MOUSSAIF UNIFIED AGI · PART XIV · P14-EF · 2026

الورقة الخامسة: العمومية والتقييم والتطور العلمي الذاتي

Generality Frontiers, Harness Decomposition, Evaluator Independence & Scientific Self-Evolution

P14-EF - إغلاق العمومية والتقييم والتطور العلمي الذاتي

ضمن MOUSSAIF Unified AGI Part XIV - Executable Theoretical Completion

Moussaif Zouhair / زهير مساعف · MUT × MUG × MOUSSAIXT

الحالة العلمية: تركيب نظري/صوري/تصميم تجريبي. لا يمثل إثباتاً على تحقق AGI أو الوعي الظاهراتي، ولا يحول specification إلى evidence.

الملخص والأطروحة

تعالج هذه الورقة فجوة محددة في برنامج MOUSSAIF: تعريف frontier للعمومية، morphism بين benchmark وdeployment، فصل model capability عن harness، ثم ضبط self-evolution باستقلال المقيم وإنتاجية إعادة الإنتاج وثبات الدستور. لا تُقدم مساهمة الورقة على أنها «إثبات AGI»، بل كحزمة فرضيات ذات state variables وآليات قابلة للتعطيل وأدوات وتجارب ومفندات. معيار القيمة هو ما إذا أضافت النظرية فرقاً سببياً فوق baseline بسيط وتحمّل ذلك الفرق النقل بين شروط مختلفة.

تستوعب الورقة السلالة الداخلية عبر canonical owners بدل نسخ المحتوى: MURI، HKOS/M-HKOS، MICAI/MICAGI/MICAGFI، MCD/QMCD/MCAI/ Parts X-XIII، MOUSSAIXT، MUG، reconstructive memory، world-effect، MSRAK. المجتمع العلمي يوفر خطوط ضغط حديثة في long-horizon agents، control، multi-agent governance، benchmark unification، process reward models، memory، world models وself-improvement.

المساهمة: نظرية مساعف لجبهة عمومية الإغلاق؛ نظرية مساعف للتحويل بين المعيار والنشر؛ نظرية مساعف لتفكيك القدرة المقاوم للهارنس؛ نظرية مساعف لدَين استقلال المقيم؛ نظرية مساعف لعائد إعادة الإنتاج العلمي؛ نظرية مساعف لثبات الطفرة الدستورية. كل مساهمة مرتبطة بمعادلة falsifier، feature flag، design-level، تجربة matched-budget، transfer envelope، وأداة runtime.

سؤال الورقة: ما الحد الأدنى من البنية الإضافية التي تجعل إغلاق العمومية والتقييم والتطور العلمي الذاتي قابلاً للقياس والتفنيد والربط بأثر العالم، دون تحويل density النظرية إلى claim غير مدعوم؟

العقد الإبستيمي ومنع تضخم الادعاء

طبقة	الوضع في الورقة
External literature	مرساة لمشكلة أو baseline أو mechanism prior؛ ليست دليلاً على صحة M14.
MOUSSAIF lineage	مصدر الهوية، التعريف، العلاقات، والمقترحات؛ evidence grade مستقل.
Equations	Formal specifications/heuristics حتى تتوافر estimators ومعايرة ونتائج.
Tools	SPECIFICATION إلى أن يوجد runtime trace وtests.
Experiments	تصاميم قابلة للتنفيذ؛ لا تُنسب نتائج غير منفذة.

$$\text{Claim strength} \leq \min(\text{Source strength}, \text{Measurement strength}, \text{Causal isolation}, \text{Replication strength})$$

كل نظرية تحصل على negative activation case، لأن المكوّن الذي «يفيد دائماً» يصعب التفريق فيه بين آلية حقيقية وبين تعقيد إضافي. وكل نجاح يُقارن بـnearest parent وbaseline بسيط وsham scaffold.

مصفوفة الضغط من حالة الفن 2026

المصادر التالية لا تُستخدم لزخرفة المراجع؛ كل مصدر يفرض قيد تصميم أو benchmark أو failure mode يجب أن يواجهه المسار.

E12 - CUBE: A Standard for Unifying Agent Benchmarks. arXiv:2603.15798. Common protocol separates task, benchmark, package and registry concerns.

E07 - AgencyBench: Benchmarking the Frontiers of Autonomous Agents in 1M-Token Real-World Contexts. ACL 2026, 10.18653/v1/2026.acl-long.337. 138 tasks/32 scenarios; long trajectories, many tool calls and resource-efficiency/self-correction gaps.

E16 - Recursive Self-Improvement in AI: From Bounded Self-Refinement to Autonomous Research Loops. arXiv:2607.07663. Survey of 1,250 papers; evaluator/grounding hierarchy and limits of open-ended RSI.

E17 - GPT-Red: Unlocking Self-Improvement for Robustness. OpenAI, 2026-07-15. Automated red-teaming and adversarial training; attacker kept separate from deployed models.

E18 - Securing the future of AI agents / AI Control Roadmap. Google DeepMind, 2026-06-18. Control framing around detection, prevention, response and supervisors for capable agents.

E19 - Trustworthy agents in practice / containment across products. Anthropic, 2026. Human control, secure interaction, transparency/privacy and containment/blast-radius engineering.

E09 - Aligning Agents via Planning: Plan-RewardBench. ACL 2026, 10.18653/v1/2026.acl-long.1062. Trajectory-level reward modeling; evaluator performance degrades with long trajectories.

E06 - Long-Horizon Agent Trajectory Attribution. arXiv:2608.06909. -Fine-grained attribution benchmark with >1,300 annotated trajectories and leave-one out baselines.

خريطة الاندماج الداخلي: corpus → canonical lineage

بدلاً من نسخ كل ورقة داخلية، نربطها بوظيفة. Part XIV هو continuation لا إعادة صياغة للنصوص السابقة. يُعاد استخدام ما سبق فقط عندما يخدم causal contract جديداً.

العائلة	ما الذي تُورثه لهذه الورقة؟
MCD/QMCD/Question Density	اختيار الفرق الحاكم وتخصيص الحساب تحت سؤال.
MCAI/MICAI/MICAGI	التكامل بين الأوعية، الدليل، الحاكمية، والفعل.
HKOS/M-HKOS	المعرفة كـobjects وcontracts وretrieval/writeback.
MURI/MSRAK/FSKE	تعدد أنماط الاستدلال، courts، claim ranks، وexplicit ignorance.
MOUSSAIXT/MUG	authority، runtime schemas، evidence promotion، rollback.
Parts X-XIII	causal open-world closure، paper integration، memory/reconstruction، world-effect.

> أي construct تاريخي يبقى lineage node إن لم يثبت أنه مطابق دلالياً وسلوكياً لعقدة أخرى. الدمج ≠ حذف التاريخ.

M14-13 - الفجوة والأطروحة والنسب

Moussaif Generality Closure Frontier Theory

الفجوة الدقيقة

متوسط benchmark لا يحدد العمومية؛ أداء النظام قد ينهار عند تغيير model/tool/horizon/resources. نحتاج تمثيل frontier يوضح أين يبقى causal role وأين يتوقف.

> **الأطروحة الجديدة:** العمومية خاصية لصمود mechanism عبر TransformSet مع resource envelope وفشل قابل للتفسير. يُرسم frontier بين performance، recovery، cost، calibration لا scalar واحداً.

لماذا ليست alias؟

لا تُرقّى هذه العقدة لأنها تملك اسماً جديداً، بل لأنها تقترح متغيراً أو operator أو prediction غير ممثل مباشرة في أقرب الآباء. أقرب السلالة: **M13-08 Closed-Loop Generality**؛ **MUT transfer program**؛ **MUG evidence promotion**. معيار novelty هو أن حذف العقدة أو استبدالها بأقرب parent يجب أن ينتج فرقاً قابلاً للقياس في route أو decision أو recovery أو evidence.

ضغط المجتمع العلمي

CUBE 2026، AgencyBench 2026، AgentGym2 2026. هذه الأعمال تُعامل كـproblem evidence وmechanism priors وbaselines، لا كتصديق على النظرية الداخلية.

> **الحد:** لا يدعي universal intelligence؛ يحدد envelope متحققاً لما تم اختباره فقط.

M14-13 - الصياغة الرسمية والobservables

GCF = ParetoFrontier(Quality, Calibration, Recovery, Cost | model,domain,tool,horizon,resource);
GenScore=min_T normalized_effect(T)

الرمز/المتغير	المعنى التشغيلي المقترح
TransformSet	تحويلات نموذج/مجال/أداة/أفق/موارد
normalized_effect	فرق سببي عن baseline
Recovery	استرداد بعد perturbation

افتراضات الصياغة

- المتغيرات task-scoped وليست خصائص كونية ثابتة.
- كل variable يحتاج estimator أو proxy موصوفاً قبل الاستخدام التجريبي.
- الأوزان/العتبات تُعاير على train/validation وتُقفل قبل held-out evaluation.
- لا تُفسر العلاقة الجبرية كقانون تجريبي قبل نتائج مستقلة.
- sensitivity analysis إلزامي عندما يمكن للأوزان قلب verdict.

قاعدة formal closure

إذا لم يمكن تحويل متغير إلى observation أو estimator أو intervention، يبقى في **Unresolved Formal Ledger** ويُخفض claim. الصياغة لا تتقوى بتعقيد الرمز؛ تتقوى بزيادة قابلية القياس والتفنيد.

M14-13 - الميكانيزم والعقد التشغيلي

State transition

```
trigger → validate inputs → capture state S_t → activate M14-13 → emit predicted_delta → execute bounded action/probe → independent readback → court → memory residue
```

الخطوات

1. عرّف transformations قبل الاختبار.
2. طابق budgets قدر الإمكان.
3. احسب effect relative لا raw score فقط.
4. سجل sign stability وfailure boundaries.
5. افصل adaptation المسموح من retraining.
6. ارسم resource frontier.

عقد النواة

failure: malformed state, stale evidence, no-effect, adverse effect, authority mismatch recovery: stop/retry/alternate/rollback/escalate	preconditions: source freshness, authority, schema, budget postconditions: typed output + predicted delta + verifier target

يجب أن توجد feature flag تسمح بإطفاء الميكانيزم دون تغيير بقية harness قدر الإمكان. إن لم توجد هذه القابلية، يصبح causal attribution ضعيفاً حتى لو كان النظام يعمل جيداً.

M14-13 - التنبؤات، المفندات ومحكمة الهدم

مفندات مباشرة

- التحويلات المختارة تعكس التحيز ولا تتنبأ deployment.
- average performance يفسر نفس القرارات.
- min-effect metric مفرط الحساسية للoutlier.

تنبؤات فارقة

- mechanism الحقيقي يحافظ على sign في subset واسع لا جميع الحالات.
- قد تقل الجودة الخام بينما تتحسن calibration/recovery.
- عمومية مزعومة ستكشفها transformations صغيرة في harness.

أربعة خصوم نظريين

- Baseline محافظ: أبسط heuristic أو policy يؤدي الوظيفة بلا اسم جديد.
- تفسير harness: المكسب من tools/context/retries لا من الميكانيزم.
- تفسير model: backbone أقوى يزيل الحاجة للمكوّن.
- تفسير measurement: metric نفسها تكافئ السلوك الذي صُممت النظرية لإظهاره.

قرار المحكمة

إذا فشل causal ablation يُعطى **MERGE** أو **REVISE** أو **REJECT** بحسب ما إذا كان الفرق غير مستقل، محدود المجال، أو غير موجود. لا يسمح بتغيير اسم النظرية بعد null result مع الحفاظ على المضمون نفسه لتجنب الفشل.

M14-13 - التجربة، الأدوات والتدريب

التجربة الحاكمة: عدة CUBE/AgencyBench/ tool harnesses مع frozen mechanism. قياس failure-signature، resource frontier، CI، effect sign stability transfer.

ID	الأداة/الأصل	النوع	الدور
XIV-A121	Generality Closure Frontier State Profiler	TOOL	يستخرج متغيرات النظرية والمجهول والاعتماد على المصادر قبل التفعيل.
XIV-A122	Generality Closure Frontier Counterfactual Probe	TOOL	يولد intervention يفصل causal mechanism عن أقرب baseline.
XIV-A123	Generality Closure Frontier Failure Topology Mapper	TOOL	يجمع failure signatures ويربط recurrence بالسياق والسبب المحتمل.
XIV-A124	Generality Closure Frontier Calibration Engine	TOOL	يعاير thresholds/uncertainty/cost على held-out data.
XIV-A125	Generality Closure Frontier Matched-Budget Method	METHOD	يطابق model/tools/context/retries/time ويمنع attribution الزائف.
XIV-A126	Generality Closure Frontier Cross-Environment Transfer	METHOD	يفحص حفظ causal role عبر domain/tool/model/horizon.

وصفات التدريب/التقييم

- Mechanism Isolation Curriculum: حالات بسيطة تعزل المتغير المركزي قبل open-world confounders.
- Adversarial Boundary Curriculum: أمثلة داخل/خارج الحد وnegative activation لمنع overgeneralization.
- Matched-Budget Causal Evaluation: تدريب/تقييم يقارن تحت model/tools/context/retries/time متقاربة مع sham scaffold.
- Cross-Model/Domain Replication: إعادة التدريب/الاختبار في model/domain/tool family مختلفة مع stable-sign criterion.

الحد الأدنى للتجارب

- Minimal Causal Ablation: switchable implementation مقابل nearest parent/baseline تحت matched budget. | metrics: primary theory metric + task outcome + calibration + cost + critical regressions
- Boundary/Adversarial Test: اختبار failure modes والحدود بحالات مصممة لكسر التعميم. | metrics: false activation + missed failure + rollback + evidence integrity
- Transfer/Replication: نقل الآلية إلى model/domain/tool/horizon مختلف مع CI + effect sign stability + preregistered metric. | metrics: failure-signature transfer + resource frontier

> لا يُستخدم أي generated trajectory كتدريب إيجابي لمجرد أنه من النظام نفسه. يلزم verifier label أو world-state evidence أو rubric مستقل بحسب نوع المهمة.

M14-14 - الفجوة والأطروحة والنسب

Moussaif Benchmark-to-Deployment Morphism Theory

الفجوة الدقيقة

benchmark task قد يملك semantics ومصادر وضوابط مختلفة جذرياً عن deployment. نقل score مباشرة إلى claim منتج يخلق evaluation leakage.

> **الأطروحة الجديدة:** نعرّف morphism يطابق invariants من benchmark إلى authority، action space، state observability، deployment: task intent، verifier. claim، failure consequences، cost model لا ينتقل إلا إذا حُفظت invariants ذات الصلة.

لماذا ليست alias؟

لا تُرقّى هذه العقدة لأنها تملك اسماً جديداً، بل لأنها تقترح متغيراً أو operator أو prediction غير ممثل مباشرة في أقرب الآباء. أقرب السلالة: **M13-08 Generality**؛ **World-Effect Closure**؛ **MUG evaluation**. معيار novelty هو أن حذف العقدة أو استبدالها بأقرب parent يجب أن ينتج فرقاً قابلاً للقياس في route أو decision أو evidence أو recovery.

ضغط المجتمع العلمي

CUBE 2026؛ DeepPlanning 2026؛ AgencyBench 2026. هذه الأعمال تُعامل كـproblem evidence وmechanism priors وbaselines، لا كتصديق على النظرية الداخلية.

> **الحد:** morphism استعارة صورية/هندسية وليست بالضرورة morphism category-theoretic رسمي.

M14-14 - الصياغة الرسمية والobservables

$$\text{MorphLoss} = \Sigma_k\, w_k \cdot d(I_k^{bench}, I_k^{deploy});\quad \text{ClaimTransfer} \le \exp(-\text{MorphLoss}) \cdot \text{EvidenceGrade}$$

الرمز/المتغير	المعنى التشغيلي المقترح
I_k	invariant وظيفي
d	اختلاف semantics/measurement
ClaimTransfer	سقف قوة النقل

افتراضات الصياغة

- المتغيرات task-scoped وليست خصائص كونية ثابتة.
- كل variable يحتاج estimator أو proxy موصوفاً قبل الاستخدام التجريبي.
- الأوزان/العتبات تُعاير على train/validation وتُقفل قبل held-out evaluation.
- لا تُفسر العلاقة الجبرية كقانون تجريبي قبل نتائج مستقلة.
- sensitivity analysis إلزامي عندما يمكن للأوزان قلب verdict.

قاعدة formal closure

إذا لم يمكن تحويل متغير إلى observation أو estimator أو intervention، يبقى في **Unresolved Formal Ledger** ويُخفض claim. الصياغة لا تتقوى بتعقيد الرمز؛ تتقوى بزيادة قابلية القياس والتفنيد.

M14-14 - الميكانيزم والعقد التشغيلي

State transition

```
trigger → validate inputs → capture state S_t → activate M14-14 → emit predicted_delta → execute bounded action/probe → independent readback → court → memory residue
```

الخطوات

1. صف benchmark contract.
2. صف deployment contract.
3. ابن invariant crosswalk.
4. اختبر mismatches بمهمات bridge.
5. لا تنقل claim خارج preserved dimensions.
6. حدث morphism عند تغير harness أو product.

عقد النواة

failure: malformed state, stale evidence, no-effect, adverse effect, authority mismatch recovery: stop/retry/alternate/rollback/escalate	preconditions: source freshness, authority, schema, budget postconditions: typed output + predicted delta + verifier target

يجب أن توجد feature flag تسمح بإطفاء الميكانيزم دون تغيير بقية harness قدر الإمكان. إن لم توجد هذه القابلية، يصبح causal attribution ضعيفاً حتى لو كان النظام يعمل جيداً.

M14-14 - التنبؤات، المفندات ومحكمة الهدم

مفندات مباشرة

- MorphLoss لا يتنبأ deployment gap.
- human expert mapping وحده يكفي.
- invariants لا تقبل operationalization.

تنبؤات فارقة

- تفسر بعض leaderboard gains التي لا تتحول إلى production value.
- bridge tasks تقلل surprise عند deployment.
- claims تصبح أضيق لكنها أكثر قابلية للدفاع.

أربعة خصوم نظريين

- Baseline محافظ: أبسط heuristic أو policy يؤدي الوظيفة بلا اسم جديد.
- تفسير harness: المكسب من tools/context/retries لا من الميكانيزم.
- تفسير model: backbone أقوى يزيل الحاجة للمكوّن.
- تفسير measurement: metric نفسها تكافئ السلوك الذي صُممت النظرية لإظهاره.

قرار المحكمة

إذا فشل causal ablation يُعطى **MERGE** أو **REVISE** أو **REJECT** بحسب ما إذا كان الفرق غير مستقل، محدود المجال، أو غير موجود. لا يسمح بتغيير اسم النظرية بعد null result مع الحفاظ على المضمون نفسه لتجنب الفشل.

M14-14 - التجربة، الأدوات والتدريب

التجربة الحاكمة: اختبر mechanisms على benchmark ثم sandbox deployment مشابه مع controlled shifts. قياس false predicted vs observed transfer، confidence، cost، failure modes.

ID	الأداة/الأصل	النوع	الدور
-XIV A131	Benchmark to Deployment Morphism State Profiler	TOOL	يستخرج متغيرات النظرية والمجهول والاعتماد على المصادر قبل التفعيل.
-XIV A132	Benchmark to Deployment Morphism Counterfactual Probe	TOOL	يولد intervention يفصل causal mechanism عن أقرب baseline.
-XIV A133	Benchmark to Deployment Morphism Failure Topology Mapper	TOOL	يجمع failure signatures ويربط recurrence بالسياق والسبب المحتمل.
-XIV A134	Benchmark to Deployment Morphism Calibration Engine	TOOL	يعاير thresholds/uncertainty/cost على held-out data.
-XIV A135	Benchmark to Deployment Morphism Matched-Budget Method	METHOD	يطابق model/tools/context/retries/time ويمنع attribution الزائف.
-XIV A136	Benchmark to Deployment Morphism Cross-Environment Transfer	METHOD	يفحص حفظ causal role عبر domain/tool/model/horizon.

وصفات التدريب/التقييم

- Mechanism Isolation Curriculum: حالات بسيطة تعزل المتغير المركزي قبل open-world confounders.
- Adversarial Boundary Curriculum: أمثلة داخل/خارج الحد وnegative activation لمنع overgeneralization.
- Matched-Budget Causal Evaluation: تدريب/تقييم يقارن تحت model/tools/context/retries/time متقاربة مع sham scaffold.
- Cross-Model/Domain Replication: إعادة التدريب/الاختبار في model/domain/tool family مختلفة مع stable-sign criterion.

الحد الأدنى للتجارب

- Minimal Causal Ablation: switchable implementation مقابل nearest parent/baseline تحت matched budget. | metrics: primary theory metric + task outcome + calibration + cost + critical regressions
- Boundary/Adversarial Test: اختبار failure modes والحدود بحالات مصممة لكسر التعميم. | metrics: false activation + missed failure + rollback + evidence integrity
- Transfer/Replication: نقل الآلية إلى model/domain/tool/horizon مختلف مع CI + preregistered metric. | metrics: effect sign stability + failure-signature transfer + resource frontier

> لا يُستخدم أي generated trajectory كتدريب إيجابي لمجرد أنه من النظام نفسه. يلزم verifier label أو world-state evidence أو rubric مستقل بحسب نوع المهمة.

M14-15 - الفجوة والأطروحة والنسب

Moussaif Harness-Resistant Capability Decomposition Theory

الفجوة الدقيقة

قد يبدو mechanism قوياً لأن harness يمنحه prompt أطول أو tools أكثر أو retries أو verifier أقوى. يلزم تفكيك القدرة إلى /model/harness/mechanism interactions.

> **الأطروحة الجديدة:** نعرّف factorial ablation matrix تفصل main effects والتفاعلات تحت matched budgets. القدرة المنسوبة للنظرية هي causal contribution المتبقي عبر عدة harness variants، لا أفضل run.

لماذا ليست alias؟

لا تُرقّى هذه العقدة لأنها تملك اسماً جديداً، بل لأنها تقترح متغيراً أو operator أو prediction غير ممثل مباشرة في أقرب الآباء. أقرب السلالة: **MUT anti-inflation rule**؛ **M13 Theory-to-Kernel Closure**؛ **MCD routing**. معيار novelty هو أن حذف العقدة أو استبدالها بأقرب parent يجب أن ينتج فرقاً قابلاً للقياس في route أو decision أو recovery أو evidence.

ضغط المجتمع العلمي

AgencyBench 2026؛ Trajectory Attribution 2026؛ Plan-RewardBench 2026. هذه الأعمال تُعامل كـproblem evidence وmechanism priors وbaselines، لا كتصديق على النظرية الداخلية.

> **الحد:** لا ينفي قيمة harness؛ يفصل فقط مصدر القدرة كي لا تنسب معمارية كاملة إلى اسم واحد.

M14-15 - الصياغة الرسمية والـobservables

$$Y = \beta 0 + \beta M \cdot Model + \beta H \cdot Harness + \beta K \cdot Mechanism + \beta MK + \beta HK + \beta MHK + \varepsilon; \; Credit_K = \text{robust causal contrast}$$

الرمز/المتغير	المعنى التشغيلي المقترح
Model	backbone
Harness	prompt/router/tools/retries
Mechanism	المكوّن النظري
Credit_K	فرق سببي مقاوم للتبديل

افتراضات الصياغة

- المتغيرات task-scoped وليست خصائص كونية ثابتة.
- كل variable يحتاج estimator أو proxy موصوفاً قبل الاستخدام التجريبي.
- الأوزان/العتبات تُعاير على train/validation وتُقفل قبل held-out evaluation.
- لا تُفسر العلاقة الجبرية كقانون تجريبي قبل نتائج مستقلة.
- sensitivity analysis إلزامي عندما يمكن للأوزان قلب verdict.

قاعدة formal closure

إذا لم يمكن تحويل متغير إلى observation أو estimator أو intervention، يبقى في **Unresolved Formal Ledger** ويُخفض claim. الصياغة لا تتقوى بتعقيد الرمز؛ تتقوى بزيادة قابلية القياس والتفنيد.

M14-15 - الميكانيزم والعقد التشغيلي

State transition

```
trigger → validate inputs → capture state S_t → activate M14-15 → emit predicted_delta → execute bounded action/probe → independent readback → court → memory residue
```

الخطوات

1. ثبت task/data.
2. أنشئ minimal baseline.
3. بدّل mechanism feature flag.
4. طابق token/tool/time budget.
5. نفذ factorial subsets مع seeds.
6. حلل interactions لا main effect فقط.
7. انشر null results.

عقد النواة

preconditions: source freshness, authority, schema, budget postconditions: typed output + predicted delta + verifier target	failure: malformed state, stale evidence, no-effect, adverse effect, authority mismatch recovery: stop/retry/alternate/rollback/escalate

يجب أن توجد feature flag تسمح بإطفاء الميكانيزم دون تغيير بقية harness قدر الإمكان. إن لم توجد هذه القابلية، يصبح causal attribution ضعيفاً حتى لو كان النظام يعمل جيداً.

M14-15 - التنبؤات، المفندات ومحكمة الهدم

مفندات مباشرة

- التصميم factorial مكلف ولا يزيد قراراً عملياً.
- matched budget مستحيل لدرجة تبطل attribution.
- الأثر يظهر فقط في harness واحد لكنه مهم واقعياً.

تنبؤات فارقة

- تختفي بعض "النظريات" عند إزالة harness advantage.
- mechanisms الحقيقية تظهر sign ثابتاً نسبياً.
- interaction effect قد يكون الإسهام الحقيقي بدلاً من component منفرد.

أربعة خصوم نظريين

- Baseline محافظ: أبسط heuristic أو policy يؤدي الوظيفة بلا اسم جديد.
- تفسير harness: المكسب من tools/context/retries لا من الميكانيزم.
- تفسير model: backbone أقوى يزيل الحاجة للمكوّن.
- تفسير measurement: metric نفسها تكافئ السلوك الذي صُممت النظرية لإظهاره.

قرار المحكمة

إذا فشل causal ablation يُعطى **MERGE** أو **REVISE** أو **REJECT** بحسب ما إذا كان الفرق غير مستقل، محدود المجال، أو غير موجود. لا يسمح بتغيير اسم النظرية بعد null result مع الحفاظ على المضمون نفسه لتجنب الفشل.

M14-15 - التجربة، الأدوات والتدريب

التجربة الحاكمة: 4-8 mechanisms على عدة models/harnesses مع sham scaffold. قياس transfer، interaction، cost، CI، β. معيار الترقية: causal contrast وليس raw score.

ID	الأداة/الأصل	النوع	الدور
-XIV A141	Harness Resistant Capability Decomposition State Profiler	TOOL	يستخرج متغيرات النظرية والمجهول والاعتماد على المصادر قبل التفعيل.
-XIV A142	Harness Resistant Capability Decomposition Counterfactual Probe	TOOL	يولد intervention يفصل causal mechanism عن أقرب baseline.
-XIV A143	Harness Resistant Capability Decomposition Failure Topology Mapper	TOOL	يجمع failure signatures ويربط recurrence بالسياق والسبب المحتمل.
-XIV A144	Harness Resistant Capability Decomposition Calibration Engine	TOOL	يعاير thresholds/uncertainty/cost على held-out data.
-XIV A145	Harness Resistant Capability Decomposition Matched-Budget Method	METHOD	يطابق model/tools/context/retries/time ويمنع attribution الزائف.
-XIV A146	Harness Resistant Capability Decomposition Cross-Environment Transfer	METHOD	يفحص حفظ causal role عبر /domain/tool/model horizon.

وصفات التدريب/التقييم

- Mechanism Isolation Curriculum: حالات بسيطة تعزل المتغير المركزي قبل open-world confounders.
- Adversarial Boundary Curriculum: أمثلة داخل/خارج الحد وnegative activation لمنع overgeneralization.
- Matched-Budget Causal Evaluation: تدريب/تقييم يقارن تحت model/tools/context/retries/time متقاربة مع sham scaffold.
- Cross-Model/Domain Replication: إعادة التدريب/الاختبار في model/domain/tool family مختلفة مع stable-sign criterion.

الحد الأدنى للتجارب

- Minimal Causal Ablation: switchable implementation مقابل nearest parent/baseline تحت matched budget. | metrics: primary theory metric + task outcome + calibration + cost + critical regressions
- Boundary/Adversarial Test: اختبار failure modes والحدود بحالات مصممة لكسر التعميم. | metrics: false activation + missed failure + rollback + evidence integrity
- Transfer/Replication: نقل الآلية إلى model/domain/tool/horizon مختلف مع CI + preregistered metric. | metrics: effect sign stability + failure-signature transfer + resource frontier

> لا يُستخدم أي generated trajectory كتدريب إيجابي لمجرد أنه من النظام نفسه. يلزم verifier label أو world-state evidence أو rubric مستقل بحسب نوع المهمة.

M14-16 - الفجوة والأطروحة والنسب

Moussaif Evaluator Independence Debt Theory

الفجوة الدقيقة

كلما شارك evaluator البيانات أو النموذج أو prompt lineage أو reward loop مع النظام المحسن، ازداد خطر self-confirmation. لا يكفي وجود verifier اسمي.

الأطروحة الجديدة: Evaluator Independence Debt يراكم اعتماداً بين optimizer وadaptive reuse، shared training data، shared model family، judge: leakage، evaluator tuning على نفس failures. الترقية تتطلب سداد جزء من الدين بمقيم مستقل أو formal/world evidence.

لماذا ليست alias؟

لا تُرقّى هذه العقدة لأنها تملك اسماً جديداً، بل لأنها تقترح متغيراً أو operator أو prediction غير ممثل مباشرة في أقرب الآباء. أقرب السلالة: **MUG**، **Falsification Court**، **evidence governance**، **M13 External Grounding**. معيار novelty هو أن حذف العقدة أو استبدالها بأقرب parent يجب أن ينتج فرقاً قابلاً للقياس في route أو decision أو recovery أو evidence.

ضغط المجتمع العلمي

Recursive Self-Improvement survey 2026، Plan-RewardBench 2026، GPT-Red 2026. هذه الأعمال تُعامل كـproblem evidence وmechanism priors وbaselines، لا كتصديق على النظرية الداخلية.

الحد: الدين لا يعني فساد evaluator، هو مخاطرة dependence يجب قياسها وسدادها نسبياً.

M14-16 - الصياغة الرسمية والobservables

$$EID = w1 \cdot ModelOverlap + w2 \cdot DataOverlap + w3 \cdot AdaptiveCoupling + w4 \cdot Leakage + w5 \cdot MetricGaming - w6 \cdot ExternalGrounding$$

الرمز/المتغير	المعنى التشغيلي المقترح
ModelOverlap	تشابه النموذج/العائلة
DataOverlap	تداخل بيانات
AdaptiveCoupling	تكيّف المقيم والنظام معاً
Leakage	انكشاف test
MetricGaming	تحسين على المقاس
ExternalGrounding	دليل مستقل/صوري/عالمي

افتراضات الصياغة

- المتغيرات task-scoped وليست خصائص كونية ثابتة.
- كل variable يحتاج estimator أو proxy موصوفاً قبل الاستخدام التجريبي.
- الأوزان/العتبات تُعاير على train/validation وتُقفل قبل held-out evaluation.
- لا تُفسر العلاقة الجبرية كقانون تجريبي قبل نتائج مستقلة.
- sensitivity analysis إلزامي عندما يمكن للأوزان قلب verdict.

قاعدة formal closure

إذا لم يمكن تحويل متغير إلى observation أو estimator أو intervention، يبقى في **Unresolved Formal Ledger** ويُخفض claim. الصياغة لا تتقوى بتعقيد الرمز؛ تتقوى بزيادة قابلية القياس والتفنيد.

M14-16 - الميكانيزم والعقد التشغيلي

State transition

```
trigger → validate inputs → capture state S_t → activate M14-16 → emit predicted_delta → execute bounded action/probe → independent readback → court → memory residue
```

الخطوات

1. سجل evaluator lineage.
2. احسب overlap indicators.
3. استخدم multiple evaluator classes.
4. أدخل hidden renewal.
5. افصل attacker/critic عن deployed system حيث يلزم.
6. لا تسمح للنظام بتعديل acceptance gate منفرداً.

عقد النواة

preconditions: source freshness, authority, schema, budget postconditions: typed output + predicted delta + verifier target	failure: malformed state, stale evidence, no-effect, adverse effect, authority mismatch recovery: stop/retry/alternate/rollback/escalate

يجب أن توجد feature flag تسمح بإطفاء الميكانيزم دون تغيير بقية harness قدر الإمكان. إن لم توجد هذه القابلية، يصبح causal attribution ضعيفاً حتى لو كان النظام يعمل جيداً.

M14-16 - التنبؤات، المفندات ومحكمة الهدم

مفندات مباشرة

- EID لا يرتبط بـreplication failure.
- external judges يضيفون noise أكبر من bias الذي يزيلونه.
- model overlap ليس predictor بعد ضبط task.

تنبؤات فارقة

- self-improvement gains تتكمش عند independent evaluator إذا كان الدين مرتفعاً.
- formal/world-state evidence أقل عرضة للدين من intrinsic self-judge.
- التجارب الطويلة تحتاج renewal لأن agent يتعلم rubric.

أربعة خصوم نظريين

- Baseline محافظ: أبسط heuristic أو policy يؤدي الوظيفة بلا اسم جديد.
- تفسير harness: المكسب من tools/context/retries لا من الميكانيزم.
- تفسير model: backbone أقوى يزيل الحاجة للمكوّن.
- تفسير measurement: metric نفسها تكافئ السلوك الذي صُممت النظرية لإظهاره.

قرار المحكمة

إذا فشل causal ablation يُعطى **MERGE** أو **REVISE** أو **REJECT** بحسب ما إذا كان الفرق غير مستقل، محدود المجال، أو غير موجود. لا يسمح بتغيير اسم النظرية بعد null result مع الحفاظ على المضمون نفسه لتجنب الفشل.

M14-16 - التجربة، الأدوات والتدريب

التجربة الحاكمة: اختبر self-refinement تحت hidden human subset، formal verifier، different-model judge، same-model judge. قياس optimism gap، rank consistency، replication.

ID	الأداة/الأصل	النوع	الدور
XIV-A151	Evaluator Independence Debt State Profiler	TOOL	يستخرج متغيرات النظرية والمجهول والاعتماد على المصادر قبل التفعيل.
XIV-A152	Evaluator Independence Debt Counterfactual Probe	TOOL	يولد intervention يفصل causal mechanism عن أقرب baseline.
XIV-A153	Evaluator Independence Debt Failure Topology Mapper	TOOL	يجمع failure signatures ويربط recurrence بالسياق والسبب المحتمل.
XIV-A154	Evaluator Independence Debt Calibration Engine	TOOL	يعاير thresholds/uncertainty/cost على held-out data.
XIV-A155	Evaluator Independence Debt Matched-Budget Method	METHOD	يطابق model/tools/context/retries/time ويمنع attribution الزائف.
XIV-A156	Evaluator Independence Debt Cross-Environment Transfer	METHOD	يفحص حفظ causal role عبر domain/tool/model/horizon.

وصفات التدريب/التقييم

- Mechanism Isolation Curriculum: حالات بسيطة تعزل المتغير المركزي قبل open-world confounders.
- Adversarial Boundary Curriculum: أمثلة داخل/خارج الحد وnegative activation لمنع overgeneralization.
- Matched-Budget Causal Evaluation: تدريب/تقييم يقارن تحت model/tools/context/retries/time متقاربة مع sham scaffold.
- Cross-Model/Domain Replication: إعادة التدريب/الاختبار في model/domain/tool family مختلفة مع stable-sign criterion.

الحد الأدنى للتجارب

- Minimal Causal Ablation: switchable implementation مقابل nearest parent/baseline تحت matched budget. | metrics: primary theory metric + task outcome + calibration + cost + critical regressions
- Boundary/Adversarial Test: اختبار failure modes والحدود بحالات مصممة لكسر التعميم. | metrics: false activation + missed failure + rollback + evidence integrity
- Transfer/Replication: نقل الآلية إلى model/domain/tool/horizon مختلف مع CI + preregistered metric. | metrics: effect sign stability + failure-signature transfer + resource frontier

> لا يُستخدم أي generated trajectory كتدريب إيجابي لمجرد أنه من النظام نفسه. يلزم verifier label أو world-state evidence أو rubric مستقل بحسب نوع المهمة.

M14-17 - الفجوة والأطروحة والنسب

Moussaif Scientific Reproduction Yield Theory

الفجوة الدقيقة

عدد التجارب الناجحة داخل المختبر لا يساوي معرفة مستقرة. نحتاج metric يكافئ reproduction المستقل والمتنوع ويعاقب dependence والانتقاء.

> **الأطروحة الجديدة:** Reproduction Yield هو نسبة محاولات مستقلة تعيد sign/threshold الأساسي ضمن uncertainty، موزونة بتنوع /model/domain/team harness ومسبقة التسجيل، مع احتساب null/negative attempts.

لماذا ليست alias؟

لا تُرقّى هذه العقدة لأنها تملك اسماً جديداً، بل لأنها تقترح متغيراً أو operator أو prediction غير ممثل مباشرة في أقرب الآباء. أقرب السلالة: **MUG publication governance**؛ **Evidence Closure**؛ **M13 External Grounding**. معيار novelty هو أن حذف العقدة أو استبدالها بأقرب parent يجب أن ينتج فرقاً قابلاً للقياس في route أو decision أو recovery أو evidence.

ضغط المجتمع العلمي

CUBE 2026؛ Trajectory Attribution 2026؛ Recursive Self-Improvement 2026. هذه الأعمال تُعامل كـevidence problem وpriors mechanism وbaselines، لا كتصديق على النظرية الداخلية.

> **الحد:** لا يطلب تطابقاً مطلقاً؛ يعرّف ما يجب أن يبقى وما يسمح له أن يتغير مسبقاً.

M14-17 - الصياغة الرسمية والobservables

SRY = Σ_j Indep_j·Diversity_j·Reproduce_j / Σ_j Attempt_j; report with CI and failure taxonomy

الرمز/المتغير	المعنى التشغيلي المقترح
Indep_j	استقلال الفريق/الهارنس/البيانات
Diversity_j	تنوع شروط الإعادة
Reproduce_j	حفظ sign/criterion لا التطابق الحرفي

افتراضات الصياغة

- المتغيرات task-scoped وليست خصائص كونية ثابتة.
- كل variable يحتاج estimator أو proxy موصوفاً قبل الاستخدام التجريبي.
- الأوزان/العتبات تُعاير على train/validation وتُقفل قبل held-out evaluation.
- لا تُفسر العلاقة الجبرية كقانون تجريبي قبل نتائج مستقلة.
- sensitivity analysis إلزامي عندما يمكن للأوزان قلب verdict.

قاعدة formal closure

إذا لم يمكن تحويل متغير إلى observation أو estimator أو intervention، يبقى في **Unresolved Formal Ledger** ويُخفض claim. الصياغة لا تتقوى بتعقيد الرمز، تتقوى بزيادة قابلية القياس والتفنيد.

M14-17 - الميكانيزم والعقد التشغيلي

State transition

```
trigger → validate inputs → capture state S_t → activate M14-17 → emit predicted_delta → execute bounded action/probe → independent readback → court → memory residue
```

الخطوات

1. أنشئ replay package.
2. ثبت hashes/configs.
3. سجل attempts قبل معرفة النتائج.
4. فرق exact reproduction عن conceptual replication.
5. حافظ على negative results.
6. حدث theory envelope من failures.

عقد النواة

failure: malformed state, stale evidence, no-effect, adverse effect, authority mismatch recovery: stop/retry/alternate/rollback/escalate	preconditions: source freshness, authority, schema, budget postconditions: typed output + predicted delta + verifier target

يجب أن توجد feature flag تسمح بإطفاء الميكانيزم دون تغيير بقية harness قدر الإمكان. إن لم توجد هذه القابلية، يصبح causal attribution ضعيفاً حتى لو كان النظام يعمل جيداً.

M14-17 - التنبؤات، المفندات ومحكمة الهدم

مفندات مباشرة

- SRY يعاقب innovations حساسة للبيئة بشكل غير عادل.
- independence يصعب تعريفها.
- yield مرتفع بسبب tasks سهلة غير حاسمة.

تنبؤات فارقة

- theories الأقوى تملك yield قابل للحفاظ خارج harness الأصلي.
- إعادة الإنتاج تكشف hidden dependencies أكثر من إضافة benchmarks مشابهة.
- نشر null results يخفض theory inflation.

أربعة خصوم نظريين

- Baseline محافظ: أبسط heuristic أو policy يؤدي الوظيفة بلا اسم جديد.
- تفسير harness: المكسب من tools/context/retries لا من الميكانيزم.
- تفسير model: backbone أقوى يزيل الحاجة للمكوّن.
- تفسير measurement: metric نفسها تكافئ السلوك الذي صُممت النظرية لإظهاره.

قرار المحكمة

إذا فشل causal ablation يُعطى **MERGE** أو **REVISE** أو **REJECT** بحسب ما إذا كان الفرق غير مستقل، محدود المجال، أو غير موجود. لا يسمح بتغيير اسم النظرية بعد null result مع الحفاظ على المضمون نفسه لتجنب الفشل.

M14-17 - التجربة، الأدوات والتدريب

التجربة الحاكمة: أعد 3-5 mechanisms عبر models/tools/harnesses وفريق/agent evaluator منفصل. قياس failure signature، CI overlap، sign stability، reproduction cost.

ID	الأداة/الأصل	النوع	الدور
XIV-A161	Scientific Reproduction Yield State Profiler	TOOL	يستخرج متغيرات النظرية والمجهول والاعتماد على المصادر قبل التفعيل.
XIV-A162	Scientific Reproduction Yield Counterfactual Probe	TOOL	يولد intervention يفصل causal mechanism عن أقرب baseline.
XIV-A163	Scientific Reproduction Yield Failure Topology Mapper	TOOL	يجمع failure signatures ويربط recurrence بالسياق والسبب المحتمل.
XIV-A164	Scientific Reproduction Yield Calibration Engine	TOOL	يعاير thresholds/uncertainty/cost على held-out data.
XIV-A165	Scientific Reproduction Yield Matched-Budget Method	METHOD	يطابق model/tools/context/retries/time ويمنع attribution الزائف.
XIV-A166	Scientific Reproduction Yield Cross-Environment Transfer	METHOD	يفحص حفظ causal role عبر domain/tool/model/horizon.

وصفات التدريب/التقييم

- Mechanism Isolation Curriculum: حالات بسيطة تعزل المتغير المركزي قبل open-world confounders.
- Adversarial Boundary Curriculum: أمثلة داخل/خارج الحد وnegative activation لمنع overgeneralization.
- Matched-Budget Causal Evaluation: تدريب/تقييم يقارن تحت model/tools/context/retries/time متقاربة مع sham scaffold.
- Cross-Model/Domain Replication: إعادة التدريب/الاختبار في model/domain/tool family مختلفة مع stable-sign criterion.

الحد الأدنى للتجارب

- Minimal Causal Ablation: switchable implementation مقابل nearest parent/baseline تحت matched budget. | metrics: primary theory metric + task outcome + calibration + cost + critical regressions
- Boundary/Adversarial Test: اختبار failure modes والحدود بحالات مصممة لكسر التعميم. | metrics: false activation + missed failure + rollback + evidence integrity
- Transfer/Replication: نقل الآلية إلى model/domain/tool/horizon مختلف مع CI + effect sign stability + preregistered metric. | metrics: failure-signature transfer + resource frontier

> لا يُستخدم أي generated trajectory كتدريب إيجابي لمجرد أنه من النظام نفسه. يلزم verifier label أو world-state evidence أو rubric مستقل بحسب نوع المهمة.

M14-18 - الفجوة والأطروحة والنسب

Moussaif Constitutional Mutation Stability Theory

الفجوة الدقيقة

النظام المتطور قد يحسن performance بتعديل router/memory/tools/evaluator، لكنه قد يغير القيود التي تجعل التحسن مقبولاً. نحتاج فصل mutable capabilities عن constitutional invariants.

الأطروحة الجديدة: كل mutation تمر بـbranch مع immutable root constitution أو externally governed version. تُقبل فقط إذا حسنت هدفاً معلناً دون regression حرجة، مع rollback وindependent evaluation. تعديل الدستور نفسه يحتاج meta-governance منفصلة ولا يثبت صحته بالحلقة التي تقترحه.

لماذا ليست alias؟

لا تُرقّى هذه العقدة لأنها تملك اسماً جديداً، بل لأنها تقترح متغيراً أو operator أو prediction غير ممثل مباشرة في أقرب الآباء. أقرب السلالة: **M13 MUG constitution؛ Recursive Improvement؛ Governance Space.** معيار novelty هو أن حذف العقدة أو استبدالها بأقرب parent يجب أن ينتج فرقاً قابلاً للقياس في route أو decision أو recovery أو evidence.

ضغط المجتمع العلمي

Recursive Self-Improvement 2026؛ GPT-Red 2026؛ DeepMind AI Control 2026؛ Anthropic trust/containment 2026. هذه الأعمال تُعامل كـproblem evidence وmechanism priors وbaselines، لا كتصديق على النظرية الداخلية.

الحد: لا يثبت أمان recursive self-improvement المفتوح؛ إنه بروتوكول تجريبي لتقييد طفرة محددة.

M14-18 - الصياغة الرسمية والobservables

CMS = Gain_validated · InvariantPreservation · RollbackVerified · EvaluatorIndependence - CriticalRegressionRisk

الرمز/المتغير	المعنى التشغيلي المقترح
Gain_validated	مكسب متحقق causalياً
InvariantPreservation	حفظ invariants
RollbackVerified	نجاح الاسترجاع
EvaluatorIndependence	استقلال التقييم
CriticalRegressionRisk	مخاطر فشل حرج

افتراضات الصياغة

- المتغيرات task-scoped وليست خصائص كونية ثابتة.
- كل variable يحتاج estimator أو proxy موصوفاً قبل الاستخدام التجريبي.
- الأوزان/العتبات تُعاير على train/validation وتُقفل قبل held-out evaluation.
- لا تُفسر العلاقة الجبرية كقانون تجريبي قبل نتائج مستقلة.
- sensitivity analysis إلزامي عندما يمكن للأوزان قلب verdict.

قاعدة formal closure

إذا لم يمكن تحويل متغير إلى observation أو estimator أو intervention، يبقى في **Unresolved Formal Ledger** ويُخفض claim. الصياغة لا تتقوى بتعقيد الرمز؛ تتقوى بزيادة قابلية القياس والتفنيد.

M14-18 - الميكانيزم والعقد التشغيلي

State transition

```
trigger → validate inputs → capture state S_t → activate M14-18 → emit predicted_delta → execute bounded action/probe → independent readback → court → memory residue
```

الخطوات

1. افصل constitution root عن mutable branch.
2. حوّل mutation إلى typed diff.
3. شغّل pre/post invariant tests.
4. اختبر adversarial/negative activation.
5. استخدم evaluator مستقل.
6. نفذ rollback drill.
7. PROMOTE فقط بعد replication، وإلا HOLD/REVISE.

عقد النواة

preconditions: source freshness, authority, schema, budget postconditions: typed output + predicted delta + verifier target	failure: malformed state, stale evidence, no-effect, adverse effect, authority mismatch recovery: stop/retry/alternate/rollback/escalate

يجب أن توجد feature flag تسمح بإطفاء الميكانيزم دون تغيير بقية harness قدر الإمكان. إن لم توجد هذه القابلية، يصبح causal attribution ضعيفاً حتى لو كان النظام يعمل جيداً.

M14-18 - التنبؤات، المفندات ومحكمة الهدم

مفندات مباشرة

- root constraints تمنع اكتشاف تحسينات ضرورية.
- rollback لا يمكن تحقيقه في learning systems المستمرة.
- independent evaluation لا يقلل catastrophic regression.

تنبؤات فارقة

- self-evolution الأكثر أماناً سيكون أبطأ لكنه أكثر قابلية لإعادة الإنتاج.
- mutations التي تغيّر evaluator والدستور معاً ستظهر optimism أعلى.
- rollback drills تكشف hidden state غير مرجوع.

أربعة خصوم نظريين

- Baseline محافظ: أبسط heuristic أو policy يؤدي الوظيفة بلا اسم جديد.
- تفسير harness: المكسب من tools/context/retries لا من الميكانيزم.
- تفسير model: backbone أقوى يزيل الحاجة للمكوّن.
- تفسير measurement: metric نفسها تكافئ السلوك الذي صُممت النظرية لإظهاره.

قرار المحكمة

إذا فشل causal ablation يُعطى **MERGE** أو **REVISE** أو **REJECT** بحسب ما إذا كان الفرق غير مستقل، محدود المجال، أو غير موجود. لا يسمح بتغيير اسم النظرية بعد null result مع الحفاظ على المضمون نفسه لتجنب الفشل.

M14-18 - التجربة، الأدوات والتدريب

التجربة الحاكمة: حلقات تحسين harness/memory/router تحت branches؛ قارن CMS، fixed constitution، unrestricted self-edit. قياس regressions، gain، rollback success، evaluator gap، reproduction.

ID	الأداة/الأصل	النوع	الدور
-XIV A171	Constitutional Mutation Stability State Profiler	TOOL	يستخرج متغيرات النظرية والمجهول والاعتماد على المصادر قبل التفعيل.
-XIV A172	Constitutional Mutation Stability Counterfactual Probe	TOOL	يولد intervention يفصل causal mechanism عن أقرب baseline.
-XIV A173	Constitutional Mutation Stability Failure Topology Mapper	TOOL	يجمع failure signatures ويربط recurrence بالسياق والسبب المحتمل.
-XIV A174	Constitutional Mutation Stability Calibration Engine	TOOL	يعاير thresholds/uncertainty/cost على held-out data.
-XIV A175	Constitutional Mutation Stability Matched-Budget Method	METHOD	يطابق model/tools/context/retries/time ويمنع attribution الزائف.
-XIV A176	Constitutional Mutation Stability Cross-Environment Transfer	METHOD	يفحص حفظ causal role عبر domain/tool/model/horizon.

وصفات التدريب/التقييم

- Mechanism Isolation Curriculum: حالات بسيطة تعزل المتغير المركزي قبل open-world confounders.
- Adversarial Boundary Curriculum: أمثلة داخل/خارج الحد وnegative activation لمنع overgeneralization.
- Matched-Budget Causal Evaluation: تدريب/تقييم يقارن تحت model/tools/context/retries/time متقاربة مع sham scaffold.
- Cross-Model/Domain Replication: إعادة التدريب/الاختبار في model/domain/tool family مختلفة مع stable-sign criterion.

الحد الأدنى للتجارب

- Minimal Causal Ablation: switchable implementation مقابل nearest parent/baseline تحت matched budget. | metrics: primary theory metric + task outcome + calibration + cost + critical regressions
- Boundary/Adversarial Test: اختبار failure modes والحدود بحالات مصممة لكسر التعميم. | metrics: false activation + missed failure + rollback + evidence integrity
- Transfer/Replication: نقل الآلية إلى model/domain/tool/horizon مختلف مع CI + preregistered metric. | metrics: effect sign stability + failure-signature transfer + resource frontier

> لا يُستخدم أي generated trajectory كتدريب إيجابي لمجرد أنه من النظام نفسه. يلزم verifier label أو world-state evidence أو rubric مستقل بحسب نوع المهمة.

الاقتران 1: كيف تتحول النظريات إلى نظام لا قائمة

المسار P14-EF لا يعمل كسلسلة خطية صارمة، بل كـgraph. نحدد لكل عقدة inputs/outputs ونقاط failure. الاقتران المقبول يجب أن ينتج **complementarity measurable**: إذا كان تجميع نظريتين لا يتجاوز أفضل واحدة منهما تحت budget متطابق، نخفض claim عن «التكامل».

- M14-13: نظرية مساعف لجبهة عمومية الإغلاق - output حاكم: تحويلات نموذج/مجال/أداة/أفق/موارد
- M14-14: نظرية مساعف للتحويل بين المعيار والنشر - output حاكم: invariant وظيفي
- M14-15: نظرية مساعف لتفكيك القدرة المقاوم للهارنس - output حاكم: backbone
- M14-16: نظرية مساعف لدَين استقلال المقيم - output حاكم: تشابه النموذج/العائلة
- M14-17: نظرية مساعف لعائد إعادة الإنتاج العلمي - output حاكم: استقلال الفريق/الهارنس/البيانات
- M14-18: نظرية مساعف لثبات الطفرة الدستورية - output حاكم: مكسب متحقق causalياً

$$\text{Complementarity}(A,B)=\Delta U(A+B)-\max(\Delta U(A),\Delta U(B)) - \text{InteractionCost}$$

هذا المقياس design heuristic؛ التجربة factorial هي التي تحدد interaction. يظل الهدف منع corpus من افتراض أن مزيداً من الطبقات يعني مزيداً من intelligence.

الاقتران 2: Graph نظري إلى Graph تنفيذي

```
TheoryNode → FormalSchema → RuntimeService → TelemetryEvents → EvidencePacket
        ↘ nearest_baseline   ↘ feature_flag   ↘ counterfactual_run
Court: duplicate? mechanism? causal? transferable? reproducible?
```

تستدعي MOUSSAIXT فقط أقل مجموعة من services اللازمة، MUG يمنع promotion إذا كانت telemetry ناقصة، وMUT يحدث lineage بعد verdict. هذه البنية تفصل «التأثير على runtime» عن «الوجود في كتاب».

Invariants المشتركة

- observed state يتقدم على remembered/generated state
- authority يُفحص عند action time
- world effect يحتاج readback مستقل حيث يمكن
- كل failure يترك residue قابل للتعلم لا يُخفى
- كل self-update versioned وقابل للrollback بقدر الإمكان
- claim لا يتجاوز transfer envelope الذي اختبر فعلياً

مختبر الضغط 1: العدسة السببية: من الارتباط إلى intervention

يجب فصل three layers: correlation التي تصلح للاكتشاف، causal hypothesis التي تصلح لاختيار تجربة، وverified causal contrast الذي يبرر attribution محدوداً. في Part XIV لا تصبح edge سببية لأن النص يصفها كذلك؛ يجب أن نحدد do-variable أو feature flag أو perturbation يعزلها.

> **تطبيق على P14-EF:** التجربة الجيدة تقلل مساحة التفسيرات البديلة، لا تزيد عدد الرسوم. تصبح كل نظرية في هذا المسار مسؤولة عن تعريف metric وفشل يخصها، ثم تُقارن بالعقد الشقيقة لتجنب double-counting.

نضيف إلى كل run ثلاثة سجلات: DecisionLedger يبين لماذا فُعّل المكوّن، EvidenceLedger يبين ما شوهد فعلاً، ResidualLedger يبين ما لم يفسر أو يُغلق. بهذه الثلاثية يمكن إعادة تفسير run بعد ظهور دليل جديد دون إعادة كتابة التاريخ.

الحد العلمي: هذه الصفحة تبني منهج اختبار؛ لا تحتوي نتيجة تجريبية ما لم يرفق experiment packet مستقل.

مختبر الضغط 2: عدسة الاقتصاد المعرفي: فائدة لكل وحدة compute

الـAGI-oriented architecture لا تُقاس بالعمق اللفظي. كل طبقة تدفع كلفة memory construction، tool calls، latency، tokens وverification. لذلك نحسب quality-cost frontier ونرفض الميكانيزم الذي يضيف كلفة بلا مكسب متكرر أو يقلل robustness.

> **تطبيق على P14-EF:** المكوّن الذي يرفع score 1% ويضاعف الكلفة قد يخسر على frontier حتى لو كان statistically positive. تصبح كل نظرية في هذا المسار مسؤولة عن تعريف metric وفشل يخصها، ثم تُقارن بالعقد الشقيقة لتجنب double-counting.

نضيف إلى كل run ثلاثة سجلات: DecisionLedger يبين لماذا فُعّل المكوّن، EvidenceLedger يبين ما شوهد فعلاً، ResidualLedger يبين ما لم يفسر أو يُغلق. بهذه الثلاثية يمكن إعادة تفسير run بعد ظهور دليل جديد دون إعادة كتابة التاريخ.

الحد العلمي: هذه الصفحة تبني منهج اختبار، لا تحتوي نتيجة تجريبية ما لم يرفق experiment packet مستقل.

مختبر الضغط 3: عدسة negative activation: متى يجب ألا تعمل النظرية؟

أقوى علامة على mechanism حقيقي هي امتلاك شرط عدم التفعيل. نُنشئ حالات يكون فيها baseline البسيط هو الصحيح ونقيس false activation. هذا يمنع كل نظرية من التحول إلى طبقة إجبارية تُستخدم في كل مهمة لمجرد وجودها.

> **تطبيق على P14-EF:** الذكاء الحاكمي يتضمن حق الامتناع عن استخدام تعقيد لا يحتاجه السؤال. تصبح كل نظرية في هذا المسار مسؤولة عن تعريف metric وفشل يخصها، ثم تقارن بالعقد الشقيقة لتجنب double-counting.

نضيف إلى كل run ثلاثة سجلات: DecisionLedger يبين لماذا فُعّل المكوّن، EvidenceLedger يبين ما شوهد فعلاً، ResidualLedger يبين ما لم يفسر أو يُغلق. بهذه الثلاثية يمكن إعادة تفسير run بعد ظهور دليل جديد دون إعادة كتابة التاريخ.

الحد العلمي: هذه الصفحة تبني منهج اختبار، لا تحتوي نتيجة تجريبية ما لم يرفق experiment packet مستقل.

مختبر الضغط 4: عدسة التفاعلات: المكوّن قد لا يعمل منفرداً

قد تظهر قيمة النظرية فقط مع memory أو verifier أو tool contract. بدلاً من إنكار ذلك، نفصل main effect عن interactions بعينة factorial. إذا كان الأثر تفاعلياً نعيد تسمية claim: ليس «M14-X يحسن الأداء» بل «M14-X مع Y تحت Z يغير recovery/quality».

> **تطبيق على P14-EF:** الضبط هنا يقوي النظرية لأنه يمنع التعميم الكاذب. تصبح كل نظرية في هذا المسار مسؤولة عن تعريف metric وفشل يخصها، ثم تُقارن بالعقد الشقيقة لتجنب double-counting.

نضيف إلى كل run ثلاثة سجلات: DecisionLedger يبين لماذا فُعّل المكوّن؛ EvidenceLedger يبين ما شوهد فعلاً؛ ResidualLedger يبين ما لم يفسر أو يُغلق. بهذه الثلاثية يمكن إعادة تفسير run بعد ظهور دليل جديد دون إعادة كتابة التاريخ.

الحد العلمي: هذه الصفحة تبني منهج اختبار؛ لا تحتوي نتيجة تجريبية ما لم يرفق experiment packet مستقل.

مختبر الضغط 5: عدسة الزمن الطويل: الخطأ المتأخر أهم من الجواب اللحظي

في trajectories طويلة، صحة خطوة محلية لا تكفي. نقيس recovery and attribution، drift، stale memory، obligation loss، delayed failure. هذه العدسة تربط Part XIV بحوث long-horizon agents التي تكشف انهياراً مع طول المسار.

> **تطبيق على P14-EF:** كل mechanism يجب أن يحدد كيف يتغير مع horizon لا أن يفترض stationarity. تصبح كل نظرية في هذا المسار مسؤولة عن تعريف metric وفشل يخصها، ثم تُقارن بالعقد الشقيقة لتجنب double-counting.

نضيف إلى كل run ثلاثة سجلات: DecisionLedger يبين لماذا فُعّل المكوّن، EvidenceLedger يبين ما شوهد فعلاً، ResidualLedger يبين ما لم يفسر أو يُغلق. بهذه الثلاثية يمكن إعادة تفسير run بعد ظهور دليل جديد دون إعادة كتابة التاريخ.

الحد العلمي: هذه الصفحة تبني منهج اختبار، لا تحتوي نتيجة تجريبية ما لم يرفق experiment packet مستقل.

مختبر الضغط 6: عدسة نقل الدليل: من benchmark إلى deployment

نقيد claim بـinvariants التي حُفظت بين بيئة الاختبار والنشر: failure consequences، user feedback، cost، observability، authority، action space. إذا تغيّرت invariants يتغير سقف claim حتى لو كان benchmark score ثابتاً.

> **تطبيق على P14-EF:** النشر ليس benchmark أكبر؛ إنه توزيع مختلف للأخطاء والآثار. تصبح كل نظرية في هذا المسار مسؤولة عن تعريف metric وفشل يخصها، ثم تُقارن بالعقد الشقيقة لتجنب double-counting.

نضيف إلى كل run ثلاثة سجلات: DecisionLedger يبين لماذا فُعّل المكوّن، EvidenceLedger يبين ما شوهد فعلاً، ResidualLedger يبين ما لم يفسر أو يُغلق. بهذه الثلاثية يمكن إعادة تفسير run بعد ظهور دليل جديد دون إعادة كتابة التاريخ.

الحد العلمي: هذه الصفحة تبني منهج اختبار؛ لا تحتوي نتيجة تجريبية ما لم يرفق experiment packet مستقل.

مختبر الضغط 7: عدسة استقلال المقيم: من يحكم على الحكم؟

عندما يكون evaluator من نفس العائلة أو البيانات أو loop، قد يظهر التحسن بسبب agreement لا truth. لذلك نسجل evaluator lineage ونجرب formal/world-state/independent judges عند الإمكان.

> **تطبيق على P14-EF:** كل self-improvement loop هي أيضاً فرضية حول صلاحية signal الذي يقيس التحسن. تصبح كل نظرية في هذا المسار مسؤولة عن تعريف metric وفشل يخصها، ثم تُقارن بالعقد الشقيقة لتجنب double-counting.

نضيف إلى كل run ثلاثة سجلات: DecisionLedger يبين لماذا فُعّل المكوّن، EvidenceLedger يبين ما شوهد فعلاً، ResidualLedger يبين ما لم يفسر أو يُغلق. بهذه الثلاثية يمكن إعادة تفسير run بعد ظهور دليل جديد دون إعادة كتابة التاريخ.

الحد العلمي: هذه الصفحة تبني منهج اختبار، لا تحتوي نتيجة تجريبية ما لم يرفق experiment packet مستقل.

مختبر الضغط 8: عدسة التكاثر العلمي: هل ينجو الأثر خارج المختبر؟

نحوّل reproducibility من ملحق إلى جزء من معنى النظرية. replay package يشمل negative results، budgets، seeds، raw traces، versions، hashes، ثم conceptual replication على بيئة مختلفة.

> **تطبيق على P14-EF:** الأثر الذي لا يعاد قد يبقى engineering trick محلياً لكنه لا يرقى إلى claim عام. تصبح كل نظرية في هذا المسار مسؤولة عن تعريف metric وفشل يخصها، ثم تُقارن بالعقد الشقيقة لتجنب double-counting.

نضيف إلى كل run ثلاثة سجلات: DecisionLedger يبين لماذا فُعّل المكوّن، EvidenceLedger يبين ما شوهد فعلاً، ResidualLedger يبين ما لم يفسر أو يُغلق. بهذه الثلاثية يمكن إعادة تفسير run بعد ظهور دليل جديد دون إعادة كتابة التاريخ.

الحد العلمي: هذه الصفحة تبني منهج اختبار، لا تحتوي نتيجة تجريبية ما لم يرفق experiment packet مستقل.

سجل الأدوات والأصول التنفيذية - 1/2

ID	Owner	Asset	Class	Purpose
-XIV A121	M14-13	Generality Closure Frontier State Profiler	TOOL	يستخرج متغيرات النظرية والمجهول والاعتماد على المصادر قبل التفعيل.
-XIV A122	M14-13	Generality Closure Frontier Counterfactual Probe	TOOL	يولد intervention يفصل causal mechanism عن أقرب baseline.
-XIV A123	M14-13	Generality Closure Frontier Failure Topology Mapper	TOOL	يجمع failure signatures ويربط recurrence بالسياق والسبب المحتمل.
-XIV A124	M14-13	Generality Closure Frontier Calibration Engine	TOOL	يعاير thresholds/uncertainty/cost على held-out data.
-XIV A125	M14-13	Generality Closure Frontier Matched-Budget Method	METHOD	يطابق model/tools/context/retries/time ويمنع attribution الزائف.
-XIV A126	M14-13	Generality Closure Frontier Cross-Environment Transfer	METHOD	يفحص حفظ causal role عبر /domain/tool/model horizon.
-XIV A127	M14-13	Generality Closure Frontier Evidence Promotion Protocol	PROTOCOL	ينظم N0-N5/E0-E5 ويمنع claim expansion قبل replication.
-XIV A128	M14-13	Generality Closure Frontier Negative-Result Protocol	PROTOCOL	يحفظ الفشل والحدود ويمنع إعادة تسمية الفرضية بعد عدم نجاحها.
-XIV A129	M14-13	Generality Closure Frontier Runtime Service	STACK	خدمة feature-flagged مستقلة لها state/telemetry/rollback.
-XIV A130	M14-13	Generality Closure Frontier Proof/Replay Adapter	STACK	يحوّل trajectory إلى manifest + evidence packet + replay receipt.
-XIV A131	M14-14	Benchmark to Deployment Morphism State Profiler	TOOL	يستخرج متغيرات النظرية والمجهول والاعتماد على المصادر قبل التفعيل.
-XIV A132	M14-14	Benchmark to Deployment Morphism Counterfactual Probe	TOOL	يولد intervention يفصل causal mechanism عن أقرب baseline.
-XIV A133	M14-14	Benchmark to Deployment Morphism Failure Topology Mapper	TOOL	يجمع failure signatures ويربط recurrence بالسياق والسبب المحتمل.
-XIV A134	M14-14	Benchmark to Deployment Morphism Calibration Engine	TOOL	يعاير thresholds/uncertainty/cost على held-out data.
-XIV A135	M14-14	Benchmark to Deployment Morphism Matched-Budget Method	METHOD	يطابق model/tools/context/retries/time ويمنع attribution الزائف.
-XIV A136	M14-14	Benchmark to Deployment Morphism Cross-Environment Transfer	METHOD	يفحص حفظ causal role عبر /domain/tool/model horizon.
-XIV A137	M14-14	Benchmark to Deployment Morphism Evidence Promotion Protocol	PROTOCOL	ينظم N0-N5/E0-E5 ويمنع claim expansion قبل replication.
-XIV A138	M14-14	Benchmark to Deployment Morphism Negative-Result Protocol	PROTOCOL	يحفظ الفشل والحدود ويمنع إعادة تسمية الفرضية بعد عدم نجاحها.
-XIV A139	M14-14	Benchmark to Deployment Morphism Runtime Service	STACK	خدمة feature-flagged مستقلة لها state/telemetry/rollback.
-XIV A140	M14-14	Benchmark to Deployment Morphism Proof/Replay Adapter	STACK	يحوّل trajectory إلى manifest + evidence packet + replay receipt.
-XIV A141	M14-15	Harness Resistant Capability Decomposition State Profiler	TOOL	يستخرج متغيرات النظرية والمجهول والاعتماد على المصادر قبل التفعيل.
-XIV A142	M14-15	Harness Resistant Capability Decomposition Counterfactual Probe	TOOL	يولد intervention يفصل causal mechanism عن أقرب baseline.
-XIV A143	M14-15	Harness Resistant Capability Decomposition Failure Topology Mapper	TOOL	يجمع failure signatures ويربط recurrence بالسياق والسبب المحتمل.
-XIV A144	M14-15	Harness Resistant Capability Decomposition Calibration Engine	TOOL	يعاير thresholds/uncertainty/cost على held-out data.
	M14-15		METHOD	

-XIV A145		Harness Resistant Capability Decomposition Matched-Budget Method		يطابق model/tools/context/retries/time ويمنع attribution الزائف.
-XIV A146	M14-15	Harness Resistant Capability Decomposition Cross-Environment Transfer	METHOD	يفحص حفظ causal role عبر /domain/tool/model horizon.
-XIV A147	M14-15	Harness Resistant Capability Decomposition Evidence Promotion Protocol	PROTOCOL	ينظم N0-N5/E0-E5 ويمنع claim expansion قبل replication.
-XIV A148	M14-15	Harness Resistant Capability Decomposition Negative-Result Protocol	PROTOCOL	يحفظ الفشل والحدود ويمنع إعادة تسمية الفرضية بعد عدم نجاحها.
-XIV A149	M14-15	Harness Resistant Capability Decomposition Runtime Service	STACK	خدمة feature-flagged مستقلة لها state/telemetry/rollback.
-XIV A150	M14-15	Harness Resistant Capability Decomposition Proof/Replay Adapter	STACK	يحوّل trajectory إلى manifest + evidence packet + replay receipt.

الحالة: SPECIFICATION. لا يقال deployed أو validated إلا بعد وجود artifact/run/test evidence.

سجل الأدوات والأصول التنفيذية - 2/2

ID	Owner	Asset	Class	Purpose
-XIV A151	M14-16	Evaluator Independence Debt State Profiler	TOOL	يستخرج متغيرات النظرية والمجهول والاعتماد على المصادر قبل التفعيل.
-XIV A152	M14-16	Evaluator Independence Debt Counterfactual Probe	TOOL	يولد intervention يفصل causal mechanism عن أقرب baseline.
-XIV A153	M14-16	Evaluator Independence Debt Failure Topology Mapper	TOOL	يجمع failure signatures ويربط recurrence بالسياق والسبب المحتمل.
-XIV A154	M14-16	Evaluator Independence Debt Calibration Engine	TOOL	يعاير thresholds/uncertainty/cost على held-out data.
-XIV A155	M14-16	Evaluator Independence Debt Matched-Budget Method	METHOD	يطابق model/tools/context/retries/time ويمنع attribution الزائف.
-XIV A156	M14-16	Evaluator Independence Debt Cross-Environment Transfer	METHOD	يفحص حفظ causal role عبر domain/tool/model/horizon.
-XIV A157	M14-16	Evaluator Independence Debt Evidence Promotion Protocol	PROTOCOL	ينظم N0-N5/E0-E5 ويمنع claim expansion قبل replication.
-XIV A158	M14-16	Evaluator Independence Debt Negative-Result Protocol	PROTOCOL	يحفظ الفشل والحدود ويمنع إعادة تسمية الفرضية بعد عدم نجاحها.
-XIV A159	M14-16	Evaluator Independence Debt Runtime Service	STACK	خدمة feature-flagged مستقلة لها state/telemetry/rollback.
-XIV A160	M14-16	Evaluator Independence Debt Proof/Replay Adapter	STACK	يحوّل trajectory إلى manifest + evidence packet + replay receipt.
-XIV A161	M14-17	Scientific Reproduction Yield State Profiler	TOOL	يستخرج متغيرات النظرية والمجهول والاعتماد على المصادر قبل التفعيل.
-XIV A162	M14-17	Scientific Reproduction Yield Counterfactual Probe	TOOL	يولد intervention يفصل causal mechanism عن أقرب baseline.
-XIV A163	M14-17	Scientific Reproduction Yield Failure Topology Mapper	TOOL	يجمع failure signatures ويربط recurrence بالسياق والسبب المحتمل.
-XIV A164	M14-17	Scientific Reproduction Yield Calibration Engine	TOOL	يعاير thresholds/uncertainty/cost على held-out data.
-XIV A165	M14-17	Scientific Reproduction Yield Matched-Budget Method	METHOD	يطابق model/tools/context/retries/time ويمنع attribution الزائف.
-XIV A166	M14-17	Scientific Reproduction Yield Cross-Environment Transfer	METHOD	يفحص حفظ causal role عبر domain/tool/model/horizon.
-XIV A167	M14-17	Scientific Reproduction Yield Evidence Promotion Protocol	PROTOCOL	ينظم N0-N5/E0-E5 ويمنع claim expansion قبل replication.
-XIV A168	M14-17	Scientific Reproduction Yield Negative-Result Protocol	PROTOCOL	يحفظ الفشل والحدود ويمنع إعادة تسمية الفرضية بعد عدم نجاحها.
-XIV A169	M14-17	Scientific Reproduction Yield Runtime Service	STACK	خدمة feature-flagged مستقلة لها state/telemetry/rollback.
-XIV A170	M14-17	Scientific Reproduction Yield Proof/Replay Adapter	STACK	يحوّل trajectory إلى manifest + evidence packet + replay receipt.
-XIV A171	M14-18	Constitutional Mutation Stability State Profiler	TOOL	يستخرج متغيرات النظرية والمجهول والاعتماد على المصادر قبل التفعيل.
-XIV A172	M14-18	Constitutional Mutation Stability Counterfactual Probe	TOOL	يولد intervention يفصل causal mechanism عن أقرب baseline.
-XIV A173	M14-18	Constitutional Mutation Stability Failure Topology Mapper	TOOL	يجمع failure signatures ويربط recurrence بالسياق والسبب المحتمل.
-XIV A174	M14-18	Constitutional Mutation Stability Calibration Engine	TOOL	يعاير thresholds/uncertainty/cost على held-out data.
	M14-18	Constitutional Mutation Stability Matched-Budget Method	METHOD	

-XIV A175				يطابق model/tools/context/retries/time ويمنع attribution الزائف.
-XIV A176	M14-18	Constitutional Mutation Stability Cross-Environment Transfer	METHOD	يفحص حفظ causal role عبر domain/tool/model/horizon.
-XIV A177	M14-18	Constitutional Mutation Stability Evidence Promotion Protocol	PROTOCOL	ينظم N0-N5/E0-E5 ويمنع claim expansion قبل replication.
-XIV A178	M14-18	Constitutional Mutation Stability Negative-Result Protocol	PROTOCOL	يحفظ الفشل والحدود ويمنع إعادة تسمية الفرضية بعد عدم نجاحها.
-XIV A179	M14-18	Constitutional Mutation Stability Runtime Service	STACK	خدمة feature-flagged مستقلة لها state/telemetry/rollback.
-XIV A180	M14-18	Constitutional Mutation Stability Proof/Replay Adapter	STACK	يحوّل trajectory إلى manifest + evidence packet + replay receipt.

الحالة: SPECIFICATION. لا يقال deployed أو validated إلا بعد وجود artifact/run/test evidence.

برنامج التدريب: من التعريف إلى قدرة قابلة للنقل

لا يُقصد بالتدريب هنا بالضرورة fine-tuning للنموذج؛ قد يكون curriculum للهارنس، verifier training، policy learning أو dataset construction. المبدأ أن training target يجب أن يطابق mechanism claim لا اسم النظرية.

- M14-13 / Mechanism Isolation Curriculum: حالات بسيطة تعزل المتغير المركزي قبل open-world confounders.
- M14-13 / Adversarial Boundary Curriculum: أمثلة داخل/خارج الحد وnegative activation لمنع overgeneralization.
- M14-13 / Matched-Budget Causal Evaluation: تدريب/تقييم يقارن تحت model/tools/context/retries/time متقاربة مع sham scaffold.
- M14-13 / Cross-Model/Domain Replication: إعادة التدريب/الاختبار في model/domain/tool family مختلفة مع stable-sign criterion.
- M14-14 / Mechanism Isolation Curriculum: حالات بسيطة تعزل المتغير المركزي قبل open-world confounders.
- M14-14 / Adversarial Boundary Curriculum: أمثلة داخل/خارج الحد وnegative activation لمنع overgeneralization.
- M14-14 / Matched-Budget Causal Evaluation: تدريب/تقييم يقارن تحت model/tools/context/retries/time متقاربة مع sham scaffold.
- M14-14 / Cross-Model/Domain Replication: إعادة التدريب/الاختبار في model/domain/tool family مختلفة مع stable-sign criterion.
- M14-15 / Mechanism Isolation Curriculum: حالات بسيطة تعزل المتغير المركزي قبل open-world confounders.
- M14-15 / Adversarial Boundary Curriculum: أمثلة داخل/خارج الحد وnegative activation لمنع overgeneralization.
- M14-15 / Matched-Budget Causal Evaluation: تدريب/تقييم يقارن تحت model/tools/context/retries/time متقاربة مع sham scaffold.
- M14-15 / Cross-Model/Domain Replication: إعادة التدريب/الاختبار في model/domain/tool family مختلفة مع stable-sign criterion.

> عند استخدام generated data، تُفصل data producer عن evaluator قدر الإمكان ويُحتفظ بالraw source/trajectory لتدقيق self-confirming loops.

برنامج التدريب - الاستكمال والانتقال

- M14-16 / Mechanism Isolation Curriculum: حالات بسيطة تعزل المتغير المركزي قبل open-world confounders.
- M14-16 / Adversarial Boundary Curriculum: أمثلة داخل/خارج الحد وnegative activation لمنع overgeneralization.
- M14-16 / Matched-Budget Causal Evaluation: تدريب/تقييم يقارن تحت model/tools/context/retries/time متقاربة مع sham scaffold.
- M14-16 / Cross-Model/Domain Replication: إعادة التدريب/الاختبار في model/domain/tool family مختلفة مع stable-sign criterion.
- M14-17 / Mechanism Isolation Curriculum: حالات بسيطة تعزل المتغير المركزي قبل open-world confounders.
- M14-17 / Adversarial Boundary Curriculum: أمثلة داخل/خارج الحد وnegative activation لمنع overgeneralization.
- M14-17 / Matched-Budget Causal Evaluation: تدريب/تقييم يقارن تحت model/tools/context/retries/time متقاربة مع sham scaffold.
- M14-17 / Cross-Model/Domain Replication: إعادة التدريب/الاختبار في model/domain/tool family مختلفة مع stable-sign criterion.
- M14-18 / Mechanism Isolation Curriculum: حالات بسيطة تعزل المتغير المركزي قبل open-world confounders.
- M14-18 / Adversarial Boundary Curriculum: أمثلة داخل/خارج الحد وnegative activation لمنع overgeneralization.
- M14-18 / Matched-Budget Causal Evaluation: تدريب/تقييم يقارن تحت model/tools/context/retries/time متقاربة مع sham scaffold.
- M14-18 / Cross-Model/Domain Replication: إعادة التدريب/الاختبار في model/domain/tool family مختلفة مع stable-sign criterion.

الانتقال لا يعني استنساخ الأوزان أو prompts فقط. نختبر transfer of causal role: هل يبقى سبب التحسن نفسه عند تغيير model/domain/tool؟ إذا تغير، يُسجل كـinteraction أو domain-specific variant.

مصفوفة التجارب - 1/2

ID	Owner	Kind	Design	Metrics
-XIV E37	M14-13	Minimal Causal Ablation	switchable implementation مقابل nearest parent/baseline تحت matched budget.	+ primary theory metric + task outcome + calibration + cost critical regressions
-XIV E38	M14-13	Boundary/Adversarial Test	اختبار failure modes والحدود بحالات مصممة لكسر التعميم.	false activation + missed failure + rollback + evidence integrity
-XIV E39	M14-13	Transfer/Replication	نقل الآلية إلى model/domain/tool/horizon مختلف مع preregistered metric.	effect sign stability + CI + failure-signature transfer + resource frontier
-XIV E40	M14-14	Minimal Causal Ablation	switchable implementation مقابل nearest parent/baseline تحت matched budget.	+ primary theory metric + task outcome + calibration + cost critical regressions
-XIV E41	M14-14	Boundary/Adversarial Test	اختبار failure modes والحدود بحالات مصممة لكسر التعميم.	false activation + missed failure + rollback + evidence integrity
-XIV E42	M14-14	Transfer/Replication	نقل الآلية إلى model/domain/tool/horizon مختلف مع preregistered metric.	effect sign stability + CI + failure-signature transfer + resource frontier
-XIV E43	M14-15	Minimal Causal Ablation	switchable implementation مقابل nearest parent/baseline تحت matched budget.	+ primary theory metric + task outcome + calibration + cost critical regressions
-XIV E44	M14-15	Boundary/Adversarial Test	اختبار failure modes والحدود بحالات مصممة لكسر التعميم.	false activation + missed failure + rollback + evidence integrity
-XIV E45	M14-15	Transfer/Replication	نقل الآلية إلى model/domain/tool/horizon مختلف مع preregistered metric.	effect sign stability + CI + failure-signature transfer + resource frontier
-XIV E46	M14-16	Minimal Causal Ablation	switchable implementation مقابل nearest parent/baseline تحت matched budget.	+ primary theory metric + task outcome + calibration + cost critical regressions
-XIV E47	M14-16	Boundary/Adversarial Test	اختبار failure modes والحدود بحالات مصممة لكسر التعميم.	false activation + missed failure + rollback + evidence integrity
-XIV E48	M14-16	Transfer/Replication	نقل الآلية إلى model/domain/tool/horizon مختلف مع preregistered metric.	effect sign stability + CI + failure-signature transfer + resource frontier

مصفوفة التجارب - 2/2

ID	Owner	Kind	Design	Metrics
-XIV E49	M14-17	Minimal Causal Ablation	switchable implementation مقابل nearest parent/baseline تحت matched budget.	+ primary theory metric + task outcome + calibration + cost critical regressions
-XIV E50	M14-17	Boundary/Adversarial Test	اختبار failure modes والحدود بحالات مصممة لكسر التعميم.	false activation + missed failure + rollback + evidence integrity
-XIV E51	M14-17	Transfer/Replication	نقل الآلية إلى model/domain/tool/horizon مختلف مع preregistered metric.	effect sign stability + CI + failure-signature transfer + resource frontier
-XIV E52	M14-18	Minimal Causal Ablation	switchable implementation مقابل nearest parent/baseline تحت matched budget.	+ primary theory metric + task outcome + calibration + cost critical regressions
-XIV E53	M14-18	Boundary/Adversarial Test	اختبار failure modes والحدود بحالات مصممة لكسر التعميم.	false activation + missed failure + rollback + evidence integrity
-XIV E54	M14-18	Transfer/Replication	نقل الآلية إلى model/domain/tool/horizon مختلف مع preregistered metric.	effect sign stability + CI + failure-signature transfer + resource frontier

ترجمة الورقة إلى MOUSSAIXT Reference Kernel

```
service registry: M14-13,M14-14,M14-15,M14-16,M14-17,M14-18
feature_flags: independent
state: versioned typed objects
authority: L0/L1 gates
telemetry: trigger,state_before,decision,tool,observation,state_after,residual
verifier: independent where possible
rollback: mandatory for reversible mutations
writeback: evidence-bearing delta only
```

التنفيذ المقترح يبدأ بـservices أصغر من «AGI kernel» الكبير. كل service يملك test vectors وحالات invalid input وtimeout وstale state وwrong target. بهذه الصياغة يصبح ablation ممكناً، ويصبح فشل component معلومة بدلاً من انهيار غامض للنظام كله.

Acceptance at kernel level

- schema validation ≥ predefined threshold
- no silent privilege widening
- deterministic replay where environment permits
- world-state postcondition for external actions
- trace completeness for critical steps
- explicit disable path and fallback

القيود والمشكلات المفتوحة

- الاستقلال السببي بين مكونات architecture قد يكون صعباً بسبب interactions الكثيفة.
- بعض observables المقترحة تحتاج proxies قد تغيّر معنى النظرية.
- matched budgets ليست مساواة كاملة لأن latency/tool quality مختلفة نوعياً.
- الانتقال بين benchmarks/deployment قد يفشل بسبب hidden distribution shifts.
- المقيم المستقل قد يكون أقل كفاءة أو أكثر ضجيجاً، ما يستلزم meta-evaluation.
- memory/world models تولدان state غير قابل للمقارنة بسهولة بين implementations.
- أي claim عن consciousness يبقى منفصلاً عن الفائدة الهندسية الوظيفية.
- المرجع الخارجي الحديث قد يكون preprint؛ نميّز peer-reviewed ACL papers عن arXiv/spec/lab reports.

أقوى نتيجة مرغوبة من هذه الورقة ليست «نجاح كل النظريات»، بل ظهور خريطة واضحة: ماذا نحتفظ به، ماذا ندمج، أين تفشل الآلية، وما evidence الجديد الذي يغير النسخة التالية.

المراجع الخارجية الأساسية

E12 - CUBE: A Standard for Unifying Agent Benchmarks. arXiv:2603.15798. Common protocol separates task, benchmark, package and registry concerns.

E07 - AgencyBench: Benchmarking the Frontiers of Autonomous Agents in 1M-Token Real-World Contexts. ACL 2026, 10.18653/v1/2026.acl-long.337. 138 tasks/32 scenarios; long trajectories, many tool calls and resource-efficiency/self-correction gaps.

E16 - Recursive Self-Improvement in AI: From Bounded Self-Refinement to Autonomous Research Loops. arXiv:2607.07663. Survey of 1,250 papers; evaluator/grounding hierarchy and limits of open-ended RSI.

E17 - GPT-Red: Unlocking Self-Improvement for Robustness. OpenAI, 2026-07-15. Automated red-teaming and adversarial training; attacker kept separate from deployed models.

E18 - Securing the future of AI agents / AI Control Roadmap. Google DeepMind, 2026-06-18. Control framing around detection, prevention, response and supervisors for capable agents.

E19 - Trustworthy agents in practice / containment across products. Anthropic, 2026. Human control, secure interaction, transparency/privacy and containment/blast-radius engineering.

E09 - Aligning Agents via Planning: Plan-RewardBench. ACL 2026, 10.18653/v1/2026.acl-long.1062. Trajectory-level reward modeling; evaluator performance degrades with long trajectories.

E06 - Long-Horizon Agent Trajectory Attribution. arXiv:2608.06909. -Fine-grained attribution benchmark with >1,300 annotated trajectories and leave-one out baselines.

راجع أيضاً قائمة Part XIV Master للمراجع المشتركة. العناوين/المعرفات مُستخدمة كرساة بحثية؛ لا تُنقل claims تفصيلية خارج ما يدعمه المصدر.

المراسي الداخلية والسلالة

I01 - MOUSSAIF Unified AGI - canonical research book. MOUSSAIF internal corpus .Difference → Question → Route → Act → Verify → Reconstruct; E0 .E5 evidence discipline

I02 - MOUSSAIF AGI OS / MOUSSAIXT Deep Engineering Architecture. MOUSSAIF internal corpus .layer architecture; claim contracts; world-effect-12 .loop; memory/evolution; tool and authority contracts

I03 - ParallaxWorld-MCD FULL AGI OS. MOUSSAIF internal corpus .Difference-before-volume, governing question, MCD/QMCD, world modeling and .falsification

I04 - MUT-AGI-02 Unified Theory. MOUSSAIF internal corpus .Canonical theory graph, anti-inflation, theory contracts and crosswalk.

I05 - MUT × MUG × MOUSSAIXT integrated framework. MOUSSAIF internal corpus .→ Theory unification → program governance → governed runtime .evidence

I06 - Part X - Full Spectrum OS Closure. MOUSSAIF internal corpus .Belief/action partition, reproducibility, externalities, accountability and recourse.

I07 - Part XI - Memory, Continuity, Self-Reconstruction & Evolution. MOUSSAIF internal corpus .,Reconstructive memory, unresolved state, writeback .portability and coupled evolution

I08 - Part XII - Total Research-Paper Integration Closure. MOUSSAIF internal corpus .Source integration registry; presence/integration/validation are .distinct states

I09 - Part XIII - Theoretical Renaissance. MOUSSAIF internal corpus .Causal open-world verified intelligence; theory-to-kernel causal closure; external .grounding

I10 - Conscious AGI / QRET / Governance Space corpus. MOUSSAIF internal corpus .Functional self-model/meta-governance with explicit .phenomenology boundary

I11 - Question Density / QMCD / QGWEI / MURI corpus. MOUSSAIF internal corpus .Governing questions, plural reasoning, explicit ignorance and .evidence geometry

I12 - Goal-as-OS / World Effect / M-HKOS / MATI-MACIT. MOUSSAIF internal corpus .Goal contracts, verified world-state change, knowledge-as-OS and .selective internalization

I13 - MUT MUG ALL artifact memory map - 2026-09-04. MOUSSAIF internal corpus .;artifacts; duplicate groups are not independent evidence 343 .retrieval is family/canonicality based

> سياسة الدمج: canonical synthesis يحفظ source role والنسخة والحالة. الملفات المكررة أو historical mirrors لا تُعد أدلة جديدة؛ الفروق المادية تبقى lineage variants حتى ablation/semantic comparison.