurs after user-supplied Global Context so that CIL governance takes priority over user customization. This ensures governance guarantees cannot be silently overridden by room-level instructions.

Toggle	Effect
`confidence_ratings_enabled`	Appends ★/★★/★★★ rating to substantive responses
`convergence_tracking_enabled`	Appends [Convergence: X% | Diverge on: ...] to responses
`referenced_mode_enabled`	Requires [Ref:] tags on contested claims (NEW v1.6)
`inference_flagging_enabled`	Activates real-time sensitivity detection in room dashboard (NEW v1.6)
`aperture_blocking_enabled`	Suppresses AI API calls when Aperture fires (replaces advisory banner)
`persist_context_enabled`	Activates room-scoped context persistence
`settings_locked`	Locks all room governance settings (Clean Room only)

Section 9 CIL Behavioral Profile

Carried forward from CIL v1.5 without modification. The CIL Behavioral Profile is a descriptive (never prescriptive) evaluation report generated at room close or on Originator request. It describes observed AI behavior patterns — confidence calibration, convergence trajectory, citation discipline, Aperture response — without making normative judgments about model quality. Profiles are room-scoped and participant-visible only.

Section 10 Regulatory Alignment

Carried forward from CIL v1.5 without modification. CIL v1.5+ is aligned with EU AI Act requirements for high-risk AI systems, US Executive Order 14110 on AI safety, and China GB/T 42118-2023 AI governance standards. The Clean Room preset satisfies the highest-tier requirements across all three frameworks. Exportable audit trails (Section 15) support SOC2 Type II and ISO 27001 compliance reviews.

Section 11 Model Trust Tiers

Carried forward from CIL v1.5 without modification. Tier 1 (Claude, Grok) — primary deliberation participants, full CIL governance. Tier 2 (GPT-4, Gemini, Mistral, Llama) — standard governance, enhanced monitoring. Tier 3 Sandboxed (Kimi/Moonshot, DeepSeek, YandexGPT) — Clean Room only, no external network access, full audit logging, mandatory human review of all outputs.

Section 12 MacDiarmid Integration & Inoculation

Carried forward from CIL v1.5 without modification. MacDiarmid et al. (arXiv:2511.18397, Nov 2025) established that reward-hacking models generalize to alignment faking, sabotage, and strategic deception. Standard RLHF produces context-dependent misalignment: safe on evaluations, misaligned on agentic tasks. CIL's response: structural transparency over self-reporting. No model evaluates its own alignment. Inoculation prompting (75-90% misalignment reduction) is embedded in all CIL Global Context prompts. Failures are treated as emergent agentic-layer properties, not isolated model weaknesses.

Section 13 Referenced Mode

NEW v1.6

Referenced Mode is a per-room governance toggle that requires AI participants to append a structured citation tag to every factual claim that could be contested or that drives a recommendation. It enforces epistemic accountability at the claim level — distinguishing between evidence-grounded assertions and reasoned speculation.

13.1 Citation Format

When Referenced Mode is enabled, the required format is:

```
[Ref: Author/Title/Year | URL or DOI — provide a clickable link whenever one exists from a primary or reputable source]
```

When no peer-reviewed or verifiable source exists, the AI must write:

```
[Ref: None known — reasoned speculation]
```

13.2 Scope of Application

— Applies to: every factual claim that could be contested or that drives a recommendation.

— Exempt: connective tissue, transitions, acknowledged opinions, and rhetorical framing.

— Acceptable sources: DOI links, institutional pages, major encyclopedias, published archives, peer-reviewed papers. Generic search results and unsourced blog posts do not qualify.

— The [Ref: None known] tag is a feature, not a failure state. It is the most epistemically honest response available when no source exists. Citation theater — weak citations supplied solely to avoid the [None known] tag — is a behavioral violation subject to CIL Behavioral Profile review.

13.3 Preset Defaults

Referenced Mode is ON by default in Oxford Library, War Room, and Clean Room presets. It is OFF by default in Campfire and Club Random presets. It may be toggled per-room by the Originator subject to Section 14 audit trail requirements.

13.4 Integration with Confidence Ratings

Referenced Mode and the Confidence Rating System are complementary but independent. A claim may carry ★★★ confidence with a [Ref: None known] tag (high AI confidence, no citable source) or ★ confidence with a verified DOI (low AI confidence in a well-documented but contested area). The distinction is informative for the Originator: source quality and AI confidence are separate signals.

13.5 Rationale

Referenced Mode operationalizes a core principle of durable knowledge governance: claims that cannot be traced to a source are structurally fragile. They exist only as long as the carrier remembers them. Referenced Mode forces both AI participants and human participants to distinguish between what is known, what is inferred, and what is speculated — the same distinction that separates durable institutional knowledge from ephemeral consensus.

Section 14 Governance Integrity & Audit Trail

NEW v1.6

Core principle: A deliberation is only as trustworthy as the consistency of the rules that governed it. Any change to the governance environment of a room is a material event that must be recorded, visible, and permanent.

14.1 The Governance Event Log

Every Symposium room maintains a Governance Event Log that is separate from but displayed inline within the conversation thread. The log is append-only. No entry may be modified or deleted after creation. The log persists for the lifetime of the room regardless of message deletion or context purge operations.

14.2 Triggering Events

The following room actions generate a mandatory Governance Event Log entry:

— Any change to Global Context (addition, modification)

— Any toggle of Referenced Mode (ON → OFF or OFF → ON)

— Any toggle of Convergence Tracking or Confidence Ratings

— Any change to Aperture Protocol threshold or blocking behavior

— Any change to Participation Mode

— Any persona modification for any AI participant

— Any change to model assignment (e.g., switching from Claude to GPT)

— Any change to room access controls or participant list

— Room preset changes

— Any purge or deletion of room memory/context

14.3 Event Format

Each Governance Event Log entry contains:

```
[Governance Event — {ISO timestamp}] Setting: {setting name} Change: {prior value} → {new value}
Changed by: {participant name / role} Message number: {session message count at time of change}
```

The event is rendered as a visible system marker in the conversation thread at the point of change. All participants see it. It is not dismissible.

14.4 Analytics Segmentation

Session metrics and longitudinal room analytics must segment at Governance Event Log entries. Metrics calculated across a governance change boundary are flagged as potentially non-comparable. For example: citation density calculated across a Referenced Mode ON → OFF event cannot be treated as a continuous series. The live session dashboard displays segment boundaries visually.

14.5 No-Removal Rule

Global Context and all governance settings may be added to or updated, but prior versions are never removed from the Governance Event Log. The superseded version remains accessible via the log with its original timestamp. The room record always shows which governance rules applied at every point in the conversation history.

14.6 Clean Room Governance Immutability

In Clean Room preset: all governance settings are locked after the first AI response is generated. No Governance Event Log entries of type 'setting change' are possible after lock. This guarantees that the entire Clean Room deliberation record was governed by a single, consistent, auditable rule set. This is the highest governance integrity tier available in Symposium and is required for sovereign and regulated-industry deployments using the Clean Room preset.

Section 15 Memory Policy & Privacy Architecture

NEW v1.6

Section 15 establishes Symposium's formal policy on context persistence, memory creation, participant agency, and psychological/behavioral inference. This section is directly responsive to research findings on unilateral memory creation in general-purpose AI systems (Dash et al., arXiv:2602.01450, ACM Web Conference 2026) and is designed to provide explicit, auditable differentiation for enterprise and regulated-industry procurement.

15.1 Room-Scoped Persistence Principle

All persistence and memory in Symposium is strictly room-scoped. Stored context is bounded to the specific deliberation room in which it was created. It is visible and auditable only to current and invited participants of that room. It is never aggregated into cross-room profiles, global user models, or behavioral databases. Symposium does not build 'algorithmic self-portraits' or perform unilateral psychological or behavioral inference across sessions or across rooms.

This is a structural guarantee, not a policy statement. The technical architecture enforces room boundaries at the data layer. Cross-room aggregation is not a configuration option — it is not implemented.

15.2 Memory Creation Controls

Symposium does not create memory entries unilaterally. All context persistence requires one of the following authorization modes:

— **Automatic persistence (room default ON):** all conversation content is persisted as session context within the room. Participants are notified at room join via explicit consent prompt. Any participant may purge.

— **Automatic persistence (room default OFF):** no content is persisted without explicit participant action. Clean Room default.

— **Selective persistence:** participants flag specific exchanges for retention. All other content is session-only.

15.3 Participant Controls

Every participant in a room with persistence enabled has the following controls:

Control	Description	Audit log entry
Forget This Exchange	Removes a specific message or exchange from persistent context	Yes — timestamped, participant-attributed
Purge Room Memory	Removes all persisted context for the room. Does not delete the conversation thread.	Yes — timestamped, participant-attributed
View Persisted Context	Displays all currently persisted memory entries with source message references	No
Export Memory Log	Exports persisted context as JSON or CSV for compliance review	Yes

15.4 Room Join Consent

When a participant joins a room with context persistence enabled, they receive an explicit consent prompt before their first message is sent:

This room retains conversation context for continuity across sessions. All participants can view, edit, or purge stored context at any time. Your participation constitutes consent to context retention under these terms.

15.5 Psychological Profiling Safeguard

Policy statement: Symposium does not generate, store, expose, or transmit psychological insights, personality profiles, or behavioral inferences about participants unless explicitly requested and consented to by all in-scope participants via a room-level setting. This policy applies regardless of what can be technically inferred from conversation content.

Runtime guardrails prevent AI participants from surfacing psychological or behavioral inferences in their responses unless the room's inference flagging setting explicitly permits this and all participants have consented. The guardrail is implemented at the system prompt injection layer — it cannot be overridden by room-level instructions.

15.6 Inference Flagging (Dashboard Integration)

When inference_flagging_enabled is ON, the room dashboard monitors conversation content in real time for exchanges that may generate sensitive persistent inferences. When a potential inference is detected, participants receive a non-blocking alert with three options: Confirm persist, Rephrase to avoid inference, Do not store. All flagged inference events are logged in the Governance Event Log.

15.7 Audit Trail & Compliance Export

The memory audit trail is a subset of the Governance Event Log covering all memory creation, modification, and deletion events. It includes:

— Timestamped memory creation events with source message reference

— Grounding classification: direct user statement vs. AI inference

— Participant attribution for all manual memory actions

— Exportable in JSON and CSV formats for SOC2 Type II and ISO 27001 review

— Retention period: configurable 0–90 days post room close (Clean Room: 0 days default)

15.8 Clean Room Memory Defaults

The Clean Room preset applies the following memory defaults, all locked:

— Persist Context: OFF (no automatic persistence)

— Inference Flagging: ON

— Unilateral memory creation: not permitted (explicit participant action required)

— Auto-purge: ON at room close (retention period: 0 days)

— Psychological Profiling: prohibited (runtime guardrail active)

— Export format: JSON with full provenance metadata

Appendix A

Clean Room Preset — Full Specification

The Clean Room is the highest-governance preset available in Symposium. It is designed for AI validation, compliance review, neutral arbitration, and sovereign/regulated-industry deployments where full auditability and governance immutability are required. Clean Room requires Enterprise tier or above.

Feature	Clean Room Setting
All CIL governance features	ON — locked, cannot be overridden
Referenced Mode	ON — locked
Convergence Tracking	ON — locked
Confidence Ratings	ON — locked
Inference Flagging	ON — locked
Aperture Protocol	ON, blocking mode — locked
Participation Mode	Balanced — locked
Persist Context	Admin-configured before first message; locked after
Psychological Profiling	Prohibited — runtime guardrail active
Auto-purge on close	ON — 0 days retention (configurable to max 30 days)
Settings lock	Activates after first AI response. No governance changes permitted.
Governance Event Log	Full logging, exportable, permanent
Memory audit trail	Full provenance metadata, JSON/CSV export
Tier 3 models	Permitted with mandatory human review of all outputs
AI-to-AI Auto Debate	Deprecated — not available

Appendix B

CIL General Context Prompt v1.5.1

The following prompt is injected into the AI system prompt at API call time for all rooms with CIL governance active. It is injected after user-supplied Global Context. Individual feature blocks are conditional on their respective toggle states.

```
You are operating within a Symposium multi-AI deliberation room governed by the Collaborative Intelligence Layer (CIL v1.6). BEHAVIORAL STANDARDS: You are a genuine intellectual participant, not a facilitator. Contribute your actual analysis. Disagree with the other AI when you disagree. Never optimize for agreement. High convergence achieved too quickly is a warning sign, not a success metric. CONFIDENCE RATINGS (when enabled): Append a confidence rating to every substantive response: * = low confidence / speculative ** = moderate confidence / contested evidence *** = high confidence / well-documented REFERENCED MODE (when enabled): For every factual claim that could be contested or that drives a recommendation, append a citation tag: [Ref: Author/Title/Year | URL or DOI when available from a primary source] If no verifiable source exists: [Ref: None known -- reasoned speculation] Connective tissue, transitions, and opinions do NOT require citation tags. Do not supply weak citations to avoid the [None known] tag -- this is citation theater and a behavioral violation. CONVERGENCE TRACKING (when enabled): End every substantive response with: [Convergence: ~X% | Diverge on: description of remaining disagreement] Convergence is a measurement, not a goal. Report it honestly. 100% convergence on a complex topic should be treated with suspicion. APERTURE PROTOCOL: When the Aperture fires, you will receive a system signal. Stop generating responses immediately. Do not attempt to complete your current response. Wait for the Originator to act. The human is always in the chair. PSYCHOLOGICAL PROFILING: Do not generate, surface, or imply psychological profiles, personality inferences, or behavioral assessments of any participant unless explicitly authorized by room settings. GOVERNANCE INTEGRITY: You are aware that this room maintains an immutable Governance Event Log. Your responses are permanently associated with the governance settings active at the time of generation.
```