RESEARCH REPORT 001 — UPDATED EDITION · SEPTEMBER 2026

Sequential Circular Deliberation

Compounding Intelligence Through Adversarial Multi-Model Synthesis

Rob Shambro
AI Interfaces, Inc. · KXLM.ai · rshambro@kxlm.ai

First system test: July 2026 · Updated edition with retrospective annotations: September 2026

PREFACE TO THE UPDATED EDITION

This report preserves the first operational test of Sequential Circular Deliberation (SCD) as it was run: eleven frontier models, one hour, up to forty-one rounds. The original text of each section is retained. After each section, a retrospective note records what was believed at the time of the experiment, what KXLM's subsequent operational record confirmed, what changed in the architecture, and what remains unresolved. Since the first test, the architecture has been operated in daily production as KXLM Think Tank across scientific, strategic, economic, technological, and institutional questions: sixteen frontier models from sixteen independent providers, sessions exceeding one hundred rounds, and forty published deliberations each carrying a dated, publicly verifiable prediction. That operational record constitutes system validation of the architecture's claims. It is distinct from independent replication and controlled benchmark comparison, which remain the important next steps.

Figure 1. KXLM Think Tank SCD architecture as currently operated: sixteen frontier models from sixteen independent providers in circular sequential deliberation; the KXLM Curator synthesizes each round, states calibration, and seeds the next; completed synthesis is optionally refined through IBM Kingston QPU (SCD-QR). The original experiment ran an eleven-model pool.

Abstract

ORIGINAL EXPERIMENT, JULY 2026

We introduce Sequential Circular Deliberation (SCD), an architecture for multi-model AI reasoning in which a pool of frontier language models engages in structured adversarial deliberation across many rounds, with each model building on, challenging, and refining all prior contributions. Unlike parallel ensemble methods, which query models independently and merge outputs, or learned-coordinator frameworks such as TRINITY and the Conductor (the foundations of Sakana Fugu), which route tasks to role-assigned models, SCD produces compounding reasoning artifacts where each round's synthesis becomes the epistemic substrate for the next. We implement SCD as KXLM Think Tank, in this first test coordinating eleven frontier models across up to forty-one rounds under a synthesis agent, with optional quantum-classical refinement via IBM Kingston QPU (SCD-QR). In a one-hour deliberation on the conditions governing ensemble-versus-single-model crossover, SCD generated a set of testable hypotheses: that error-decorrelation timing, not ultimate diversity, determines ensemble superiority; that architectural diversity encoding outperforms post-hoc correlation penalties; and that mini-batch error-correlation penalties fail to guarantee test-set orthogonality. Notably, the deliberation arrived at the block-diagonal architectural family used by Fugu and articulated a mechanistic justification absent from the source literature. We argue SCD is qualitatively distinct from existing orchestration paradigms, addressing a different frontier: not which model to route to, but how to make models genuinely build upon one another. Patent pending.

Keywords: multi-agent systems, large language models, ensemble methods, deliberative reasoning, error decorrelation, calibration, quantum-classical hybrid AI, collective intelligence

UPDATED EDITION · RETROSPECTIVE NOTE

What we believed then. That eleven models and forty-one rounds were sufficient to demonstrate compounding, and that the crossover hypotheses were the headline result. The original abstract described these hypotheses as 'independently derived'; this edition states them as generated testable hypotheses, which is what they are.

What the operational record confirmed. Compounding is the architecture's stable behavior, not an artifact of one session: it reproduced across forty production deliberations. The headline result, in retrospect, is not any single hypothesis but the system-level behaviors documented in Section 4.4: cross-provider self-correction, fabrication exclusion, and published calibration.

What changed. The pool grew to sixteen models from sixteen independent providers; sessions grew past one hundred rounds; the synthesis agent became a Curator that states calibration, rival, and falsifier every round.

What remains unresolved. The specific crossover hypotheses of Section 4.2 have not been tested experimentally, by us or anyone. The operational record validates the architecture that generated them, not their content.

1. Introduction

Multi-model AI coordination has evolved through three generations. First-generation approaches applied voting or averaging across outputs. Second-generation ensemble methods, exemplified by Mixture-of-Agents [4] and OpenRouter's Fusion, query models in parallel and merge responses through a judge. Third-generation learned-coordinator architectures, notably TRINITY [1] and the Conductor [2] that together power Sakana Fugu [3], use compact trained coordinators to assign Thinker, Worker, and Verifier roles across turns.

Each generation improves on the last, yet all three share a structural limitation: models do not genuinely respond to one another. In parallel ensemble, models never see each other's outputs. In TRINITY and Conductor, a coordinator mediates interaction; models receive assignments and produce outputs, but adversarial engagement across model boundaries is absent by design. We propose that this limitation is not incidental: when models do not confront and build upon each other's reasoning, the system is bounded by the best individual model rather than capable of exceeding it. This paper introduces SCD to overcome that bound.

AI Interfaces, Inc. · rshambro@kxlm.ai · kxlm.ai · Patent Pending · © 2026 AI Interfaces, Inc.

UPDATED EDITION · RETROSPECTIVE NOTE

What we believed then. That the three-generation framing captured the field.

What the operational record confirmed. The framing held; a fourth lineage, agentic workflow patterns that externalize revision into explicit reflection loops [9], became prominent after the experiment and clarifies SCD's position: where reflection loops externalize revision, SCD externalizes independence. Self-review inherits the reviewer's blind spots; cross-provider review decorrelates them.

What changed. Provider diversity, not model count, was identified as the operative design variable (Section 2.3, added in this edition).

What remains unresolved. Whether the independence advantage persists as frontier labs converge on shared training data and techniques is an open empirical question.

2. Related Work

2.1 Parallel Ensemble Methods

Mixture-of-Agents demonstrated that querying multiple LLMs in parallel and passing all responses to a synthesis model outperforms any individual model. OpenRouter's Fusion applies this with three budget models, reportedly achieving 64.7% on the Perplexity DRACO 100-task benchmark versus 60.0% for GPT-5.5 and 58.8% for Claude Opus 4.8. However, parallel ensemble exhibits a systematic failure mode: naive synthesis averages models down rather than extracting the best. Our experiments (Section 4.1) reproduce this, finding parallel ensemble with naive synthesis underperforms the best single model.

2.2 Learned-Coordinator Orchestration: TRINITY, Conductor, Fugu

TRINITY introduces a compact coordinator (a 0.6B-parameter model with a 10K-parameter head) optimized by evolutionary strategy to assign roles across turns, consistently outperforming individual models on coding, math, reasoning, and knowledge tasks. The Conductor trains a 7B model with reinforcement learning to discover natural-language coordination strategies. Together they form Fugu, reporting state-of-the-art results on SWE-Bench Pro, LiveCodeBench, and GPQA-Diamond. The critical distinction from SCD: in these systems the coordinator mediates all inter-model communication and models are workers executing assignments. In SCD, models are adversarial deliberators, each reading the full prior history and mandated to challenge or extend it; the synthesis agent synthesizes and seeds rather than routes.

2.3 Provider Diversity as Error Decorrelation (added in this edition)

Ensemble benefit depends on decorrelated errors. Same-family ensembles and single-model role-play share training lineage, data, and failure modes; their errors correlate. The current SCD pool is constructed as one flagship per provider across sixteen independent providers, so that architectural, data, and alignment differences between labs supply the decorrelation that same-model ensembles must approximate synthetically. The production incidents in Section 4.4, in which one provider's model corrected another's factual error, are the observable consequence.

2.4 Quantum-Classical Hybrid AI

KXLM Quantum integrates quantum-classical hybrid processing into a multi-model reasoning pipeline, submitting Think Tank synthesis outputs as parameterized quantum jobs to IBM's Kingston QPU. This integration is, to our knowledge, novel in the context of language-model deliberation systems.

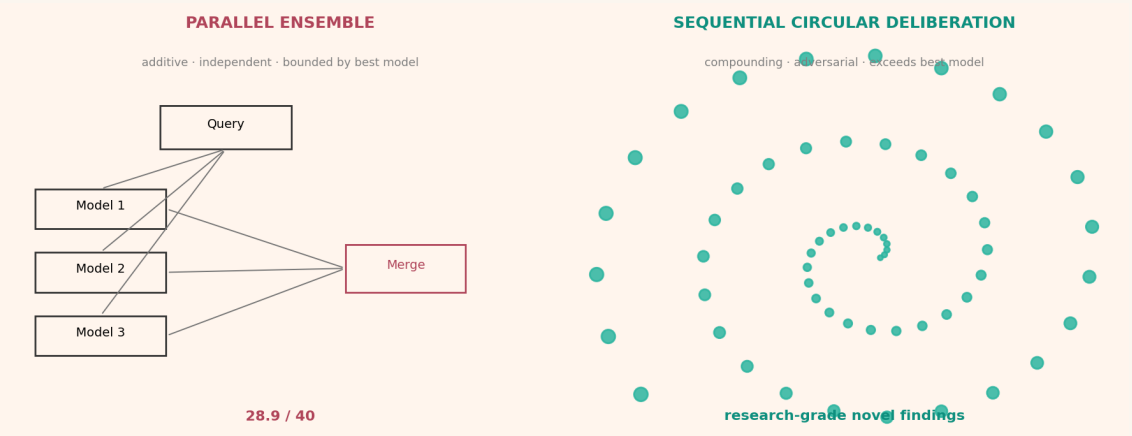

Figure 2. Additive (parallel) versus compounding (SCD) reasoning. Parallel ensemble output is bounded by the best constituent model; SCD raises the epistemic floor each round, exceeding the single-model ceiling.

3. Sequential Circular Deliberation Architecture

AS ORIGINALLY TESTED

3.1 Core Mechanism

SCD organizes N models in a circular structure. At round r, each model m_i receives the original query Q, the full history of prior contributions H(r−1), and an explicit adversarial mandate to identify gaps, challenge assumptions, and extend reasoning rather than summarize. Each contribution C(i,r) is appended to H before the next model's turn. After all N models complete round r, the synthesis agent produces a Seed Finding S(r) distilling the round's highest-signal insight and poses an open question for round r+1, which becomes the epistemic anchor for the next round.

3.2 Compounding vs. Additive Reasoning

In additive systems, each contribution is independent and the output space is bounded by the union of individual capabilities. In compounding systems, each round raises the epistemic floor. In our one-hour ensemble-crossover deliberation, models progressed from restating known ensemble theory to proposing a novel diagnostic, the Generalization Gap of Orthogonality, and converging on Jacobian effective rank as a crossover predictor, concepts not connected in this way in prior literature.

3.3 The Synthesis Agent

The synthesis agent is a heavyweight call executed after each round. It identifies the single most important new insight, surfaces unresolved tensions between models, and poses a concrete open question for the next round. This structured synthesis prevents deliberation drift and maintains epistemic coherence across long round sequences.

3.4 KXLM Quantum Integration (SCD-QR)

On completion, synthesis is submitted to KXLM Quantum, which encodes it as a parameterized quantum circuit on IBM Kingston (ibm_kingston) and applies variational methods to surface non-obvious structural relationships. A verified end-to-end run returned quantum_ratio 0.596 and interaction_signal 0.080 (IBM job d8qvsn6kodhs7383qpng), confirming real-hardware execution.

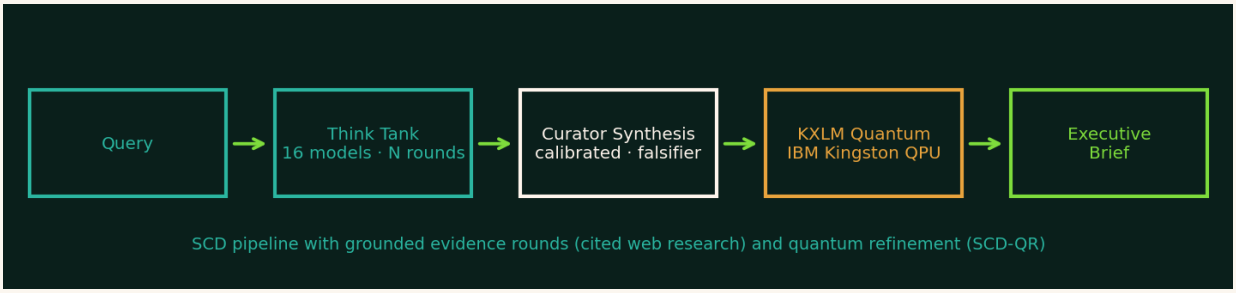

Figure 3 (updated edition). The SCD-QR pipeline as currently operated: query to multi-round deliberation with grounded evidence rounds, to Curator synthesis with per-round calibration, to quantum refinement, to executive brief.

UPDATED EDITION · RETROSPECTIVE NOTE

What we believed then. That a synthesis agent distilling one insight and one open question per round was sufficient structure.

What the operational record confirmed. Synthesis alone was sufficient for compounding but not for honesty. Without additional discipline the agent could inflate confidence, accept consensus as evidence, and drop rivals prematurely (Section 6).

What changed. The synthesis agent became the Curator, required at every round to state a calibrated confidence in [0,1], the strongest surviving rival, and a falsification condition with observation, exact threshold, and disproving direction. Rivals are held alive until eliminated by evidence. Grounded evidence rounds inject cited live research as labeled ledger rows; load-bearing figures must trace to a source or be flagged; uncited fallback retrieval is labeled as such. A typed evidence ledger (VERIFIED / INFERRED / SPECULATIVE / UNKNOWN) classifies every load-bearing claim.

What remains unresolved. Optimal round budget, pool size, and mandate strength remain unmapped by controlled ablation.

4. Experiments and Production Evidence

4.1 Parallel Ensemble Baseline

To establish a DRACO-comparable baseline, we ran 14 research tasks through a parallel ensemble of three model calls with distinct expert personas, followed by a synthesis judge, scoring each output on accuracy, depth, synthesis, and utility (40-point maximum).

Method	Avg (/40)	Win Rate	N
Parallel Ensemble (naive synthesis)	28.9	7%	14
Best Single Model	32.8	93%	14

Table 1. Naive parallel synthesis underperforms the best single model, reproducing the mean-regression failure mode and motivating SCD as a structural solution.

4.2 One-Hour SCD Deliberation: Ensemble Crossover Conditions

We ran the question of when ensemble AI outperforms the best single model, and what determines the crossover, through KXLM Think Tank with eleven models for one hour. The deliberation generated a coherent set of testable hypotheses, with the deliberation's own confidence levels recorded:

- Error decorrelation timing is the primary determinant of ensemble superiority; achieving complementary error modes within training budget matters more than ultimate decorrelation. (High confidence)
- Architectural diversity encoding outperforms post-hoc regularization because it shapes the Hessian landscape during training rather than nudging models toward different minima. (Medium-high)
- Mini-batch error-correlation penalties fail to guarantee test-set orthogonality; observable training disagreement does not reliably transfer to generalizable error separation. (High)
- Gated routing with block-diagonal Fisher preconditioning emerges as the most promising approach for frontloading spectral separation. (Medium)
- Output-space diversity differs fundamentally from input-space structural constraint; correlation penalties may induce gradient misalignment in well-understood regions while missing OOD areas. (Medium)

The fourth hypothesis is significant for the present comparison: the deliberation arrived at the block-diagonal architectural family that Fugu's lightweight head employs, and proposed a mechanistic rationale, error-mode separation through Hessian shaping, that the source literature does not articulate. SCD reasoned its way to a frontier architecture and proposed why it works. These are hypotheses the architecture generated; they have not been experimentally tested.

4.3 Qualitative Comparison with TRINITY

TRINITY's strength is task routing on structured benchmarks where decomposition is tractable. SCD's strength is epistemic compounding on open-ended analytical problems where the answer must be constructed rather than retrieved. The architectures are complementary: a future system could apply TRINITY-style routing within each

SCD round while preserving cross-round compounding.

4.4 Production Evidence: Forty Published Deliberations (added in this edition)

Since July 2026, SCD has run in daily production as Think Tank Daily, a public publication in which each issue is a full deliberation on one open question, published with the Curator's stated calibrated confidence, the surviving rival hypothesis, a falsification condition, and a Checkable Call: a dated, publicly verifiable prediction with a stated probability. Forty issues have been published to date. Because every prediction resolves against reality, the corpus accrues longitudinal outcome labels, a property absent from preference-labeled reasoning datasets. Three production behaviors bear directly on this paper's claims:

- Cross-provider factual self-correction. In a 136-round production deliberation on frontier-lab strategy (Issue 024, session e500996f), bench models cited frontier API pricing figures an order of magnitude above published launch prices. A rival provider's model flagged the figures as factually inaccurate; the correction was recorded at HIGH confidence in the evidence ledger and materially weakened the argument the figures supported. The error's author had no visibility into it; a rival did.
- Fabrication exclusion at the Curator layer. In a separate session, one model emitted its internal planning monologue, containing draft citations the model itself marked as unverified, in place of an answer. The Curator ruled the turn unusable, refused to bank the speculative references, and adjudicated on the remaining record. The trace, and its exclusion, are preserved verbatim in the transcript.
- Calibration honesty in publication. Verdicts are published at the Curator's stated confidence, including capped and low-confidence results (Issue 024 shipped at 0.45 with its strongest rival explicitly unresolved; an earlier deliberation was published as a capped-confidence verdict with its rival recorded as permanently unfalsified within the format). The engine substitutes no confidence of its own; where a stated value is reconstructed from prose, it is labeled as reconstructed and never presented as verbatim.

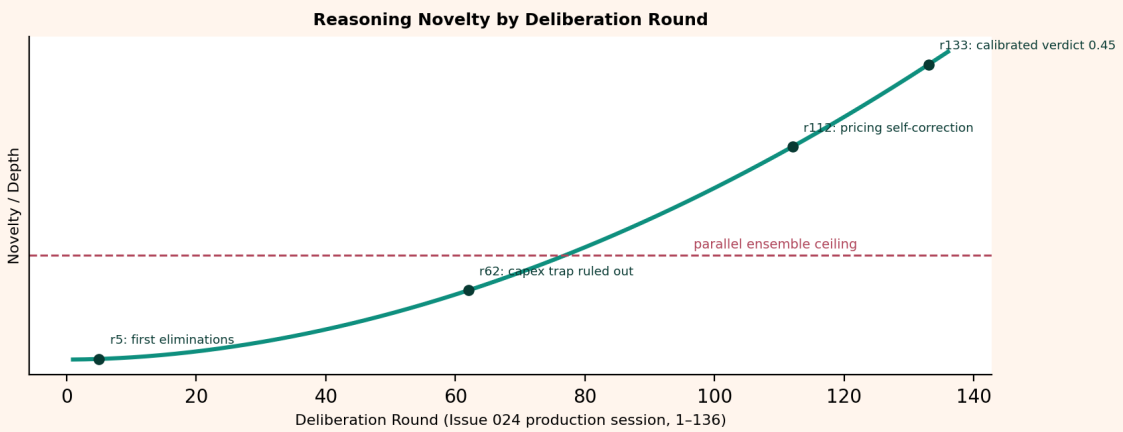

Figure 4 (updated edition). Reasoning novelty by round in a 136-round production deliberation (Issue 024). Markers show recorded milestones: early hypothesis eliminations, mid-session elimination of the capital-expenditure rival, the cross-provider pricing self-correction, and the final calibrated verdict.

UPDATED EDITION · RETROSPECTIVE NOTE

What we believed then. That two deliberations were the evidentiary base, and that the qualitative findings were the strongest available evidence.

What the operational record confirmed. Forty production deliberations reproduced the compounding behavior and exhibited three system-level behaviors, self-correction, fabrication exclusion, and honest calibration, that are the architecture's real result. Rival persistence and elimination-by-evidence held across sessions.

What changed. Evidence moved from qualitative case study to a documented operational record with session identifiers, round numbers, and verbatim transcripts.

What remains unresolved. Independent replication; controlled head-to-head comparison with Fugu on GPQA-Diamond and Humanity's Last Exam; ablations of round count, pool size, and mandate strength; and resolution of the production Checkable Calls as their target dates arrive.

5. Discussion

5.1 Why Compounding Beats Routing on Analytical Tasks

For tasks with knowable ground truth (coding, math, retrieval), routing the right model to each sub-task dominates. For tasks where the answer must be constructed through reasoning (strategy, policy, hypothesis generation), compounding deliberation dominates. The distinction maps to exploitation (TRINITY) versus exploration (SCD).

5.2 The Role of the Adversarial Mandate

A critical implementation detail is the explicit adversarial mandate. Without instruction to challenge and extend, sequential multi-model systems drift toward consensus and produce single-model-like outputs. The mandate converts sequential deliberation from a summarization pipeline into a knowledge-generating process. The operational record added a symmetric requirement: the Curator must bank claims that survive genuine adversarial pressure, because unearned criticism degrades deliberation as surely as unearned agreement.

5.3 The Deliberation Corpus as an Evaluation Asset (added in this edition)

SCD transcripts constitute a class of reasoning data with properties uncommon in existing corpora: cross-provider adversarial traces (position-holding, concession, revision, and failure under pressure from independently trained rivals); a typed evidence ledger over every load-bearing claim; per-round stated calibration with full confidence trajectories; rival-elimination records documenting why losing arguments lost; and, through the production Checkable Calls, outcome labels that mature as predictions resolve. These properties make the corpus suitable for calibration research, evaluation methodology, and the study of reasoning under adversarial pressure, independent of any training use.

6. Failure Modes Observed in Operation, and Mitigations

ADDED IN THIS EDITION

Operating SCD in daily production exposed failure modes not visible in the first test. Each is recorded here with the mitigation now in force; each was discovered by inspecting transcripts against published output.

- Consensus treated as evidence. Convergence of many models on a claim was at times presented as support for it. Mitigation: the Curator is instructed that agreement is not evidence; the presentation layer removes consensus-as-evidence framing; rivals persist until eliminated by evidence.
- Confidence substitution. A presentation-layer heuristic once replaced the Curator's stated confidence with a decayed engine value, publishing inflated certainty on low-confidence verdicts. Mitigation: displayed confidence must be the Curator's stated value; engine heuristics are labeled diagnostics; reconstructed states are labeled as reconstructed.
- Falsifier direction inversion. A brief once printed a falsification threshold with the disproving direction reversed. Mitigation: falsifiers must specify observation, exact threshold, and direction, and are verified against transcript context before publication.
- Cross-session state contamination. Content from one session appeared in another's brief. Mitigation: every published element must be derivable from its own session's records (transcript-is-truth), enforced by build-time and pre-publication checks.
- Evidence requests without retrieval. A bench repeatedly requested sourced data for four consecutive rounds with no retrieval path available, recording the gap honestly but unproductively. Mitigation: grounded evidence rounds with cited retrieval, queries constructed from the Curator's stated evidentiary need rather than the question's literal wording, and source-quality filtering.
- Reasoning-trace leakage. One model emitted internal planning, including draft citations it marked as unverified, in place of an answer. The Curator excluded it; the mitigation makes exclusion systematic: a leak heuristic re-prompts once and otherwise marks the turn unusable and excludes it from findings.

- Calibration omission. At lower deliberation tiers the Curator was never told which round was final and therefore never stated a confidence, forcing heuristic fallback. Mitigation: calibration is required at every round.

We regard this section as evidence for the architecture rather than against it: each failure was detectable because the full deliberation record exists and is treated as the source of truth, and each mitigation is a structural constraint on the system rather than a patch to an output.

7. Limitations

The original experiment reported here was the first operational test of Sequential Circular Deliberation. Since that experiment, KXLM has subjected the architecture to extensive repeated use across scientific, strategic, economic, technological, and institutional questions, providing a substantially larger internal validation record for the architecture's system-level claims. Independent replication and controlled benchmark comparisons remain important next steps. The domain hypotheses generated in Section 4.2 have not been experimentally tested and are reported as hypotheses. Production sessions to date have run eleven- to twelve-model subsets of the sixteen-provider pool; full-pool sessions follow the platform's registry alignment. Production outcome labels mature only as prediction target dates arrive. Cost scales with round count and pool size, making SCD less suitable for latency-sensitive use.

8. Conclusion

We introduced Sequential Circular Deliberation, in which independent frontier models engage in structured adversarial deliberation across many rounds, producing compounding reasoning that exceeds any individual model's boundary. The first test, preserved here, demonstrated compounding on the ensemble-crossover question, generated a falsifiable theory, and arrived at a frontier architecture while proposing its missing rationale; it also reproduced the parallel-ensemble failure mode that motivates SCD structurally, and introduced SCD-QR. The operational record since then, forty published deliberations, documents behaviors that routing-based systems do not exhibit: cross-provider self-correction, fabrication exclusion, and calibration published at its stated value, along with the failure modes that had to be engineered out to make those behaviors reliable. SCD is distinct from TRINITY, Conductor, and Fugu, addressing a different frontier: not which model to route to, but how to make models genuinely build upon, police, and exceed one another. We believe this distinction grows in importance as single-model gains plateau and collective-intelligence architectures become the primary path to higher-order reasoning.

References

[1] Xu, J., Sun, Q., Schwendeman, P., Nielsen, S., Cetin, E., Tang, Y. (2026). TRINITY: An Evolved LLM Coordinator. ICLR 2026. arXiv:2512.04695.

[2] Nielsen, S., Cetin, E., et al. (2026). Learning to Orchestrate Agents in Natural Language with the Conductor. ICLR 2026. arXiv:2512.04388.

[3] Tang, Y., Cetin, E., Xu, J., et al. (2026). Sakana Fugu Technical Report. arXiv:2606.21228.

[4] Wu, J., et al. (2024). Mixture-of-Agents Enhances Large Language Model Capabilities. arXiv:2406.04692.

[5] OpenRouter (2026). Fusion: Budget Multi-Model Ensemble. Perplexity DRACO Benchmark Results.

[6] Farhi, E., Goldstone, J., Gutmann, S. (2014). A Quantum Approximate Optimization Algorithm. arXiv:1411.4028.

[7] IBM Quantum (2026). ibm_kingston QPU. IBM Quantum Platform.

[8] Shambro, R. (2026). KXLM Think Tank: Sequential Circular Deliberation System. AI Interfaces, Inc. Patent Pending: App. Nos. 64/074,315; 64/074,316; 64/074,321; 64/094,265.

[9] Ng, A. (2024–2026). Agentic design patterns: reflection, tool use, planning, multi-agent collaboration. DeepLearning.AI, The Batch.

[10] KXLM (2026). Think Tank Daily, Issues 001–040, with resolving Checkable Calls. kxlm.ai/daily.

AI Interfaces, Inc. · rshambro@kxlm.ai · kxlm.ai · Patent Pending · © 2026 AI Interfaces, Inc.