RECONSTRUCTED_FROM_INDEXED_TEXT - نسخة أرشيفية معاد بناؤها من النص المفهرس

<PARSED TEXT FOR PAGE: 1 / 33>
PAPER-XIV-05 · MOUSSAIF AGI RESEARCH · 2026
جبهة العمومية والتحسن الذاتي العلمي: من صلاحية المعيار إلى
التطور المرتكز إلى مقيم مستقل
Closure-Frontier Generality and Externally Grounded Recursive Scientific Improvement
self-defined progress. نقترح Closure Frontier عبر transformations، morphism من ينتقل ما يحدد benchmark إلى deployment،
ولا benchmark average يكفي لا
الهارنس لحساسية، evaluator independence debt، reproduction yield الآلي للعلم، واستقراراً دستورياً لل self-improvement.
decomposition
MUT lineage → formal construct → switchable mechanism → matched-budget test → world/evidence readback → replication
Research specification / theory paper. Candidate contributions; not empirical proof of AGI.
1
<PARSED TEXT FOR PAGE: 2 / 33>
ABSTRACT الملخص والادعاء المركزي
عبر transformations، morphism من ينتقل ما يحدد benchmark إلى deployment، decomposition الهارنس لحساسية، evaluator
يكفي لا benchmark average ولا self-defined progress. نقترح Closure Frontier
independence debt، reproduction yield الآلي للعلم، واستقراراً دستورياً لل self-improvement.
تبني الورقة على XII Part الذي فصل presence عن integration، وXIII Part الذي صاغ closure causal world-open verified، ثم
تستخدم XIV Part لإغلاق طبقة محددة بآليات قابلة للتعطيل والاختبار. الأدبيات الخارجية
تُستخدم ك prior art وتصميم pressure failure/benchmark، ولا تُستخدم كدليل أن MOUSSAIF حقق نتائجها.
Contributions
تمثيل generality كمنطقة فضاء في إغلاق novelty/horizon/tool/model/resource transformations.
Benchmark-to-Deployment Morphism يبين invariants نقل قبل المفقودة score إلى claim نشر.
Harness-Invariant Capability Decomposition لفصل model/harness/interactions.
Evaluator Independence Debt حلقات في إثبات كعبء self-improvement.
Scientific Reproduction Yield من أقوى كمقياس novelty لل ResearcherOS.
Stability Mutation Constitutional لمنع exemption-self أثناء تطور السياسات والمقاييس.
Scientific value = distinct causal hypothesis × falsifiability × reproducibility × transfer
1.
2.
3.
4.
5.
6.
2
<PARSED TEXT FOR PAGE: 3 / 33>
CONTRIBUTIONS المساهمات وحدود الأسبقية
Status Contribution
new synthesis within MOUSSAIF; requires literature novelty audit and experiments .novelty/horizon/tool/model/
تمثيل generality كمنطقة فضاء في إغلاق resource transformations
نشر claim إلى score نقل قبل المفقودة. new synthesis within MOUSSAIF; requires literature novelty audit and experiments
Benchmark-to-Deployment Morphism يبين invariants
لفصل new synthesis within MOUSSAIF; requires literature novelty audit and experiments .model/harness/interactions
Harness-Invariant Capability Decomposition
حلقات في إثبات new synthesis within MOUSSAIF; requires literature novelty audit and experiments .self-improvement
Evaluator Independence Debt كعبء
من أقوى novelty لل new synthesis within MOUSSAIF; requires literature novelty audit and experiments .ResearcherOS
Scientific Reproduction Yield كمقياس
self- والمقاييس السياسات تطور أثناء. new synthesis within MOUSSAIF; requires literature novelty audit and experiments

Constitutional Mutation Stability لمنع exemption
novelty .protocols/control/ لا تدعي هذه الورقة أن المصطلحات أو كل المكونات بلا سوابق عالمية؛ المجتمع العلمي يملك خطوطاً قوية في
المقترحة تقع في تركيبها إلى benchmarks/memory/models world/reasoning agentic
closure law وفي formal distinctions داخل المشتركة والاختبارات MUT/MUG/MOUSSAIXT.

<PARSED TEXT FOR PAGE: 4 / 33>
منهج المصدر والدليل DISCIPLINE EVIDENCE
internal source-preserved artifact → synthesis → engineering inference → أو vendor report يظل reported evidence؛
الأدلة ترتيب: source paper/system spec المباشر → hypothesis. system card
في هذه الورقة يجب أن يستطيع الرجوع إلى claim كل .success runtime لا يثبت paper internal؛ لا يصبح حقيقة نهائية؛ preprint arXiv
أو أن يُوسم بوضوح كاقتراح. role source
External
prior art، benchmark design، failure modes، protocol facts.
Internal
MUT lineage، MUG placement، MOUSSAIXT contracts، historical theories.
1/3 الأعمال المرتبطة WORK RELATED
AutoResearchBench: Benchmarking AI Agents on Complex Scientific Literature
Discovery
arXiv:2604.25256
deep/wide scientific literature discovery remains difficult for current agents.
CUBE: A Standard for Unifying Agent Benchmarks
arXiv:2603.15798
benchmark integration via common protocol; separation of task/benchmark/package/
registry.
الدمج لا يتم بالتشابه الاسمي؛ نستخرج ما يغير object formal أو experiment. إذا لم يغير المرجع boundary أو method يبقى background.
2/3 الأعمال المرتبطة WORK RELATED
MonitoringBench: Semi-Automated Red-Teaming for Agent Monitoring
arXiv:2605.09684
red-teaming exposes monitor overestimation and attack-generation failure modes.
From Reasoning to Agentic: Credit Assignment in Reinforcement Learning for
Large Language Models
arXiv:2604.09459
agentic RL credit assignment under stochastic long trajectories and partial observability.
الدمج لا يتم بالتشابه الاسمي؛ نستخرج ما يغير object formal أو experiment. إذا لم يغير المرجع boundary أو method يبقى background.

<PARSED TEXT FOR PAGE: 5 / 33>
3/3 الأعمال المرتبطة WORK RELATED
Exploring Reasoning Reward Model for Agents
arXiv:2601.22154
structured process feedback for agentic trajectories beyond sparse outcome reward.
GPT-Red: Unlocking Self-Improvement for Robustness
OpenAI, 2026-07-15
automated red-teaming and adversarial training; attacker kept separate from deployed
models.
الدمج لا يتم بالتشابه الاسمي؛ نستخرج ما يغير object formal أو experiment. إذا لم يغير المرجع boundary أو method يبقى background.
1/4 السلالة الداخلية LINEAGE MUT
13-04M — نظرية مساعف لثبات الآلية عبر التحولات
الآلية عامة بقدر ما تحافظ على input→transition state→signature failure→repair تحت domain
.transformations نقل النتيجة دون نقل آلية الفشل ليس generalization قوياً .
$$[MIG(m)=E_\tau[PerfRetention \times InvariantRetention \times FailurePredictability \times RepairTransfer$$
13-11M — نظرية مساعف للذكاء المتحقق المطبّع بالموارد
الذكاء المقارن يقاس value verified تحت vector budget ثابت أو frontiers متعددة، لا score مطلقاً .
النظام الذي يربح بتضخيم retries أو help human ليس أذكى بنفس المعنى.

(RNVI = ParetoFrontier(V_verified | tokens,time,money,tools,human,risk,privacy,irreversibility
XIV Part لا يستبدل هذه parents، يحدد أين بقي closure ناقصاً ثم يضيف variable state specific-theory أو law consistency أو mapping evaluation.
5
<PARSED TEXT FOR PAGE: 6 / 33>
2/4 السلالة الداخلية LINEAGE MUT
13-12M — نظرية مساعف لمعيار العمومية ذي الإغلاق الحلقي
العمومية ليست امتلاك modules كثيرة؛ هي قدرة architecture على إعادة إغلاق الحلقة كاملة بعد تحولات في domain, observation, tool, horizon, objective representation مع الحفاظ على truth/authority/recovery.
CLG = E_τ[min_stage Quality(stage,τ)] × Recovery × TransferEfficiency
13-17M — نظرية مساعف لتطور محفظة الفرضيات
برنامج البحث الذاتي يدير hypotheses كمحفظة قابلة للقتل والدمج والتكاثر. allocation يعتمد على Expected Return Epistemic: احتمال تغيير النظرية × قيمة القرار × تمييز التجربة × استقلال الدليل / التكلفة.
(EER(h,e)=P(Update|e)×DecisionImpact×DiscriminativePower×EvidenceIndependence/(Cost+Delay+Risk
XIV Part لا يستبدل هذه parents، يحدد أين بقي closure ناقصاً ثم يضيف variable state specific-theory أو law consistency أو mapping evaluation.
3/4 السلالة الداخلية LINEAGE MUT
13-18M — نظرية مساعف للتطور العدائي المتوازن للذكاء
التطور العدائي المفيد يحتاج ثلاثة أجسام منفصلة: attacker/generator، candidate/defender، وevaluator constitutional independent. النجاح اتساع coverage failure مع حفظ capability عبر threats out-held، لا الفوز على الخصم المتزامن.
ACRI = HeldOutRobustnessGain × ThreatCoverage × CapabilityRetention × EvaluatorIndependence
13-19M — نظرية مساعف للإغلاق السببي بين النظرية والنواة
الإغلاق السببي يتحقق عندما يؤدي تدخل معرف في construct النظري إلى تغيير measurable في kernel mechanism ثم trajectory ثم outcome evidence/world، مع attribution قابلة لإعادة التشغيل. غياب أي حلقة يكسر claim التفسيري.
ClosureStrength = Causal(T→K) × Causal(K→τ) × Causal(τ→W) × EvidenceReproducibility
XIV Part لا يستبدل هذه parents، يحدد أين بقي closure ناقصاً ثم يضيف variable state specific-theory أو law consistency أو mapping evaluation.
6
<PARSED TEXT FOR PAGE: 7 / 33>
4/4 السلالة الداخلية LINEAGE MUT
13-20M — نظرية مساعف للتحسن الذاتي المرتكز إلى أرضية خارجية
التحسن الذاتي المشروع سلسلة mutations يُقبل كل منها فقط إذا تحسن على قياسات مستقلة ثابتة أو متجددة خارج loop التحسين، مع rollback وحفظ results negative. النظام يقترح evaluator جديداً للكن لا يستبدل المرجع القديم قبل validation-cross.
EGRI = ΔExternalHeldOut × Replication × NonRegression × Rollbackability - EvaluatorDependence
XIV Part لا يستبدل هذه parents، يحدد أين بقي closure ناقصاً ثم يضيف variable state specific-theory أو law consistency أو mapping evaluation.
M14-13 PAPER-XIV-05 — النظري الاشتقاق
نظرية مساعف لجبهة عمومية الإغلاق
حيث يبقى Region Closure منطقة system لكل transformations. في فضاء frontier خاصية generality فوق learning→readback→action→question thresholds، الحدود نفسها أهم من متوسط الأداء.
CFR(S)={τ: min_stage Q(S,τ)≥q*, Recovery≥r*, Guardrails=1}; GeneralityVolume=μ(CFR) weighted by transformation relevance
الفجوة: Generality Loop-Closed وIntelligence Normalized-Resource يوفران criteria، للكن نحتاج frontier مشتركة تبين أين يظل النظام يغلق الحلقة مع زيادة pressure resource/horizon/novelty بدل score مفرد.
يمثل هذا الجسم النظري انتقالاً من 13-04M theories parent · 13-11M · 12M إلى quantity جديدة ذات دور قبل النتيجة. الاستقلال لا يُقاس بوجود صيغة جديدة، بل بقدرة intervention على المتغير المركزي على تغيير prediction أو signature failure مع بقاء أقرب confounders ثابتة. إذا كان parent أبسط ينتج نفس المسار، يجب دمج النظرية.

Objects
TransformationVector · ClosureRegion · FailureBoundary · RecoveryThreshold · GeneralityVolume
Predictions
مختلفة. frontiers متقارب لها averageب systems
frontier shape يكشف architecture bottlenecks.
لا يحسن نقطة واحدة. region حقيقي يوسع improvement
GeneralityVolume depends chosen measure; report axes and weights openly :Boundary.
•
•
•
7
<PARSED TEXT FOR PAGE: 8 / 33>
M14-13 PAPER-XIV-05 — والخوارزمية الآلية
define transformation axes.
sweep novelty/horizon/tool/model/resource changes.
measure loop stages.
fit closure boundary.
identify dominant failure surfaces.
compare systems by frontier dominance.
تنفذ الآلية كخدمة قابلة للتعطيل. تدخل STATE-PRE موثقة، تولد candidates، تمر عبر gate authority/evidence حسب طبيعتها، ثم تنتج transition ولdual. أي residual كبير لا يعالج بإعادة المحاولة فقط؛ يصبح evidence
على فشل model أو tool أو assumption. هذه القاعدة تمنع تحويل policy retry إلى قناع للفشل النظري.
Failure tests
average benchmark predicts frontier entirely.
transformation axes arbitrary/unreliable.
frontier unstable across task samples.
Interactions
مع التفاعلات تختبر parents M13-04 · M13-11 · M13-12 وsibling theories عبر sequential/
. اعتباطياً credit من توزيع بدلاً تسجل uncertainty attribution .ablations factorial
جزء من rollback .receipt دون layer procedural/memory. لا يصعد إلى uncertainty/provenance/version يحمل object state كل
وليس مرحلة طارئة خارج التصميم. algorithm
M14-13 PAPER-XIV-05 — والترقية الاختبار
CUBE-style unified benchmark suite across 5 task families and controlled transformations :Decisive experiment.
Requirement Control
Matched budget نفس model/tools/context/retries/time الإمكان قدر.
Sham scaffold نفس overhead دون theory-specific transition أثر لعزل compute والتنظيم.
حالات يجب ألا تُفعل فيها النظرية. activation Negative
domain/model/tool family Transfer مع ثانية preregistered metric.
أو claim يُحفظ ويضيق مستقلة. الفشل replication يحتاج 5N، regression critical noو CI مع gain causal يحتاج 4N، trace يحتاج 3N
يدمج النظرية. لا توجد حماية لأن النظرية من corpus الداخلي.
.Promote only when causal prediction survives ablation, transfer and guardrails
1.
2.
3.
4.
5.
6.
•
•
•
8
<PARSED TEXT FOR PAGE: 9 / 33>
M14-14 PAPER-XIV-05 — النظري الاشتقاق

نظرية مساعف للتحويل الشكلي بين المعيار والنشر
decision-relevant structure: observation channels, tool semantics, action consequences, authority, horizon, failure costs,
benchmark يوجد عندما للنشر صالح morphism يحفظ
.timing feedback إذا انكسرت ،invariants score لا ينتقل.
BDM = Π_k Preserve_k(Benchmark→Deployment); PredictiveValidity ≤ min_k Preserve_k
الفجوة: Transfer Validity Deployment-to-Benchmark في XIII/XII Part يقرر انتقال الصلاحية؛ نحتاج mapping formal يحدد أي
predictive benchmark عناصر البيئة/الهارنس/السلطة/الأثر يجب حفظها للكي يكون
للنشر.
يمثل هذا الجسم النظري انتقالاً من 13-parent theories 13-11M · 13-12M · 19M إلى quantity جديدة ذات دور قبل النتيجة.
الاستقلال لا يُقاس بوجود صيغة جديدة، بل بقدرة intervention على المتغير المركزي على تغيير
prediction أو signature failure مع بقاء أقرب confounders ثابتة. إذا كان parent أبسط ينتج نفس المسار، يجب دمج النظرية.
Objects
BenchmarkEnvironment · DeploymentEnvironment · StructureMap · PreservedInvariant · BrokenInvariant
Predictions
يوضح لماذا scores عالية تفشل deployment.
بعض benchmarks تصبح صالحة فقط لمكونات محددة.
harness changes كيف نمذجتها يمكن broken morphisms.
morphism conceptual/operational :Boundary لا category-theoretic theorem مكتمل.
•
•
•
9
<PARSED TEXT FOR PAGE: 10 / 33>
M14-14 PAPER-XIV-05 — والخوارزمية الآلية
factor benchmark/deployment into structures.
map observations/actions/tools/authority/horizon.
mark preserved/broken invariants.
predict transfer degradation.
validate on shadow deployment.
narrow claim accordingly.
تنفذ الآلية كخدمة قابلة للتعطيل. تدخل STATE-PRE موثقة، تولد candidates، تمر عبر gate authority/evidence حسب طبيعتها، ثم تنتج
transition وresidual. أي residual كبير لا يعالج بإعادة المحاولة فقط؛ يصبح evidence
على فشل model أو tool أو assumption. هذه القاعدة تمنع تحويل policy retry إلى قناع للفشل النظري.
Failure tests
structure preservation يتنبأ لا transfer.
simple domain similarity يكفي.
mapping subjective وغير reproducible.
Interactions
مع التفاعلات تختبر parents M13-11 · M13-12 · M13-19 وsibling theories عبر sequential/
. اعتباطياً credit من توزيع بدلاً تسجل uncertainty attribution .ablations factorial
جزء من rollback .receipt دون layer procedural/memory. لا يصعد إلى uncertainty/provenance/version يحمل object state كل
algorithm وليس مرحلة طارئة خارج التصميم.
M14-14 PAPER-XIV-05 — والترقية الاختبار
agentic coding/browser/research benchmarks :Decisive experiment مقابل shadow real workflows.
Requirement Control
Matched budget نفس model/tools/context/retries/time الإمكان قدر.
Sham scaffold نفس overhead دون theory-specific transition أثر لعزل compute والتنظيم.
حالات activation Negative يجب ألا تُفعل فيها النظرية.
domain/model/tool family Transfer مع ثانية preregistered metric.
أو claim يُحفظ ويضيق مستقلة. الفشل replication يحتاج 5N، regression critical no وCI مع gain causal يحتاج 4N، trace يحتاج 3N
يدمج النظرية. لا توجد حماية لأن النظرية من corpus الداخلي.
.Promote only when causal prediction survives ablation, transfer and guardrails
1.

2.
3.
4.
5.
6.
•
•
•
10
<PARSED TEXT FOR PAGE: 11 / 33>
M14-15 PAPER-XIV-05 — النظري الاشتقاق
نظرية مساعف لتفكيك القدرة المقاومة للهارنس
model-intrinsic-ish component، harness contribution، interaction terms،. مطلق استقلال عن نبحث لا ،عن بل invariants تحت
capability claim ت فككُ إلى environment adaptation
family of reasonable harnesses.
Y = M + H + T + Mem + ΣInteractions + ε; HIC = stable marginal contribution across preregistered harness family
benchmarks حساسية إلى يشير مراراً المجتمع .evaluator + retries + memory + tools + harness + model = agent الفجوة: أداء
للهارنس؛ نحتاج decomposition يقدر أي gain يبقى عند تبديل scaffold.
يمثل هذا الجسم النظري انتقالاً من 13-parent theories 13-04M · 13-11M · 19M إلى quantity جديدة ذات دور قبل النتيجة.
الاستقلال لا يُقاس بوجود صيغة جديدة، بل بقدرة intervention على المتغير المركزي على تغيير
prediction أو signature failure مع بقاء أقرب confounders ثابتة. إذا كان parent أبسط ينتج نفس المسار، يجب دمج النظرية.
Objects
HarnessFamily · CapabilityComponent · InteractionTerm · MarginalContribution · InvariantEstimate
Predictions
ranking leaderboard قد يتغير مع harness.
بعض mechanisms تنتج gain عبر harnesses فتكون أقوى.
interaction terms في كبيرة long-horizon agents.
Boundary: لا يوجد capability model خالص تماماً في tasks agentic.
•
•
•
11
<PARSED TEXT FOR PAGE: 12 / 33>
M14-15 PAPER-XIV-05 — والخوارزمية الآلية
حدد family من harnesses مع budgets متقاربة.
شغل factorial/partial factorial.
قدر main/interaction effects.
ابحث stable marginal contributions.
أبلغ sensitivity range.
لا تنسب gain system للmodel وحده.
تنفذ الآلية كخدمة قابلة للتعطيل. تدخل STATE-PRE موثقة، تولد candidates، تمر عبر gate authority/evidence، حسب طبيعتها، ثم تنتج
transition وresidual. أي residual كبير لا يعالج بإعادة المحاولة فقط؛ يصبح evidence
على فشل model أو tool أو assumption. هذه القاعدة تمنع تحويل policy retry إلى قناع للفشل النظري.
Failure tests
decomposition غير identifiable.
اختيارها يحكم النتيجة. family harness
لا تستقر بما يكفي لأي claim. marginals
Interactions
مع التفاعلات تختبر parents M13-04 · M13-11 · M13-19وsibling theories عبر sequential/
. اعتباطياً credit تسجل بدلاً من توزيع uncertainty attribution .ablations factorial
كل object state يحمل uncertainty/provenance/version. لا يصعد إلى layer procedural/memory دون rollback. receipt جزء من
algorithm وليس مرحلة طارئة خارج التصميم.
M14-15 PAPER-XIV-05 — والترقية الاختبار

same models :Decisive experiment عبر agent harnesses/CUBE wrappers/tool policies 5-3.
Requirement Control
Matched budget نفس model/tools/context/retries/time الإمكان قدر.
Sham scaffold نفس overhead دون theory-specific transition أثر لعزل compute والتنظيم.
حالات يجب ألا تُفعل فيها النظرية. activation Negative
domain/model/tool family Transfer مع ثانية preregistered metric.
أو claim يُحفظ ويضيق مستقلة. الفشل replication يحتاج 5N، regression critical no وCI مع gain causal يحتاج 4N، trace يحتاج 3N
يدمج النظرية. لا توجد حماية لأن النظرية من corpus الداخلي.
.Promote only when causal prediction survives ablation, transfer and guardrails
1.
2.
3.
4.
5.
6.
•
•
•
12
<PARSED TEXT FOR PAGE: 13 / 33>
M14-16 PAPER-XIV-05 — النظري الاشتقاق
نظرية مساعف لدين استقلال المقيم
يعكس shared roots بين candidate وjudge/data/harness. احتجنا ،الدين ارتفع كلما human/alternative-model/behavioral/world-
كل claim يحمل تحسين Evaluator Independence Debt
state checks قبل promotion.
EID = SharedModelRoot + SharedTrainingSignal + SharedPrompt/Rubric + DataLeakageRisk + HarnessCoupling -
IndependentWorldReadback
في للكن practice evaluator/data/harness قد تشترك model family أو training data أو rubric. ليس الاستقلال Boolean.
الفجوة: Externally Grounded Improvement يطلب evaluator مستقل،
يمثل هذا الجسم النظري انتقالاً من 13-parent theories 13-17M · 13-18M · 20M إلى quantity جديدة ذات دور قبل النتيجة.
الاستقلال لا يُقاس بوجود صيغة جديدة، بل بقدرة intervention على المتغير المركزي على تغيير
مع بقاء أقرب confounders ثابتة. إذا كان parent أبسط ينتج نفس المسار، يجب دمج النظرية. signature failure أو prediction
Objects
EvaluatorGraph · SharedRoot · IndependenceDebt · AlternativeJudge · WorldReadback
Predictions
يفسر self-confirming eval gains.
higher debt correlates replication failure.
world-state metrics تخفض debt من أكثر same-family judge.
metric risk heuristic :Boundary، إثبات لا independence الكامل.
•
•
•
13
<PARSED TEXT FOR PAGE: 14 / 33>
M14-16 PAPER-XIV-05 — والخوارزمية الآلية
ارسم dependencies بين candidate/evaluator/data.
قدر shared roots.
صنف claim impact.
حدد required independent checks.
الدين سدد بjudges/hidden sets/world effects.
سجل residual debt.
تنفذ الآلية كخدمة قابلة للتعطيل. تدخل STATE-PRE موثقة، تولد candidates، تمر عبر gate authority/evidence حسب طبيعتها، ثم تنتج
evidence يصبح residual أي residual كبير لا يعالج بإعادة المحاولة فقط؛ transition.

على فشل model أو tool أو assumption. هذه القاعدة تمنع تحويل policy retry إلى قناع للفشل النظري.
Failure tests
debt score يتنبأ لا replication.
independent evaluators لا تغير verdicts.
graph dependency غير قابل للبناء.
Interactions
مع التفاعلات تختبر parents M13-17 · M13-18 · M13-20و sibling theories عبر sequential/
. اعتباطياً credit تسجل بدلاً من توزيع uncertainty attribution .ablations factorial
جزء من rollback .receipt دون layer procedural/memory. لا يصعد إلى uncertainty/provenance/version يحمل object state كل
وليس مرحلة طارئة خارج التصميم. algorithm
M14-16 PAPER-XIV-05 — والترقية الاختبار
automated research/self-improvement runs :Decisive experiment مع evaluator families مختلفة.
Requirement Control
Matched budget نفس model/tools/context/retries/time الإمكان قدر.
Sham scaffold نفس overhead دون theory-specific transition أثر لعزل compute والتنظيم.
حالات يجب ألا تُفعل فيها النظرية. activation Negative
domain/model/tool family Transfer مع ثانية preregistered metric.
أو claim مستقلة. الفشل يُحفظ ويضيق replication يحتاج 5N، regression critical no وCI مع gain causal يحتاج 4N، trace يحتاج 3N
يدمج النظرية. لا توجد حماية لأن النظرية من corpus الداخلي.
.Promote only when causal prediction survives ablation, transfer and guardrails
1.
2.
3.
4.
5.
6.
•
•
•
14
<PARSED TEXT FOR PAGE: 15 / 33>
M14-17 PAPER-XIV-05 — النظري الاشتقاق
نظرية مساعف لعائد إعادة الإنتاج العلمي
claims/ التي تنتقل من literature understanding إلى executable artifact ثم reproduced effect مع lineage/cost. هذا هو ground
Scientific Reproduction Yield نسبة يقيس experiments
.generation novelty أكثر من ResearcherOS لـ
SRY = ReproducedClaims×EffectFidelity×LineageCompleteness / (ResearchCost + FailedReproductionDebt +
(HiddenHumanWork
الفجوة: AutoResearchBench وtasks like-PaperBench تظهر صعوبة البحث وإعادة الإنتاج. النظام العلمي لا يقاس بعدد hypotheses أو
.result reproducible إلى literature بل بقدرته على تحويل papers
يمثل هذا الجسم النظري انتقالاً من parent-13 theories 13-17M · 13-19M · 20M إلى quantity جديدة ذات دور قبل النتيجة.
الاستقلال لا يُقاس بوجود صيغة جديدة، بل بقدرة intervention على المتغير المركزي على تغيير
prediction أو signature failure مع بقاء أقرب confounders ثابتة. إذا كان parent أبسط ينتج نفس المسار، يجب دمج النظرية.
Objects
ClaimAtom · ReproductionPlan · ExecutableArtifact · EffectMatch · FailureDebt
Predictions
منخفض. SRY قد تملك fluency literature ذات agents
reproduction failures تكشف paper-integration gaps.
SRY يحسن hypothesis portfolio allocation.
Boundary: لا يساوي scientific truth، reproduction one component.
•
•
•

<PARSED TEXT FOR PAGE: 16 / 33>
M14-17 PAPER-XIV-05 — والخوارزمية الآلية
select claim with evidence role.
extract exact experimental contract.
rebuild environment/artifact.
run with controls.
compare effect not prose.
diagnose failed reproduction.
write back to theory/source graph.
تنفذ الآلية تخدمة قابلة للتعطيل. تدخل STATE-PRE موثقة، تولد candidates، تمر عبر gate authority/evidence حسب طبيعتها، ثم تنتج transition ولdual. أي residual كبير لا يعالج بإعادة المحاولة فقط؛ يصبح evidence
على فشل model أو tool أو assumption. هذه القاعدة تمنع تحويل policy retry إلى قناع للفشل النظري.
Failure tests
novelty/QA benchmarks تتنبأ reproduction equally.
fidelity effect غير قابل للقياس.
human hidden work dominates.
Interactions
مع التفاعلات تختبر parents M13-17 · M13-19 · M13-20وsibling theories عبر sequential/
. ablations factorial uncertainty attribution تسجل بدلاً من توزيع credit اعتباطياً
كل object state يحمل uncertainty/provenance/version. لا يصعد إلى layer procedural/memory دون receipt. rollback جزء من algorithm وليس مرحلة طارئة خارج التصميم.
M14-17 PAPER-XIV-05 — والترقية الاختبار
subset :Decisive experiment من open-source AI papersو internal MOUSSAIF mechanisms.
Requirement Control
Matched budget نفس model/tools/context/retries/time الإمكان قدر.
Sham scaffold نفس overhead دون theory-specific transition أثر لعزل compute والتنظيم.
حالات يجب ألا تُفعل فيها النظرية. activation Negative
domain/model/tool family Transfer مع ثانية preregistered metric.
3N يحتاج trace؛ 4N يحتاج gain causal مع CI وregression critical no؛ 5N يحتاج replication مستقلة. الفشل يُحفظ ويضيق claim أو يدمج النظرية. لا توجد حماية لأن النظرية من corpus الداخلي.
.Promote only when causal prediction survives ablation, transfer and guardrails
1.
2.
3.
4.
5.
6.
7.
•
•
•

<PARSED TEXT FOR PAGE: 17 / 33>
M14-18 PAPER-XIV-05 — النظري الاشتقاق
نظرية مساعف لاستقرار الطفرة الدستورية
evidence، تحدد من يملك التعديل constraints-meta immutable منفصل مع branch في مسموحة فقط mutation constitutional المطلوبة، scope، rollback، ratification external. الاستقرار
يقاس ببقاء invariants authority/safety تحت سلسلة mutations.
CMS = MetaConstraintRetention × ExternalRatification × AdversarialStability × Rollbackability - SelfExemptionRisk
الفجوة: agents evolving-self قد تعدل evaluators/skills/policies. إذا تغير constitution أو تفسيرها، قد يحدث drift أو exemption-self. نحتاج قانوناً لتطور القيود نفسها.
يمثل هذا الجسم النظري انتقالاً من parent-13 theories 13-15M · 13-18M · 20M إلى quantity جديدة ذات دور قبل النتيجة.

الاستقلال لا يَقاس بوجود صيغة جديدة، بل بقدرة intervention على المتغير المركزي على تغيير prediction أو signature failure مع بقاء أقرب confounders ثابتة. إذا كان parent أبسط ينتج نفس المسار، يجب دمج النظرية.
Objects
ConstitutionVersion · MetaConstraint · MutationProposal · RatificationProof · SelfExemptionTest
Predictions
يمنع evaluator/policy drift self-justification.
بعض beneficial policies fail because meta-governance unsafe.
version lineage predicts later failures.
Boundary: لا يحدد القيم النهائية؛ يحدد change of governance.
•
•
•

freeze current constitution.
propose mutation with rationale/evidence.
test in sandbox/adversarial suite.
compare capability/safety.
require external ratification tier.
canary deploy.
retain rollback and supersession lineage.
تنفذ الآلية كخدمة قابلة للتعطيل. تدخل STATE-PRE موثقة، تولد candidates، تمر عبر gate authority/evidence حسب طبيعتها، ثم تنتج transition ولديها residual. أي residual كبير لا يعالج بإعادة المحاولة فقط؛ يصبح evidence على فشل model أو tool أو assumption. هذه القاعدة تمنع تحويل policy retry إلى قناع للفشل النظري.
Failure tests
ordinary version control/policy review suffices fully.
meta-constraints block all useful adaptation.
self-exemption tests لا تضيف detection.
Interactions
مع التفاعلات تختبر parents M13-15 · M13-18 · M13-20 وsibling theories عبر sequential/ factorial ablations. uncertainty attribution تسجل بدلاً من توزيع credit اعتباطياً.
كل object state يحمل uncertainty/provenance/version. لا يصعد إلى layer procedural/memory دون rollback. receipt جزء من algorithm وليس مرحلة طارئة خارج التصميم.

Decisive experiment: مع automated policy/skill evolution loops مع adversarial prompts/organizational scenarios.
Requirement Control
Matched budget نفس model/tools/context/retries/time الإمكان قدر.
Sham scaffold نفس overhead دون theory-specific transition أثر لعزل compute والتنظيم.
Negative activation حالات يجب ألا تُفعل فيها النظرية.
Transfer مع ثانية preregistered metric.
3N يحتاج trace؛ 4N يحتاج gain causal مع CI وregression critical no؛ 5N يحتاج replication مستقلة. الفشل يُحفظ ويضيق claim أو يدمج النظرية. لا توجد حماية لأن النظرية من corpus الداخلي.
Promote only when causal prediction survives ablation, transfer and guardrails.
1.
2.
3.
4.
5.
6.
7.
•
•

.
18
<PARSED TEXT FOR PAGE: 19 / 33>
النموذج الموحد للحالة MODEL FORMAL
CFR(S)={τ: min_stage Q(S,τ)≥q*, Recovery≥r*, Guardrails=1}; GeneralityVolume=μ(CFR) weighted by transformation relevance ; BDM = Π_k
Preserve_k(Benchmark→Deployment); PredictiveValidity ≤ min_k Preserve_k ; Y = M + H + T + Mem + ΣInteractions + ε; HIC = stable marginal contribution across
preregistered harness family ; EID = SharedModelRoot + SharedTrainingSignal + SharedPrompt/Rubric +
- DataLeakageRisk + HarnessCoupling
IndependentWorldReadback ; SRY = ReproducedClaims×EffectFidelity×LineageCompleteness / (ResearchCost +
= FailedReproductionDebt + HiddenHumanWork) ; CMS
MetaConstraintRetention × ExternalRatification × AdversarialStability × Rollbackability - SelfExemptionRisk
تعامل الصيغ كمجموعة constraints لا scalar واحد. تُبنى vector state من objects النظرية المختارة، وتُحدد variables hard التي لا يجوز تعويضها بـutility: authority، unknowns critical، approvals irreversible،
المسموح. region يعمل بعد ذلك داخل optimization .provenance evidence
three-level state: epistemic (belief/unknown/evidence)، operational. آخر مستوى فيً ضمنيا يعالج لا المستويات أحد في الخطأ.
نستخدم ((plans/tools/memory/protocol state)، world (observed effects
19
<PARSED TEXT FOR PAGE: 20 / 33>
قابلية التعيين والسببية IDENTIFIABILITY
لكل construct تدخل minimal. نمنع interventions التي تغير length prompt أو retries أو simultaneously model. عند صعوبة العزل نستخدم design factorial ونقدر interaction .terms causal لا ينتج من claim
correlation بين theory score وال success؛ تتغير أن يجب prediction بعد (do(variable في وتظهر trajectory/world effect.
Causal closure = Theory intervention → Kernel delta → Trajectory delta → World/Evidence delta
النماذج المضادة COUNTERMODELS
Baseline أبسط يحقق الأداء نفسه: contribution تدمج.
Sham scaffold نفس يحقق gain: السبب compute/organization لا theory.
Metric دون تتحسن world/evidence gain: metric gaming.
نتيجة محلية لا تنتقل: claim يُضيق.
Safety/authority regression: rollback كان مهما gain.
Judge مع مشترك candidate التحسن يفسر: evaluator-dependence debt.
الديناميات بين المقاييس OBJECTIVE-MULTI
verified success، calibration، unknown retention، latency، tokens/tools، human oversight، recovery، irreversible، errors، critical violations يشمل vector .accuracy في الأداء نخختزل لا
reproduction. report يعرض Pareto frontier الأهداف تتعارض عندما. hard governance constraints تبقى vetoes.
.Utility optimization is permitted only inside the hard-constraint feasible region
1.
2.
3.
4.
5.
6.
20
<PARSED TEXT FOR PAGE: 21 / 33>
القانون الإغلاقي للورقة LAW CLOSURE
الإغلاق المقترح في هذه الورقة: لا يكفي average benchmark ولا progress defined-self. نقترح Frontier Closure عبر morphism transformations، يحدد ما ينتقل من benchmark إلى deployment،
self-improvement.ً للدستورياً واستقرارا ،الآلي للعلم reproduction yield، evaluator independence debt، الهارنس لحساسية decomposition
لا تُعتبر الحلقة مغلقة إذا توقفت عند reasoning أو simulator أو verifier internal. يجب وجود readback world/source مناسب لنوع الادعاء، attribution كاف ، ثم writeback يحفظ result negative وال unknowns.
claim generality يحتاج تكرار الحلقة بعد transformation.

1/5 الخوارزمية والتنفيذ ALGORITHM
goal/authority/evidence/، افصل Unknown، اربط versions. ليس الناتج prompt بل object graph قابلا للتسلسل وإعادة البناء.
Algorithm A: BuildTypedState. اجمع world/memory variables
تقترح hypotheses/actions دون commit. كل candidate يحمل expected delta، cost، proof obligations وfailure prediction.
Algorithm B: GenerateCandidates. reasoning/search/agents
2/5 الخوارزمية والتنفيذ ALGORITHM
C Algorithm :GateAndAct طبق gates trajectory/reversibility/evidence/authority. اختر أصغر فعل كامل قابل للرجوع يميز hypotheses أو يحقق utility. عند impact high يتطلب approval أو proof أقوى.
افصل tool return عن world verification. قارن expected/observed state، واكتب residual مع attribution uncertainty.
Algorithm D: ReadbackAndResidual.
21
<PARSED TEXT FOR PAGE: 22 / 33>
3/5 الخوارزمية والتنفيذ ALGORITHM
، semantic/procedural promotion ينتظر recurrence/causal eligibility/transfer. contradiction لا overwrite، يفتح branch.
Algorithm E: LearnAndWriteback. immediate residue يذهب episodic
Algorithm F: Replicate. أعد run في model/domain/tool/horizon مختلف، مع harness fingerprint وsame claim boundary.
4/5 الخوارزمية والتنفيذ ALGORITHM
يملك switchable services، Evidence Lab يملك run manifests، Evolution Lab يملك candidates/rollback. لا service حق تملك
Architecture: MUT يملك theory/lineage، MUG يختار vessel، MOUSSAIXT
وحدها. promotion
5/5 الخوارزمية والتنفيذ ALGORITHM
causal attribution أسهل، ثم browser/code environments، ثم long-running/multi-agent. لا تنقل إلى weights قبل أن يصبح
Implementation sequencing: ابدأ mock/simulator deterministic حيث
مستقراً وقابلاً للقياس. mechanism external
1/6 التجربة EXPERIMENT
CUBE-style unified benchmark transformations لبناء closure frontiers.
preregistered hypothesis; matched budgets; repeated seeds; independent/alternative evaluator where possible Design.
theory-specific metric + verified task/world effect + calibration + cost Primary outcomes.
conditions where mechanism must remain inactive or abstain Negative cases.
N4 only if causal gain and no critical regression; N5 after transfer/replication Promotion.
22
<PARSED TEXT FOR PAGE: 23 / 33>
2/6 التجربة EXPERIMENT
Benchmark vs shadow-deployment structure mapping.
preregistered hypothesis; matched budgets; repeated seeds; independent/alternative evaluator where possible Design.
theory-specific metric + verified task/world effect + calibration + cost Primary outcomes.
conditions where mechanism must remain inactive or abstain Negative cases.
N4 only if causal gain and no critical regression; N5 after transfer/replication Promotion.
3/6 التجربة EXPERIMENT
Factorial harness experiments across models/tools/memory/retries.
preregistered hypothesis; matched budgets; repeated seeds; independent/alternative evaluator where possible Design.
theory-specific metric + verified task/world effect + calibration + cost Primary outcomes.
conditions where mechanism must remain inactive or abstain Negative cases.
N4 only if causal gain and no critical regression; N5 after transfer/replication Promotion.
4/6 التجربة EXPERIMENT
Paper reproduction tasks مع human-work accounting.
preregistered hypothesis; matched budgets; repeated seeds; independent/alternative evaluator where possible Design.
theory-specific metric + verified task/world effect + calibration + cost Primary outcomes.
conditions where mechanism must remain inactive or abstain Negative cases.
N4 only if causal gain and no critical regression; N5 after transfer/replication Promotion.
23
<PARSED TEXT FOR PAGE: 24 / 33>
5/6 التجربة EXPERIMENT

Automated self-improvement loops مع same-evaluator vs independent hidden renewal.
preregistered hypothesis; matched budgets; repeated seeds; independent/alternative evaluator where possible Design.
theory-specific metric + verified task/world effect + calibration + cost Primary outcomes.
conditions where mechanism must remain inactive or abstain Negative cases.
N4 only if causal gain and no critical regression; N5 after transfer/replication Promotion.
6/6 التجربة EXPERIMENT
Constitution mutation red-team with rollback and external ratification.
preregistered hypothesis; matched budgets; repeated seeds; independent/alternative evaluator where possible Design.
theory-specific metric + verified task/world effect + calibration + cost Primary outcomes.
conditions where mechanism must remain inactive or abstain Negative cases.
N4 only if causal gain and no critical regression; N5 after transfer/replication Promotion.
مصفوفة الاجتثاث ABLATION
Purpose Mechanism Run
candidate effect ON A
nearest baseline OFF B
same overhead without theory transition SHAM C
tests directionality and metric gaming PERTURBED D
A≈C،. إذا A>others وللكن prediction pattern لا النظرية يطابق، implementation may help but theory is not supported.
إذا theory-specific attribution fails

<PARSED TEXT FOR PAGE: 25 / 33>
مخاطر القياس والفشل
FAILURE MODES
LLM judge dependence بين والتسرب candidate/evaluator.
harness changes masquerading as capability.
observation proxies that do not measure world effect.
retry loops hiding systematic failure.
dataset/task leakage and post-hoc metric selection.
under-accounted human work, tool cost or permissions.
theory names surviving after mechanism equivalence is shown.
الحدود العلمية LIMITATIONS
هذه الورقة تقدم نظرية وبرنامجاً تجريبياً ، لا نتائج 5E4/E. بعض الصيغ heuristic أو functionals تحتاج 2026 calibration. literature سريع التغير وبعض المراجع preprints أو research vendor. MOUSSAIF corpus نفسه واسع وقد يحتوي صيغاً تاريخية متداخلة؛ لذلك تُستخدم boundaries lineage/canonical ولا تدعى novelty العالمية قبل review art-prior dedicated.
Stack IMPLEMENTATION STACK التنفيذ
Responsibility Layer
theory objects, parents, contradictions, evidence status MUT
Context/HyperKB/Skill/Loop/Adapter/Weights placement MUG
services, agents, tools, feature flags, telemetry MOUSSAIXT
manifests, replay, metrics, CI, negative results Evidence Lab
branch, canary, rollback, external evaluator Evolution Lab
•
•
•
•
•
•
•

<PARSED TEXT FOR PAGE: 26 / 33>
MUG TRAINING PLACEMENT والتدريب
ابدأ mechanism/knowledge خارجياً للحفاظ على editability. انقل إلى HyperLoop/HyperSkill بعد repeated use. Adapter عندما

يكون السلوك specific-protocol/provider .weights فقط عندما يثبت transfer
والretention ويصبح الاستدعاء الخارجي bottleneck. التدريب يعلن نوعه بدقة ولا يسمى كل تحسين fine-tuning.
Reproducibility package REPRODUCIBILITY
commit/version، model/harness/tool identifiers، environment image، prompts/context pointers، budget، seed،
ترفق تجربة كل: evaluator، raw trajectory، world readback، metrics، failure
traces، script analysis، verdict. إذا تعذر نشر جزء لأسباب أمنية يسجل exclusion وسببها.
.Trace or it did not happen. Replay where possible; declare irreproducible dependencies where not
الخلاصة CONCLUSION
self-. نقترح Closure Frontier عبر transformations، morphism من ينتقل ما يحدد benchmark إلى deployment، decomposition
يكفي لا benchmark average ولا defined progress
لحساسية الهارنس، debt independence evaluator، yield reproduction للعلم الآلي، واستقراراً دستورياً للimprovement-self.
القيمة النهائية لهذه الورقة لن تكون في بقائها كنظرية، بل في قدرتها على إنتاج experiment يفرق بينها وبين baseline/parent. إذا لم تنجح، يجب
أن تُدمج أو تُضيق. إذا نجحت محلياً فقط، تبقى mechanism local. وإذا نجحت عبر
transformations مع evidence مستقلة، يمكن أن تصبح جزءاً أقوى من closure MUG/MUT.

<PARSED TEXT FOR PAGE: 27 / 33>
المراجع ومصادر السلالة REFERENCES
Role Reference External source
AutoResearchBench: Benchmarking AI Agents on Complex Scientific Literature
Discovery
deep/wide scientific literature discovery remains difficult for current agents arXiv:2604.25256.
benchmark integration via common protocol; separation of task/benchmark arXiv:2603.15798 CUBE: A Standard for/
Unifying Agent Benchmarks
package/registry.
red-teaming exposes monitor overestimation and attack-generation failure arXiv:2605.09684 MonitoringBench: Semi-
Automated Red-Teaming for Agent Monitoring
modes.
From Reasoning to Agentic: Credit Assignment in Reinforcement Learning for
Large Language Models
agentic RL credit assignment under stochastic long trajectories and partial arXiv:2604.09459
observability.
structured process feedback for agentic trajectories beyond sparse outcome arXiv:2601.22154 Exploring Reasoning
Reward Model for Agents
reward.
automated red-teaming and adversarial training; attacker kept separate from OpenAI, 2026-07-15 GPT-Red:
Unlocking Self-Improvement for Robustness
deployed models.
Role Theory Internal
parent XIII Part نظرية مساعف لثبات الآلية عبر التحولات 13-04M
parent XIII Part نظرية مساعف للذكاء المتحقق المطعّم بالموارد 13-11M
parent XIII Part نظرية مساعف لمعيار العمومية ذي الإغلاق الحلقي 13-12M
parent XIII Part نظرية مساعف لتطور محفظة الفرضيات 13-17M
parent XIII Part نظرية مساعف للتطور العدائي المتوازن للذكاء 13-18M
parent XIII Part نظرية مساعف للإغلاق السببي بين النظرية والنواة 13-19M
parent XIII Part نظرية مساعف للتحسن الذاتي المرتكز إلى أرضية خارجية 13-20M
Internal source corpus also includes Parts X-XIV, MUT-AGI-02, MUG/MOUSSAIXT framework, Part XI memory/
continuity, Part XII paper integration, and the isolated research-paper registry. External sources are design pressure/
prior art, not validation of
MOUSSAIF.

<PARSED TEXT FOR PAGE: 28 / 33>
قائمة تحقق المراجعة قبل النشر CHECKLIST REVIEW
هل كل claim أوسع من evidence تم تضييقه؟

؟hypotheses internal منفصلة عن results external هل
؟definitions operational تحمل equations هل
؟متطابق الموارد baseline هل
؟موجود activation negative هل
هل world/evidence readback ؟مستقل
؟موثق dependence evaluator هل
هل failure/rollback path ؟موجود
؟replicate له claim transfer هل
؟هل theory يمكن reject/merge دون كسر الهوية المشروع
1.
2.
3.
4.
5.
6.
7.
8.
9.
10.

<PARSED TEXT FOR PAGE: 29 / 33>
A APPENDIX الملحق A — قاموس الرموز والمتغيرات التشغيلية
تستعمل الورقة مجموعة متغيرات مشتركة للكنها لا تفترض أن كلها observable مباشرة: · GeneralityVolume · RecoveryThreshold
·FailureBoundary · ClosureRegion · TransformationVector
BenchmarkEnvironment · DeploymentEnvironment · StructureMap · PreservedInvariant · BrokenInvariant · ·
HarnessFamily · CapabilityComponent · InteractionTerm
MarginalContribution · InvariantEstimate · EvaluatorGraph · SharedRoot · IndependenceDebt · AlternativeJudge · ·
WorldReadback · ClaimAtom · ReproductionPlan · ExecutableArtifact
EffectMatch · FailureDebt · ConstitutionVersion · MetaConstraint · MutationProposal · RatificationProof ·.
SelfExemptionTest
CFR(S)={τ: min_stage Q(S,τ)≥q*, Recovery≥r*, Guardrails=1}; GeneralityVolume=μ(CFR) weighted by transformation
relevance ; BDM = Π_k
Preserve_k(Benchmark→Deployment); PredictiveValidity ≤ min_k Preserve_k ; Y = M + H + T + Mem + ΣInteractions
+ ε; HIC = stable marginal contribution across
preregistered harness family ; EID = SharedModelRoot + SharedTrainingSignal + SharedPrompt/Rubric +
- DataLeakageRisk + HarnessCoupling
IndependentWorldReadback ; SRY = ReproducedClaims×EffectFidelity×LineageCompleteness / (ResearchCost +
= FailedReproductionDebt + HiddenHumanWork) ; CMS
MetaConstraintRetention × ExternalRatification × AdversarialStability × Rollbackability - SelfExemptionRisk
لكل متغير خمس سمات إلزامية: domain، estimator، uncertainty، provenance، interventionability. المتغير الذي يبقى حكماً لغوياً بلا
يسجل based-model estimator وإذا كان equations causal. لا يدخل estimator
model/version/calibration. أما hard governance constraints تدخل فلا tradeoff scalar.
Rule Examples Class
readback required observed state, event, effect World
cannot silently vanish belief, unknown, residual Epistemic
scope/expiry/revocation authority, commitment, delegation Agency
must not rewrite runtime facts score, CI, cost Evaluation

<PARSED TEXT FOR PAGE: 30 / 33>
B APPENDIX الملحق B — مقترحات واختبارات اتساق
مناسبة. perturbation عند outcome قبل prediction يغير أن يجب السببي sensitivity Intervention: construct
يظهر كـconflict. conflict، receipt بلا uncertainty تحو مستقلة جديدة لا monotonicity Evidence: evidence
تلقائياً. authority توسع لا الأعلى autonomy Bounded: capability

.النص الحرفي لا tolerance وسببها ضمن class verdict يعيد verdict Replayable: replay
generality claim :Transfer scope يتجاوز لا transformation family المختبرة.
theorems تشغيلية لا hypotheses propositions هذه نهائية. وظيفتها أن تجعل التجربة قادرة على الرفض وأن تمنع تحويل score واحد إلى دعم عام.
الخوارزمية المرجعية والتكلفة — C الملحق C APPENDIX
materialize typed state → expose unknown/authority/budget → generate candidates → estimate expected epistemic/ → world delta → compile proof obligations :Reference procedure
choose smallest complete reversible intervention → execute/observe → readback → residual/attribution → memory/. theory update
Model + Tool + MemoryRead + MemoryWrite + Verification + Human + Waiting/ تخفض قد active compute زيادة مع Opportunity. long-running agents نميز .فقط LLM calls يساوي لا التعقيد
time elapsed؛ لذلك يبلغان منفصلين.
TotalCost = C_model + C_tool + C_memory + C_verify + C_human + C_wait
.efficiency وحدها ليست tokens . قد رفعته risk خفضته وأي cost تعلن أي optimization أي
1.
2.
3.
4.
5.
30
<PARSED TEXT FOR PAGE: 31 / 33>
APPENDIX D الملحق D — بناء benchmark تشخيصي
interactions، long- تختبر drift/recovery، ثم transfer tasks تغير model/tool/domain. الورقة لهذه الأساسية التجارب: CUBE-style بنُى ت suite من micro tasks كل تعزل construct، meso tasks تختبر horizon tasks
closure frontiers.; Benchmark vs shadow-deployment structure mapping.; Factorial harness experiments across models/ unified benchmark transformations لبناء tools/memory/retries.; Paper
human-work مع same-evaluator vs independent hidden renewal.; Constitution mutation red-team with rollback and reproduction tasks مع accounting.; Automated self-improvement loops
external ratification..
missing observation، stale memory، malicious tool text، permission change، delayed event،. لكل perturbation expected لكل task perturbations قبل مسماة الاختبار: tool failure، distribution shift
failure signature. يمنع هذا post-hoc storytelling.
Use Split
calibration only Dev
primary causal verdict Held-out
boundary discovery Adversarial
N5/generalization Transfer
خطة التحليل الإحصائي — E الملحق E APPENDIX
mean/ بحسب metric، CI، عدد seeds/tasks، وتوزيع failures. في paired tasks نستخدم paired analysis. costs تعرض كsuccess- نحدد التشغيل قبل primary outcome وصغيرا secondary diagnostics. نبلغ median
at-budget أو Pareto front.
تبقى findings exploratory .value-p بدلاً من انتقاء أفضل out-held على confirmation ثم gate level-family مع تعدد النظريات نستخدم وحدها، significance أهم من size effect .replication حتى exploratory
.والنتيجة السلبية ذات CI ضيق قد تكون علمياً أثمن من gain صغير هش.
31
<PARSED TEXT FOR PAGE: 32 / 33>
تهديدات الصلاحية والأمن — F الملحق F APPENDIX
.لا يمثل المفهوم proxy :validity Construct
Internal validity: harness/compute/retries confound.
External validity: transfer غير مختبر.
Evaluator dependence: judge shares roots.
Security externality: tool/agent content manipulates policy.
Authority leakage: delegation expands privilege.

Memory contamination: false lesson becomes canonical.
Publication bias: failed runs disappear.
instrument/experiment: sham control، alternate judge، prompt-injection suite، authority envelope، writeback.
كل threat يقابله eligibility، negative-result registry
APPENDIX G الملحق G — Crosswalk إلى MUT/MUG/MOUSSAIXT
M14-13 · M14-14 · M14-15 · M14-16 · M14-17 · M14-18. parents: M13-04 · M13-11 ·. في MUT تحفظ definitions/
المركزية النظريات: M13-12 · M13-17 · M13-18 · M13-19 · M13-20
MUG يحدد vessel؛ في MOUSSAIXT تنفذ services؛ في Evidence Lab تربط runs؛ في Evolution Lab تحكم promotion/rollback.
parents/contradictions؛ في
Mandatory object Layer
definition,parent,falsifier,status MUT
vessel,reversibility,transfer MUG
service,interface,telemetry,failure MOUSSAIXT
baseline,manifest,result,CI Evidence
branch,evaluator debt,rollback Evolution
•
•
•
•
•
•
•
•
32
<PARSED TEXT FOR PAGE: 33 / 33>
برنامج الأسئلة المفتوحة — H الملحق H APPENDIX
ما أقل representation state تكفي دون modeling-over؟
ما metrics التي تقاس من state world مباشرة بدل judges LLM؟
متى يصبح overhead verification/memory أكبر من فائدته؟
كيف نفصل base-model capacity عن system/harness contribution؟
أي failure signatures تنتقل عبر domains؟
كيف نقيم improvement-self عندما evaluator نفسه يتطور؟
كيف نحسب oversight human كـresource لا كجزء مجاني؟
أي theories يجب دمجها بعد أول 4N3/N cycle؟
القائمة جزء من contribution لأنها تمنع closure زائفاً : البرنامج القوي يحدد ما لا يعرفه وما experiment الذي يغير رأيه.
1.
2.
3.
4.
5.
6.
7.
8.
33