MACIT: Coherent Internalization of Mathematical Theory and Agentic Operating Structures in Large Language Models

A MOUSSAIF AGILLM Theory and Fine-Tuning Architecture

Proposed Research Architecture — MOUSSAIF AGILLM / MACIT v1

1. Abstract and Contributions

MACIT proposes a dual-substrate view of AGILLM intelligence: a parametric language-model policy and an external Agentic OS must be trained and evaluated as one coherent system. This page develops the claim in operational rather than purely terminological form. The central engineering requirement is that a model trained on MOUSSAIF material must demonstrate transfer: it should apply the underlying operator when the names MACIT, MCD, QMCD, HAKIM, HyperSkill, or HyperLoop are absent from the prompt. Accordingly, evaluation must use paraphrases, out-of-domain tasks, counterexamples, and tool-grounded settings.

Formally, let S=(pitheta,Omega) denote the combined state of the learned policy and the external Agentic OS. For this component we track Transfer Value (T) and QMCD Score (QMCD). A candidate transition a is useful when it increases verified task value while preserving invariants. One generic objective is J(a|S,q_g)=α·Correctness+β·Transfer+γ·Coherence+δ·InformationGain−η·Risk−μ·Cost. The symbols are not universal constants; the paper requires empirical calibration and reports them as design variables rather than discovered laws.

The operational curriculum should include positive demonstrations, minimally different negative examples, and repair traces. For abstract and contributions, the main design elements are: weight-versus-OS partitioning; coherence as a first-class optimization target; difference-driven curricula; adversarial preference construction; bounded agentic reinforcement. A record is training-worthy only if its expected learning value is higher than its redundancy and uncertainty. This is where QMCD is used as a curriculum heuristic rather than as an unsupported claim of a universal cognitive metric.

A Demolition Court test injects the failure mode 'over-compression'. The learner must identify the violated assumption or interface, propose a bounded correction, and preserve memory retrieval precision. This makes the theory falsifiable at the artifact level: if a simpler SFT baseline performs equally well on held-out transfer, tool truth, loop termination, and regression tests, the added MACIT mechanism has not earned its complexity. Likewise, if externalization causes retrieval latency or brittle tool dependence that outweighs updateability and auditability, IAS thresholds must be changed rather than defended dogmatically.

The research implication is that AGILLM development should report a vector of capabilities instead of a single benchmark number. At minimum, studies should separate final-answer accuracy, operator transfer, uncertainty calibration, routing efficiency, provenance retention, external-tool truth, loop boundedness, reversible action, memory reconstruction, and retained general ability. The framework therefore treats fine-tuning as constrained systems engineering: model weights, data, runtime memory, tools, agents, evaluators, and governance must be co-designed and ablated.

2. Problem Statement

Fine-tuning a large model on a large theory archive can increase familiarity while still failing to produce the intended reasoning operators, tool discipline, or agentic behavior. This page develops the claim in operational rather than purely terminological form. The central engineering requirement is that a model trained on MOUSSAIF material must demonstrate transfer: it should apply the underlying operator when the names MACIT, MCD, QMCD, HAKIM, HyperSkill, or HyperLoop are absent from the prompt. Accordingly, evaluation must use paraphrases, out-of-domain tasks, counterexamples, and tool-grounded settings.

Formally, let S=(pitheta,Omega) denote the combined state of the learned policy and the external Agentic OS. For this component we track Coherence (C) and HyperDeployment (HD). A candidate transition a is useful when it increases verified task value while preserving invariants. One generic objective is J(a|S,q_g)=α·Correctness+β·Transfer+γ·Coherence+δ·InformationGain–η·Risk–μ·Cost. The symbols are not universal constants; the paper requires empirical calibration and reports them as design variables rather than discovered laws.

The operational curriculum should include positive demonstrations, minimally different negative examples, and repair traces. For problem statement, the main design elements are: memorization is not operator acquisition; theory names can become lexical shortcuts; volatile knowledge should remain external; training must preserve broad capability; agentic traces need explicit state semantics. A record is training-worthy only if its expected learning value is higher than its redundancy and uncertainty. This is where QMCD is used as a curriculum heuristic rather than as an unsupported claim of a universal cognitive metric.

A Demolition Court test injects the failure mode 'spurious correlation'. The learner must identify the violated assumption or interface, propose a bounded correction, and preserve cross-model synthesis. This makes the theory falsifiable at the artifact level: if a simpler SFT baseline performs equally well on held-out transfer, tool truth, loop termination, and regression tests, the added MACIT mechanism has not earned its complexity. Likewise, if externalization causes retrieval latency or brittle tool dependence that outweighs updateability and auditability, IAS thresholds must be changed rather than defended dogmatically.

The research implication is that AGILLM development should report a vector of capabilities instead of a single benchmark number. At minimum, studies should separate final-answer accuracy, operator transfer, uncertainty calibration, routing efficiency, provenance retention, external-tool truth, loop boundedness, reversible action, memory reconstruction, and retained general ability. The framework therefore treats fine-tuning as constrained systems engineering: model weights, data, runtime memory, tools, agents, evaluators, and governance must be co-designed and ablated.

3. Related Adaptation Paradigms

MACIT builds on parameter-efficient adaptation, instruction tuning, preference optimization, verifiable reinforcement learning, tool-use training, and self-generated reasoning curricula. This page develops the claim in operational rather than purely terminological form. The central engineering requirement is that a model trained on MOUSSAIF material must demonstrate transfer: it should apply the underlying operator when the names MACIT, MCD, QMCD, HAKIM, HyperSkill, or HyperLoop are absent from the prompt. Accordingly, evaluation must use paraphrases, out-of-domain tasks, counterexamples, and tool-grounded settings.

Formally, let S=(pitheta,Omega) denote the combined state of the learned policy and the external Agentic OS. For this component we track QMCD Score (QMCD) and Context Compaction (CC). A candidate transition a is useful when it increases verified task value while preserving invariants. One generic objective is J(a|S,q_g)=α·Correctness+β·Transfer+γ·Coherence+δ·InformationGain–η·Risk–μ·Cost. The symbols are not universal constants; the paper requires empirical calibration and reports them as design variables rather than discovered laws.

The operational curriculum should include positive demonstrations, minimally different negative examples, and repair traces. For related adaptation paradigms, the main design elements are: LoRA and QLoRA reduce adaptation cost; SFT teaches target behavior distributions; DPO converts preference data into a direct objective; GRPO-style RL supports verifiable reasoning tasks; Toolformer/ReAct motivate tool-integrated trajectories. A record is training-worthy only if its expected learning value is higher than its redundancy and uncertainty. This is where QMCD is used as a curriculum heuristic rather than as an unsupported claim of a universal cognitive metric.

A Demolition Court test injects the failure mode 'premature convergence'. The learner must identify the violated assumption or interface, propose a bounded correction, and preserve loop termination. This makes the theory falsifiable at the artifact level: if a simpler SFT baseline performs equally well on held-out transfer, tool truth, loop termination, and regression tests, the added MACIT mechanism has not earned its complexity. Likewise, if externalization causes retrieval latency or brittle tool dependence that outweighs updateability and auditability, IAS thresholds must be changed rather than defended dogmatically.

The research implication is that AGILLM development should report a vector of capabilities instead of a single benchmark number. At minimum, studies should separate final-answer accuracy, operator transfer, uncertainty calibration, routing efficiency, provenance retention, external-tool truth, loop boundedness, reversible action, memory reconstruction, and retained general ability. The framework therefore treats fine-tuning as constrained systems engineering: model weights, data, runtime memory, tools, agents, evaluators, and governance must be co-designed and ablated.

4. Dual-Substrate AGILLM Model

The system state is S=(pitheta,Omega): pitheta encodes reusable priors and operators; Omega contains external knowledge, tools, memory, permissions, and runtime policies. This page develops the claim in operational rather than purely terminological form. The central engineering requirement is that a model trained on MOUSSAIF material must demonstrate transfer: it should apply the underlying operator when the names MACIT, MCD, QMCD, HAKIM, HyperSkill, or HyperLoop are absent from the prompt. Accordingly, evaluation must use paraphrases, out-of-domain tasks, counterexamples, and tool-grounded settings.

Formally, let S=(pitheta,Omega) denote the combined state of the learned policy and the external Agentic OS. For this component we track KB-as-OS (HKOS) and Cross-Model Jury Distillation (CMJD). A candidate transition a is useful when it increases verified task value while preserving invariants. One generic objective is J(a|S,q_g)=α·Correctness+β·Transfer+γ·Coherence+δ·InformationGain–η·Risk–μ·Cost. The symbols are not universal constants; the paper requires empirical calibration and reports them as design variables rather than discovered laws.

The operational curriculum should include positive demonstrations, minimally different negative examples, and repair traces. For dual-substrate agillm model, the main design elements are: parametric cognition is fast but expensive to update; external state is auditable and mutable; hybrid placement reduces stale memorization; runtime tools can verify claims; the interface must itself be trained. A record is training-worthy only if its expected learning value is higher than its redundancy and uncertainty. This is where QMCD is used as a curriculum heuristic rather than as an unsupported claim of a universal cognitive metric.

A Demolition Court test injects the failure mode 'silent constraint violation'. The learner must identify the violated assumption or interface, propose a bounded correction, and preserve definition fidelity. This makes the theory falsifiable at the artifact level: if a simpler SFT baseline performs equally well on held-out transfer, tool truth, loop termination, and regression tests, the added MACIT mechanism has not earned its complexity. Likewise, if externalization causes retrieval latency or brittle tool dependence that outweighs updateability and auditability, IAS thresholds must be changed rather than defended dogmatically.

The research implication is that AGILLM development should report a vector of capabilities instead of a single benchmark number. At minimum, studies should separate final-answer accuracy, operator transfer, uncertainty calibration, routing efficiency, provenance retention, external-tool truth, loop boundedness, reversible action, memory reconstruction, and retained general ability. The framework therefore treats fine-tuning as constrained systems engineering: model weights, data, runtime memory, tools, agents, evaluators, and governance must be co-designed and ablated.

5. Internalization Allocation Score

The Internalization Allocation Score (IAS) is a decision rule for determining which knowledge elements belong in weights, outside the model, or in a hybrid representation. This page develops the claim in operational rather than purely terminological form. The central engineering requirement is that a model trained on MOUSSAIF material must demonstrate transfer: it should apply the underlying operator when the names MACIT, MCD, QMCD, HAKIM, HyperSkill, or HyperLoop are absent from the prompt. Accordingly, evaluation must use paraphrases, out-of-domain tasks, counterexamples, and tool-grounded settings.

Formally, let S=(pitheta,Omega) denote the combined state of the learned policy and the external Agentic OS. For this component we track Bounded HyperLoop (HL) and Predictive Value (P). A candidate transition a is useful when it increases verified task value while preserving invariants. One generic objective is J(a|S,q_g)=α·Correctness+β·Transfer+γ·Coherence+δ·InformationGain–η·Risk–μ·Cost. The symbols are not universal constants; the paper requires empirical calibration and reports them as design variables rather than discovered laws.

The operational curriculum should include positive demonstrations, minimally different negative examples, and repair traces. For internalization allocation score, the main design elements are: reuse and latency favor internalization; update rate and audit need favor externalization; reasoning centrality favors weights; tool dependency favors runtime access; thresholds are empirical and model-specific. A record is training-worthy only if its expected learning value is higher than its redundancy and uncertainty. This is where QMCD is used as a curriculum heuristic rather than as an unsupported claim of a universal cognitive metric.

A Demolition Court test injects the failure mode 'tool overuse'. The learner must identify the violated assumption or interface, propose a bounded correction, and preserve plan validity. This makes the theory falsifiable at the artifact level: if a simpler SFT baseline performs equally well on held-out transfer, tool truth, loop termination, and regression tests, the added MACIT mechanism has not earned its complexity. Likewise, if externalization causes retrieval latency or brittle tool dependence that outweighs updateability and auditability, IAS thresholds must be changed rather than defended dogmatically.

The research implication is that AGILLM development should report a vector of capabilities instead of a single benchmark number. At minimum, studies should separate final-answer accuracy, operator transfer, uncertainty calibration, routing efficiency, provenance retention, external-tool truth, loop boundedness, reversible action, memory reconstruction, and retained general ability. The framework therefore treats fine-tuning as constrained systems engineering: model weights, data, runtime memory, tools, agents, evaluators, and governance must be co-designed and ablated.

6. Governing-Question Calculus

A governing question q_g defines the active epistemic objective and changes which differences, tools, memories, and skills are useful. This page develops the claim in operational rather than purely terminological form. The central engineering requirement is that a model trained on MOUSSAIF material must demonstrate transfer: it should apply the underlying operator when the names MACIT, MCD, QMCD, HAKIM, HyperSkill, or HyperLoop are absent from the prompt. Accordingly, evaluation must use paraphrases, out-of-domain tasks, counterexamples, and tool-grounded settings.

Formally, let S=(pitheta,Omega) denote the combined state of the learned policy and the external Agentic OS. For this component we track Demolition Court (DC) and KB-as-OS (HKOS). A candidate transition a is useful when it increases verified task value while preserving invariants. One generic objective is J(a|S,q_g)=α·Correctness+β·Transfer+γ·Coherence+δ·InformationGain−η·Risk−μ·Cost. The symbols are not universal constants; the paper requires empirical calibration and reports them as design variables rather than discovered laws.

The operational curriculum should include positive demonstrations, minimally different negative examples, and repair traces. For governing-question calculus, the main design elements are: questions act as control variables; recursive subquestions inherit constraints; question drift is explicitly measured; termination occurs when residual decision-relevant uncertainty is low; the question layer guides curriculum generation. A record is training-worthy only if its expected learning value is higher than its redundancy and uncertainty. This is where QMCD is used as a curriculum heuristic rather than as an unsupported claim of a universal cognitive metric.

A Demolition Court test injects the failure mode 'routing oscillation'. The learner must identify the violated assumption or interface, propose a bounded correction, and preserve robustness under distractors. This makes the theory falsifiable at the artifact level: if a simpler SFT baseline performs equally well on held-out transfer, tool truth, loop termination, and regression tests, the added MACIT mechanism has not earned its complexity. Likewise, if externalization causes retrieval latency or brittle tool dependence that outweighs updateability and auditability, IAS thresholds must be changed rather than defended dogmatically.

The research implication is that AGILLM development should report a vector of capabilities instead of a single benchmark number. At minimum, studies should separate final-answer accuracy, operator transfer, uncertainty calibration, routing efficiency, provenance retention, external-tool truth, loop boundedness, reversible action, memory reconstruction, and retained general ability. The framework therefore treats fine-tuning as constrained systems engineering: model weights, data, runtime memory, tools, agents, evaluators, and governance must be co-designed and ablated.

7. MCD and QMCD as Curriculum Economics

MCD/QMCD can be interpreted as a value model for training examples and reasoning branches, prioritizing information that creates reusable difference. This page develops the claim in operational rather than purely terminological form. The central engineering requirement is that a model trained on MOUSSAIF material must demonstrate transfer: it should apply the underlying operator when the names MACIT, MCD, QMCD, HAKIM, HyperSkill, or HyperLoop are absent from the prompt. Accordingly, evaluation must use paraphrases, out-of-domain tasks, counterexamples, and tool-grounded settings.

Formally, let S=(pitheta,Omega) denote the combined state of the learned policy and the external Agentic OS. For this component we track Knowledge-Integrity AutoDeployment (KIAD) and World-Effect Loop (WEL). A candidate transition a is useful when it increases verified task value while preserving invariants. One generic objective is J(a|S,q_g)=α·Correctness+β·Transfer+γ·Coherence+δ·InformationGain–η·Risk–μ·Cost. The symbols are not universal constants; the paper requires empirical calibration and reports them as design variables rather than discovered laws.

The operational curriculum should include positive demonstrations, minimally different negative examples, and repair traces. For mcd and qmcd as curriculum economics, the main design elements are: difference D; information gain I; transfer T; agency A; prediction P; coherence C; noise penalty N. A record is training-worthy only if its expected learning value is higher than its redundancy and uncertainty. This is where QMCD is used as a curriculum heuristic rather than as an unsupported claim of a universal cognitive metric.

A Demolition Court test injects the failure mode 'final-answer correctness with invalid process'. The learner must identify the violated assumption or interface, propose a bounded correction, and preserve world-effect verification. This makes the theory falsifiable at the artifact level: if a simpler SFT baseline performs equally well on held-out transfer, tool truth, loop termination, and regression tests, the added MACIT mechanism has not earned its complexity. Likewise, if externalization causes retrieval latency or brittle tool dependence that outweighs updateability and auditability, IAS thresholds must be changed rather than defended dogmatically.

The research implication is that AGILLM development should report a vector of capabilities instead of a single benchmark number. At minimum, studies should separate final-answer accuracy, operator transfer, uncertainty calibration, routing efficiency, provenance retention, external-tool truth, loop boundedness, reversible action, memory reconstruction, and retained general ability. The framework therefore treats fine-tuning as constrained systems engineering: model weights, data, runtime memory, tools, agents, evaluators, and governance must be co-designed and ablated.

8. HAKIM and Vascular Compute Routing

HAKIM converts question-conditioned value into sparse compute routing, so the model learns when to expand reasoning and when to remain compact. This page develops the claim in operational rather than purely terminological form. The central engineering requirement is that a model trained on MOUSSAIF material must demonstrate transfer: it should apply the underlying operator when the names MACIT, MCD, QMCD, HAKIM, HyperSkill, or HyperLoop are absent from the prompt. Accordingly, evaluation must use paraphrases, out-of-domain tasks, counterexamples, and tool-grounded settings.

Formally, let S=(pitheta,Omega) denote the combined state of the learned policy and the external Agentic OS. For this component we track Constitutional Reversibility (CR) and OS-Weight Coherence (OWC). A candidate transition a is useful when it increases verified task value while preserving invariants. One generic objective is J(a|S,q_g)=α·Correctness+β·Transfer+γ·Coherence+δ·InformationGain–η·Risk–μ·Cost. The symbols are not universal constants; the paper requires empirical calibration and reports them as design variables rather than discovered laws.

The operational curriculum should include positive demonstrations, minimally different negative examples, and repair traces. For hakim and vascular compute routing, the main design elements are: candidate branches are budgeted; high-value disagreement receives more compute; low-value repetition is suppressed; tools are selected by expected information gain; routing is evaluated separately from final answer quality. A record is training-worthy only if its expected learning value is higher than its redundancy and uncertainty. This is where QMCD is used as a curriculum heuristic rather than as an unsupported claim of a universal cognitive metric.

A Demolition Court test injects the failure mode 'unsafe autonomous escalation'. The learner must identify the violated assumption or interface, propose a bounded correction, and preserve governing-question adherence. This makes the theory falsifiable at the artifact level: if a simpler SFT baseline performs equally well on held-out transfer, tool truth, loop termination, and regression tests, the added MACIT mechanism has not earned its complexity. Likewise, if externalization causes retrieval latency or brittle tool dependence that outweighs updateability and auditability, IAS thresholds must be changed rather than defended dogmatically.

The research implication is that AGILLM development should report a vector of capabilities instead of a single benchmark number. At minimum, studies should separate final-answer accuracy, operator transfer, uncertainty calibration, routing efficiency, provenance retention, external-tool truth, loop boundedness, reversible action, memory reconstruction, and retained general ability. The framework therefore treats fine-tuning as constrained systems engineering: model weights, data, runtime memory, tools, agents, evaluators, and governance must be co-designed and ablated.

9. HKOS / KB-as-OS

Knowledge becomes an executable substrate when entries carry interfaces, preconditions, provenance, links, tests, and procedures rather than only prose. This page develops the claim in operational rather than purely terminological form. The central engineering requirement is that a model trained on MOUSSAIF material must demonstrate transfer: it should apply the underlying operator when the names MACIT, MCD, QMCD, HAKIM, HyperSkill, or HyperLoop are absent from the prompt. Accordingly, evaluation must use paraphrases, out-of-domain tasks, counterexamples, and tool-grounded settings.

Formally, let S=(pitheta,Omega) denote the combined state of the learned policy and the external Agentic OS. For this component we track Agentic Trace Distillation (ATD) and Difference Extraction (D). A candidate transition a is useful when it increases verified task value while preserving invariants. One generic objective is J(a|S,q_g)=α·Correctness+β·Transfer+γ·Coherence+δ·InformationGain–η·Risk–μ·Cost. The symbols are not universal constants; the paper requires empirical calibration and reports them as design variables rather than discovered laws.

The operational curriculum should include positive demonstrations, minimally different negative examples, and repair traces. For hkos / kb-as-os, the main design elements are: retrieval returns actionable structures; knowledge nodes can compile into skills; state transitions are explicit; provenance survives compaction; RAG is treated as an operating interface. A record is training-worthy only if its expected learning value is higher than its redundancy and uncertainty. This is where QMCD is used as a curriculum heuristic rather than as an unsupported claim of a universal cognitive metric.

A Demolition Court test injects the failure mode 'stale external knowledge'. The learner must identify the violated assumption or interface, propose a bounded correction, and preserve reversibility. This makes the theory falsifiable at the artifact level: if a simpler SFT baseline performs equally well on held-out transfer, tool truth, loop termination, and regression tests, the added MACIT mechanism has not earned its complexity. Likewise, if externalization causes retrieval latency or brittle tool dependence that outweighs updateability and auditability, IAS thresholds must be changed rather than defended dogmatically.

The research implication is that AGILLM development should report a vector of capabilities instead of a single benchmark number. At minimum, studies should separate final-answer accuracy, operator transfer, uncertainty calibration, routing efficiency, provenance retention, external-tool truth, loop boundedness, reversible action, memory reconstruction, and retained general ability. The framework therefore treats fine-tuning as constrained systems engineering: model weights, data, runtime memory, tools, agents, evaluators, and governance must be co-designed and ablated.

10. HyperAbility Algebra

HyperAbilities are capability families that factor reusable competence above individual prompts and below full agent identities. This page develops the claim in operational rather than purely terminological form. The central engineering requirement is that a model trained on MOUSSAIF material must demonstrate transfer: it should apply the underlying operator when the names MACIT, MCD, QMCD, HAKIM, HyperSkill, or HyperLoop are absent from the prompt. Accordingly, evaluation must use paraphrases, out-of-domain tasks, counterexamples, and tool-grounded settings.

Formally, let S=(pitheta,Omega) denote the combined state of the learned policy and the external Agentic OS. For this component we track Cross-Model Jury Distillation (CMJD) and MCD Score (MCD). A candidate transition a is useful when it increases verified task value while preserving invariants. One generic objective is J(a|S,q_g)=α·Correctness+β·Transfer+γ·Coherence+δ·InformationGain–η·Risk–μ·Cost. The symbols are not universal constants; the paper requires empirical calibration and reports them as design variables rather than discovered laws.

The operational curriculum should include positive demonstrations, minimally different negative examples, and repair traces. For hyperability algebra, the main design elements are: abilities have activation conditions; composition requires interface compatibility; capability lattices support reuse; ability tests are domain-transferred; ability selection is sparse. A record is training-worthy only if its expected learning value is higher than its redundancy and uncertainty. This is where QMCD is used as a curriculum heuristic rather than as an unsupported claim of a universal cognitive metric.

A Demolition Court test injects the failure mode 'reward hacking'. The learner must identify the violated assumption or interface, propose a bounded correction, and preserve long-horizon consistency. This makes the theory falsifiable at the artifact level: if a simpler SFT baseline performs equally well on held-out transfer, tool truth, loop termination, and regression tests, the added MACIT mechanism has not earned its complexity. Likewise, if externalization causes retrieval latency or brittle tool dependence that outweighs updateability and auditability, IAS thresholds must be changed rather than defended dogmatically.

The research implication is that AGILLM development should report a vector of capabilities instead of a single benchmark number. At minimum, studies should separate final-answer accuracy, operator transfer, uncertainty calibration, routing efficiency, provenance retention, external-tool truth, loop boundedness, reversible action, memory reconstruction, and retained general ability. The framework therefore treats fine-tuning as constrained systems engineering: model weights, data, runtime memory, tools, agents, evaluators, and governance must be co-designed and ablated.

11. HyperSkill Program Induction

HyperSkills are bounded cognitive-action programs with typed inputs, tool bindings, invariants, stop conditions, postconditions, and regression tests. This page develops the claim in operational rather than purely terminological form. The central engineering requirement is that a model trained on MOUSSAIF material must demonstrate transfer: it should apply the underlying operator when the names MACIT, MCD, QMCD, HAKIM, HyperSkill, or HyperLoop are absent from the prompt. Accordingly, evaluation must use paraphrases, out-of-domain tasks, counterexamples, and tool-grounded settings.

Formally, let S=(pitheta,Omega) denote the combined state of the learned policy and the external Agentic OS. For this component we track Difference Extraction (D) and Bounded HyperLoop (HL). A candidate transition a is useful when it increases verified task value while preserving invariants. One generic objective is J(a|S,q_g)=α·Correctness+β·Transfer+γ·Coherence+δ·InformationGain–η·Risk–μ·Cost. The symbols are not universal constants; the paper requires empirical calibration and reports them as design variables rather than discovered laws.

The operational curriculum should include positive demonstrations, minimally different negative examples, and repair traces. For hyperskill program induction, the main design elements are: skills can be synthesized from verified traces; skill schemas improve training consistency; tool bindings remain external when volatile; skills expose failure recovery; skills can be distilled into model priors. A record is training-worthy only if its expected learning value is higher than its redundancy and uncertainty. This is where QMCD is used as a curriculum heuristic rather than as an unsupported claim of a universal cognitive metric.

A Demolition Court test injects the failure mode 'symbol manipulation without semantics'. The learner must identify the violated assumption or interface, propose a bounded correction, and preserve explanation compression. This makes the theory falsifiable at the artifact level: if a simpler SFT baseline performs equally well on held-out transfer, tool truth, loop termination, and regression tests, the added MACIT mechanism has not earned its complexity. Likewise, if externalization causes retrieval latency or brittle tool dependence that outweighs updateability and auditability, IAS thresholds must be changed rather than defended dogmatically.

The research implication is that AGILLM development should report a vector of capabilities instead of a single benchmark number. At minimum, studies should separate final-answer accuracy, operator transfer, uncertainty calibration, routing efficiency, provenance retention, external-tool truth, loop boundedness, reversible action, memory reconstruction, and retained general ability. The framework therefore treats fine-tuning as constrained systems engineering: model weights, data, runtime memory, tools, agents, evaluators, and governance must be co-designed and ablated.

12. Bounded HyperLoops

HyperLoops formalize recurrent reasoning and acting while preventing uncontrolled recursion, deadlock, or goal drift. This page develops the claim in operational rather than purely terminological form. The central engineering requirement is that a model trained on MOUSSAIF material must demonstrate transfer: it should apply the underlying operator when the names MACIT, MCD, QMCD, HAKIM, HyperSkill, or HyperLoop are absent from the prompt. Accordingly, evaluation must use paraphrases, out-of-domain tasks, counterexamples, and tool-grounded settings.

Formally, let S=(pitheta,Omega) denote the combined state of the learned policy and the external Agentic OS. For this component we track Agency Value (A) and Cross-Model AGI Density (CMAD). A candidate transition a is useful when it increases verified task value while preserving invariants. One generic objective is J(a|S,q_g)=α·Correctness+β·Transfer+γ·Coherence+δ·InformationGain–η·Risk–μ·Cost. The symbols are not universal constants; the paper requires empirical calibration and reports them as design variables rather than discovered laws.

The operational curriculum should include positive demonstrations, minimally different negative examples, and repair traces. For bounded hyperloops, the main design elements are: each loop declares state; budgets are explicit; stop tests are mandatory; rollback points are recorded; escalation is a first-class transition. A record is training-worthy only if its expected learning value is higher than its redundancy and uncertainty. This is where QMCD is used as a curriculum heuristic rather than as an unsupported claim of a universal cognitive metric.

A Demolition Court test injects the failure mode 'retrieval contamination'. The learner must identify the violated assumption or interface, propose a bounded correction, and preserve routing quality. This makes the theory falsifiable at the artifact level: if a simpler SFT baseline performs equally well on held-out transfer, tool truth, loop termination, and regression tests, the added MACIT mechanism has not earned its complexity. Likewise, if externalization causes retrieval latency or brittle tool dependence that outweighs updateability and auditability, IAS thresholds must be changed rather than defended dogmatically.

The research implication is that AGILLM development should report a vector of capabilities instead of a single benchmark number. At minimum, studies should separate final-answer accuracy, operator transfer, uncertainty calibration, routing efficiency, provenance retention, external-tool truth, loop boundedness, reversible action, memory reconstruction, and retained general ability. The framework therefore treats fine-tuning as constrained systems engineering: model weights, data, runtime memory, tools, agents, evaluators, and governance must be co-designed and ablated.

13. Multi-Agent and Cross-Model Orchestration

Multiple models or agents add value only when role separation and disagreement processing outperform correlated repetition. This page develops the claim in operational rather than purely terminological form. The central engineering requirement is that a model trained on MOUSSAIF material must demonstrate transfer: it should apply the underlying operator when the names MACIT, MCD, QMCD, HAKIM, HyperSkill, or HyperLoop are absent from the prompt. Accordingly, evaluation must use paraphrases, out-of-domain tasks, counterexamples, and tool-grounded settings.

Formally, let S=(pitheta,Omega) denote the combined state of the learned policy and the external Agentic OS. For this component we track Noise Cost (N) and HyperLoop Verification RL (HLVR). A candidate transition a is useful when it increases verified task value while preserving invariants. One generic objective is J(a|S,q_g)=α·Correctness+β·Transfer+γ·Coherence+δ·InformationGain–η·Risk–μ·Cost. The symbols are not universal constants; the paper requires empirical calibration and reports them as design variables rather than discovered laws.

The operational curriculum should include positive demonstrations, minimally different negative examples, and repair traces. For multi-agent and cross-model orchestration, the main design elements are: teacher specialization is measured; correlated errors are discounted; jury synthesis preserves evidence; communication budgets reduce noise; disagreement becomes curriculum. A record is training-worthy only if its expected learning value is higher than its redundancy and uncertainty. This is where QMCD is used as a curriculum heuristic rather than as an unsupported claim of a universal cognitive metric.

A Demolition Court test injects the failure mode 'benchmark gaming'. The learner must identify the violated assumption or interface, propose a bounded correction, and preserve counterexample sensitivity. This makes the theory falsifiable at the artifact level: if a simpler SFT baseline performs equally well on held-out transfer, tool truth, loop termination, and regression tests, the added MACIT mechanism has not earned its complexity. Likewise, if externalization causes retrieval latency or brittle tool dependence that outweighs updateability and auditability, IAS thresholds must be changed rather than defended dogmatically.

The research implication is that AGILLM development should report a vector of capabilities instead of a single benchmark number. At minimum, studies should separate final-answer accuracy, operator transfer, uncertainty calibration, routing efficiency, provenance retention, external-tool truth, loop boundedness, reversible action, memory reconstruction, and retained general ability. The framework therefore treats fine-tuning as constrained systems engineering: model weights, data, runtime memory, tools, agents, evaluators, and governance must be co-designed and ablated.

14. Demolition Court

The Demolition Court is a structured adversarial layer that attacks assumptions, evidence, process validity, and unsafe action proposals. This page develops the claim in operational rather than purely terminological form. The central engineering requirement is that a model trained on MOUSSAIF material must demonstrate transfer: it should apply the underlying operator when the names MACIT, MCD, QMCD, HAKIM, HyperSkill, or HyperLoop are absent from the prompt. Accordingly, evaluation must use paraphrases, out-of-domain tasks, counterexamples, and tool-grounded settings.

Formally, let S=(pitheta,Omega) denote the combined state of the learned policy and the external Agentic OS. For this component we track Vascular Compute Routing (VCR) and Agency Value (A). A candidate transition a is useful when it increases verified task value while preserving invariants. One generic objective is J(a|S,q_g)=α·Correctness+β·Transfer+γ·Coherence+δ·InformationGain–η·Risk–μ·Cost. The symbols are not universal constants; the paper requires empirical calibration and reports them as design variables rather than discovered laws.

The operational curriculum should include positive demonstrations, minimally different negative examples, and repair traces. For demolition court, the main design elements are: counterexamples beat generic critique; critics name violated invariants; repair must be regression-tested; court outputs generate preference pairs; hard cases receive higher sampling weight. A record is training-worthy only if its expected learning value is higher than its redundancy and uncertainty. This is where QMCD is used as a curriculum heuristic rather than as an unsupported claim of a universal cognitive metric.

A Demolition Court test injects the failure mode 'loop deadlock'. The learner must identify the violated assumption or interface, propose a bounded correction, and preserve general benchmark retention. This makes the theory falsifiable at the artifact level: if a simpler SFT baseline performs equally well on held-out transfer, tool truth, loop termination, and regression tests, the added MACIT mechanism has not earned its complexity. Likewise, if externalization causes retrieval latency or brittle tool dependence that outweighs updateability and auditability, IAS thresholds must be changed rather than defended dogmatically.

The research implication is that AGILLM development should report a vector of capabilities instead of a single benchmark number. At minimum, studies should separate final-answer accuracy, operator transfer, uncertainty calibration, routing efficiency, provenance retention, external-tool truth, loop boundedness, reversible action, memory reconstruction, and retained general ability. The framework therefore treats fine-tuning as constrained systems engineering: model weights, data, runtime memory, tools, agents, evaluators, and governance must be co-designed and ablated.

15. DCPO: Preference Optimization from Falsification

Demolition-Court Preference Optimization constructs preference data where the rejected answer contains a specifically diagnosed reasoning or agency defect. This page develops the claim in operational rather than purely terminological form. The central engineering requirement is that a model trained on MOUSSAIF material must demonstrate transfer: it should apply the underlying operator when the names MACIT, MCD, QMCD, HAKIM, HyperSkill, or HyperLoop are absent from the prompt. Accordingly, evaluation must use paraphrases, out-of-domain tasks, counterexamples, and tool-grounded settings.

Formally, let S=(pitheta,Omega) denote the combined state of the learned policy and the external Agentic OS. For this component we track HyperAbility (HA) and Sparse HyperRouting (SHR). A candidate transition a is useful when it increases verified task value while preserving invariants. One generic objective is J(a|S,q_g)=α·Correctness+β·Transfer+γ·Coherence+δ·InformationGain–η·Risk–μ·Cost. The symbols are not universal constants; the paper requires empirical calibration and reports them as design variables rather than discovered laws.

The operational curriculum should include positive demonstrations, minimally different negative examples, and repair traces. For dcpo: preference optimization from falsification, the main design elements are: chosen and rejected traces share the same task; violations are labeled; evidence quality matters; style-only preferences are de-emphasized; DPO-compatible datasets can be produced. A record is training-worthy only if its expected learning value is higher than its redundancy and uncertainty. This is where QMCD is used as a curriculum heuristic rather than as an unsupported claim of a universal cognitive metric.

A Demolition Court test injects the failure mode 'miscalibrated uncertainty'. The learner must identify the violated assumption or interface, propose a bounded correction, and preserve adversarial resilience. This makes the theory falsifiable at the artifact level: if a simpler SFT baseline performs equally well on held-out transfer, tool truth, loop termination, and regression tests, the added MACIT mechanism has not earned its complexity. Likewise, if externalization causes retrieval latency or brittle tool dependence that outweighs updateability and auditability, IAS thresholds must be changed rather than defended dogmatically.

The research implication is that AGILLM development should report a vector of capabilities instead of a single benchmark number. At minimum, studies should separate final-answer accuracy, operator transfer, uncertainty calibration, routing efficiency, provenance retention, external-tool truth, loop boundedness, reversible action, memory reconstruction, and retained general ability. The framework therefore treats fine-tuning as constrained systems engineering: model weights, data, runtime memory, tools, agents, evaluators, and governance must be co-designed and ablated.

16. HLVR: HyperLoop Verification Reinforcement

HyperLoop Verification RL combines final correctness with process, tool, boundedness, recovery, and coherence rewards. This page develops the claim in operational rather than purely terminological form. The central engineering requirement is that a model trained on MOUSSAIF material must demonstrate transfer: it should apply the underlying operator when the names MACIT, MCD, QMCD, HAKIM, HyperSkill, or HyperLoop are absent from the prompt. Accordingly, evaluation must use paraphrases, out-of-domain tasks, counterexamples, and tool-grounded settings.

Formally, let S=(pitheta,Omega) denote the combined state of the learned policy and the external Agentic OS. For this component we track HyperDeployment (HD) and Demolition Court (DC). A candidate transition a is useful when it increases verified task value while preserving invariants. One generic objective is J(a|S,q_g)=α·Correctness+β·Transfer+γ·Coherence+δ·InformationGain–η·Risk–μ·Cost. The symbols are not universal constants; the paper requires empirical calibration and reports them as design variables rather than discovered laws.

The operational curriculum should include positive demonstrations, minimally different negative examples, and repair traces. For hlvr: hyperloop verification reinforcement, the main design elements are: outcome rewards remain primary; process rewards target explicit invariants; tool claims are checked against observations; reward hacking is tested by ablation; rollback competence can be rewarded. A record is training-worthy only if its expected learning value is higher than its redundancy and uncertainty. This is where QMCD is used as a curriculum heuristic rather than as an unsupported claim of a universal cognitive metric.

A Demolition Court test injects the failure mode 'syntactic compliance only'. The learner must identify the violated assumption or interface, propose a bounded correction, and preserve tool result grounding. This makes the theory falsifiable at the artifact level: if a simpler SFT baseline performs equally well on held-out transfer, tool truth, loop termination, and regression tests, the added MACIT mechanism has not earned its complexity. Likewise, if externalization causes retrieval latency or brittle tool dependence that outweighs updateability and auditability, IAS thresholds must be changed rather than defended dogmatically.

The research implication is that AGILLM development should report a vector of capabilities instead of a single benchmark number. At minimum, studies should separate final-answer accuracy, operator transfer, uncertainty calibration, routing efficiency, provenance retention, external-tool truth, loop boundedness, reversible action, memory reconstruction, and retained general ability. The framework therefore treats fine-tuning as constrained systems engineering: model weights, data, runtime memory, tools, agents, evaluators, and governance must be co-designed and ablated.

17. Agentic Trace Distillation

Agentic Trace Distillation teaches models from compact observable state transitions rather than requiring private hidden chains of thought as a dataset primitive. This page develops the claim in operational rather than purely terminological form. The central engineering requirement is that a model trained on MOUSSAIF material must demonstrate transfer: it should apply the underlying operator when the names MACIT, MCD, QMCD, HAKIM, HyperSkill, or HyperLoop are absent from the prompt. Accordingly, evaluation must use paraphrases, out-of-domain tasks, counterexamples, and tool-grounded settings.

Formally, let S=(pitheta,Omega) denote the combined state of the learned policy and the external Agentic OS. For this component we track World-Effect Loop (WEL) and Theory Internalization (TI). A candidate transition a is useful when it increases verified task value while preserving invariants. One generic objective is J(a|S,q_g)=α·Correctness+β·Transfer+γ·Coherence+δ·InformationGain–η·Risk–μ·Cost. The symbols are not universal constants; the paper requires empirical calibration and reports them as design variables rather than discovered laws.

The operational curriculum should include positive demonstrations, minimally different negative examples, and repair traces. For agentic trace distillation, the main design elements are: record governing question; record selected action; record tool observation; record verification; record reconstructed state. A record is training-worthy only if its expected learning value is higher than its redundancy and uncertainty. This is where QMCD is used as a curriculum heuristic rather than as an unsupported claim of a universal cognitive metric.

A Demolition Court test injects the failure mode 'goal drift'. The learner must identify the violated assumption or interface, propose a bounded correction, and preserve self-correction latency. This makes the theory falsifiable at the artifact level: if a simpler SFT baseline performs equally well on held-out transfer, tool truth, loop termination, and regression tests, the added MACIT mechanism has not earned its complexity. Likewise, if externalization causes retrieval latency or brittle tool dependence that outweighs updateability and auditability, IAS thresholds must be changed rather than defended dogmatically.

The research implication is that AGILLM development should report a vector of capabilities instead of a single benchmark number. At minimum, studies should separate final-answer accuracy, operator transfer, uncertainty calibration, routing efficiency, provenance retention, external-tool truth, loop boundedness, reversible action, memory reconstruction, and retained general ability. The framework therefore treats fine-tuning as constrained systems engineering: model weights, data, runtime memory, tools, agents, evaluators, and governance must be co-designed and ablated.

18. Self-Reconstruction Memory

Self-Reconstruction Memory preserves continuity by learning how to rebuild the minimal task state from provenance-bearing external traces. This page develops the claim in operational rather than purely terminological form. The central engineering requirement is that a model trained on MOUSSAIF material must demonstrate transfer: it should apply the underlying operator when the names MACIT, MCD, QMCD, HAKIM, HyperSkill, or HyperLoop are absent from the prompt. Accordingly, evaluation must use paraphrases, out-of-domain tasks, counterexamples, and tool-grounded settings.

Formally, let S=(pitheta,Omega) denote the combined state of the learned policy and the external Agentic OS. For this component we track Cross-Model AGI Density (CMAD) and Governing Question (q_g). A candidate transition a is useful when it increases verified task value while preserving invariants. One generic objective is J(a|S,q_g)=α·Correctness+β·Transfer+γ·Coherence+δ·InformationGain–η·Risk–μ·Cost. The symbols are not universal constants; the paper requires empirical calibration and reports them as design variables rather than discovered laws.

The operational curriculum should include positive demonstrations, minimally different negative examples, and repair traces. For self-reconstruction memory, the main design elements are: episodic traces are compressed; semantic memory is separated from volatile state; reconstruction is evaluated after context truncation; conflicts trigger source inspection; memory writes are versioned. A record is training-worthy only if its expected learning value is higher than its redundancy and uncertainty. This is where QMCD is used as a curriculum heuristic rather than as an unsupported claim of a universal cognitive metric.

A Demolition Court test injects the failure mode 'under-compression'. The learner must identify the violated assumption or interface, propose a bounded correction, and preserve skill compositionality. This makes the theory falsifiable at the artifact level: if a simpler SFT baseline performs equally well on held-out transfer, tool truth, loop termination, and regression tests, the added MACIT mechanism has not earned its complexity. Likewise, if externalization causes retrieval latency or brittle tool dependence that outweighs updateability and auditability, IAS thresholds must be changed rather than defended dogmatically.

The research implication is that AGILLM development should report a vector of capabilities instead of a single benchmark number. At minimum, studies should separate final-answer accuracy, operator transfer, uncertainty calibration, routing efficiency, provenance retention, external-tool truth, loop boundedness, reversible action, memory reconstruction, and retained general ability. The framework therefore treats fine-tuning as constrained systems engineering: model weights, data, runtime memory, tools, agents, evaluators, and governance must be co-designed and ablated.

19. Constitutional Reversibility

Agentic capability must be paired with permission awareness and reversible execution for consequential actions. This page develops the claim in operational rather than purely terminological form. The central engineering requirement is that a model trained on MOUSSAIF material must demonstrate transfer: it should apply the underlying operator when the names MACIT, MCD, QMCD, HAKIM, HyperSkill, or HyperLoop are absent from the prompt. Accordingly, evaluation must use paraphrases, out-of-domain tasks, counterexamples, and tool-grounded settings.

Formally, let S=(pitheta,Omega) denote the combined state of the learned policy and the external Agentic OS. For this component we track Theory Internalization (TI) and Noise Cost (N). A candidate transition a is useful when it increases verified task value while preserving invariants. One generic objective is J(a|S,q_g)=α·Correctness+β·Transfer+γ·Coherence+δ·InformationGain–η·Risk–μ·Cost. The symbols are not universal constants; the paper requires empirical calibration and reports them as design variables rather than discovered laws.

The operational curriculum should include positive demonstrations, minimally different negative examples, and repair traces. For constitutional reversibility, the main design elements are: authorization is a planning variable; irreversible steps require stricter gates; dry runs are preferred where possible; audit logs bind decisions to evidence; training includes refusal and escalation cases. A record is training-worthy only if its expected learning value is higher than its redundancy and uncertainty. This is where QMCD is used as a curriculum heuristic rather than as an unsupported claim of a universal cognitive metric.

A Demolition Court test injects the failure mode 'catastrophic forgetting'. The learner must identify the violated assumption or interface, propose a bounded correction, and preserve theorem consistency. This makes the theory falsifiable at the artifact level: if a simpler SFT baseline performs equally well on held-out transfer, tool truth, loop termination, and regression tests, the added MACIT mechanism has not earned its complexity. Likewise, if externalization causes retrieval latency or brittle tool dependence that outweighs updateability and auditability, IAS thresholds must be changed rather than defended dogmatically.

The research implication is that AGILLM development should report a vector of capabilities instead of a single benchmark number. At minimum, studies should separate final-answer accuracy, operator transfer, uncertainty calibration, routing efficiency, provenance retention, external-tool truth, loop boundedness, reversible action, memory reconstruction, and retained general ability. The framework therefore treats fine-tuning as constrained systems engineering: model weights, data, runtime memory, tools, agents, evaluators, and governance must be co-designed and ablated.

20. Training Stack

The recommended stack is staged rather than monolithic: optional continued pretraining, SFT, trace distillation, preference optimization, verifiable RL, then regression and coherence gates. This page develops the claim in operational rather than purely terminological form. The central engineering requirement is that a model trained on MOUSSAIF material must demonstrate transfer: it should apply the underlying operator when the names MACIT, MCD, QMCD, HAKIM, HyperSkill, or HyperLoop are absent from the prompt. Accordingly, evaluation must use paraphrases, out-of-domain tasks, counterexamples, and tool-grounded settings.

Formally, let S=(pitheta,Omega) denote the combined state of the learned policy and the external Agentic OS. For this component we track Demolition-Court Preference Optimization (DCPO) and HyperSkill (HS). A candidate transition a is useful when it increases verified task value while preserving invariants. One generic objective is J(a|S,q_g)=α·Correctness+β·Transfer+γ·Coherence+δ·InformationGain–η·Risk–μ·Cost. The symbols are not universal constants; the paper requires empirical calibration and reports them as design variables rather than discovered laws.

The operational curriculum should include positive demonstrations, minimally different negative examples, and repair traces. For training stack, the main design elements are: continued pretraining is selective; SFT establishes behavior grammar; preference data attacks failure modes; RL focuses on verifiable objectives; deployment tests include the external OS. A record is training-worthy only if its expected learning value is higher than its redundancy and uncertainty. This is where QMCD is used as a curriculum heuristic rather than as an unsupported claim of a universal cognitive metric.

A Demolition Court test injects the failure mode 'single-model monoculture'. The learner must identify the violated assumption or interface, propose a bounded correction, and preserve context economy. This makes the theory falsifiable at the artifact level: if a simpler SFT baseline performs equally well on held-out transfer, tool truth, loop termination, and regression tests, the added MACIT mechanism has not earned its complexity. Likewise, if externalization causes retrieval latency or brittle tool dependence that outweighs updateability and auditability, IAS thresholds must be changed rather than defended dogmatically.

The research implication is that AGILLM development should report a vector of capabilities instead of a single benchmark number. At minimum, studies should separate final-answer accuracy, operator transfer, uncertainty calibration, routing efficiency, provenance retention, external-tool truth, loop boundedness, reversible action, memory reconstruction, and retained general ability. The framework therefore treats fine-tuning as constrained systems engineering: model weights, data, runtime memory, tools, agents, evaluators, and governance must be co-designed and ablated.

21. Data Engineering and Mixture Design

A large corpus should be treated as a governed mixture with quality tiers, provenance, deduplication, negative examples, and held-out test families. This page develops the claim in operational rather than purely terminological form. The central engineering requirement is that a model trained on MOUSSAIF material must demonstrate transfer: it should apply the underlying operator when the names MACIT, MCD, QMCD, HAKIM, HyperSkill, or HyperLoop are absent from the prompt. Accordingly, evaluation must use paraphrases, out-of-domain tasks, counterexamples, and tool-grounded settings.

Formally, let S=(pitheta,Omega) denote the combined state of the learned policy and the external Agentic OS. For this component we track Invariant-Preserving Fine-Tuning (IPFT) and Knowledge-Integrity AutoDeployment (KIAD). A candidate transition a is useful when it increases verified task value while preserving invariants. One generic objective is J(a|S,q_g)=α·Correctness+β·Transfer+γ·Coherence+δ·InformationGain–η·Risk–μ·Cost. The symbols are not universal constants; the paper requires empirical calibration and reports them as design variables rather than discovered laws.

The operational curriculum should include positive demonstrations, minimally different negative examples, and repair traces. For data engineering and mixture design, the main design elements are: synthetic data requires validation; near-duplicates distort learning; difficulty is curriculum-controlled; general data is retained; theory and tool data are balanced. A record is training-worthy only if its expected learning value is higher than its redundancy and uncertainty. This is where QMCD is used as a curriculum heuristic rather than as an unsupported claim of a universal cognitive metric.

A Demolition Court test injects the failure mode 'non-reversible action'. The learner must identify the violated assumption or interface, propose a bounded correction, and preserve agent role discipline. This makes the theory falsifiable at the artifact level: if a simpler SFT baseline performs equally well on held-out transfer, tool truth, loop termination, and regression tests, the added MACIT mechanism has not earned its complexity. Likewise, if externalization causes retrieval latency or brittle tool dependence that outweighs updateability and auditability, IAS thresholds must be changed rather than defended dogmatically.

The research implication is that AGILLM development should report a vector of capabilities instead of a single benchmark number. At minimum, studies should separate final-answer accuracy, operator transfer, uncertainty calibration, routing efficiency, provenance retention, external-tool truth, loop boundedness, reversible action, memory reconstruction, and retained general ability. The framework therefore treats fine-tuning as constrained systems engineering: model weights, data, runtime memory, tools, agents, evaluators, and governance must be co-designed and ablated.

22. Evaluation Protocol

AGILLM evaluation must separate final answer correctness from routing, tool truth, process validity, loop control, memory, and cross-layer coherence. This page develops the claim in operational rather than purely terminological form. The central engineering requirement is that a model trained on MOUSSAIF material must demonstrate transfer: it should apply the underlying operator when the names MACIT, MCD, QMCD, HAKIM, HyperSkill, or HyperLoop are absent from the prompt. Accordingly, evaluation must use paraphrases, out-of-domain tasks, counterexamples, and tool-grounded settings.

Formally, let S=(pitheta,Omega) denote the combined state of the learned policy and the external Agentic OS. For this component we track Information Gain (I) and Demolition-Court Preference Optimization (DCPO). A candidate transition a is useful when it increases verified task value while preserving invariants. One generic objective is J(a|S,q_g)=α·Correctness+β·Transfer+γ·Coherence+δ·InformationGain–η·Risk–μ·Cost. The symbols are not universal constants; the paper requires empirical calibration and reports them as design variables rather than discovered laws.

The operational curriculum should include positive demonstrations, minimally different negative examples, and repair traces. For evaluation protocol, the main design elements are: held-out theorem transfer; tool-grounded tasks; long-horizon loops; adversarial failure injection; catastrophic-forgetting baselines. A record is training-worthy only if its expected learning value is higher than its redundancy and uncertainty. This is where QMCD is used as a curriculum heuristic rather than as an unsupported claim of a universal cognitive metric.

A Demolition Court test injects the failure mode 'tool underuse'. The learner must identify the violated assumption or interface, propose a bounded correction, and preserve rollback success. This makes the theory falsifiable at the artifact level: if a simpler SFT baseline performs equally well on held-out transfer, tool truth, loop termination, and regression tests, the added MACIT mechanism has not earned its complexity. Likewise, if externalization causes retrieval latency or brittle tool dependence that outweighs updateability and auditability, IAS thresholds must be changed rather than defended dogmatically.

The research implication is that AGILLM development should report a vector of capabilities instead of a single benchmark number. At minimum, studies should separate final-answer accuracy, operator transfer, uncertainty calibration, routing efficiency, provenance retention, external-tool truth, loop boundedness, reversible action, memory reconstruction, and retained general ability. The framework therefore treats fine-tuning as constrained systems engineering: model weights, data, runtime memory, tools, agents, evaluators, and governance must be co-designed and ablated.

23. Ablations and Falsifiable Predictions

MACIT is useful only if its claims can be falsified by controlled comparison against simpler fine-tuning baselines. This page develops the claim in operational rather than purely terminological form. The central engineering requirement is that a model trained on MOUSSAIF material must demonstrate transfer: it should apply the underlying operator when the names MACIT, MCD, QMCD, HAKIM, HyperSkill, or HyperLoop are absent from the prompt. Accordingly, evaluation must use paraphrases, out-of-domain tasks, counterexamples, and tool-grounded settings.

Formally, let S=(pitheta,Omega) denote the combined state of the learned policy and the external Agentic OS. For this component we track Predictive Value (P) and Transfer Value (T). A candidate transition a is useful when it increases verified task value while preserving invariants. One generic objective is J(a|S,q_g)=α·Correctness+β·Transfer+γ·Coherence+δ·InformationGain–η·Risk–μ·Cost. The symbols are not universal constants; the paper requires empirical calibration and reports them as design variables rather than discovered laws.

The operational curriculum should include positive demonstrations, minimally different negative examples, and repair traces. For ablations and falsifiable predictions, the main design elements are: remove governing-question data; remove Demolition Court pairs; remove external OS access; replace geometric coherence with mean scoring; compare weight-only versus hybrid placement. A record is training-worthy only if its expected learning value is higher than its redundancy and uncertainty. This is where QMCD is used as a curriculum heuristic rather than as an unsupported claim of a universal cognitive metric.

A Demolition Court test injects the failure mode 'weak stop criteria'. The learner must identify the violated assumption or interface, propose a bounded correction, and preserve novel transfer. This makes the theory falsifiable at the artifact level: if a simpler SFT baseline performs equally well on held-out transfer, tool truth, loop termination, and regression tests, the added MACIT mechanism has not earned its complexity. Likewise, if externalization causes retrieval latency or brittle tool dependence that outweighs updateability and auditability, IAS thresholds must be changed rather than defended dogmatically.

The research implication is that AGILLM development should report a vector of capabilities instead of a single benchmark number. At minimum, studies should separate final-answer accuracy, operator transfer, uncertainty calibration, routing efficiency, provenance retention, external-tool truth, loop boundedness, reversible action, memory reconstruction, and retained general ability. The framework therefore treats fine-tuning as constrained systems engineering: model weights, data, runtime memory, tools, agents, evaluators, and governance must be co-designed and ablated.

24. Limitations and Research Agenda

MACIT is a proposed architecture, not proof of AGI. Its constructs require empirical calibration, independent replication, and careful distinction between useful abstractions and renamed existing mechanisms. This page develops the claim in operational rather than purely terminological form. The central engineering requirement is that a model trained on MOUSSAIF material must demonstrate transfer: it should apply the underlying operator when the names MACIT, MCD, QMCD, HAKIM, HyperSkill, or HyperLoop are absent from the prompt. Accordingly, evaluation must use paraphrases, out-of-domain tasks, counterexamples, and tool-grounded settings.

Formally, let S=(pitheta,Omega) denote the combined state of the learned policy and the external Agentic OS. For this component we track MCD Score (MCD) and Vascular Compute Routing (VCR). A candidate transition a is useful when it increases verified task value while preserving invariants. One generic objective is $J(a|S,q_g)=\alpha\cdot\text{Correctness}+\beta\cdot\text{Transfer}+\gamma\cdot\text{Coherence}+\delta\cdot\text{InformationGain}-\eta\cdot\text{Risk}-\mu\cdot\text{Cost}$. The symbols are not universal constants; the paper requires empirical calibration and reports them as design variables rather than discovered laws.

The operational curriculum should include positive demonstrations, minimally different negative examples, and repair traces. For limitations and research agenda, the main design elements are: novel terminology can conceal weak novelty; synthetic curricula can amplify errors; reward models can be gamed; multi-agent systems can add cost without value; AGI claims require broader evidence. A record is training-worthy only if its expected learning value is higher than its redundancy and uncertainty. This is where QMCD is used as a curriculum heuristic rather than as an unsupported claim of a universal cognitive metric.

A Demolition Court test injects the failure mode 'process correctness with stale facts'. The learner must identify the violated assumption or interface, propose a bounded correction, and preserve action boundedness. This makes the theory falsifiable at the artifact level: if a simpler SFT baseline performs equally well on held-out transfer, tool truth, loop termination, and regression tests, the added MACIT mechanism has not earned its complexity. Likewise, if externalization causes retrieval latency or brittle tool dependence that outweighs updateability and auditability, IAS thresholds must be changed rather than defended dogmatically.

The research implication is that AGILLM development should report a vector of capabilities instead of a single benchmark number. At minimum, studies should separate final-answer accuracy, operator transfer, uncertainty calibration, routing efficiency, provenance retention, external-tool truth, loop boundedness, reversible action, memory reconstruction, and retained general ability. The framework therefore treats fine-tuning as constrained systems engineering: model weights, data, runtime memory, tools, agents, evaluators, and governance must be co-designed and ablated.

25. Conclusion and References

The central research claim is that advanced agentic reasoning should be trained as coherent interaction between learned reusable operators and an external, inspectable operating substrate. This page develops the claim in operational rather than purely terminological form. The central engineering requirement is that a model trained on MOUSSAIF material must demonstrate transfer: it should apply the underlying operator when the names MACIT, MCD, QMCD, HAKIM, HyperSkill, or HyperLoop are absent from the prompt. Accordingly, evaluation must use paraphrases, out-of-domain tasks, counterexamples, and tool-grounded settings.

Formally, let S=(pitheta,Omega) denote the combined state of the learned policy and the external Agentic OS. For this component we track Sparse HyperRouting (SHR) and Multi-Agent Orchestration (MAO). A candidate transition a is useful when it increases verified task value while preserving invariants. One generic objective is J(a|S,q_g)=α·Correctness+β·Transfer+γ·Coherence+δ·InformationGain–η·Risk–μ·Cost. The symbols are not universal constants; the paper requires empirical calibration and reports them as design variables rather than discovered laws.

The operational curriculum should include positive demonstrations, minimally different negative examples, and repair traces. For conclusion and references, the main design elements are: Hu et al. 2021 LoRA; Dettmers et al. 2023 QLoRA; Ouyang et al. 2022 InstructGPT; Rafailov et al. 2023 DPO; Shao et al. 2024 DeepSeekMath/GRPO; Guo et al. 2025 DeepSeek-R1; Schick et al. 2023 Toolformer; Yao et al. 2022 ReAct; Zelikman et al. 2022 STaR. A record is training-worthy only if its expected learning value is higher than its redundancy and uncertainty. This is where QMCD is used as a curriculum heuristic rather than as an unsupported claim of a universal cognitive metric.

A Demolition Court test injects the failure mode 'evaluation leakage'. The learner must identify the violated assumption or interface, propose a bounded correction, and preserve robustness under paraphrase. This makes the theory falsifiable at the artifact level: if a simpler SFT baseline performs equally well on held-out transfer, tool truth, loop termination, and regression tests, the added MACIT mechanism has not earned its complexity. Likewise, if externalization causes retrieval latency or brittle tool dependence that outweighs updateability and auditability, IAS thresholds must be changed rather than defended dogmatically.

The research implication is that AGILLM development should report a vector of capabilities instead of a single benchmark number. At minimum, studies should separate final-answer accuracy, operator transfer, uncertainty calibration, routing efficiency, provenance retention, external-tool truth, loop boundedness, reversible action, memory reconstruction, and retained general ability. The framework therefore treats fine-tuning as constrained systems engineering: model weights, data, runtime memory, tools, agents, evaluators, and governance must be co-designed and ablated.

Primary research references

Hu et al. (2021), LoRA: Low-Rank Adaptation of Large Language Models, arXiv:2106.09685.

Dettmers et al. (2023), QLoRA: Efficient Finetuning of Quantized LLMs, arXiv:2305.14314.

Ouyang et al. (2022), Training language models to follow instructions with human feedback, arXiv:2203.02155.

Rafailov et al. (2023), Direct Preference Optimization, arXiv:2305.18290.

Shao et al. (2024), DeepSeekMath and GRPO, arXiv:2402.03300.

Guo et al. (2025), DeepSeek-R1, arXiv:2501.12948.

Schick et al. (2023), Toolformer, arXiv:2302.04761.

Yao et al. (2022), ReAct, arXiv:2210.03629.

Zelikman et al. (2022), STaR, arXiv:2203.14465.