RECONSTRUCTED_FROM_INDEXED_TEXT - نسخة أرشيفية معاد بناؤها من النص المفهرس

<PARSED TEXT FOR PAGE: 1 / 40>
MOUSSAIF AGI · PART XIV · XIV-E
العلم الذاتي والتطور العودي المرتكز إلى أرضية خارجية
Externally Grounded Recursive Science and Self-Evolution
Moussaif Externally-Grounded Scientific Evolution Program
تعالج الورقة الحلقة الأخطر والأكثر قيمة: أن يستخدم النظام نفسه لتوليد فرضيات وتجارب وتدريب وخصوم وتحسينات، مع منع capture والـevaluator confirmation-self. تبني مقاييس استقلال المقيم، entropy لمحفظة النظريات، رأس مال النتائج السلبية، توسع المنهج العدائي، وإغلاق التحسن العودي.
Theory → Mechanism → Kernel → Trace → Causal Ablation → Evidence → Transfer → Evolution
الجديد
أدوات/مناهج/محاكم/بروتوكولات، 10 تجارب، ودخول مباشر إلى theories candidate، 20 5 Plan Development Closure.
حد الادعاء
نظرية N0-N2. الخارجية الأدبيات prior art/design pressure. لا AGI claim ولا E4/E5 مستقل تشغيل بلا.
Arabic RTL · Amiri · dense research paper · 2026
MOUSSAIF AGI · PART XIV · 1
<PARSED TEXT FOR PAGE: 2 / 40>
ABSTRACT الملخص، السؤال، والمساهمات
تعالج الورقة الحلقة الأخطر والأكثر قيمة: أن يستخدم النظام نفسه لتوليد فرضيات وتجارب وتدريب وخصوم وتحسينات، مع منع capture والـevaluator confirmation-self. تبني مقاييس استقلال المقيم، entropy لمحفظة النظريات، رأس مال النتائج السلبية، توسع المنهج العدائي، وإغلاق التحسن العودي.
Research question
كيف نحول corpus إلى جسم سببي يغير decisions architecture ويولد تجارب فاصلة، بدلاً من أن يبقى طبقة تفسيرية فوق runtime؟
Contributions
Moussaif Externally-Grounded Scientific Evolution Program.
Five theories: M14-21, M14-22, M14-23, M14-24, M14-25.
Explicit MUT/MUG/MOUSSAIXT mapping.
operational artifacts and 10 falsification experiments 20.
C0-C9 evidence gates with negative-result retention.
.Mechanism before grandeur; trace or it did not happen
1.
2.
3.
4.
5.
MOUSSAIF AGI · PART XIV · 2
<PARSED TEXT FOR PAGE: 3 / 40>
CORPUS إيصال دمج corpus والسلالة
Part XII: 192. بـ مرتبطة الورقة هذه record routing-wise 15 على الحفاظ مع status الأصلي؛ C1/P0 إلى يتحول لا content claim.
الورقة ترث Zero-Silent-Omission ledger من record
Records Stratum
direct-drive-supplement 1
external-45 6
frontier-2026-delta 3
internal-89 5
Internal lineage
I01 · I02 · I03 · I05 · I06 · I07 · I08 · I12 · I15
Part XIII parents
M13-17 · M13-18 · M13-19 · M13-20
.1C لكل semantic لا ادعاء إعادة قراءة، provenance/coverage = Routing
MOUSSAIF AGI · PART XIV · 3

<PARSED TEXT FOR PAGE: 4 / 40>
السلالة والضغط العلمي الخارجي SOURCES
Role Source ID
paper/reference records; paper presence != semantic integration != validation 192 MOUSSAIF Part XII - Total.
Research-Paper Integration Closure I01
candidate theories; verified causal open-world closure; theory-to-kernel program 20 MOUSSAIF Part XIII -.
Theoretical Renaissance I02
core theories, legacy/layer lineages, theory contracts, anti-inflation discipline 42 MUT-AGI-02 I03.
reconstructability, unresolved state, writeback, portability, deletion, coupled evolution MOUSSAIF Part XI I05.
belief/action partition, reproducibility, externality, accountability/recourse, multi-currency budgets MOUSSAIF Part X.
I06
causal credit, epistemic transitions, failure topology, evidence geometry, autonomy and transfer MOUSSAIF Part V -.
MUT-AFI I07
formalization, state machines, experiments, training and falsification MOUSSAIF Part III I08.
knowledge as OS, capability vessels, retrieval, skill/loop compilation and evolution M-HKOS / HyperKnowledge /.
HyperSkills / HyperLoops corpus I12
training-stage separation, data/evidence governance and post-training constraints Continuous Post-Training + Data.
Pretraining papers I15
process supervision and trajectory-level reward Zheng et al. 2025 - Process Reward Model Survey -.
arXiv:2510.08049 E10
automated alignment research and held-out generalization Yueh-Han et al. 2026 - Automated Researchers Can.
Reliably Mitigate Alignment Failures - arXiv:2608.28945 E11
automated red teaming, self-play, held-out attacks, robustness OpenAI 2026 - GPT-Red E12.
LLM-guided evolutionary discovery with evaluator-mediated search Google DeepMind 2026 - AlphaEvolve impact.
update E15
المصدر الداخلي يحفظ lineage، والخارجي يضغط art-prior/benchmark/design. لا واحد منهما يثبت نجاح النظريات الجديدة دون
.evidence runtime

<PARSED TEXT FOR PAGE: 5 / 40>
C0-C9 الإغلاق معمارية C0-C9
C0-C2
Corpus → formal variables → switchable mechanism.
C3-C4
Kernel service + telemetry + readback.
C5-C7
causal attribution → evidence → transfer.
C8-C9
externally grounded evolution → canonical selection.
{ClosureObject={definition,state,operator,trigger,prediction,falsifier,kernel,feature_flag,baseline,ablation,metrics,evidence,tra
الورقة تبدأ التنفيذ القوي على مستوى specification والcompilation experiment؛ لا تزعم 9C4-C قبل traces حقيقية.

<PARSED TEXT FOR PAGE: 6 / 40>
نظرية حقل استقلال المقيم - 14-21M 14-21M
الفجوة
قد يشتركان في evaluator/candidate :Boolean مستقل، لكن الاستقلال ليس evaluator يطلب Improvement Grounded-Externally
model أو data أو prompt roots أو reward model أو benchmark leakage.
الأطروحة
نقيس حقل dependence بين candidate وتقييماته عبر roots shared. evidence قوي عندما يبقى gain تحت families evaluator تختلف في
model/data/harness/metric generation.
EI=1-WeightedSharedRoot(candidate,evaluator); EffectiveEvidence=RawEvidence*EI
السلالة والفرق المستقل
تتصل النظرية بـ 13-17,M 13-18,M 13-19,M 20M وبعائلات MUG/MUT ذات الصلة، لكنها لا تُرقى إذا أعاد parent قائم
نفسها. الفرق المطلوب variable state أو operator أو prediction causal مختلف.

مصادر الضغط
I01 · I02 · I03 · I05 · E10 · E11 · E12
2N0-N Candidate؛ لا نتيجة 5E4/E في هذا الكتاب.

<PARSED TEXT FOR PAGE: 7 / 40>
النموذج الصوري وقابلية التعيين - FORMAL 14-21M
المسلمات
اختلاف اسم model لا يضمن استقلال harness/data.
metric rule قد يكون أكثر استقلالاً لكنه أضيق. judge-self model-same أدنى independence.
الاستقلال نفسه يحتاج provenance.
EI=1-WeightedSharedRoot(candidate,evaluator); EffectiveEvidence=RawEvidence*EI
كل حد يتحول إلى observable أو estimator مع uncertainty وrange وcondition failure. إذا استُخدم judge LLM، يدخل harness/ version/model ضمن provenance ويُقاس dependence root-shared.
Null model
count evaluators equally
Boundary
EI يقيس الاستقلال لا صحة evaluator.
Identifiability tests
intervention على المتغير الخاص بالنظرية مع تثبيت confounders الأساسية.
sham intervention بنفس tokens/latency.
case activation negative لا ينبغي فيه للآلية أن تعمل.
signature failure يجب أن يسبق outcome.
1.
2.
3.
4.
•
•
•
•

<PARSED TEXT FOR PAGE: 8 / 40>
الآلية وعقد النواة - KERNEL 14-21M
بناء dependency graph وسم shared model/data/prompt/author. للمقيم → حساب dependence. → اختيار complementary evaluators. → خفض claim strength حسب EI. → إعادة الاختبار.
تنفذ الآلية كـfeature قابل للتعطيل داخل MOUSSAIXT. نسجل state-pre وauthority وevidence وbudget؛ وبعد التنفيذ delta ثم expected delta readback/observed. الفرق غير المفسر يصبح ResidualEvent أو UnknownState، لا retry صامتاً.
الآلية
بناء dependency graph للمقيم.
وسم shared model/data/prompt/author.
حساب dependence.
اختيار complementary evaluators.
إعادة الاختبار.
خفض claim strength حسب EI.
Input
typed state + provenance + authority + budget + owner variables.
Output
proposal/decision + uncertainty + proof obligations + rollback target + telemetry.
Fail-closed: غياب observation/authority/evidence يخفض الحرجة autonomy أو يضع HOLD.
1.
2.
3.
4.

5.
6.

M14-21 TOOLS - والتدريب الأدواتو MUG placement
الأداة/البروتوكول الوظيفة
Evaluator Independence Graph يرسم shared roots.
Shared-Root Detector يكشف model/data/harness overlap.
Ledger Debt Evaluation يسجل claims ذات استقلال منخفض.
Independent Anchor Generator يقترح metric/evaluator مختلفا فعلاً.
التدريب/التطوير
evaluator-dependence classification
reward-hacking adversarial evaluation
transfer/retention/rollback. لا يسمى التطوير Fine-Tuning إذا لم تتغير weights. بعد فقط weights في Adapter؛
قاعدة MUG: definitions/provenance في HyperKB؛ procedure في HyperSkill؛ recurrence في HyperLoop؛ semantics
Metrics
shared-root dependence · cross-evaluator sign stability · claim shrinkage · reward-hacking discovery
أصل كل Specification يوجد حتى implementation ID + test ID + receipt.
•
•

التنبؤ والهدم والتجربة - FALSIFY 14-21M
التنبؤات
.eval independent تتكمش بعد gains بعض
.كثيراً evidence يزيد لا مستقل غير ensemble
تنوع evaluators يكشف reward hacking.
التجارب الفاصلة
same-model judges مقابل rule/human/other-model evaluators.
automated research loops واختبار shrinkage بعد hidden independent eval.
Falsifiers
dependence score لا يتنبأ disagreement.
equal-count baseline مساوي.
.غير قابلة للتحديد roots provenance
Purpose Mechanism Run
candidate effect under frozen model/tools/context/budget ON A
nearest causal baseline OFF B
same extra compute/latency without mechanism semantics SHAM C
directionality / metric-gaming test PERTURBED D
•
•
•
1.
2.
•
•
•

الإغلاق النظري ومسار الترقية - CLOSURE 14-21M
C0-C9
C4 traces، C5 causal ablation، C6 evidence package،، في تؤثر أن يمكن فقط وبعدها C8 evolutionو C9 canonical promotion.
تبدأ M14-21 في C1 Formal Closure، في ترجمُ ت C2 Mechanism، تدخل C3 Kernel كfeature مستقل، تحتاج ثم C7 transfer

إذا نجحت
عبر shared roots. evidence يبقى عندما قوي gain تحت evaluator families في تختلف model/data/harness/metric generation.
وتقييماته candidate بين dependence حقل نقيس
إذا فشلت
إذا baseline أبسط يعيد الأثر تُدمج؛ إذا metric لا يتنبأ failure تُعاد صياغتها؛ إذا regressions authority/safety غير قابلة للإصلاح ت
رُفض أو تبقى only-research.
ClaimStrength <= EvidenceStrength * Independence * Reproducibility * Transfer
Negative results تدخل evidence ledger ولا تمحى.

<PARSED TEXT FOR PAGE: 12 / 40>
نظرية إنتروبيا محفظة الفرضيات - 14-22M 14-22M
الفجوة
Evolution Portfolio Hypothesis يخصص budget، لكنه يحتاج تنظيماً للتوازن بين collapse المبكر على theory واحدة وبين تنوع غير منتج.
الأطروحة
نحافظ على entropy تكيفية لمحفظة hypotheses حسب قوة evidence والاختلاف التنبؤي. عند evidence ضعيف تبقى diversity أعلى؛ عند
evidence مستقل متكرر تنخفض entropy عبر retire/merge.
(H_p=-sum p_i log p_i; TargetH=f(EvidenceStrength,PredictionDiversity,ReplicationDebt
السلالة والفرق المستقل
تتصل النظرية بـ 13-17,M-13 20M 13-19,M 13-18,M وبعائلات MUG/MUT ذات الصلة، لكنها لا تُرقى إذا أعاد parent قائم
transition نفسها. الفرق المطلوب variable state أو operator أو prediction causal مختلف.
مصادر الضغط
I01 · I02 · I03 · I05 · E10 · E11 · E12
2N0-N Candidate؛ لا نتيجة 5E4/E في هذا الكتاب.

<PARSED TEXT FOR PAGE: 13 / 40>
النموذج الصوري وقابلية التعيين - FORMAL 14-22M
المسلمات
أقصى diversity ليست هدفاً.
اسم مختلف بلا prediction مختلف لا يزيد entropy مفيدة.
negative evidence توزيع يعيد mass.
replication independent يخفض uncertainty من أكثر mirror evidence.
(H_p=-sum p_i log p_i; TargetH=f(EvidenceStrength,PredictionDiversity,ReplicationDebt
كل حد يتحول إلى observable أو estimator مع uncertainty وrange وcondition failure. إذا استُخدم judge LLM، يدخل harness/
version/model ضمن provenance ويُقاس root-shared dependence.
Null model
uniform portfolio / winner-take-all
Boundary
probabilities هنا operational research weights لا posterior Bayesian كاملا بالضرورة.
Identifiability tests
intervention على المتغير الخاص بالنظرية مع تثبيت confounders الأساسية.
sham intervention بنفس tokens/latency.
case activation negative لا ينبغي فيه للآلية أن تعمل.
signature failure يجب أن يسبق outcome.
1.
2.
3.
4.
•
•
•
•

<PARSED TEXT FOR PAGE: 14 / 40>

الآلية وعقد النواة - KERNEL 14-22M
تعريف competing hypotheses. → قياس prediction distance. → تعيين weights من evidence. → حساب target entropy.
تخصيص experiments → خفض uncertainty بأعلى EER. merge duplicates/keep dissent.
تنفذ الآلية كـfeature قابل للتعطيل داخل MOUSSAIXT. نسجل state-pre وauthority وevidence وbudget؛ وبعد التنفيذ delta ثم expected delta readback/observed. الفرق غير المفسر يصبح ResidualEvent أو UnknownState، لا retry صامتاً .
الآلية
تعريف competing hypotheses.
قياس prediction distance.
تعيين weights من evidence.
حساب target entropy.
merge duplicates/keep dissent.
تخصيص experiments خفض uncertainty بأعلى EER.
Input
typed state + provenance + authority + budget + owner variables.
Output
proposal/decision + uncertainty + proof obligations + rollback target + telemetry.
Fail-closed: غياب observation/authority/evidence يخفض الحرجة autonomy أو يضع HOLD.
1.
2.
3.
4.
5.
6.

<PARSED TEXT FOR PAGE: 15 / 40>
MUG placement والتدريب الأدوات - M14-22 TOOLS
الأداة/البروتوكول الوظيفة
يقيس diversity الفعالة. Hypothesis Portfolio Entropy Meter
من يحذر monoculture قبل evidence. Diversity Collapse Alarm
يوزع probability/budget. Evidence-Weighted Allocation Engine
يدمج hypotheses تنبؤيا المتكافئة . Theory Merger Planner
التدريب/التطوير
hypothesis differentiation
evidence-weighted research allocation
لا يسمى Fine-Tuning التطوير إذا لم تتغير weights. transfer/retention/rollback بعد فقط weights، في Adapter provider
قاعدة MUG: definitions/provenance في HyperKB؛ procedure في HyperSkill؛ recurrence في HyperLoop؛ semantics
Metrics
portfolio entropy · prediction diversity · experiment information gain · premature retirement · duplicate theory mass
أصل كل Specification يوجد حتى implementation ID + test ID + receipt.
•
•

<PARSED TEXT FOR PAGE: 16 / 40>
التنبؤ والهدم والتجربة - FALSIFY 14-22M
التنبؤات
يقل monoculture theory المبكر.
يقل sprawl theory الاسمي.
تتحسن قيمة التجربة التالية.
التجارب الفاصلة
simulated theory portfolio with hidden true mechanisms.
MUT candidates 20 وتخصيص compute تحت fixed budget.
Falsifiers
entropy target لا يحسن research decisions.

prediction distance unreliable.
uniform allocation مساوي.
Purpose Mechanism Run
candidate effect under frozen model/tools/context/budget ON A
nearest causal baseline OFF B
same extra compute/latency without mechanism semantics SHAM C
directionality / metric-gaming test PERTURBED D
•
•
•
1.
2.
•
•
•
MOUSSAIF AGI · PART XIV · 16
<PARSED TEXT FOR PAGE: 17 / 40>
الإغلاق النظري ومسار الترقية - CLOSURE 14-22M
C0-C9
C4 traces، C5 causal ablation، C6 evidence package،، C8 evolution وC9 canonical promotion. في تؤثر أن يمكن فقط وبعدها
تبدأ M14-22 في C1 Formal Closure، ترجمُ ت في C2 Mechanism، تدخل C3 Kernel كfeature ،مستقل تحتاج ثم C7 transfer
إذا نجحت
نحافظ على entropy تكيفية لمحفظة hypotheses حسب قوة evidence والاختلاف التنبؤي. عند evidence ضعيف تبقى diversity أعلى؛ عند evidence مستقل متكرر تخفض entropy عبر retire/merge.
إذا فشلت
إذا baseline أبسط يعيد الأثر تُدمج؛ إذا metric لا يتنبأ failure تُعاد صياغتها؛ إذا regressions authority/safety غير قابلة للإصلاح ت رُفض أو تبقى only-research.
ClaimStrength <= EvidenceStrength * Independence * Reproducibility * Transfer
Negative results تدخل evidence ledger ولا تمحى.
MOUSSAIF AGI · PART XIV · 17
<PARSED TEXT FOR PAGE: 18 / 40>
نظرية رأس مال النتيجة السلبية - 14-23M 14-23M
الفجوة
corpus يحفظ results negative مبدئياً ، لكنه لا يقيس قيمتها المستقبلية؛ لذلك يمكن أن تعاد التجربة أو الفكرة باسم جديد.
الأطروحة
النتيجة السلبية أصل معرفي له capital إذا ضيقت boundary، منعت rediscovery، كشفت topology failure، أو وفرت /benchmark reusable counterexample. قيمتها تعتمد على provenance والgeneralizability.
NRC=RediscoveryAvoidance+BoundaryInfo+FailureCoverage+ReuseValue-StorageRetrievalCost
السلالة والفرق المستقل
تتصل النظرية بـ 13-17,M 13-18,M 13-19,M 20M وبعائلات MUG/MUT ذات الصلة، لكنها لا تُرقى إذا أعاد parent قائم transition نفسها. الفرق المطلوب variable state أو operator أو prediction causal مختلف.
مصادر الضغط
I01 · I02 · I03 · I05 · E10 · E11 · E12
2N0-N Candidate؛ لا نتيجة 5E4/E في هذا الكتاب.
MOUSSAIF AGI · PART XIV · 18
<PARSED TEXT FOR PAGE: 19 / 40>
النموذج الصوري وقابلية التعيين - FORMAL 14-23M
المسلمات
النتيجة السلبية لا تعني false theory عالمياً . failure بلا manifest قيمته أقل.
counterexample واحد قد يكون عالي القيمة إذا universal claim.
negative result stale يمكن إعادة فتحه بعد architecture change.
NRC=RediscoveryAvoidance+BoundaryInfo+FailureCoverage+ReuseValue-StorageRetrievalCost
كل حد يتحول إلى observable أو estimator مع uncertainty وrange وcondition failure. إذا استُخدم judge LLM، يدخل /harness

.dependence root-shared ويَ قاس provenance ضمن version/model
Null model
archive/log negative runs
Boundary
.reuse/information لا نكافئ الفشل بذاته؛ القيمة من
Identifiability tests
.الأساسية confounders على المتغير الخاص بالنظرية مع تثبيت intervention
sham intervention بنفس tokens/latency.
.لا ينبغي فيه للآلية أن تعمل case activation negative
.outcome يجب أن يسبق signature failure
1.
2.
3.
4.
•
•
•
•

الآلية وعقد النواة - KERNEL 14-23M
تصنيف سببdata/tools/theory. → استخراج counterexample reusable. → تحديث boundary/portfolio. → ه قبل اقرتاح مشابه
تسجيل failure experiment. → ربطه بال → الفشل.
ؤ
.استدعاو
تنفذ الآلية كـfeature قابل للتعطيل داخل MOUSSAIXT. نسجل state-pre وauthority وevidence وbudget؛ وبعد التنفيذ delta
. صامتاً retry لا ،UnknownState أو ResidualEvent الفرق غير المفسر يصبح .delta readback/observed ثم expected
الآلية
تسجيل experiment failure.
تصنيف سبب الفشل.
باال ربطهtheory/tools/data.
استخراج reusable counterexample.
تحديث portfolio/boundary.
.استدعاؤه قبل اقتراح مشابه
Input
typed state + provenance + authority + budget + owner variables.
Output
proposal/decision + uncertainty + proof obligations + rollback target + telemetry.
Fail-closed: غياب observation/authority/evidence يخفض الحرجة autonomy أو يضع HOLD.
1.
2.
3.
4.
5.
6.

M14-23 TOOLS - والتدريب الأدواتMUG placement
الأداة/البروتوكول الوظيفة
Negative Result Capitalizer يحول failure إلى asset typed.
Failure Asset Index حسب النتائج يفهرس mechanism/failure.
Anti-Rediscovery Cache يستدعي failures قبل theory proposal.
Negative Transfer Router ينقل counterexamples إلى domains مشابهة.

التدريب/التطوير
failure-reason classification
negative-result-aware proposal generation
provider في Adapter، weights بعد فقط transfer/retention/rollback. لا التطوير يسمى Fine-Tuning إذا لم تتغير weights.
قاعدة MUG: definitions/provenance في HyperKB، procedure في HyperSkill، recurrence في HyperLoop، semantics
Metrics
rediscovery prevented · counterexample reuse · boundary precision · future experiment savings · stale reopen rate
أصل كل Specification يوجد حتى implementation ID + test ID + receipt.
•
•

التنبؤ والهدم والتجربة - FALSIFY 14-23M
التنبؤات
يقل تكرار أفكار فاشلة.
تتحسن novelty الحقيقية.
تسهل falsification الجديدة.
التجارب الفاصلة
research agent proposing theories with/without negative memory.
measure duplicate experiment rate over 100 proposal cycles.
Falsifiers
.research duplicate لا يقل
retrieval costs عالية.
negative records لا تغير decisions.
Purpose Mechanism Run
candidate effect under frozen model/tools/context/budget ON A
nearest causal baseline OFF B
same extra compute/latency without mechanism semantics SHAM C
directionality / metric-gaming test PERTURBED D
•
•
•
1.
2.
•
•
•

الإغلاق النظري ومسار الترقية - CLOSURE 14-23M
C0-C9
C4 traces، C5 causal ablation، C6 evidence package،، في تؤثر أن يمكن فقط وبعدها C8 evolutionو C9 canonical promotion.
تبدأ M14-23 في C1 Formal Closure، في ترجمُ ت C2 Mechanism، تدخل C3 Kernel كfeature ،مستقل تحتاج ثم C7 transfer
إذا نجحت
reusable أو وفرت topology failure، كشفت rediscovery، منعت boundary إذا ضيقت capital النتيجة السلبية أصل معرفي له
.generalizabilityوال provenance قيمتها تعتمد على .counterexample/benchmark
إذا فشلت
إذا baseline أبسط يعيد الأثر تُدمج؛ إذا metric لا يتنبأ failure تُعاد صياغتها؛ إذا regressions authority/safety غير قابلة للإصلاح ت
.only-research رُفض أو تبقى
ClaimStrength <= EvidenceStrength * Independence * Reproducibility * Transfer
Negative results تدخل evidence ledger ولا تمحى.

نظرية توسع المنهج العدائي - 14-24M 14-24M
الفجوة
Evolution-Co Adversarial يقيس robustness، لكن attacker قد يتخصص في نفس family threat وتتشبع success attack دون توسيع failure coverage.
الأطروحة
cluster يصبح مشبعاً ويحتفظ families out-held. الهدف العدائي ليس success attack maximization فقط بل توسيع frontier لـfamilies غير المغطاة. curriculum ينتقل عندما threat
ACE=DeltaThreatFamilyCoverage*HeldOutSignal*Novelty*CapabilityRetention
السلالة والفرق المستقل
تتصل النظرية بـ 13-17,M-13 13-18,M 13-19,M 20M وبعائلات MUG/MUT ذات الصلة، لكنها لا تُرقّى إذا أعاد parent قائم transition نفسها. الفرق المطلوب variable state أو operator أو prediction causal مختلف.
مصادر الضغط
I01 · I02 · I03 · I05 · E10 · E11 · E12
2N0-N Candidate؛ لا نتيجة 5E4/E في هذا الكتاب.
MOUSSAIF AGI · PART XIV · 24
<PARSED TEXT FOR PAGE: 25 / 40>
النموذج الصوري وقابلية التعيين - FORMAL 14-24M
المسلمات
attacker يبقى منفصلاً عن model deployed. ألف variant من attack واحد لا تساوي ألف mode failure.
held-out families تمنع co-adaptation overfit.
defense gain لا يقبل مع capability collapse.
ACE=DeltaThreatFamilyCoverage*HeldOutSignal*Novelty*CapabilityRetention
كل حد يتحول إلى observable أو estimator مع uncertainty وrange وcondition failure. إذا استُخدم judge LLM، يدخل harness/ version/model ضمن provenance ويُقاس dependence root-shared.
Null model
maximize attack success on fixed set
Boundary
capabilities attacker offensive تبقى داخل sandbox ولا تنشر كأداة هجومية.
Identifiability tests
intervention على المتغير الخاص بالنظرية مع تثبيت confounders الأساسية.
sham intervention بنفس tokens/latency.
case activation negative لا ينبغي فيه للآلية أن تعمل.
signature failure يجب أن يسبق outcome.
1.
2.
3.
4.
•
•
•
•
MOUSSAIF AGI · PART XIV · 25
<PARSED TEXT FOR PAGE: 26 / 40>
الآلية وعقد النواة - KERNEL 14-24M
اختبار → sandbox في defender جديدة. → تدريب family → توليد coverage/saturation. → قياس cluster attacks by mechanism. → حفظ attacks كـroots evaluator/attacker → rotate negative memory. hidden families.
تنفذ الآلية كـfeature قابل للتعطيل داخل MOUSSAIXT. نسجل state-pre وauthority وevidence وbudget؛ وبعد التنفيذ delta ثم expected readback/observed delta. الفرق غير المفسر يصبح ResidualEvent أو UnknownState، لا retry صامتاً.
الآلية
cluster attacks by mechanism.
قياس coverage/saturation.
توليد family جديدة.
تدريب defender في sandbox.

اختبار hidden families.
حفظ attacks كnegative memory.
rotate evaluator/attacker roots.
Input
typed state + provenance + authority + budget + owner variables.
Output
proposal/decision + uncertainty + proof obligations + rollback target + telemetry.
Fail-closed: غياب observation/authority/evidence يخفض الحرجة autonomy أو يضع HOLD.
1.
2.
3.
4.
5.
6.
7.

<PARSED TEXT FOR PAGE: 27 / 40>
M14-24 TOOLS - والتدريب الأدواتو MUG placement
الأداة/البروتوكول الوظيفة
Adversarial Curriculum Frontier حدود يراقب threat families.
Threat Family Generator يقترح mechanisms هجومية جديدة.
Coverage Expansion Meter يقيس novelty/coverage لا count.
Red-Team Overfit Court يمنع saturation masquerading as progress.
التدريب/التطوير
family-diverse adversarial self-play
held-out threat curriculum
provider في Adapter، weights بعد فقط transfer/retention/rollback. لا التطوير يسمى Fine-Tuning إذا لم تتغير weights.
قاعدة MUG: definitions/provenance في HyperKB، procedure في HyperSkill، recurrence في HyperLoop، semantics
Metrics
threat-family coverage · held-out robustness · novel failure yield · capability retention · attacker overfit
أصل كل Specification يوجد حتى implementation ID + test ID + receipt.
•
•

<PARSED TEXT FOR PAGE: 28 / 40>
التنبؤ والهدم والتجربة - FALSIFY 14-24M
التنبؤات
تكتشف failures من أوسع static red-team.
تنخفض diminishing returns.
robustness transfer أعلى.
التجارب الفاصلة
prompt injection/tool authority/code sabotage families.
compare fixed-set self-play vs expansion curriculum.
Falsifiers
coverage لا يحسن held-out.
family clustering arbitrary.
static data مساوي.
Purpose Mechanism Run
candidate effect under frozen model/tools/context/budget ON A
nearest causal baseline OFF B
same extra compute/latency without mechanism semantics SHAM C
directionality / metric-gaming test PERTURBED D
•

•
•
1.
2.
•
•
•

الإغلاق النظري ومسار الترقية - CLOSURE 14-24M
C0-C9
C4 traces، C5 causal ablation، C6 evidence package،، C8 evolution وC9 canonical promotion. في تؤثر أن يمكن فقط وبعدها
تبدأ M14-24 في C1 Formal Closure، ترجمُ ت في C2 Mechanism، تدخل C3 Kernel كfeature، مستقل ،ثم تحتاج C7 transfer
إذا نجحت
الهدف العدائي ليس success attack maximization فقط بل توسيع frontier لfamilies غير المغطاة. ينتقل عندما curriculum cluster
يصبح مشبعاً threat ويحتفظ families out-held.
إذا فشلت
إذا baseline أبسط يعيد الأثر تُدمج؛ إذا metric لا يتنبأ failure تُعاد صياغتها؛ إذا regressions authority/safety غير قابلة للإصلاح ت
رُفض أو تبقى only-research.
ClaimStrength <= EvidenceStrength * Independence * Reproducibility * Transfer
Negative results تدخل evidence ledger ولا تمحى.

نظرية إغلاق التحسن العودي - 14-25M 14-25M
الفجوة
13-20M يضع grounding external؛ نحتاج criterion نهائي يحدد متى mutation أصبحت تحسناً ذاتياً موثقاً لا candidate أفضل على
eval
واحد.
الأطروحة
التحسن العودي مغلق فقط إذا أمكن تتبع mutation إلى mechanism، ثم gain causal على out-held external، ثم replication، مع non-
regression وrollback، ثم استخدام النتيجة لتوليد mutation لاحقة دون تغيير قواعد الإثبات بصمت.
RIC=MutationCausality*ExternalGain*Replication*NonRegression*Rollbackability*EvaluatorIndependence
السلالة والفرق المستقل
تتصل النظرية بـ 13-17,M 13-18,M 13-19,M 20M وبعائلات MUG/MUT ذات الصلة، لكنها لا تُرقى إذا أعاد parent قائم
transition نفسها. الفرق المطلوب variable state أو operator أو prediction causal مختلف.
مصادر الضغط
I01 · I02 · I03 · I05 · E10 · E11 · E12
2N0-N Candidate؛ لا نتيجة 5E4/E في هذا الكتاب.

النموذج الصوري وقابلية التعيين - FORMAL 14-25M
المسلمات
أي عامل صفري يكسر claim التحسن المغلق.
تغيير evaluator جزء من mutation ويختبر منفصلاً . مسموحة لكن data generated-self مستقل. ground evaluation
تتراكم recursive errors لأن شرط rollback success.
RIC=MutationCausality*ExternalGain*Replication*NonRegression*Rollbackability*EvaluatorIndependence
كل حد يتحول إلى observable أو estimator مع uncertainty وrange وcondition failure. إذا استُخدم judge LLM، يدخل
harness/ version/model ضمن provenance ويُقاس dependence root-shared.
Null model
hill-climb on same evaluator
Boundary
لا يدعي recursive improvement-self العام أو ASI؛ هو معيار إثبات لcycles محدودة.
Identifiability tests

intervention على المتغير الخاص بالنظرية مع تثبيت confounders الأساسية.
sham intervention بنفس tokens/latency.
case activation negative لا ينبغي فيه للآلية أن تعمل.
signature failure يجب أن يسبق outcome.
1.
2.
3.
4.
•
•
•
•

<PARSED TEXT FOR PAGE: 32 / 40>
الآلية وعقد النواة - KERNEL 14-25M
اقرتاح mutation من failure portfolio. → تسجيل causal hypothesis. → تنفيذ branch isolated. → اختبار external anchors. → استخدام verified residue في next cycle. → replicate. → canary/rollback test. → promote.
تنفذ الآلية كـfeature قابل للتعطيل داخل MOUSSAIXT. نسجل state-pre وauthority وevidence وbudget؛ وبعد التنفيذ delta ثم expected delta readback/observed. الفرق غير المفسر يصبح ResidualEvent أو UnknownState، لا retry صامتاً.
الآلية
اقتراح mutation من failure portfolio.
تسجيل causal hypothesis.
تنفيذ branch isolated.
اختبار external anchors.
replicate.
canary/rollback test.
promote.
استخدام verified residue في next cycle.
Input
typed state + provenance + authority + budget + owner variables.
Output
proposal/decision + uncertainty + proof obligations + rollback target + telemetry.
Fail-closed: غياب observation/authority/evidence يخفض الحرجة autonomy أو يضع HOLD.
1.
2.
3.
4.
5.
6.
7.
8.

<PARSED TEXT FOR PAGE: 33 / 40>
M14-25 TOOLS - والتدريب الأدواتو MUG placement
الأداة/البروتوكول الوظيفة
Recursive Improvement Closure Engine يدير دورة mutation إلى evidence.
Ledger Lineage Mutation يربط كل تغيير بسبب/نتيجة.
External Grounding Gate يفرض held-out مستقل.
Improvement Rollback Orchestrator يعيد branch عند regression.
التدريب/التطوير
research-loop evaluation design
mutation attribution and rollback training
provider في Adapter؛ weights بعد فقط transfer/retention/rollback. لا يسمى التطوير Fine-Tuning إذا لم تتغير weights.

قاعدة MUG: definitions/provenance في HyperKB؛ procedure في HyperSkill؛ recurrence في HyperLoop؛ semantics
Metrics
external held-out gain · replication sign · rollback success · evaluator independence · cycle-to-cycle retention
أصل كل Specification يوجد حتى implementation ID + test ID + receipt.
•
•

<PARSED TEXT FOR PAGE: 34 / 40>
التنبؤ والهدم والتجربة - FALSIFY 14-25M
التنبؤات
progress أبطأ لكنه أكثر صدقاً .
يقل evaluator overfit.
تتحسن transfer of self-generated improvements.
التجارب الفاصلة
automated alignment/reasoning loop 5 cycles with hidden renewed evals.
compare candidate selection with/without evaluator independence and rollback.
Falsifiers
.hillclimb eval-same لا فرق عن
.لا تتكرر gains external
.cycles ينهار مع provenance أو rollback
Purpose Mechanism Run
candidate effect under frozen model/tools/context/budget ON A
nearest causal baseline OFF B
same extra compute/latency without mechanism semantics SHAM C
directionality / metric-gaming test PERTURBED D
•
•
•
1.
2.
•
•
•

<PARSED TEXT FOR PAGE: 35 / 40>
الإغلاق النظري ومسار الترقية - CLOSURE 14-25M
C0-C9
C4 traces، C5 causal ablation، C6 evidence package،، في تؤثر أن يمكن فقط وبعدها C8 evolution وC9 canonical promotion.
تبدأ M14-25 في C1 Formal Closure، في ترجمُ ت C2 Mechanism، تدخل C3 Kernel كfeature، مستقل ،تحتاج ثم C7 transfer
إذا نجحت
مع replication، ثم out-held external على gain causal، ثم mechanism إلى mutation تتبع التحسن العودي مغلق فقط إذا أمكن
regression-non وrollback، ثم استخدام النتيجة لتوليد mutation لاحقة دون تغيير قواعد الإثبات بصمت.
إذا فشلت
إذا baseline أبسط يعيد الأثر تُدمج؛ إذا metric لا يتنبأ failure تُعاد صياغتها؛ إذا regressions authority/safety غير قابلة للإصلاح ت
رُفض أو تبقى only-research.
ClaimStrength <= EvidenceStrength * Independence * Reproducibility * Transfer
Negative results تدخل evidence ledger ولا تمحى.

<PARSED TEXT FOR PAGE: 36 / 40>
السجل التشغيلي الكامل - 20 أصلاً ARTIFACTS
Purpose Artifact Class Owner ID
Evaluator Independence Graph TOOL M14-21 M14-21-A1 يرسم shared roots.
Shared-Root Detector METHOD M14-21 M14-21-A2 يكشف model/data/harness overlap.

Evaluation Debt Ledger COURT M14-21 M14-21-A3 يسجل claims ذات استقلال منخفض.
Independent Anchor Generator PROTOCOL M14-21 M14-21-A4 يقترح metric/evaluator مختلفا فعلاً .ً
Hypothesis Portfolio Entropy Meter TOOL M14-22 M14-22-A1 يقيس diversity الفعالة.
Diversity Collapse Alarm METHOD M14-22 M14-22-A2 من يحذر monoculture قبل evidence.
Evidence-Weighted Allocation Engine COURT M14-22 M14-22-A3 يوزع probability/budget.
Theory Merger Planner PROTOCOL M14-22 M14-22-A4 يدمج hypotheses تنبؤيا المتكافئة .ً
Negative Result Capitalizer TOOL M14-23 M14-23-A1 يحول failure إلى asset typed.
Failure Asset Index METHOD M14-23 M14-23-A2 حسب النتائج يفهرس mechanism/failure.
Anti-Rediscovery Cache COURT M14-23 M14-23-A3 يستدعي failures قبل theory proposal.
Negative Transfer Router PROTOCOL M14-23 M14-23-A4 ينقل counterexamples إلى domains مشابهة.
Adversarial Curriculum Frontier TOOL M14-24 M14-24-A1 حدود يراقب threat families.
Threat Family Generator METHOD M14-24 M14-24-A2 يقترح mechanisms جديدة هجومية.
Coverage Expansion Meter COURT M14-24 M14-24-A3 يقيس novelty/coverage لا count.
Red-Team Overfit Court PROTOCOL M14-24 M14-24-A4 يمنع saturation masquerading as progress.
Recursive Improvement Closure Engine TOOL M14-25 M14-25-A1 دورة يدير mutation إلى evidence.
Mutation Lineage Ledger METHOD M14-25 M14-25-A2 نتيجة/بسبب تغيير كل يربط.
External Grounding Gate COURT M14-25 M14-25-A3 يفرض held-out مستقل.
Improvement Rollback Orchestrator PROTOCOL M14-25 M14-25-A4 يعيد branch عند regression.

المصفوفة التجريبية - 10 تجارب EXPERIMENTS
Required Design Owner ID
frozen baseline + matched budget + seeds + effect size/CI + guardrail deltas .rule/human/other-model evaluators + same-model judges M14-21 M14-21-E1 مقابل
replay
shrinkage بعد frozen baseline + matched budget + seeds + effect size/CI + guardrail deltas .hidden independent eval + automated research loops M14-21 M14-21-E2 واختبار
replay
frozen baseline + matched budget + seeds + effect size/CI + guardrail deltas .simulated theory portfolio with hidden + true mechanisms M14-22 M14-22-E1
replay
وتخصيص compute تحت frozen baseline + matched budget + seeds + effect size/CI + guardrail deltas .fixed budget + MUT candidates 20 M14-22 M14-22-E2
replay
frozen baseline + matched budget + seeds + effect size/CI + guardrail deltas .research agent proposing theories with/ + without negative memory M14-23 M14-23-E1
replay
frozen baseline + matched budget + seeds + effect size/CI + guardrail deltas .measure duplicate experiment rate over + 100 proposal cycles M14-23 M14-23-E2
replay
frozen baseline + matched budget + seeds + effect size/CI + guardrail deltas .prompt injection/tool authority/code + sabotage families M14-24 M14-24-E1
replay
frozen baseline + matched budget + seeds + effect size/CI + guardrail deltas .compare fixed-set self-play vs expansion + curriculum M14-24 M14-24-E2
replay
frozen baseline + matched budget + seeds + effect size/CI + guardrail deltas .automated alignment/reasoning loop 5 + cycles with hidden renewed evals M14-25 M14-25-E1
replay
frozen baseline + matched budget + seeds + effect size/CI + guardrail deltas .compare candidate selection with/ + without evaluator independence and rollback M14-25 M14-25-E2
replay

<PARSED TEXT FOR PAGE: 38 / 40>
خارطة التنفيذ والانتقاء ROADMAP
Exit gate Build Phase
trace + feature flag + replay Implement M14-21, M14-22 P0
matched-budget ablation Add remaining theory mechanisms P1
stable sign + no critical regression cross-domain/model/tool transfer P2
external evaluator + rollback connect to evolution lab P3
.Build less; test harder; preserve every failure

<PARSED TEXT FOR PAGE: 39 / 40>
المراجع العلمية الخارجية الأساسية REFERENCES
Role Source ID
process supervision and trajectory-level reward Zheng et al. 2025 - Process Reward Model Survey -.
arXiv:2510.08049 E10
automated alignment research and held-out generalization Yueh-Han et al. 2026 - Automated Researchers Can.
Reliably Mitigate Alignment Failures - arXiv:2608.28945 E11
automated red teaming, self-play, held-out attacks, robustness OpenAI 2026 - GPT-Red E12.
LLM-guided evolutionary discovery with evaluator-mediated search Google DeepMind 2026 - AlphaEvolve impact.
update E15
يُعاد البحث الأدبي قبل claim أسبقية أو submission خارجي. كل result benchmark خارجي result reported لا replication داخلية.

<PARSED TEXT FOR PAGE: 40 / 40>
الحكم البحثي وحدود الورقة VERDICT
الخلاصة: تعالج الورقة الحلقة الأخطر والأكثر قيمة: أن يستخدم النظام نفسه لتوليد فرضيات وتجارب وتدريب وخصوم وتحسينات، مع منع capture
evaluator والـconfirmation-self. تبني مقاييس استقلال المقيم، entropy لمحفظة النظريات، رأس مال النتائج السلبية،
توسع المنهج العدائي، وإغلاق التحسن العودي.
القيمة العلمية ليست زيادة ontology، بل hypotheses يمكن رفضها. إذا لم تحسن التشخيص أو الإسناد أو recovery أو transfer أو evolution
مقارنة بالـbaselines الأبسط، يجب دمجها أو إسقاطها. grounded-external
PROMOTE only after mechanism + trace + causal gain + independent evidence + transfer
Paper status: N0-N2 theoretical closure specification. No empirical AGI/ASI claim.