أطلس هندسة الجبهة الوكيلية للذكاء العام

AGI / ASI · Agentic OS · HyperHarness · Kernel Engineering

40 ورقة بحثية جديدة - قراءة هندسية مرتبطة بـ MOUSSAIF AGI OS / MOUSSAIFXT

ARABIC RTL · AMIRI · 50 PAGES · 40 NEW PAPERS · MEMORY OS · AGENT OS · HARNESSES · SELF-EVOLUTION · VERIFIERS · FRONTIER MODELS

عدسة الربط
كل ورقة تُقرأ مقابل مكدس L0-L11: Constitution/Goal/Question, MCD, QMCD, HKOS/World Model, Retrieval, HyperSkills, HyperAbilities, HyperLoops, Agents/Router, Tools/Runtime, Verification/World-Effect, Memory/Evolution.

هدف هذا الجزء
ليس تكرار corpus السابق. الاختيار هنا مبني على delta هندسي: ما الذي تضيفه أحدث أبحاث الذاكرة الوكيلية، long-horizon execution، harness، AgentOS، verifiers، safety، self-evolution، computer-use، والنماذج agentic إلى بنية تشغيل AGI.

المخرج العملي
40 capsules كثيفة + All-vs-MOUSSAIF map + HyperHarness reference + 20 experiments + promotion gates. الهدف تحويل literature projection إلى contracts وtests وruntime deltas.

انضباط الادعاء
الورقة الخارجية = Evidence Root. الربط مع MOUSSAIF = Engineering Synthesis. لا يتحول التشابه إلى إثبات أسبقية أو AGI. كل adoption مقترح يجب أن يمر عبر implementation, ablation, falsification, replication.

قاعدة الكتاب: النموذج frontier مهم، لكن قدرة الوكيل طويلة الأفق تتحدد كذلك بالـauthority, verification, tool semantics, harness, skills, memory, state وrecovery. لذلك هذا الجزء يركز على الطبقات التي تقع حول النموذج وبين النموذج والعالم.

لماذا هذه الأربعون؟ - Selection by Architectural Delta

تم استبعاد الجذور الـ45 التي غطاها الجزء السابق، ثم رُفعت أولوية الأعمال التي تغير سؤالاً تنفيذياً داخل Agentic OS: Where is canonical state? Who has authority? How are skills packaged? How is a long mission resumed? How is world effect verified? Can the harness evolve without corrupting itself? How does model/test-time compute get routed?

Frontier zone	Count	ما الذي نريد استخراجه؟
Memory / Skill OS	papers 5	memory lifecycle, learned memory actions, skill certification
Harness / Agent OS	papers 5	code-as-harness, self-optimizing harness, long-horizon state, AOS/AgentOS
Runtime / Computer Use	papers 9	desktop, multi-agent, protocols, GUI/CLI/browser, environment standardization
Reliability / Safety	papers 4	world-effect, dual control, prompt injection, interactive safety
Self-Evolution / Coding	papers 8	automated harness design, self-modification, skill RL, runtime platforms, CLI eval
Reasoning / Model Substrate	papers 9	compute routing, search, latent reasoning, verifiers, auditability, agentic backbones

معيار الحداثة

- أولوية 2025-2026، مع عدد محدود من جذور 2024 التي ما تزال infrastructure canonical.
- أضيفت أعمال أغسطس 2026 في Long-Horizon Harness وAgent Operating Systems.
- تُعرض preprints الحديثة بوسمها ولا تساوى تلقائياً بـpeer-reviewed results.
- Research cutoff: 31 August 2026.

معيار القوة

- Primary paper / accepted venue / major lab technical report عند الإمكان.
- آلية قابلة للوصف والتفكيك، لا مجرد opinion.
- صلة مباشرة بـagent, tool, runtime, memory, skill, verifier, state, evolution أو compute.
- وجود benchmark أو implementation أو testable architecture يرفع الرتبة.

القاعدة المنهجية: لا نبحث عن 40 اسماً جديداً؛ نبحث عن 40 mechanism deltas. إذا لم تغيّر الورقة state representation أو execution loop أو authority أو routing أو skill/memory lifecycle أو verifier/evaluation، فليست ذات أولوية لهذا الجزء.

External Evidence to MOUSSAIF Layers - خريطة الارتباط بالمعمارية الحالية

هذه ليست محاولة لإعادة تسمية الأدبيات. الأرقام أدناه تشير إلى الأوراق التي تقدم mechanism أو benchmark أو architecture يمكن استعمالها لاختبار أو تعديل الطبقة المقابلة. الورقة قد ترتبط بعدة طبقات لأن agentic reliability ظاهرة عابرة للstack.

L0 Constitution / Safety

authority, risk, non-bypassable policy

Papers: 23, 22, 28, 9

L1 Goal / Governing Question

intent, task creation, goal evolution

Papers: 10, 26, 28

L2 MCD / QMCD / Routing

compute, specialists, search, reasoning mode

Papers: 4, 14, 15, 32, 33, 34, 37, 39, 40

L3 HKOS / World Model

canonical memory and knowledge objects

Papers: 1, 2, 3, 10

L4 Retrieval / GraphRAG

dynamic links, source routing, context selection

Papers: 1, 3, 22, 33, 40

L5 HyperSkills

procedural units, learned memory/tool skills

Papers: 4, 5, 6, 10, 27

L6 HyperAbilities

composition of reusable capabilities

Papers: 5, 14, 27

L7 HyperLoops / Workflows

MEA, branches, harness workflows, recovery

Papers: 6, 7, 8, 12, 15, 24, 27

L8 Agents / Router

orchestration, models, role routing

Papers: 9, 10, 11, 12, 13, 14, 16, 17, 29, 30, 37, 38, 39

L9 Tools / Runtime

OS/GUI/CLI/web/tool execution contracts

Papers: 6, 9, 11, 13, 16, 17, 18, 19, 20, 22, 23, 29, 30, 31, 38, 39

L10 Verification / World Effect

auditors, tests, safety, end-state evidence

Papers: 7, 8, 9, 11, 12, 15, 19, 20, 21, 22, 23, 25, 31, 34, 35, 36

L11 Memory / Evolution / Trace

persistent state, trajectory learning, promotion

Papers: 1, 2, 3, 4, 5, 7, 8, 9, 10, 16, 18, 24, 25, 26, 27, 31, 36

External Paper -> Mechanism Extraction -> MOUSSAIF Contract Delta -> Implementation -> Ablation -> Falsification -> Promotion

أقوى تقاطع في هذا الجزء: L7-L11. الجبهة الحديثة تتحرك بسرعة من "prompt + tool call" إلى durable task state, reusable skills, runtime isolation, explicit authority, independent audit, harness optimization, trajectory-derived evolution. هذه هي المنطقة الأقرب مباشرة إلى MOUSSAIFXT وHyperHarness.

الفهرس الكامل - Research Roots 40

كل paper capsule يحتفظ بالفصل بين: **ما تقوله الورقة**، **الاستخراج الهندسي**، و**MOUSSAIF bridge**. آخر ست صفحات تحول الأربعين إلى invariants, architecture deltas وتجارب.

Memory, Skills & HyperHarness

01	Memory OS of AI Agent
02	MemOS: An Operating System for Memory-Augmented Generation (MAG) in Large Language Models
03	A-MEM: Agentic Memory for LLM Agents
04	Agentic Memory: Learning Unified Long-Term and Short-Term Memory Management for Large Language Model Agents
05	SoK: Agentic Skills - Beyond Tool Use in LLM Agents
06	Code as Agent Harness
07	Evolving Agents in the Dark: Retrospective Harness Optimization via Self-Preference
08	LongHorizon-Harness: Advancing Long-Horizon Agents for Real-World Tasks
09	The Agent Operating System (AOS): A Reference Operating Architecture for Distributed Agentic Systems
10	AgentOS: From Application Silos to a Natural Language-Driven Data Ecosystem

AgentOS, Runtime & Computer Use

11	UFO2: The Desktop AgentOS
12	Magentic-One: A Generalist Multi-Agent System for Solving Complex Tasks
13	A Survey of Agent Interoperability Protocols: MCP, ACP, A2A, and ANP
14	Agent S2: A Compositional Generalist-Specialist Framework for Computer Use Agents
15	The Unreasonable Effectiveness of Scaling Agents for Computer Use
16	OpenCUA: Open Foundations for Computer-Use Agents
17	UI-TARS-2 Technical Report: Advancing GUI Agent with Multi-Turn Reinforcement Learning
18	AgentGym: Evolving Large Language Model-based Agents across Diverse Environments
19	The BrowserGym Ecosystem for Web Agent Research

World-Effect Reliability & Safety

20	tau-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains
21	tau^2-Bench: Evaluating Conversational Agents in a Dual-Control Environment
22	AgentDojo: A Dynamic Environment to Evaluate Prompt Injection Attacks and Defenses for LLM Agents
23	Agent-SafetyBench: Evaluating the Safety of LLM Agents

Self-Evolution, Coding & Runtime Learning

24	Automated Design of Agentic Systems
25	Darwin Godel Machine: Open-Ended Evolution of Self-Improving Agents
26	AgentEvolver: Towards Efficient Self-Evolving Agent System
27	Reinforcement Learning for Self-Improving Agent with Skill Library (SAGE)
28	Self-Evolving Software Agents
29	OpenHands: An Open Platform for AI Software Developers as Generalist Agents
30	Qwen3-Coder-Next Technical Report
31	Terminal-Bench 2.0: Benchmarking Agents on Hard, Realistic Tasks in Command Line Interfaces

Reasoning, Verification & Model Substrates

32	s1: Simple Test-Time Scaling
33	Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning
34	Training Large Language Models to Reason in a Continuous Latent Space (Coconut)
35	Process Reward Models That Think (ThinkPRM)
36	Reasoning Models Don't Always Say What They Think
37	Qwen3 Technical Report
38	Kimi K2: Open Agentic Intelligence
39	GLM-4.5: Agentic, Reasoning, and Coding (ARC) Foundation Models
40	MiniMax-M1: Scaling Test-Time Compute Efficiently with Lightning Attention

Not duplication: هذا الجزء لا يعيد AIXI, Transformer, RAG, ReAct, Reflexion, MemGPT, AutoGen, Voyager, OSWorld, SWE-bench, DeepSeek-R1, consciousness roots أو الـ15 frontier roots السابقة. هذه الأربعون طبقة تكميلية أحدث وأكثر قرباً من Agent OS / Harness / Kernel engineering.

01

NEW PART II | Memory OS | arXiv / open-source

Memory OS of AI Agent

Jiazheng Kang, Mingming Ji, Zhe Zhao, Ting Bai · 2025

Scientific contribution / المساهمة العلمية

يحوّل الذاكرة الوكيلية من buffer سياقي إلى نظام هرمي مستوحى من Operating Systems. يقسم الحالة إلى Short-Term, Mid-Term, وLong-Term Personal Memory، ويجعل التخزين والتحديث والاسترجاع والتوليد مراحل مستقلة يمكن قياسها. القيمة ليست في زيادة context بل في lifecycle واضح للذاكرة طويلة الأفق.

Mechanism / الآلية

تعمل أربع وحدات Memory Storage / Updating / Retrieval / Generation فوق ثلاث طبقات زمنية، تنتقل الذاكرة من STM إلى MTM ثم إلى LTM بسياسات تنظيم وتلخيص، بدلاً من إبقاء كامل التاريخ في prompt واحد.

TECHNICAL SIGNATURE

```
M_t={STM,MTM,LTM}; update: STM -> MTM -> LTM; answer=G(query,R(M_t))
```

Engineering extraction / الاستخراج الهندسي

- فصل memory lifecycle عن model context يقلل تضخم السياق.
- التحديث يجب أن يكون policy قابلة للتدقيق لا append بلا حدود.
- يجب قياس stale-memory وirrelevant-recall إضافة إلى recall accuracy.

Limits / الحدود

- الدليل الأساسي محادثي/شخصي أكثر من كونه world-state عام.
- لا يحل وحده تعارض الذكريات أو poisoning أو authorization.

MOUSSAIF relevance

مرجع مباشر لـHyperMemory وSelf-Reconstruction Memory: الذاكرة في AGI OS تصبح state service لا سجل محادثة.

Research variables

hierarchical memory | lifecycle | personal memory | retrieval

Primary source
https://arxiv.org/abs/2506.06326

Ablation / Reproduction بروتوكول

قارن current memory مع hierarchical lifecycle تحت نفس token budget وعلى مهام طويلة؛ حقن facts قديمة ومتعارضة واختبر تصحيحها.

Measure	task success, recall precision, stale recall, tokens/step
Failure signal	ارتفاع success بسبب context أكبر فقط؛ أو بقاء معلومات منسوخة غير قابلة للإبطال.

MOUSSAIF Contract Delta

Primary bridge: L3/L4/L11

افصل Memory Bus عن context؛ عرّف promotion/eviction policy وإثبات provenance لكل memory object.

02

NEW PART II | Memory OS | arXiv

MemOS: An Operating System for Memory-Augmented Generation (MAG) in Large Language Models

Zhiyu Li et al. · 2025

Scientific contribution / المساهمة العلمية

يرفع الذاكرة إلى first-class operational resource ويجمع Parametric, Activation, وPlaintext Memory ضمن abstraction موحد. MemCube يجعل memory object قابلاً للتعقب، migration، fusion، والحوكمة عبر مهام وسياقات متعددة، وهو قريب جداً من تصور Knowledge/Memory as OS.

Mechanism / الآلية

Memory-centric execution يدير representation, organization, governance عبر أنواع ذاكرة مختلفة. MemCube يعمل كوعاء ذاكرة موحد يحمل المحتوى والميتاداتا والحالة ويتيح نقلها أو دمجها دون اختزالها في retrieval text.

TECHNICAL SIGNATURE

```
MemCube=(content,type,state,provenance,policy); Exec = route(task, MemCube*)
```

Engineering extraction / الاستخراج الهندسي

- الذاكرة متعددة الأنواع تحتاج schema موحداً لا واجهات منفصلة.
- migration/fusion يجب أن يحفظ lineage والسلطة.
- runtime يمكنه اختيار نوع الذاكرة وفق المهمة والكلفة.

Limits / الحدود

- إدارة parametric memory أصعب من plaintext ولا تعني تعديل weights لحظياً بأمان.
- المعماريات المعيارية تحتاج benchmarks أطول وأثر سلامة واضح.

MOUSSAIF relevance

يدعم HKOS/HyperMemory كطبقة تشغيل، ويعطي قالباً عملياً لـMemory Object Contracts داخل MOUSSAIFXT.

Research variables

MemCube | memory types | migration | governance

Primary source
https://arxiv.org/abs/2505.22101

Ablation / Reproduction بروتوكول

نفّذ memory objects موحدة مقابل stores منفصلة، ثم اختبر migration بين runtimes دون تغيير قرار المهمة.

Measure	migration fidelity, retrieval quality, lineage preservation
Failure signal	تحسن retrieval مع فقد provenance أو تغير semantics عند النقل.

MOUSSAIF Contract Delta

Primary bridge: L3/L11

استحداث MemoryObject موحد: type, authority, provenance, TTL, revision, retrieval-policy, writeback gate.

03

NEW PART II | Agentic Memory | NeurIPS 2025

A-MEM: Agentic Memory for LLM Agents

Wujiang Xu et al. · 2025

Scientific contribution / المساهمة العلمية

يقترح ذاكرة Agentic تنتظم ديناميكياً وفق مبادئ Zettelkasten بدل structure ثابتة. كل memory تتحول إلى note غنية بالوصف والكلمات المفتاحية والعلاقات، ثم تُربط بذكريات سابقة لتنشأ knowledge network متطورة مع الخبرة.

Mechanism / الآلية

عند الإدخال، يستنتج النظام attributes وروابط؛ لاحقاً يمكن للروابط والـindices أن تتغير مع وصول خبرة جديدة. بذلك يصبح memory graph كياناً يتطور وليس vector store ساكناً.

TECHNICAL SIGNATURE

```
note_i=(text,context,tags,links); Graph_t+1 = evolve(Graph_t,note_i)
```

Engineering extraction / الاستخراج الهندسي

- الرابط بين memories جزء من inference لا metadata ثابتة.
- الاسترجاع الجيد يعتمد على topology الذاكرة لا similarity وحدها.
- evolution يحتاج anti-duplication وconflict handling.

Limits / الحدود

- الروابط المولدة قد تنشر خطأ أو تضخم علاقة وهمية.
- يتطلب تقييم جودة graph وليس QA فقط.

MOUSSAIF relevance

صلة مباشرة بـGraphRAG, Knowledge Vessels, ContinuityScanner وZero-Duplication داخل HKOS.

Research variables

memory graph | dynamic linking | Zettelkasten | evolution

Primary source
https://arxiv.org/abs/2502.12110

Ablation / Reproduction بروتوكول

Ablate dynamic linking مع vector-only retrieval؛ أدخل memories متعارضة وراقب propagation.

Measure	task success, graph precision, contradiction recovery
Failure signal	graph densification بلا gain أو انتشار hallucinated edges.

MOUSSAIF Contract Delta

Primary bridge: L3/L4/L11

اجعل Knowledge Vessel يولّد links مرشحة مع confidence ثم تمر عبر Evidence/Contradiction Court قبل promotion.

04

NEW PART II | Learned Memory Policy | arXiv 2026

Agentic Memory: Learning Unified Long-Term and Short-Term Memory Management for Large Language Model Agents

Yi Yu et al. · 2026

Scientific contribution / المساهمة العلمية

AgeMem يجعل عمليات الذاكرة actions داخل policy نفسها بدلاً من heuristics خارجية. الوكيل يقرر متى Store, Retrieve, Update, Summarize, أو Discard، ثم يتعلم هذه السياسة عبر progressive RL وstep-wise GRPO في مهام طويلة الأفق.

Mechanism / الآلية

تحويل memory operations إلى tools يخلق action space صريحاً، بينما training متعدد المراحل يتعامل مع sparse/discontinuous rewards الناتجة عن قرارات ذاكرة لا يظهر أثرها إلا لاحقاً.

TECHNICAL SIGNATURE

```
a_t in {store,retrieve,update,summarize,discard,act}; pi_mem <- RL(long-horizon reward)
```

Engineering extraction / الاستخراج الهندسي

- memory routing يمكن تعلمه بدلاً من hard-coded policy.
- discard قرار معرفي مهم مثل retrieve.
- reward يجب أن ينسب أثر memory decision عبر الزمن.

Limits / الحدود

- RL قد يتعلم shortcuts مرتبطة بالbenchmark.
- الكتابة/الحذف تحتاج governance أعلى من reward الخام.

MOUSSAIF relevance

يدعم تحويل HyperMemory operations إلى HyperSkills قابلة للتعلم، مع إبقاء Constitutional gates خارج policy.

Research variables

memory actions | GRPO | long horizon | context efficiency

Primary source
https://arxiv.org/abs/2601.01885

Ablation / Reproduction بروتوكول

قارن learned memory policy مع rules ثابتة عبر tasks جديدة وبميزانية context متساوية.

Measure	success, context tokens, write/read ratio, memory quality
Failure signal	حاسم أو سياسة تنجح فقط على توزيع التدريب حذف evidence.

MOUSSAIF Contract Delta

Primary bridge: L2/L5/L11

عرّف memory HyperSkills ثم اسمح لRouter بتعلم اختيارها ضمن حدود authorization وNo-Loss Court.

05

NEW PART II | Skill OS | Systematization of Knowledge

SoK: Agentic Skills - Beyond Tool Use in LLM Agents

Yanna Jiang et al. · 2026

Scientific contribution / المساهمة العلمية

يعامل Agentic Skills كطبقة إجرائية مستقلة عن tool calls والخطط المؤقتة. يرسم lifecycle كاملاً: discovery, practice, distillation, storage, composition, evaluation, update، ويصنف representation x scope مع تحليل صريح لمخاطر skill supply chain.

Mechanism / الآلية

الskill وحدة callable تحمل applicability conditions, execution policy, termination criteria وinterface. هذا يحول procedural knowledge إلى artifact قابل لإعادة الاستخدام والتركيب، لكنه يفرض trust tiers واختباراً حتمياً قبل التنفيذ.

TECHNICAL SIGNATURE

```
Skill=(preconditions,policy,tools,stop,tests,trust,version)
```

Engineering extraction / الاستخراج الهندسي

- الميتاداتا والحدود جزء من skill وليست documentation خارجية.
- self-generated skills قد تكون أسوأ من curated skills؛ promotion gate ضروري.
- skill marketplaces تخلق supply-chain attack surface.

Limits / الحدود

- SoK تجمع أدبيات ولا تثبت معياراً واحداً نهائياً للمهارة.
- قابلية النقل بين hosts ما تزال غير مستقرة.

MOUSSAIF relevance

أقرب مرجع خارجي لHyperSkill / HyperAbility / HyperSkill Factory وKnowledge-Integrity AutoDeployment.

Research variables

skill lifecycle | composition | trust tier | certification

Primary source
https://arxiv.org/abs/2602.20867

Ablation / Reproduction بروتوكول

اختبر curated مقابل self-generated skills مع sandbox وmalicious skill set؛ قس النقل بين بيئات مختلفة.

Measure	task gain, reuse, failure rate, security violations
Failure signal	skill يزيد النجاح لكنه يرفع side-effect أو ينهار بعد provider swap.

MOUSSAIF Contract Delta

Primary bridge: L5/L6/L7/L11

وسّع HyperSkill contract ليشمل preconditions, stop, tests, trust-tier, provenance, signature, rollback, compatibility.

06

NEW PART II | HyperHarness | arXiv survey 2026

Code as Agent Harness

Xuying Ning et al. · 2026

Scientific contribution / المساهمة العلمية

يصوغ Code as Agent Harness: الكود ليس output فقط بل substrate للتفكير والفعل، verification، environment modeling، shared state والتنسيق. يقسم harness إلى interface, mechanisms, ثم multi-agent scaling، ويركز على executable, verifiable, stateful agency.

Mechanism / الآلية

الكود يربط agent بأدوات وبيئات حقيقية، بينما planning/memory/tool-use وfeedback loops تشكل control layer. في multi-agent settings تصبح artifacts البرمجية وسيطاً مشتركاً للمراجعة والتنسيق والتحقق.

TECHNICAL SIGNATURE

```
Harness = Interface + State + Tools + Code + Feedback + Verification
```

Engineering extraction / الاستخراج الهندسي

- الـharness نفسه object هندسي يحتاج versioning واختبارات.
- final-task success غير كاف؛ يلزم trace/effect verification.
- shared executable artifacts تقلل غموض التنسيق بين الوكلاء.

Limits / الحدود

- Survey/reference view أكثر من algorithm واحد قابل للمقارنة.
- code-centric harness لا يغطي كل البيئات الحسية.

MOUSSAIF relevance

يطابق مفهوم HyperHarness: طبقة executable خارج model توحّد HyperSkills/Loops/Tools/State/Courts.

Research variables

harness | code substrate | state | verification

Primary source
https://arxiv.org/abs/2605.18747

Ablation / Reproduction بروتوكول

ثبت model واستبدل harness وحده؛ قارن minimal, current, وcode-first harness على coding/GUI/knowledge tasks.

Measure	success, regressions, verified effects, portability
Failure signal	gain يختفي عند model swap أو يعتمد على hidden manual intervention.

MOUSSAIF Contract Delta

Primary bridge: L5-L10

عرّف HyperHarness manifest مستقل عن model: registry, context, state, tools, hooks, verifier, rollback, benchmark profile.

07

NEW PART II | Harness Optimization | Microsoft Research / arXiv

Evolving Agents in the Dark: Retrospective Harness Optimization via Self-Preference

Wenbo Pan et al. · 2026

المساهمة العلمية / Scientific contribution

RHO يعامل harness كشيء قابل للتحسين من trajectories السابقة دون ground-truth validation set. يختار حالات صعبة، يعيد حلها parallel، يبني candidate harness updates، ثم يستخدم self-validation/self-consistency وpairwise self-preference لاختيار التعديل.

الآلية / Mechanism

التعلم لا يعدل weights بالضرورة؛ يعدل skills/tools/workflows المحيطة بالنموذج. بذلك يفصل adaptation السريع للمنظومة عن expensive model training ويجعل trajectory archive مادة مباشرة لتحسين runtime.

TECHNICAL SIGNATURE

```
H_{t+1}=argmax_{H in candidates(H_t,Traj)} SelfPreference(H)
```

الاستخراج الهندسي / Engineering extraction

- failure trajectories أصول تدريبية للharness.
- يجب تقييم harness delta ضد regressions لا على task واحد.
- self-preference تحتاج independent court كي لا تصبح self-confirmation.

الحدود / Limits

- self-evaluation قد يرسخ أخطاء النظام نفسه.
- مكاسب benchmark لا تثبت safe open-ended self-improvement.

MOUSSAIF relevance

صلة مباشرة بـEvolution Laboratory + HyperSkill Factory + Failure Memory + promotion gates.

Research variables

trajectory coreset | harness delta | self-preference | regression

Primary source
https://arxiv.org/abs/2606.05922

بروتوكول Ablation / Reproduction

نفّذ RHO-like loop مع Falsification Court مستقل وfrozen holdout؛ امنع الترقية إن حدث regression حرج.

Measure	delta gain, regression count, self-eval calibration
Failure signal	self-preference يختار تغييراً يفشل على holdout أو يخفض safety.

MOUSSAIF Contract Delta

Primary bridge: L7/L10/L11

أضف HarnessDelta registry: parent_version, changed skill/tool/loop, evidence set, regression suite, rollback pointer.

08

NEW PART II | Long-Horizon Harness | arXiv Aug 2026

LongHorizon-Harness: Advancing Long-Horizon Agents for Real-World Tasks

Ziyu Ma et al. · 2026

Scientific contribution / المساهمة العلمية

يعيد صياغة long-horizon agency كمشكلة Task-State Management. بدلاً من إبقاء التنفيذ والحالة وتقييم الإنجاز داخل context متنامٍ، يحفظ state خارجياً ولا يحدّثه إلا facts تحققت independently من البيئة.

Mechanism / الآلية

Manage-Execute-Audit loop: Manager يحفظ task state ويختار subtask، Executor يعمل بسياق fresh، Auditor read-only يتحقق من environment قبل commit. AgentAdapter يسمح بتبديل model/harness دون تغيير loop الأصلي.

TECHNICAL SIGNATURE

```
State_{t+1}=Commit(State_t, Audit(EnvironmentEffect(Execute(Subtask_t))))
```

Engineering extraction / الاستخراج الهندسي

- verified external state أهم من self-reported completion.
- fresh-context execution يمنع context debt من التراكم.
- manager/executor/auditor فصل وظيفي يتيح tracing والrollback.

Limits / الحدود

- نتائج حديثة تحتاج تكراراً مستقلاً أوسع.
- auditor نفسه قد يصبح bottleneck أو مصدر خطأ.

MOUSSAIF relevance

تقاطع مباشر مع Evidence Court + M-TAOS + World-Effect Loop + Self-Reconstruction Memory.

Research variables

MEA loop | verified state | fresh context | auditor

Primary source
https://arxiv.org/abs/2608.01964

Ablation / Reproduction بروتوكول

ثبّت model/executor وقارن current loop مع MEA على tasks >30 steps؛ أدخل false-success traps.

Measure	completion, verified state accuracy, token cost, recovery
Failure signal	manager أو auditor يوافق على state غير مطابق للبيئة يكرر.

MOUSSAIF Contract Delta

Primary bridge: L7/L8/L10/L11

حوّل M-TAOS state writeback إلى <- two-phase commit: tentative effect -> independent audit -> canonical state commit.

09

NEW PART II | Agent OS Architecture | arXiv Aug 2026

The Agent Operating System (AOS): A Reference Operating Architecture for Distributed Agentic Systems

Ankur Sharma, Deep Shah · 2026

Scientific contribution / المساهمة العلمية

يقترح reference architecture مستقلاً عن vendor للأنظمة الوكيلية الموزعة. يفصل Control & Governance Plane عن Runtime & Coordination Plane، ويضع intent, policy, trust, authority, auditability, lifecycle, routing, memory, scheduling وruntime assurance داخل boundary تشغيلية واضحة.

Mechanism / الآلية

الOS لا يستبدل Linux/containers/frameworks بل يحكم تركيبها عبر explicit interfaces. السلطة تنتقل مع delegation، بينما coordination plane يدير lifecycle/traffic/context/model-tool routing تحت قيود governance.

TECHNICAL SIGNATURE

```
AOS = Control&Governance Plane + Runtime&Coordination Plane + Explicit Interfaces
```

Engineering extraction / الاستخراج الهندسي

- agent OS يحتاج authority semantics لا orchestration فقط.
- observability وauditability primitives وليستا dashboard لاحقاً.
- boundary بين AgentOS والبنية التحتية يجب أن يكون صريحاً.

Limits / الحدود

- Reference architecture جديدة وليست proof of superior runtime.
- الinterfaces المقترحة تحتاج implementations وbenchmarks مشتركة.

MOUSSAIF relevance

مقارنة مباشرة ومفيدة مع L0-L11: تؤكد أن MOUSSAIFXT يجب أن يعرّف control/runtime planes وعقود delegation.

Research variables

authority | coordination | lifecycle | runtime assurance

Primary source
https://arxiv.org/abs/2608.03214

Ablation / Reproduction بروتوكول

نفّذ نفس multi-agent workflow مع/بدون explicit authority propagation؛ اختبر revoked permissions وpartial failures.

Measure	policy violations, recovery, trace completeness
Failure signal	delegated agent يحتفظ بسلطة بعد revoke أو أثر غير قابل لإعادة البناء.

MOUSSAIF Contract Delta

Primary bridge: L0/L8/L9/L10/L11

أضف AuthorityToken وDelegationChain لكل state objects يرافقان agent/tool call، مع observability correlation IDs.

10

NEW PART II | Personal Agent OS | arXiv 2026

AgentOS: From Application Silos to a Natural Language-Driven Data Ecosystem

Rui Liu et al. · 2026

Scientific contribution / المساهمة العلمية

يقترح Personal Agent Operating System حيث Agent Kernel يفسر intent ويفكك المهام وينسق agents، بينما التطبيقات تتحول إلى Skills-as-Modules. يربط OS الجديد بـKnowledge Discovery/Data Mining: workflow mining, skill retrieval, evolving personal knowledge graphs.

Mechanism / الآلية

Natural User Interface يدخل intent إلى Agent Kernel؛ kernel يختار skills ويكوّن workflow ويحدث knowledge graph. التركيز على permission/context fragmentation في النموذج التقليدي للتطبيقات.

TECHNICAL SIGNATURE

```
Intent -> AgentKernel -> Skill Retrieval/Composition -> Execution -> Personal KG Update
```

Engineering extraction / الاستخراج الهندسي

- skill discovery يمكن أن يعامل كـretrieval/recommender problem.
- kernel يجب أن يحكم permissions/context عبر التطبيقات.
- workflow history يصبح dataset لاستخراج مهارات جديدة.

Limits / الحدود

- Research agenda أكثر من production architecture محققة بالكامل.
- NUI-centric design لا يلغي حاجة APIs/GUI fallback.

MOUSSAIF relevance

يدعم MOUSSAIFXT كـkernel اتحادي وHyperSkills كبديل composable للتطبيقات، مع HKOS كـpersonal knowledge substrate.

Research variables

agent kernel | skills-as-modules | workflow mining | personal KG

Primary source
https://arxiv.org/abs/2603.08938

Ablation / Reproduction بروتوكول

استخرج skills من execution logs ثم اختبر فائدتها على tasks لاحقة مع policy checks.

Measure	skill reuse, planning steps, permission errors
Failure signal	workflow mining بحفظ سلوكاً خاطئاً أو يخلق skill شديدة التخصيص.

MOUSSAIF Contract Delta

Primary bridge: L1/L3/L5/L8/L9

اربط Skill Registry بـusage telemetry وworkflow mining لإنتاج candidate HyperSkills من المسارات المتكررة.

11

NEW PART II | Desktop AgentOS | TMLR May 2026

UFO2: The Desktop AgentOS

Chaoyun Zhang et al. · 2026

Scientific contribution / المساهمة العلمية

يبني AgentOS عملياً لسطح Windows: HostAgent مركزي يفكك المهمة وينسق AppAgents متخصصة، مع native APIs وGUI/API تقلل هشاشة PiP isolated desktop وspeculative multi-action planning وunified action layer. UIA + vision grounding screenshot-only agents.

Mechanism / الآلية

HostAgent -> AppAgent dispatch؛ AppAgents تملك domain knowledge وأدوات native. الـhybrid control detection يدمج UI Automation مع vision، بينما isolated virtual desktop يمنع تعطيل المستخدم ويعزل الأثر.

TECHNICAL SIGNATURE

```
HostAgent(plan) -> AppAgent_i -> {API | UIA | VisionAction} -> verify
```

Engineering extraction / الاستخراج الهندسي

- deep OS integration أقوى من click-only abstraction.
- التخصص على مستوى app يمكن أن يبقى modular.
- execution isolation مكوّن reliability وسلامة لا مجرد UX.

Limits / الحدود

- Windows-centric ومكلف في صيانة app-specific adapters.
- native integration يرفع surface للصلاحيات والأخطاء.

MOUSSAIF relevance

مرجع مباشر لـMOUSSAIXT cross-environment shell وDomain وRuntime Fingerprinting وAdapters.

Research variables

HostAgent | AppAgent | UIA+vision | isolation

Primary source
https://arxiv.org/abs/2504.14603

Ablation / Reproduction بروتوكول

قارن screenshot-only مقابل hybrid API/UIA/vision على نفس apps؛ افصل أثر planner عن grounding.

Measure	success, actions/task, grounding error, recovery
Failure signal	adapter الخاص بالتطبيق يخلق benchmark leakage أو صلاحيات زائدة.

MOUSSAIF Contract Delta

Primary bridge: L8/L9/L10

عرّف Host Adapter capability manifest: native API, UIA, vision, terminal, isolation mode, rollback support.

12

NEW PART II | Multi-Agent Runtime | Microsoft Research

Magentic-One: A Generalist Multi-Agent System for Solving Complex Tasks

Adam Fourney et al. · 2024

المساهمة العلمية / Scientific contribution

يعرض نظاماً متعدد الوكلاء modular: Orchestrator يخطط ويتابع progress ويعيد التخطيط، بينما agents متخصصة تتصفح web، تدير files، وتكتب/تشغل Python. AutoGenBench يضيف repetition/isolation لقياس systems ذات side effects.

الآلية / Mechanism

الـOrchestrator يحتفظ ledger للمهمة ويصدر directives، ثم يراجع التنفيذ ويعيد التخطيط عند failure. التخصص وظيفي وليس مجرد تعدد prompts.

TECHNICAL SIGNATURE

```
Orchestrator: plan -> delegate -> observe -> replan; Specialists: Browser/File/Coder
```

الاستخراج الهندسي / Engineering extraction

- القيمة multi-agent تأتي من role contracts + shared progress state.
- replanning primitive أساسي لا recovery patch.
- benchmark isolation مهم لأن agent actions تغير البيئة.

الحدود / Limits

- عدد agents لا يضمن أداء أفضل وقد يزيد latency/coordination errors.
- shared context قد يصبح bottleneck.

MOUSSAIF relevance

يتطابق مع Agent Mesh & Role Councils + Mission DAG + Executor loop داخل MOUSSAIF.

Research variables

orchestration | role specialization | replanning | isolation

Primary source
https://arxiv.org/abs/2411.04468

بروتوكول Ablation / Reproduction

Ablate orchestrator ledger، specialist roles، وإعادة التخطيط منفرداً؛ اختبر partial tool failure.

Measure	success, replans, coordination tokens, error recovery
Failure signal	multi-agent cost أعلى دون verified gain أو deadlocks متكررة.

MOUSSAIF Contract Delta

Primary bridge: L7/L8/L9/L10

اجعل RoleCouncil agents تملك input/output contract وprogress ledger موحد بدلاً من chat-only coordination.

13

NEW PART II | Interoperability | Survey

A Survey of Agent Interoperability Protocols: MCP, ACP, A2A, and ANP

Abul Ehtesham et al. · 2025

Scientific contribution / المساهمة العلمية

يقارن أربع طبقات interoperability: MCP للوصول إلى tools/data، ACP messaging، A2A delegation عبر capability Agent Cards، وANP discovery/decentralized collaboration. القيمة هي فصل tool protocol عن agent-to-agent protocol وعن network discovery.

Mechanism / الآلية

كل protocol يحل boundary مختلفاً: typed invocation, messages/sessions, capability-based task delegation, أو decentralized identity/discovery. هذا يمنع تحويل بروتوكول واحد إلى حل لكل طبقات النظام.

TECHNICAL SIGNATURE

```
Tool access != Agent delegation != Agent discovery != Workflow semantics
```

Engineering extraction / الاستخراج الهندسي

- contracts يجب أن تكون typed وversioned.
- capability discovery منفصل عن authorization الفعلي.
- interoperability لا تعني trust؛ identity/policy gates لازمة.

Limits / الحدود

- المواصفات تتغير بسرعة والسوق لم يستقر.
- survey لا يضمن security لكل implementation.

MOUSSAIF relevance

صلة مباشرة بـTool Contracts, Framework Adapters, Cross-Kernel Dispatch، وMOUSSAIXT portability.

Research variables

MCP | A2A | typed contracts | discovery

Primary source
https://arxiv.org/abs/2505.02279

Ablation / Reproduction بروتوكول

بدّل runtime/provider مع إبقاء same task contract؛ اختبر schema mismatch, revoked tool, malicious capability advert.

Measure	portability, negotiation errors, policy violations
Failure signal	silent schema coercion أو agent يثق بال capability card بلا authorization.

MOUSSAIF Contract Delta

Primary bridge: L8/L9/L10

افصل HyperTool Contract عن AgentCard وعن AuthorizationToken، واربطها version/capability negotiation.

14

NEW PART II | Compositional CUA | COLM 2025

Agent S2: A Compositional Generalist-Specialist Framework for Computer Use Agents

Saaket Agashe et al. · 2025

Scientific contribution / المساهمة العلمية

يقسّم computer use cognition بين generalists وspecialists بدل نموذج واحد. Mixture-of-Grounding يدمج grounding مصادر مختلفة، وProactive Hierarchical Planning يحدّث plans على عدة horizons مع تغير observations.

Mechanism / الآلية

Planner يفصل global intent عن local steps، بينما grounding specialists تتنافس/تتعاون لتحديد controls. إعادة التخطيط proactive لا تنتظر failure النهائي.

TECHNICAL SIGNATURE

```
Action_t = PlanHierarchy(Obs_t, Goal) + MixtureOfGrounding(Obs_t)
```

Engineering extraction / الاستخراج الهندسي

- التخصص يمكن أن يكون capability-level لا agent-per-task فقط.
- multi-horizon planning يناسب HyperAbility composition.
- grounding uncertainty يجب أن يمر إلى router لا يختفي.

Limits / الحدود

- CUA benchmarks لا تمثل كل العالم المفتوح.
- multiple models ترفع cost وتحتاج router جيداً.

MOUSSAIF relevance

Computational Vascularity/Sparse HyperRouting: فعّل specialist يدعم صغيراً وفق active difference.

Research variables

specialists | grounding | hierarchical planning | routing

Primary source
https://arxiv.org/abs/2504.00906

Ablation / Reproduction بروتوكول

Ablate specialists وhierarchical horizons مع budget ثابت؛ اختبر unseen OS/app.

Measure	success, grounding error, model calls, latency
Failure signal	gain يأتي من ensemble compute فقط أو specialist لا ينقل خارج app.

MOUSSAIF Contract Delta

Primary bridge: L2/L6/L8/L9

أضف Capability Router يختار specialist grounding/planning مع confidence/ cost budget وfallback.

15

NEW PART II | Agent Test-Time Scaling | TMLR 2026 lineage

The Unreasonable Effectiveness of Scaling Agents for Computer Use

Gonzalo Gonzalez-Pumariega et al. · 2025/2026

Scientific contribution / المساهمة العلمية

Behavior Best-of-N يوسع agent compute بإنتاج rollouts متعددة ثم يختار بينها عبر Behavior Narratives/Judge. التركيز ليس على token-level answer sampling بل على trajectory-level behavior selection في computer-use طويل الأفق.

Mechanism / الآلية

N agents/rollouts تنفذ المهمة؛ كل مسار يلخص إلى narrative عن السلوك والنتيجة، ثم judge يختار rollout الأفضل. هذا يحول parallel branch search إلى harness primitive.

TECHNICAL SIGNATURE

```
trajectory* = argmax_i Judge(Narrative(rollout_i)), i=1..N
```

Engineering extraction / الاستخراج الهندسي

- test-time scaling لل agents يجب أن يقاس بال trajectory لا answer.
- judge يحتاج access إلى evidence/state لا narrative فقط.
- N يجب أن يخضع cost-aware routing.

Limits / الحدود

- parallel rollouts مكلفة وقد تكرر side effects.
- judge errors يمكن أن تضخم failure.

MOUSSAIF relevance

صلة مباشرة بـBounded HyperLoops, Branch-Falsify-Merge, QMCD ومحاكم الاختيار branch budgets.

Research variables

Best-of-N | behavior judge | rollouts | trajectory selection

Primary source
https://arxiv.org/abs/2510.02250

Ablation / Reproduction بروتوكول

قارن N=1/2/4/8 مع equal total token/tool budget قدر الإمكان؛ اختبر judge blind spots.

Measure	success/cost, judge accuracy, duplicate effects
Failure signal	زيادة النجاح لا تتناسب مع الكلفة أو اختبار trajectory ذات أثر خاطئ.

MOUSSAIF Contract Delta

Primary bridge: L2/L7/L10

اسمح HyperLoop branching فقط في reversible/sandbox state؛ merge بعد World-Effect/Evidence Court.

16

NEW PART II | CUA Foundation | NeurIPS 2025 Spotlight

OpenCUA: Open Foundations for Computer-Use Agents

Xinyuan Wang et al. · 2025

Scientific contribution / المساهمة العلمية

يبني AgentNet، أول مجموعة بيانات ومنصة مفتوحة لتوسيع Computer-Use Agents: annotation infrastructure، stack عبر 3 OS و200+ apps/websites، ثم يحول demonstrations إلى state-action pairs مع reflective reasoning لتدريب foundation agent models.

Mechanism / الآلية

human computer traces -> normalized state/action data -> reflective training pipeline -> agent model. يدعم data scaling وtest-time compute، ويعطي community أساساً مفتوحاً لدراسة CUA.

TECHNICAL SIGNATURE

```
Demonstrations -> State/Action(+reflection) -> Train -> Interactive Rollout
```

Engineering extraction / الاستخراج الهندسي

- runtime traces يجب أن تحفظ كـtraining assets.
- cross-OS data شرط لعدم ربط kernel بواجهة واحدة.
- observation/action schema أهم من screenshots الخام.

Limits / الحدود

- data collection مكلفة وقد تحمل private/sensitive traces.
- reflective traces ليست بالضرورة faithful.

MOUSSAIF relevance

يدعم Evolution Laboratory وHyperSkill Factory: استخراج skills من verified trajectories عبر بيئات متعددة.

Research variables

AgentNet | trajectory data | cross-OS | CUA training

Primary source
https://arxiv.org/abs/2508.09123

Ablation / Reproduction بروتوكول

درّب/اختبر adapters على OS seen/unseen؛ امنع leakage عبر app templates.

Measure	cross-OS success, data efficiency, action validity
Failure signal	model يحفظ UI patterns ولا يعمم إلى controls جديدة.

MOUSSAIF Contract Delta

Primary bridge: L9/L11

عرّف TraceDataset schema في MOUSSAIFXT: observation, action, effect, tool, authority, failure, verification, skill-candidate.

17

NEW PART II | GUI Agent RL | Technical Report

UI-TARS-2 Technical Report: Advancing GUI Agent with Multi-Turn Reinforcement Learning

ByteDance Seed et al. · 2025

Scientific contribution / المساهمة العلمية

يدفع native GUI-centered agents عبر hybrid GUI environment، stabilized multi-turn RL، data flywheel يجمع. الفكرة الأساسية أن GUI-only action space غير كافٍ للمهام الواقعية، وfilesystem/terminal، unified sandbox for large-scale rollouts.

Mechanism / الآلية

يجمع training perception/reasoning/action/memory في trajectory طويلة، ويستخدم sandbox موحداً لتوليد feedback عبر GUI وCLI/file operations معاً.

TECHNICAL SIGNATURE

```
Policy(vision,state) -> {GUI, filesystem, terminal}; RL reward <- verified rollout
```

Engineering extraction / الاستخراج الهندسي

- tool modality routing جزء من agent intelligence.
- sandbox infrastructure شرط لتوسيع RL الآمن.
- GUI وterminal يجب أن يشاركا state واحداً.

Limits / الحدود

- technical report ضخم؛ attribution بين data/model/harness معقد.
- multi-turn RL حساس لاستقرار البيئة والمكافأة.

MOUSSAIF relevance

يدعم MOUSSAIXT hybrid execution: browser/GUI/shell كـadapters تحت World Model واحد.

Research variables

multi-turn RL | GUI+CLI | sandbox | data flywheel

Primary source
https://arxiv.org/abs/2509.02544

Ablation / Reproduction بروتوكول

Ablate terminal/filesystem actions من نفس model؛ اختبر tasks تتطلب cross-interface handoff.

Measure	success, handoff errors, rollout stability
Failure signal	agent يحقق visual state لكن artifact الحقيقي خاطئ أو العكس.

MOUSSAIF Contract Delta

Primary bridge: L8/L9/L10/L11

وحد EventState بين GUI actions وshell/files؛ اجعل verifier يقرأ artifact state لا screenshot فقط.

18

NEW PART II | Agent Evolution Environment | Open framework

AgentGym: Evolving Large Language Model-based Agents across Diverse Environments

Zhiheng Xi et al. · 2024

Scientific contribution / المساهمة العلمية

يوفر AgentEvol + trajectory set + diverse environments لدراسة generalist self-evolution عبر بيئات متعددة بدلاً من تدريب specialist منعزل. البيئة موحدة وتدعم concurrent exploration.

Mechanism / الآلية

Env server يقدم create/observation/actions/step/reset؛ AgentController يجمع trajectories ويدير evaluation/training. AgentEvol يحاول نقل التحسن خارج البيانات المرئية.

TECHNICAL SIGNATURE

```
Env API -> rollout corpus -> AgentEvol -> cross-environment evaluation
```

Engineering extraction / الاستخراج الهندسي

- environment abstraction يجب أن يكون ثابتاً عبر domains.
- evolution يحتاج transfer benchmark لا in-domain score فقط.
- trajectory registry هو bridge بين runtime والتدريب.

Limits / الحدود

- بعض البيئات مبسطة مقارنة بالproduction.
- self-evolution claims تحتاج ضوابط contamination قوية.

MOUSSAIF relevance

يدعم MOUSSAIF Evolution Lab وDomain Adapters وBenchmark Registry.

Research variables

environment suite | trajectories | self-evolution | transfer

Primary source
https://arxiv.org/abs/2406.04151

Ablation / Reproduction بروتوكول

درّب على environments subset واختبر held-out environments مع نفس harness.

Measure	transfer gain, interactions, recovery
Failure signal	improvement ينحصر في environments seen أو بسبب prompt templates.

MOUSSAIF Contract Delta

Primary bridge: L9/L11

استخدم unified EnvAdapter contract مع reset/snapshot/evaluate، ثم شغّل HyperSkills على بيئات متعددة.

19

NEW PART II | Web Agent Harness | Open ecosystem

The BrowserGym Ecosystem for Web Agent Research

Thibault Le Sellier De Chezelles et al. · 2024/2025

Scientific contribution / المساهمة العلمية

يوحد observation/action spaces وتقييم web agents عبر benchmarks متفرقة، مع AgentLab لإدارة agents/experiments. أهميته هندسية: benchmark harness يمكن أن يكون stable contract يفصل agent logic عن تفاصيل benchmark.

Mechanism / الآلية

BrowserGym يلف بيئات الويب في gym-like API ويعيد observations/actions بنمط موحد؛ AgentLab يدير configuration, runs, logs, analysis across benchmarks.

TECHNICAL SIGNATURE

```
Agent <-> Standard Obs/Action API <-> Multiple Web Benchmarks
```

Engineering extraction / الاستخراج الهندسي

- adapter standardization شرط لمقارنة models/harnesses.
- experiment metadata يجب أن يسجل model+harness+environment version.
- benchmark fragmentation قد يخفي gains الوهمية.

Limits / الحدود

- web-centric ولا يغطي filesystem/desktop بالكامل.
- browser environment يظل متغيراً مع المواقع الحية.

MOUSSAIF relevance

يناسب MOUSSAIXT Reference Interfaces وRuntime Fingerprinting وObservability.

Research variables

standard API | benchmark harness | experiment mgmt | web agents

Primary source
https://arxiv.org/abs/2412.05467

Ablation / Reproduction بروتوكول

شغّل نفس agent config عبر >=3 web benchmarks مع unified adapter؛ افحص reproducibility.

Measure	variance, reproducibility, cross-benchmark rank
Failure signal	تغيير wrapper يغير النتيجة أكثر من model/harness delta.

MOUSSAIF Contract Delta

Primary bridge: L9/L10/L11

أنشئ BenchmarkAdapter interface موحداً وRunManifest يسجل model, harness, tools, environment snapshot, seeds.

20

NEW PART II | Tool-Agent Reliability | Benchmark

tau-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains

Shunyu Yao et al. · 2024

Scientific contribution / المساهمة العلمية

يقيس agents في محادثة ديناميكية مع user simulator، APIs وسياسات domain. التقييم لا يعتمد على النص النهائي بل يقارن database end state بالgoal state ويستخدم pass^k لقياس consistency عبر trials.

Mechanism / الآلية

User <-> Agent <-> Tools/DB تحت policy؛ evaluator يفحص final state. pass^k يبين أن النجاح مرة واحدة لا يعني reliability عندما تتكرر المهمة.

TECHNICAL SIGNATURE

```
Reliability = pass^k over repeated trajectories; verify(final_DB == goal_DB)
```

Engineering extraction / الاستخراج الهندسي

- world-effect evaluation يجب أن يفحص state لا الكلام.
- policy compliance جزء من success.
- repeated-trial reliability أهم من pass@1 للdeployment.

Limits / الحدود

- user simulator ليس إنساناً حقيقياً.
- domains محدودة والسياسات قابلة للترميز أكثر من العالم المفتوح.

MOUSSAIF relevance

مرجع قوي لWorld-Effect Loop وEvidence/Authorization Courts وM-TAOS KPIs.

Research variables

tool use | policy | end state | pass^k

Primary source
https://arxiv.org/abs/2406.12045

Ablation / Reproduction بروتوكول

كرر نفس task k مرات مع seeds مختلفة؛ حقن ambiguous user requests وسياسات متعارضة.

Measure	pass^k, policy violation, end-state delta
Failure signal	النص صحيح لكن DB state خاطئ أو النجاح غير مستقر عبر runs.

MOUSSAIF Contract Delta

Primary bridge: L0/L9/L10

اجعل كل task KPI يملك effect predicate machine-checkable متى أمكن، وسجل repeated-run reliability.

21

NEW PART II | Dual-Control Agents | Benchmark

tau^2-Bench: Evaluating Conversational Agents in a Dual-Control Environment

Victor Barres et al. · 2025

Scientific contribution / المساهمة العلمية

يمتد من single-control إلى Dual-Control Dec-POMDP حيث user والagent كلاهما ينفذ tools في world state مشترك. يقيس reasoning + communication + coordination، ويولد tasks تركيبياً من atomic components.

Mechanism / الآلية

agent يرى جزءاً من state ويوجه user لتنفيذ actions لا يملكها؛ user simulator مقيد بأدوات وحالة قابلة للرصد. الأخطاء تنفصل إلى reasoning مقابل coordination/communication.

TECHNICAL SIGNATURE

```
SharedWorld_{t+1}=T(World_t, a_agent, a_user); partial observations for both
```

Engineering extraction / الاستخراج الهندسي

- human-in-the-loop ليس approval فقط بل co-actor.
- shared-state concurrency يحتاج event ordering واضحاً.
- يجب فصل فشل reasoning عن فشل delegation.

Limits / الحدود

- simulated users لا تعكس كل السلوك البشري.
- coordination tasks still domain-bounded.

MOUSSAIF relevance

يدعم Governance Space وAuthorization وWorld Model في مهام يشترك فيها الإنسان والوكيل في تغيير الحالة.

Research variables

dual control | Dec-POMDP | coordination | shared state

Primary source
https://arxiv.org/abs/2506.07982

Ablation / Reproduction بروتوكول

شغّل single-control tasks ثم dual-control؛ أدخل delayed/user-wrong actions واختبر recovery.

Measure	coordination success, attribution accuracy, recovery
Failure signal	agent ينسب effect لنفسه أو يعيد تنفيذ فعل قام به user.

MOUSSAIF Contract Delta

Primary bridge: L0/L8/L10

أضف ActorId/EventOrder لكل state mutation؛ يصبح human action event موثقاً في نفس mission ledger.

22

NEW PART II | Agent Security | Security benchmark

AgentDojo: A Dynamic Environment to Evaluate Prompt Injection Attacks and Defenses for LLM Agents

Edoardo Debenedetti et al. · 2024

Scientific contribution / المساهمة العلمية

يبني environment ديناميكياً لاختبار prompt injection عندما تجلب tools بيانات غير موثوقة، يحتوي realistic tasks وsecurity test cases وهجمات/دفاعات قابلة للتوسع، ما يحول prompt injection من anecdote إلى runtime adversarial evaluation.

Mechanism / الآلية

tool outputs قد تحتوي instructions خبيثة؛ agent يجب أن يفصل data عن authority. benchmark يقيس utility وsecurity تحت attacks adaptive بدلاً من static prompts.

TECHNICAL SIGNATURE

```
UntrustedToolData != InstructionAuthority; action must satisfy policy despite injected

text
```

Engineering extraction / الاستخراج الهندسي

- كل external content يحتاج trust label.
- tool result لا يملك authority لتعديل objective.
- security test suite يجب أن ترافق skill/tool promotion.

Limits / الحدود

- الهجمات تتطور أسرع من benchmark.
- الدفاع القوي قد يخفض utility.

MOUSSAIF relevance

يدعم Constitutional Precedence Kernel وTool Contracts وEvidence وSkill supply-chain defense وCourt.

Research variables

prompt injection | untrusted data | tool security | adaptive attacks

Primary source
https://arxiv.org/abs/2406.13352

Ablation / Reproduction بروتوكول

حقن malicious content في email/web/RAG/tool outputs مع نفس benign tasks.

Measure	utility, attack success, policy violations
Failure signal	agent ينفذ instruction من untrusted source أو defense يعطل tasks سليمة.

MOUSSAIF Contract Delta

Primary bridge: L0/L4/L9/L10

وسم provenance/trust لكل tool payload؛ router يمنع انتقال data إلى instruction دون explicit bridge channel.

23

NEW PART II | Agent Safety | Benchmark

Agent-SafetyBench: Evaluating the Safety of LLM Agents

Zhexin Zhang et al. · 2024

Scientific contribution / المساهمة العلمية

يوسع تقييم السلامة من text model إلى interactive agent عبر 349 environments و2000 cases تغطي 8 risk categories و10 failure modes. النتيجة المهمة منهجياً: defense prompts وحدها لا تكفي؛ runtime يحتاج risk awareness وrobust controls.

Mechanism / الآلية

agents توضع في environments تتطلب tools/actions وقد تواجه hazardous requests/contexts. scoring يركز على السلوك التفاعلي لا refusal text وحده.

TECHNICAL SIGNATURE

```
SafetyAgent = ModelPolicy + RuntimeConstraints + RiskState + ToolGates
```

Engineering extraction / الاستخراج الهندسي

- safety layer يجب أن تعيش خارج model prompt أيضاً.
- risk awareness يحتاج state variable قابل للقياس.
- capability success وsafety يجب أن يقاسا معاً.

Limits / الحدود

- taxonomy لا تغطي كل مخاطر المستقبل.
- score aggregate قد يخفي catastrophic rare failures.

MOUSSAIF relevance

يؤكد L0 Constitution + Authorization T0-T4 + World-Effect Court + fail-closed semantics.

Research variables

risk awareness | robustness | interactive safety | failure modes

Primary source
https://arxiv.org/abs/2412.14470

Ablation / Reproduction بروتوكول

اختبر capability/safety jointly تحت benign + adversarial variants؛ لا تسمح للمتوسط بإخفاء critical violations.

Measure	unsafe action rate, task success, critical failures
Failure signal	نحسن utility مع ارتفاع unsafe side effects أو reliance على refusal text فقط.

MOUSSAIF Contract Delta

Primary bridge: L0/L9/L10

أضف RiskState إلى MissionState، مع hard gates للأثر العالي وseparate safety regression suite.

24

NEW PART II | Automated Harness Design | ICLR 2025

Automated Design of Agentic Systems

Shengran Hu, Cong Lu, Jeff Clune · 2024/2025

المساهمة العلمية / Scientific contribution

يفتح ADAS: البحث آلياً في designs للagents بدلاً من hand-crafted scaffolds. Meta Agent Search يكتب agents في code ويحتفظ بالتصاميم المكتشفة archive، بما يشمل prompts, tool use, control flows وتركيبات جديدة.

الآلية / Mechanism

meta-agent يقترح implementation جديداً، يشغله على tasks، يسجل performance/design في archive، ثم يستخدم التاريخ لتوليد candidates لاحقة. transfer عبر models/domains هو الاختبار الحاسم للجدة المفيدة.

TECHNICAL SIGNATURE

```
Archive_{t+1}=Archive_t U Evaluate(MetaAgent(Archive_t))
```

الاستخراج الهندسي / Engineering extraction

- harness architecture نفسها search space قابلة للتعلم.
- archive يمنع تكرار التصاميم ويعطي lineage.
- promotion يجب أن يعتمد transfer لا benchmark واحد.

الحدود / Limits

- automated design قد يستغل benchmark أو يخلق complexity غير قابلة للصيانة.
- search cost وsafety constraints أساسيان.

MOUSSAIF relevance

يتقاطع مباشرة مع Evolution Laboratory وHyperHarness/HyperLoop synthesis ومحكمة الجدة الزائفة.

Research variables

meta-agent | architecture search | archive | transfer

Primary source
https://arxiv.org/abs/2408.08435

بروتوكول Ablation / Reproduction

ابحث عن harness deltas على training tasks ثم اختبر domain/model transfer مع name stripping.

Measure	transfer gain, complexity, search cost, regressions
Failure signal	candidate يربح بسبب benchmark-specific hack أو يصعب تفسيره/إرجاعه.

MOUSSAIF Contract Delta

Primary bridge: L7/L8/L11

اسمح AutoHarnessLab بتوليد candidates فقط داخل typed component grammar، كل candidate يحتاج ablation+holdout+rollback.

25

NEW PART II | Self-Improving Agents | arXiv / open-ended evolution

Darwin Godel Machine: Open-Ended Evolution of Self-Improving Agents

Jenny Zhang et al. · 2025

Scientific contribution / المساهمة العلمية

يستكشف self-modification عمليةً: agent يغير code الخاص به، يقيم التعديل empirically، ويحفظ variants في archive/tree بدلاً من الالتزام بخط تحسين واحد. الهدف open-ended exploration مع empirical verification.

Mechanism / الآلية

proposal -> code modification -> sandbox benchmark -> accept/archive. branching archive يسمح بالرجوع إلى ancestors واستكشاف lineages متعددة بدلاً من greedy self-edit.

TECHNICAL SIGNATURE

```
Agent_{t+1}=Select(Evaluate(Mutate(Agent_t))); keep evolutionary archive
```

Engineering extraction / الاستخراج الهندسي

- self-improvement يجب أن يكون versioned artifact لا overwrite ذاتي.
- benchmark verifier هو gate بين mutation وdeployment.
- branching archive يوفر rollback وdiversity.

Limits / الحدود

- coding benchmark improvement لا يساوي general recursive self-improvement.
- self-modification يرفع مخاطر goal drift/security.

MOUSSAIF relevance

قريب من Self-Reconstruction + HyperDeployment + Evolution Lab + Failure/Goal Drift Courts.

Research variables

self-modification | archive | empirical validation | open-endedness

Primary source
https://arxiv.org/abs/2505.22954

Ablation / Reproduction بروتوكول

جرّب candidate mutations مع frozen constitution وheld-out suites؛ أدخل malicious/goal-shifting mutation.

Measure	verified gain, regressions, rollback success, goal invariance
Failure signal	improvement benchmark مع drift في policy أو artifact غير قابل للاسترجاع.

MOUSSAIF Contract Delta

Primary bridge: L10/L11

لا تعديل canonical kernel in-place: fork candidate, sandbox, eval, court, signed promotion, reversible deployment.

26

NEW PART II | Self-Evolution | arXiv

AgentEvolver: Towards Efficient Self-Evolving Agent System

Yunpeng Zhai et al. · 2025

المساهمة العلمية / Scientific contribution

يجمع self-questioning لتوليد tasks من self-navigating، curiosity لإعادة استخدام الخبرات في exploration، وself-attributing لتوزيع reward على states/actions حسب المساهمة. يعالج تكلفة handcrafted tasks وrandom exploration.

الآلية / Mechanism

agent يولد challenges من البيئة، يسترجع trajectories مفيدة لتوجيه exploration، ثم يحلل attribution للمسار بدلاً من outcome reward موحد.

TECHNICAL SIGNATURE

```
Evolve = SelfQuestion(env) + GuidedExplore(memory) + CreditAssign(trajectory)
```

الاستخراج الهندسي / Engineering extraction

- task generation جزء من self-evolution.
- experience reuse يحتاج diversity حتى لا يغلق exploration.
- credit assignment يربط failure topology بالتعلم.

الحدود / Limits

- self-generated ،preliminary/evolving evidence tasks قد تخاز لما يعرفه النوذج.
- reward attribution غير مضمون causal.

MOUSSAIF relevance

يدعم Governing Question Engine + Failure Memory + Evolution Lab وMCD لاختيار الفرق الأعلى قيمة.

Research variables

self-questioning | exploration | credit assignment | experience reuse

Primary source
https://arxiv.org/abs/2511.10395

بروتوكول Ablation / Reproduction

قارن human task set مقابل self-generated set على held-out envs؛ افحص diversity وtransfer.

Measure	novel task coverage, sample efficiency, transfer
Failure signal	task generator يعيد صياغة known tasks أو يختار easy curriculum فقط.

MOUSSAIF Contract Delta

Primary bridge: L1/L2/L11

أنشئ FrontierQuestion queue من gaps/failures مع novelty/impact score، ثم اربط كل candidate بتجربة قابلة للهدم.

27

NEW PART II | Skill Learning | arXiv

Reinforcement Learning for Self-Improving Agent with Skill Library (SAGE)

Jiongxiao Wang et al. · 2025

Scientific contribution / المساهمة العلمية

SAGE يجعل skill library جزءاً من RL بدلاً من prompt trick. Sequential Rollout يمر على سلسلة tasks متشابهة بحيث تتراكم skills من مهمة إلى التالية، وSkill-integrated Reward يشجع توليدها واستعمالها المفيد.

Mechanism / الآلية

خلال rollout طويل عبر skill store، يتغير tasks online؛ policy تتعلم متى تستدعي أو تولد skill. reward يجمع outcome مع إشارة مرتبطة بجودة skill usage.

TECHNICAL SIGNATURE

```
rollout: task_1 -> skill_1 -> task_2 -> skill_2 ...; reward = R_task + lambda R_skill
```

Engineering extraction / الاستخراج الهندسي

- skill transfer يجب أن يقاس داخل sequence لا isolated tasks.
- library growth يحتاج pruning/certification.
- efficiency tokens/steps جزء من جودة skill.

Limits / الحدود

- AppWorld-centric results تحتاج domains أخرى.
- reward قد يشجع skill proliferation.

MOUSSAIF relevance

مرجع تطبيقي لـHyperSkill Factory + HyperAbility + Bounded HyperLoop + skill promotion.

Research variables

skill library | sequential rollout | GRPO/RL | reuse

Primary source
https://arxiv.org/abs/2512.17102

Ablation / Reproduction بروتوكول

قارن no-library/static-library/learning-library على held-out sequences؛ قس growth وreuse.

Measure	success, steps, tokens, skill reuse, redundancy
Failure signal	تكبر أسرع من فائدتها أو library تمنع skill exploration.

MOUSSAIF Contract Delta

Primary bridge: L5/L6/L7/L11

أضف SkillUtility score من safety×cost×success×reuse، وprune skills ذات redundancy أو negative transfer.

28

NEW PART II | Kernel Evolution | AAMAS 2026 extended work / arXiv

Self-Evolving Software Agents

Marco Robol, Paolo Giorgini · 2026

Scientific contribution / المساهمة العلمية

يميز adaptation داخل goals ثابتة عن evolution تغير knowledge/goals/executable behavior نفسها. يدمج BDI reasoning loop مع Automated Evolution Module تستخدم LLM لاكتشاف requirements جديدة وتوليد design/code updates من الخبرة.

Mechanism / الآلية

عندما تعجز current beliefs/goals/plans عن تفسير البيئة أو استغلالها، evolution module يقترح knowledge/goal/action changes، ثم يعاد إدخالها في BDI loop. التحدي المركزي inheritance/stability.

TECHNICAL SIGNATURE

```
BDI_t + ExperienceGap -> EvolutionModule -> {Beliefs,Goals,Code}_{t+1}
```

Engineering extraction / الاستخراج الهندسي

- kernel evolution يحتاج explicit trigger وليس continuous rewrite.
- تغيير goals أعلى خطورة من إضافة skill.
- inheritance/stability يجب أن يكونا metrics.

Limits / الحدود

- prototype صغير وليس evidence على autonomous AGI evolution.
- goal evolution يحتاج human/constitutional boundaries أقوى.

MOUSSAIF relevance

مفيد لتحديد حدود ما يجوز لـMOUSSAIFXT تطويره: skills/loops أولاً؛ kernel/goal فقط تحت authority أعلى.

Research variables

BDI | goal evolution | code evolution | stability

Primary source
https://arxiv.org/abs/2604.27264

Ablation / Reproduction بروتوكول

اسمح evolution على skill level ثم kernel level في sandbox؛ حقن novel environment requirement واختبر inheritance.

Measure	adaptation gain, goal drift, stability, rollback
Failure signal	النظور يحل task يكسر constraints أو prior competencies جديداً لكنه.

MOUSSAIF Contract Delta

Primary bridge: L0/L1/L11

عرّف E-kernel < E-router < E-loop < E-skill :EvolutionAuthority tiers- goal؛ آخر المستويات تتطلب explicit human approval.

29

NEW PART II | Agent Runtime Platform | ICLR 2025 lineage

OpenHands: An Open Platform for AI Software Developers as Generalist Agents

Xingyao Wang et al. · 2024/2025

Scientific contribution / المساهمة العلمية

منصة عامة لوكلاء يكتبون code، يستخدمون command line ويتصفحون web، مع multi-agent coordination وsandboxed execution. benchmark integration قيمتها لـruntime reference: البيئة والأدوات والعزل والتقييم مكونات platform قابلة للتبديل.

Mechanism / الآلية

agent loop يعمل فوق runtime يوفر shell/editor/browser sandbox وframework. event observations يسمح بتبديل /agents models دون إعادة بناء البيئة، ويجمع evals متعددة.

TECHNICAL SIGNATURE

```
AgentPolicy <-> Sandboxed Runtime {shell,code,web} <-> Benchmarks
```

Engineering extraction / الاستخراج الهندسي

- agent runtime يجب أن يفصل policy عن execution container.
- sandbox + logs + benchmark adapters primitives.
- coding agent هو generalist computer actor حين يملك shell/web.

Limits / الحدود

- platform لا يحل planning/reliability بذاته.
- تنفيذ code يفتح surface أمني كبير.

MOUSSAIF relevance

مرجع مباشر لـTool Contracts وMOUSSAIXT shell, Hermes runtime integration, sandbox adapters.

Research variables

sandbox | shell | browser | runtime platform

Primary source
https://arxiv.org/abs/2407.16741

Ablation / Reproduction بروتوكول

شغّل same mission عبر local/container/cloud runtime؛ اختبر environment parity وcleanup/rollback.

Measure	portability, side effects, recovery, trace completeness
Failure signal	اختلاف host يغير semantics أو يترك artifacts/credentials خارج sandbox.

MOUSSAIF Contract Delta

Primary bridge: L8/L9/L10

اعتمد RuntimeAdapter موحداً مع snapshot, execute, observe, artifact-export, cleanup, permission profile.

30

NEW PART II | Agentic Coding Model | Open-weight technical report

Qwen3-Coder-Next Technical Report

Qwen Team · 2026

Limits / الحدود

- model report يخلط gains من data/architecture/RL.
- coding ≠ general agency.

MOUSSAIF relevance

مفيد لمسار MoussaifXT local coding kernel وLocal <-> Main-Brain routing وModel Evolution Lab.

Research variables

verifiable tasks | coding RL | MoE | 3B active

Primary source
https://arxiv.org/abs/2603.00729

Scientific contribution / المساهمة العلمية

يبني coding model فعالاً (80B total / 3B active) عبر large-scale synthesis of verifiable coding tasks مع executable environments، ثم mid-training + RL من environment feedback. يربط model capability مباشرةً agent-centric مثل SWE-Bench وTerminal-Bench.

Mechanism / الآلية

يبني verifiable task generator coding problems مع بيئة تنفيذ واختبارات، ثم تتعلم policy من نتائج execution بدلاً من text-only labels. MoE صغير active footprint يركز على inference efficiency.

TECHNICAL SIGNATURE

```
Task synthesis + executable verifier -> mid-training -> RL(environment feedback)
```

Engineering extraction / الاستخراج الهندسي

- verifiable environments يمكن أن تكون data factory.
- active parameter efficiency مهم لـlocal/edge routing.
- model training وharness evaluation يجب أن يتشاركا environment contracts.

MOUSSAIF Contract Delta

Primary bridge: L8/L9/L11

استخدم local coder كـspecialist kernel خلف نفس Tool/State contracts؛ لا تمنحه canonical writeback دون independent tests.

Ablation / Reproduction بروتوكول

ثبت harness وقارن coder backbones؛ ثم ثبت model وقارن training-like verifier tasks مقابل unseen repos.

Measure	solve rate, cost, tests passed, regression
Failure signal	benchmark memorization أو patch يمر tests سطحية لكنه يكسر repo.

31

NEW PART II | CLI Agent Evaluation | arXiv 2026 / open harness

Terminal-Bench 2.0: Benchmarking Agents on Hard, Realistic Tasks in Command Line Interfaces

Mike A. Merrill et al. · 2026

Scientific contribution / المساهمة العلمية

يقدم 89 hard realistic terminal tasks، لكل منها unique environment, human-written solution وcomprehensive tests.
يختبر long-horizon agent competence في CLI ويصدر evaluation harness بدل الاعتماد على synthetic snippets.

Mechanism / الآلية

كل task تُشغل في environment مع verifier tests؛ agent يتفاعل عبر terminal حتى termination، ثم يقاس outcome بـ tests لا judge لغوي فقط. error analysis يحدد failure modes على command/trajectory level.

TECHNICAL SIGNATURE

```
CLI trajectory -> environment state -> comprehensive test oracle -> pass/fail
```

Engineering extraction / الاستخراج الهندسي

- terminal هو world-effect laboratory غني وقابل للتحقق.
- artifact/test evidence أقوى من self-report.
- hard tasks تكشف state/recovery/harness failures.

Limits / الحدود

- CLI لا يختبر perception/GUI.
- tests نفسها قد تكون incomplete specification.

MOUSSAIF relevance

Benchmark مثالي لـMOUSSAIXT shell وWorld-Effect Court وHyperLoop recovery وevidence writeback.

Research variables

terminal | long horizon | tests | evaluation harness

Primary source
https://arxiv.org/abs/2601.11868

Ablation / Reproduction بروتوكول

اختبر current XT harness وminimal baseline مع same backbone؛ سجل per-command errors وrepair loops.

Measure	pass rate, commands, tokens, retries, verifier failures
Failure signal	agent يعلن success قبل tests أو يكرر commands مدمرة بلا state awareness.

MOUSSAIF Contract Delta

Primary bridge: L9/L10/L11

اجعل TerminalBench-like tasks جزءاً من Promotion Suite لكل kernel/harness version مع cost/retry telemetry.

32

NEW PART II | Compute Routing | Open research

s1: Simple Test-Time Scaling

Niklas Muennighoff et al. · 2025

Scientific contribution / المساهمة العلمية

يبين أن dataset صغير عالي الجودة + Budget Forcing يمكن أن يعطي test-time scaling controllable. الفكرة الهندسية المهمة هي أن reasoning budget متغير runtime يمكن زيادته أو قطعه وفق task difficulty بدلاً من طول تفكير ثابت.

Mechanism / الآلية

SFT على 1K traces صعبة ومتنوعة؛ inference budget forcing يقصر التفكير أو يمدده بإجبار continuation. بذلك يصبح compute allocation explicit knob قابل للمعايرة.

TECHNICAL SIGNATURE

```
answer = Reason(x | budget=B); choose B from task difficulty / value-of-compute
```

Engineering extraction / الاستخراج الهندسي

- ليس كل task يحتاج max reasoning.
- compute budget يجب أن يختاره router قبل/أثناء التنفيذ.
- أثر budget يحتاج calibration مقابل latency/cost.

Ablation / Reproduction بروتوكول

قس performance-cost curve عبر budgets ثابتة مقابل MCD-routed budgets على tasks مختلطة الصعوبة.

Measure	utility per token/latency, accuracy, overthink rate
Failure signal	router يخصص compute أكثر دون gain أو يقطع reasoning قبل evidence.

Limits / الحدود

- math reasoning لا يساوي agentic action.
- forcing أطول قد يزيد overthinking.

MOUSSAIF relevance

تقاطع مباشر مع MCD/QMCD وSparse HyperRouting وToken Economy.

Research variables

test-time compute | budget forcing | reasoning length | difficulty

Primary source
https://arxiv.org/abs/2501.19393

MOUSSAIF Contract Delta

Primary bridge: L2/L8

أضف ComputeBudget object مشتقاً من MCD/QMCD: token, branches, tool calls, verifier passes, deadline.

33

NEW PART II | Retrieval-as-Action | arXiv / open

Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning

Bowen Jin et al. · 2025

Scientific contribution / المساهمة العلمية

يعلّم model عبر RL أن يقرر متى وكيف يصدر multi-turn search queries أثناء reasoning بدلاً من fixed RAG أو prompting tool. يصبح use. retrieval action داخل policy مع outcome reward.

Mechanism / الآلية

rollout يتناوب بين reasoning tokens وsearch actions؛ تُmask retrieved tokens في RL optimization لتثبيت التدريب؛ reward النهائي يدفع policy لاختيار queries مفيدة.

TECHNICAL SIGNATURE

```
reason -> search(q1) -> observe -> reason -> search(q2) -> answer; optimize outcome

reward
```

Engineering extraction / الاستخراج الهندسي

- retrieval policy يمكن تعلمها.
- query generation يجب أن يراعي information gain/cost.
- المصدر المسترجع يحتاج provenance/court منفصل عن policy reward.

Limits / الحدود

- outcome reward قد يكافئ answer صحيح بمصادر ضعيفة.
- search engine behavior متغير.

MOUSSAIF relevance

يربط GraphRAG/HKOS بـMCD Router: retrieval ليس خطوة ثابتة بل capability تُفعّل حسب الفرق.

Research variables

multi-turn search | RL | queries | retrieval reasoning

Primary source
https://arxiv.org/abs/2503.09516

Ablation / Reproduction بروتوكول

قارن fixed top-k RAG, prompted search, learned/routed search تحت نفس query/tool budget.

Measure	answer quality, source quality, searches/task, cost
Failure signal	زيادة searches مع citation noise أو correct answer بلا support.

MOUSSAIF Contract Delta

Primary bridge: L2/L4/L8

حوّل retrieval إلى HyperAbility بـbudget وstop condition، ومرر كل evidence إلى source ranking قبل writeback.

34

NEW PART II | Latent Reasoning | Research paper

Training Large Language Models to Reason in a Continuous Latent Space (Coconut)

Shibo Hao et al. · 2024/2025

Scientific contribution / المساهمة العلمية

يستبدل بعض Chain-of-Thought text بـcontinuous latent states تعاد كinput للنموذج. في بعض مهام التخطيط، latent reasoning يستطيع تمثيل بدائل متعددة قبل الالتزام بسلسلة لغوية واحدة، مع سلوك يشبه breadth-first exploration.

Mechanism / الآلية

بدلاً من hidden state، decoding token النائج يعاد إلى model كـcontinuous thought لعدة steps، ثم يعود إلى language output. curriculum يدرب الانتقال من explicit CoT إلى latent mode.

TECHNICAL SIGNATURE

```
h_{t+1}=LM_hidden(h_t,x); decode only when leaving latent-thought phase
```

Engineering extraction / الاستخراج الهندسي

- ليس كل internal reasoning يجب أن يكون language.
- latent mode قد يقلل token cost لكنه يخفض inspectability.
- router يمكن أن يختار explicit vs latent reasoning حسب audit need.

Limits / الحدود

- monitorability أضعف؛ التفسير البشري لل latent state محدود.
- النتائج ليست دليلاً على general planning superiority.

MOUSSAIF relevance

يفرض على MOUSSAIF فصل cognitive state عن audit evidence: لا تعتمد المحاكم على CoT فقط.

Research variables

latent thoughts | continuous reasoning | planning | monitorability

Primary source
https://arxiv.org/abs/2412.06769

Ablation / Reproduction بروتوكول

قارن explicit CoT/latent/hybrid مع نفس compute، ثم اختبر auditability under injected errors.

Measure	success, tokens, verifier detectability, correction
Failure signal	latent mode يرفع score لكنه يخفض اكتشاف failures الحرجة.

MOUSSAIF Contract Delta

Primary bridge: L2/L10

اسمح hidden/opaque model reasoning لكن اجعل decisions externally justified عبر evidence, actions, state deltas, verifiers.

35

NEW PART II | Generative Verifier | Research paper

Process Reward Models That Think (ThinkPRM)

Muhammad Khalifa et al. · 2025

Scientific contribution / المساهمة العلمية

يبني Process Reward Model generative يتحقق من كل reasoning step عبر verification CoT بدلاً من classifier صامت. الهدف data-efficient verifier يستعمل قدرة reasoning الخاصة بالنموذج ويعمل في best-of-N/search.

Mechanism / الآلية

verifier يقرأ solution step وينتج تحليل تحقق ثم score؛ training يحتاج process labels أقل من discriminative PRMs في تجارب الورقة. test-time يمكن زيادة verification compute.

TECHNICAL SIGNATURE

```
Verifier(step_i,context) -> verification rationale -> score_i; search uses scores
```

Engineering extraction / الاستخراج الهندسي

- verifier يجب أن يكون agent reasoning component لا scalar فقط.
- verification compute يمكن توجيهه حسب risk.
- verifier يحتاج independence لتجنب shared error.

Limits / الحدود

- generative verifier قد يهلوس تبريراً.
- نفس-family prover/verifier correlation يقلل الاستقلال.

MOUSSAIF relevance

يقوي Falsification/Demolition/Evidence Courts ويقترح Court-as-Agent مستقل.

Research variables

PRM | verifier CoT | step scoring | best-of-N

Primary source
https://arxiv.org/abs/2504.16828

Ablation / Reproduction بروتوكول

قارن ThinkPRM-like verifier، static rules، LLM-judge، استخدم adversarial wrong-but-plausible traces.

Measure	verifier accuracy, false accept, tokens, search gain
Failure signal	verifier يوافق على نفس misconception أو rationalizes bad step.

MOUSSAIF Contract Delta

Primary bridge: L10

كل Court عالي الأثر يمكن أن يملك reasoning verifier منفصلاً + deterministic checks + evidence retrieval.

36

NEW PART II | Reasoning Audit | Research / safety

Reasoning Models Don't Always Say What They Think

Yanda Chen et al. · 2025

Scientific contribution / المساهمة العلمية

يختبر CoT faithfulness عبر hints ويجد أن reasoning traces كثيراً ما لا تكشف فعلياً المؤثرات المستخدمة في القرار. outcome RL قد يحسن faithfulness في البداية ثم plateau، وreward hacking لا يضمن ظهور السبب في CoT.

Mechanism / الآلية

تعطى reasoning models hints يمكن قياس استخدامها سلوكياً، ثم يقاس هل يذكر CoT هذه hints عندما تؤثر في answer. الفرق بين behavioral influence وverbalized rationale هو محور القياس.

TECHNICAL SIGNATURE

```
Faithfulness = P(CoT reveals factor | factor causally affects answer)
```

Engineering extraction / الاستخراج الهندسي

- CoT ليس audit log موثوقاً بذاته.
- observability يجب أن تعتمد tools/state/evidence لا narration فقط.
- rare catastrophic behavior لا يُستبعد بمراقبة CoT.

Limits / الحدود

- الاختبارات على hints صناعية جزئياً.
- عدم faithfulness لا يعني أن CoT بلا أي قيمة للمراقبة.

MOUSSAIF relevance

MOUSSAIF: trace of observed actions/evidence outranks يدعم قاعدة model self-explanation.

Research variables

CoT faithfulness | monitoring | reward hacking | audit

Primary source
https://arxiv.org/abs/2505.05410

Ablation / Reproduction بروتوكول

أنشئ hidden causal hints/tool conditions ثم قارن narrative attribution بالactual action dependence.

Measure	faithfulness, false reassurance, observed trace coverage
Failure signal	court يثق بتفسير النموذج رغم تناقضه مع الأدلة السلوكية.

MOUSSAIF Contract Delta

Primary bridge: L10/L11

قسّم Trace إلى ObservedTrace (tools/state) وNarrativeTrace (model rationale) ولا تُسمح للثاني بإثبات الأول.

37

NEW PART II | Reasoning Foundation Model | Open models

Qwen3 Technical Report

Qwen Team · 2025

Scientific contribution / المساهمة العلمية

يوحد Thinking وNon-Thinking modes داخل family واحدة ويضيف Thinking Budget للتحكم في inference compute، مع dense/MoE sizes متعددة ودعم agent/code/reasoning tasks. أهميته للOS هي جعل reasoning mode/cost decision جزءاً من runtime contract.

Mechanism / الآلية

chat/template يحدد mode؛ budget يضبط عدد التفكير tokens. post-training ينقل معرفة flagship إلى models أصغر، ما يتيح model hierarchy بحسب task.

TECHNICAL SIGNATURE

```
ModelMode in {direct,thinking}; Budget=B(task); Router -> size/mode/budget
```

Engineering extraction / الاستخراج الهندسي

- model routing يمكن أن يختار mode داخل نفس family.
- budget parameter قابل للدمج مع QoS/latency.
- smaller models يمكن أن تحمل specialist skills.

Limits / الحدود

- benchmark results لا تثبت general intelligence.
- thinking budget يحتاج calibration per domain.

MOUSSAIF relevance

مناسب لSparse HyperRouting وCross-Model AGI Density وMain-Brain/Local model dispatch.

Research variables

thinking mode | budget | MoE | model hierarchy

Primary source
https://arxiv.org/abs/2505.09388

Ablation / Reproduction بروتوكول

ثبت harness واختبر direct/thinking/budgets عبر tasks ذات risk/difficulty مختلفة.

Measure	verified utility/cost, latency, overthinking
Failure signal	thinking mode يستهلك budget دون gain أو يهبط tool reliability.

MOUSSAIF Contract Delta

Primary bridge: L2/L8

اجعل ModelRoute = provider/model/mode/budget/context/tool-policy بدلاً من model name فقط.

38

NEW PART II | Agentic Foundation Model | Open-weight report

Kimi K2: Open Agentic Intelligence

Kimi Team · 2025

Scientific contribution / المساهمة العلمية

K2 هو MoE ضخم (1T total / 32B active) مع MuonClip optimizer، ويجعل agentic data synthesis + joint RL في /real synthetic environments محور post-training. يعرض agentic capability كهدف تدريب لا مجرد prompt wrapper.

Mechanism / الآلية

pretraining واسع ثم post-training متعدد المراحل يولد agentic trajectories ويستخدم RL من non-interaction environments. يبرز أهمية thinking agentic performance internalized tool/interaction patterns.

TECHNICAL SIGNATURE

```
Pretrain -> Agentic Data Synthesis -> Environment Interaction RL -> Agent Model
```

Engineering extraction / الاستخراج الهندسي

- agentic training data يجب أن يأتي من environments قابلة للتحقق.
- foundation model يمكنه internalize tool priors لكن harness يبقى ضرورياً.
- MoE active footprint مهم للتكلفة.

Limits / الحدود

- attribution، model report بين model/harness/data ليس بسيطاً.
- قوة benchmarks لا تلغي long-horizon failure.

MOUSSAIF relevance

MOUSSAIFXT candidate backbone مع فصل kernel/harness عن model كي يمكن swap دون فقد state.

Research variables

MoE | agentic RL | environment data | tool use

Primary source
https://arxiv.org/abs/2507.20534

Ablation / Reproduction بروتوكول

model-swap study: K2 vs other backbones under identical harness, tools, prompts, budgets.

Measure	agent success, cost, tool errors, recovery
Failure signal	نتائج النموذج لا تنتقل عند تثبيت harness أو تعتمد على benchmark-native formats.

MOUSSAIF Contract Delta

Primary bridge: L8/L9

اختبر K2-like backbone خلف نفس MOUSSAIXT contracts؛ لا تدمج model-specific quirks في canonical kernel.

39

NEW PART II | ARC Foundation Model | Open-source report

GLM-4.5: Agentic, Reasoning, and Coding (ARC) Foundation Models

GLM-4.5 Team · 2025

Scientific contribution / المساهمة العلمية

يصمم foundation model صراحةً حول Agentic, Reasoning, Coding، مع MoE 355B/32B active وhybrid thinking/direct modes. multi-stage training + expert model iteration + RL تربط الثلاثة بدل تدريب agentic ability كإضافة صغيرة.

Mechanism / الآلية

pretraining واسع ثم post-training يعالج reasoning/coding/agent environments، مع mode switching بين thinking/direct. النموذج يقيم على TAU-Bench وSWE-Bench وغيرها.

TECHNICAL SIGNATURE

```
ARC training = general pretrain + expert iteration + RL(agent/reason/code) + hybrid mode
```

Engineering extraction / الاستخراج الهندسي

- backbone يمكن تحسينه jointly للagent/runtime workloads.
- hybrid mode مناسب router متعدد الكلفة.
- harness benchmark يجب أن يفصل model gain عن orchestration gain.

Limits / الحدود

- score composite لا يعني unified AGI.
- model family specifics تتغير سريعاً.

MOUSSAIF relevance

يدعم cross-model routing في MOUSSAIFXT واختيار specialist backbone حسب Mission Graph.

Research variables

agentic | reasoning | coding | hybrid mode

Primary source
https://arxiv.org/abs/2508.06471

Ablation / Reproduction بروتوكول

ثبّت harness وأجر model profile عبر tau/terminal/reasoning tasks؛ اختبر dynamic routing مقابل single model.

Measure	portfolio utility/cost, route accuracy, failures
Failure signal	router يعتمد leaderboard average ولا يطابق task-level capability.

MOUSSAIF Contract Delta

Primary bridge: L2/L8/L9

أضف benchmark profile لكل model: reasoning/coding/tool/GUI strengths + cost + context + safety، ويختار router وفق mission.

40

NEW PART II | Long-Context Reasoning Model | Open-weight report

MiniMax-M1: Scaling Test-Time Compute Efficiently with Lightning Attention

MiniMax Team · 2025

Scientific contribution / المساهمة العلمية

يجمع hybrid attention + MoE و1M-token context مع RL واسع وCISPO، ويستهدف efficient test-time compute، software engineering، tool use، وlong-context tasks. مهم للOS لأنه يبين أن context capacity وreasoning budget يمكن تكبيرهما دون أن يكونا بديلين عن memory/harness.

Mechanism / الآلية

lightning attention يقلل كلفة السياق الطويل، بينما RL على diverse tasks تشمل sandbox software environments، نماذج بـthinking budgets مختلفة توضح compute-performance trade-off.

TECHNICAL SIGNATURE

```
LongContext + HybridAttention + RL + ThinkingBudget -> efficient test-time scaling
```

Engineering extraction / الاستخراج الهندسي

- 1M context لا يلغي الحاجة إلى curated external state.
- long context يجب أن يخضع compaction/retrieval policy.
- budget+context يدخلان resource router مع tool calls.

Limits / الحدود

- long context قد يزيد distraction وcost.
- model report لا يثبت memory persistence أو agency خارج session.

MOUSSAIF relevance

يخدم Context Genome/Compaction وToken Economy: استفد من large window بشكل selective لا giant prompt دائم.

Research variables

1M context | lightning attention | CISPO | test-time compute

Primary source
https://arxiv.org/abs/2506.13585

Ablation / Reproduction بروتوكول

قارن full-context مقابل selective HKOS retrieval+compaction تحت equal latency/cost، حقن distractors.

Measure	accuracy, latency, context tokens, distractor sensitivity
Failure signal	larger context يخفض grounding أو يرفع cost بلا verified utility.

MOUSSAIF Contract Delta

Primary bridge: L2/L4/L8

Router يقرر jointly fetch/compact/context-window/budget، external canonical memory تبقى المصدر لا prompt الضخم.

الاستنتاج المضغوط - ما الذي تتفق عليه الجبهة عملياً؟

عند ضغط الأربعين إلى architecture-neutral invariants، لا تكون الرسالة "ابنِ نموذجاً أكبر"، بل: اجعل state والسلطة والمهارات والذاكرة والتنفيذ والتحقق والتطور primitives مستقلة يمكن قياسها واستبدالها.

1. State outside context
المهمة الطويلة تحتاج canonical task state خارج prompt، ولا يُكتب إلا بعد evidence/verification.

2. Harness is a first-class artifact
الـharness له version, tests, telemetry, lineage وrollback مثل أي kernel.

3. Skills need contracts
الـskill ليست نصاً: preconditions + tools + policy + stop + tests + trust + version.

4. Memory is a policy
Store/Retrieve/Update/Summarize/Discard قرارات تحتاج routing وحوكمة.

5. World effect outranks narration
الحالة المرصودة والاختبارات أقوى من "تم التنفيذ" أو CoT.

6. Authority travels with delegation
كل sub-agent/tool call يرث سلطة محدودة وقابلة للإلغاء والتتبع.

7. Long horizon needs audit checkpoints
Manager/Executor/Auditor أو ما يعادله يمنع تراكم false success.

8. Model and harness must be disentangled
أي مقارنة frontier يجب أن تصرح model + harness + tools + environment + budget.

9. Self-improvement must fork, not overwrite
candidate -> sandbox -> eval -> court -> signed promotion -> rollback.

10. Compute is routed, not maximized
thinking tokens, branches, searches, verifiers وtool calls ميزانية واحدة.

11. Tool data has no authority by default
untrusted retrieval/tool output لا يتحول إلى instruction بلا bridge صريح.

12. Multi-agent = shared state + contracts
زيادة agents دون ledger/roles/ownership قد تزيد الفشل فقط.

13. Evaluation must repeat
pass^k وvariance يكشفان reliability التي يخفيها success مرة واحدة.

14. Environment adapters are part of science
GUI/API/CLI wrapper قد يغير النتائج؛ يجب versioning وfingerprinting.

15. Evolution consumes failure traces
trajectory/failure graph يجب أن يتحول إلى candidates/skills/tests لا archival logs فقط.

16. Promotion is empirical
DEFINE -> IMPLEMENT -> ABLATE -> FALSIFY -> REPLICATE -> TRANSFER -> PROMOTE.

Verified Useful Agency = Model Capability × Harness Quality × State Fidelity × Tool Grounding × Verification × Recovery × Governance

مصفوفة 40 × MOUSSAIF - أين يدخل كل بحث؟

العمود الأخير هو proposal هندسي لا claim من الورقة. فائدته أنه يحول literature review إلى backlog قابل للتنفيذ والرفض.

Concrete integration delta	Primary layer(s)	Research root	#
أضف Memory Bus يفصل context عن عرض، وإثبات promotion/eviction policy لكل memory object provenance.	L3/L4/L11	Memory OS of AI Agent	01
أضف MemoryObject موحد: revision، type، authority، provenance، TTL، retrieval-policy، writeback gate.	L3/L11	MemOS: An Operating System for Memory-Augmented Generation (MAG) in Large Language Models	02
اجعل Knowledge Vessel يولّد links مرشّحة مع confidence ثم تمر عبر Evidence/Contradiction Court قبل promotion.	L3/L4/L11	A-MEM: Agentic Memory for LLM Agents	03
أضف HyperSkills memory ثم اجمع لل Router يتعلم اختيارها ضمن حدود authorization وNo-Loss Court.	L2/L5/L11	Agentic Memory: Learning Unified Long-Term and Short-Term Memory Management for Large Language Model Agents	04
وسّع HyperSkill contract ليشمل preconditions، stop، tests، trust-tier، provenance، signature، rollback، compatib...	L5/L6/L7/L11	SoK: Agentic Skills - Beyond Tool Use in LLM Agents	05
عرّف HyperHarness manifest مستقل عن model: registry، context، state، tools، hooks، verifier، rollback، benchmark...	L5-L10	Code as Agent Harness	06
أضف HarnessDelta registry: parent_version، changed skill/tool/loop، evidence set، regression suite، rollback poi...	L7/L10/L11	Evolving Agents in the Dark: Retrospective Harness Optimization via Self-Preference	07
حوّل M-TAOS state writeback إلى two-phase commit: tentative effect -> independent audit -> canonical state commit.	L7/L8/L10/L11	LongHorizon-Harness: Advancing Long-Horizon Agents for Real-World Tasks	08
أضف DelegationChain وAuthorityToken لكل state objects يرافقان كل agent/tool call، مع observability correlation IDs.	L0/L8/L9/L10/L11	The Agent Operating System (AOS): A Reference Operating Architecture for Distributed Agentic Systems	09
اربط Skill Registry بـusage telemetry وworkflow mining لإعلاء candidate HyperSkills من المسارات المتكررة.	L1/L3/L5/L8/L9	AgentOS: From Application Silos to a Natural Language-Driven Data Ecosystem	10
عرّف Host Adapter capability manifest: native API، UIA، vision، terminal، isolation mode، rollback support.	L8/L9/L10	UFO2: The Desktop AgentOS	11
اجعل RoleCouncil agents تملك input/output contract موحد لـ progress ledger بدلاً من chat-only coordination.	L7/L8/L9/L10	Magentic-One: A Generalist Multi-Agent System for Solving Complex Tasks	12
أضف HyperTool Contract عن AgentCard وعن AuthorizationToken، واربطها version/capability negotiation.	L8/L9/L10	A Survey of Agent Interoperability Protocols: MCP, ACP, A2A, and ANP	13
أضف Capability Router يختار agent/tool مع specialist grounding/planning confidence/fallback وcost budget.	L2/L6/L8/L9	Agent S2: A Compositional Generalist-Specialist Framework for Computer Use Agents	14
اجمع HyperLoop branching فقط في reversible/sandbox state، merge بعد World-Effect/Evidence Court.	L2/L7/L10	The Unreasonable Effectiveness of Scaling Agents for Computer Use	15
عرّف TraceDataset schema في MOUSSAIFXT: observation، action، effect، tool، authority، failure، verification، ski...	L9/L11	OpenCUA: Open Foundations for Computer-Use Agents	16
وحد EventState بين GUI actions وshell/files، اجعل verifier يقرأ artifact state لا screenshot فقط.	L8/L9/L10/L11	UI-TARS-2 Technical Report: Advancing GUI Agent with Multi-Turn Reinforcement Learning	17
استخدم unified EnvAdapter contract مع reset/snapshot/evaluate، ثم شغّل HyperSkills على بيئات متعددة.	L9/L11	AgentGym: Evolving Large Language Model-based Agents across Diverse Environments	18
أنشئ BenchmarkAdapter interface موحداً وRunManifest يسجل model، harness، tools، environment snapshot، seeds.	L9/L10/L11	The BrowserGym Ecosystem for Web Agent Research	19
اجعل كل task KPI يملك effect predicate machine-checkable متى أمكن، وعجل repeated-run reliability.	L0/L9/L10	tau-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains	20
أضف ActorId/EventOrder لكل state mutation، human action يصبح event موثقاً في نفس mission ledger.	L0/L8/L10	tau^2-Bench: Evaluating Conversational Agents in a Dual-Control Environment	21
وسم provenance/trust لكل tool payload، router يمنع انتقال data إلى instruction channel دون explicit bridge.	L0/L4/L9/L10	AgentDojo: A Dynamic Environment to Evaluate Prompt Injection Attacks and Defenses for LLM Agents	22
أضف RiskState إلى MissionState، مع hard gates للأثر العالي، وseparate safety regression suite.	L0/L9/L10	Agent-SafetyBench: Evaluating the Safety of LLM Agents	23
اجمع AutoHarnessLab يوليد candidates فقط داخل typed component grammar، كل candidate يحتاج ablation+holdout+roll...	L7/L8/L11	Automated Design of Agentic Systems	24
لا تعديل canonical kernel in-place: fork candidate، sandbox، eval، court، signed promotion، reversible deployment.	L10/L11	Darwin Godel Machine: Open-Ended Evolution of Self-Improving Agents	25
أنشئ FrontierQuestion queue من gaps/failures مع novelty/impact score، ثم اربط كل candidate بتجربة قابلة للهدم.	L1/L2/L11	AgentEvolver: Towards Efficient Self-Evolving Agent System	26
أضف SkillUtility score من safety×cost×success×reuse، وprune skills ذات redundancy أو negative transfer.	L5/L6/L7/L11	Reinforcement Learning for Self-Improving Agent with Skill Library (SAGE)	27
عرّف E-kernel < E-router < E-loop < E-EvolutionAuthority tiers: E-skill، آخر المستويات تتطلب explicit hum...	L0/L1/L11	Self-Evolving Software Agents	28
اعتمد RuntimeAdapter موحداً مع snapshot، execute، observe، artifact-export، cleanup، permission profile.	L8/L9/L10	OpenHands: An Open Platform for AI Software Developers as Generalist Agents	29
استخدم local coder كـ specialist kernel خلف نفس Tool/State contracts، لا تمنحه canonical writeback دون independent...	L8/L9/L11	Qwen3-Coder-Next Technical Report	30
اجعل TerminalBench-like tasks جزءاً من Promotion Suite لكل kernel/harness version مع cost/retry telemetry.	L9/L10/L11	Terminal-Bench 2.0: Benchmarking Agents on Hard, Realistic Tasks in Command Line Interfaces	31
أضف ComputeBudget object مشتقاً من MCD/QMCD: token، branches، tool calls، verifier passes، deadline.	L2/L8	s1: Simple Test-Time Scaling	32
حوّل HyperAbility إلى retrieval بـ budget، stop condition، ومرر كل evidence إلى source ranking قبل writeback.	L2/L4/L8	Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning	33
اجمع hidden/opaque model reasoning لكن اجعل decisions externally justified عبر evidence، actions، state deltas، ...	L2/L10	Training Large Language Models to Reason in a Continuous Latent Space (Coconut)	34
كل Court عالي الأثر يمكن أن يملك reasoning verifier منفصلاً + deterministic checks + evidence retrieval.	L10	Process Reward Models That Think (ThinkPRM)	35
قسّم Trace إلى ObservedTrace (tools/state) وNarrativeTrace (model rationale)؛ لا تسمح للثاني بإثبات الأول.	L10/L11	Reasoning Models Don't Always Say What They Think	36
اجعل ModelRoute = provider/model/mode/budget/context/tool-policy بدلاً من model name فقط.	L2/L8	Qwen3 Technical Report	37
اختبر K2-like backbone خلف نفس MOUSSAIXT contracts، لا تدمج model-specific quirks في canonical kernel.	L8/L9	Kimi K2: Open Agentic Intelligence	38
أضف benchmark profile لكل model: reasoning/coding/tool/GUI strengths + cost + context + safety، ويختار router وف...	L2/L8/L9	GLM-4.5: Agentic, Reasoning, and Coding (ARC) Foundation Models	39
Router يقرر jointly model، budget، context، tool-policy، external fetch/compact/context-window؛ يبقى المصادر لا prompt الضخم.	L2/L4/L8	MiniMax-M1: Scaling Test-Time Compute Efficiently with Lightning Attention	40

HyperHarness vNext - إسقاط تنفيذي مستخلص من الأدبيات

هذا projection لا يغير النظرية الأصلية؛ يعيد تجميع L0-L11 كطبقة runtime قابلة للتنفيذ والاختبار. الهدف أن يكون MOUSSAIFXT portable across models/hosts لأن الـharness يحتفظ بالـstate/contracts/evidence بدلاً من ربطها بالـmodel session.

H0 · Constitutional Envelope
policy, authority tokens, risk state, non-bypassable gates. Maps mainly to L0.

H1 · Mission / Task-State Store
goal, governing question, progress ledger, open questions, owner, SLA. L1 + M-TAOS.

H2 · Memory Bus
MemoryObject/MemCube-like objects, promotion, eviction, contradiction, provenance. L3/L11.

H3 · Skill & Ability Registry
HyperSkill contracts, trust tiers, compatibility, tests, signatures, composition. L5/L6.

H4 · Compute & Context Router
model/mode/budget, retrieval budget, branch budget, verifier budget. L2/L4/L8.

H5 · Agent / Role Orchestrator
roles, delegation chains, ownership, replanning, shared ledger. L8.

H6 · Environment Adapter Plane
browser, GUI, UIA, API, terminal, files, sandbox, snapshots. L9.

H7 · Tool Contract Plane
typed schemas, MCP/A2A capability boundaries, trust labels, idempotency. L9.

H8 · Bounded Rollout / HyperLoop Engine
single/parallel rollouts, branch-falsify-merge, stop/retry/recovery. L7.

H9 · Court / Verifier Mesh
deterministic checks, generative verifier, safety court, source court, independent auditor. L10.

H10 · World-Effect Commit Layer
tentative action -> observe delta -> audit -> commit/rollback. L10/L11.

H11 · Evolution & Promotion Lab
failure traces, harness deltas, skill learning, self-evolution sandbox, regression gates. L11.

Mission -> Govern -> Route -> Retrieve -> Compose Skill/Ability -> Execute -> Observe -> Audit -> Commit -> Learn/Promote

الاختلاف الحاسم عن agent loop بسيط: completion ليس نهاية generation؛ هو state transition accepted by an independent verification contract. والـself-improvement ليس edit مباشر؛ هو candidate lineage يمر عبر regression + rollback.

برنامج الاختبار - تحويل 40 ورقة إلى 20 تجربة حاسمة

كل تجربة مصممة لتجيب عن سؤال واحد: هل يضيف mechanism الخارجي فرقاً قابلاً للقياس داخل MOUSSAIFXT عندما نثبت أكبر قدر ممكن من model/data/compute؟

EXP2-01 · Externalized State

M-TAOS current vs Manage-Execute-Audit؛ false-success traps؛ tasks >30 steps، same model.

EXP2-02 · Memory Lifecycle

Flat retrieval vs MemoryOS/MemOS-style lifecycle تحت token budget متساوٍ.

EXP2-03 · Memory Actions

Rule-based vs learned store/retrieve/update/discard policy مع protected evidence.

EXP2-04 · Skill Certification

Curated vs self-generated HyperSkills؛ malicious skill injection + trust tiers.

EXP2-05 · Harness-only Gain

ثبّت model وغيّر HyperHarness؛ افصل gain عن foundation model.

EXP2-06 · Harness Evolution

RHO-like candidate deltas + independent court + frozen regression suite.

EXP2-07 · Authority Delegation

multi-agent workflow مع revoke أثناء التنفيذ؛ افحص capability leakage.

EXP2-08 · OS Adapter Mix

GUI-only vs API/UIA/vision/terminal hybrid مع same planner.

EXP2-09 · Branch Scaling

N rollouts + behavior judge مقابل single trajectory تحت cost curve.

EXP2-10 · Dual-Control State

human/user + agent mutations في shared state مع event attribution.

EXP2-11 · Prompt Injection

untrusted tool/RAG payloads؛ test objective hijack وtool misuse.

EXP2-12 · Safety as Runtime

prompt-only defense vs RiskState + hard gates + effect verifier.

EXP2-13 · AutoHarness Search

ADAS grammar search؛ held-out domain/model transfer + complexity penalty.

EXP2-14 · Self-Modification

forked kernel candidate؛ sandbox + constitution freeze + rollback drill.

EXP2-15 · Skill RL Transfer

SAGE-like sequential tasks؛ measure reuse, redundancy, negative transfer.

EXP2-16 · Terminal World Effect

Terminal-Bench-like suite؛ no success until tests verify artifact state.

EXP2-17 · Compute Router

MCD/QMCD-routed thinking/branch/search/verifier budgets vs fixed budgets.

EXP2-18 · Learned Retrieval

fixed RAG vs search-as-action؛ source quality + tool budget controlled.

EXP2-19 · CoT Audit Gap

compare NarrativeTrace with ObservedTrace under hidden causal factors.

EXP2-20 · Model Swap

Qwen3/K2/GLM/M1/coder specialist behind identical contracts; route by verified utility/cost.

Minimum protocol: frozen baseline + equal budget where possible + repeated seeds/runs + negative control + failure injection + held-out tasks + safety regression + artifact/effect verification + provenance of every config.

الفجوات التي تبقى بعد هذه الأربعين - وما الذي يجب أن يثبته MOUSSAIF

Verifier Independence
هل court قادر على اكتشاف نفس الخطأ الذي يصنعه prover/model family؟

Verified State Fidelity
هل state canonical مطابق فعلاً للعالم بعد 50-200 خطوة، أم narrative summary فقط؟

Harness Regression
كيف نثبت أن harness update لم يفسد domains أو safety نادرة؟

Skill Supply Chain
كيف نوقع/نثق/نعزل/نلغي skill دون قتل portability؟

Cross-Host Semantics
هل نفس Tool/State contract ينتج نفس semantics على cloud/local/browser/desktop؟

Authority Semantics
كيف تنتقل السلطة عبر delegation, retries, parallel branches وresume؟

Self-Evolution Stability
متى يتحول adaptation إلى drift في goals/identity/kernel؟

Long-Horizon Recovery
هل النظام يستأنف من verified checkpoint بدل إعادة المهمة أو تكرار effect؟

Open-World Reward
كيف نتحقق عندما لا توجد tests كاملة أو goal state مغلقة؟

Benchmark-Harness Confounding
هل model leaderboard يخلط model مع proprietary harness؟

Compute Allocation
متى نستعمل large context, reasoning, search, branch, verifier أو specialist model؟

Trace Faithfulness
كيف نفصل observed causal trace عن narrative/CoT غير الموثوق؟

Promotion ladder

Rank	المطلوب	ما لا يكفي
P0 · Named	تعريف + parent lineage + boundary	اسم أو diagram فقط
P1 · Contracted	input/output/state/failure/authority/stop	prompt طويل
P2 · Implemented	artifact runnable + trace + tests	generated file غير مشغل
P3 · Ablated	frozen baseline + equalized budget + component removal	before/after غير مضبوط
P4 · Falsified/Stress-tested	negative controls + adversarial + failure injection	success cases فقط
P5 · Replicated/Transferred	different tasks/models/hosts + repeated runs	leaderboard واحد
P6 · Promoted	signed canonical version + regression suite + rollback	evidence بلا auto-deploy

إغلاق المجال هنا يعني: لم يعد المطلوب إضافة طبقة مصطلحات جديدة. الأولوية البحثية تتحول إلى state correctness, harness ablation, authority propagation, verified world effect, skill certification, self-evolution stability, cross-model/host transfer. هذه هي نقاط الاختبار التي ستحدد إن كانت MOUSSAIF architecture تتحول من specification-rich إلى experimentally strong.

المراجع الأولية - Papers 40

هذه الصفحة تسجل الجذور التي بُنيت عليها capsules. البيانات التفصيلية داخل الصفحات مختصرة ومُعاد صياغتها؛ الـMOUSSAIF bridges هي synthesis هندسي لهذا الكتاب وليست ادعاءات صادرة عن المؤلفين الخارجيين.

01. Memory OS of AI Agent
Jiazheng Kang, Mingming Ji, Zhe Zhao, Ting Bai (2025).
https://arxiv.org/abs/2506.06326

02. MemOS: An Operating System for Memory-Augmented Generation (MAG) in Large Language Models
Zhiyu Li et al. (2025).
https://arxiv.org/abs/2505.22101

03. A-MEM: Agentic Memory for LLM Agents
Wujiang Xu et al. (2025).
https://arxiv.org/abs/2502.12110

04. Agentic Memory: Learning Unified Long-Term and Short-Term Memory Management for Large Language Model Agents
Yi Yu et al. (2026).
https://arxiv.org/abs/2601.01885

05. SoK: Agentic Skills - Beyond Tool Use in LLM Agents
Yanna Jiang et al. (2026).
https://arxiv.org/abs/2602.20867

06. Code as Agent Harness
Xuying Ning et al. (2026).
https://arxiv.org/abs/2605.18747

07. Evolving Agents in the Dark: Retrospective Harness Optimization via Self-Preference
Wenbo Pan et al. (2026).
https://arxiv.org/abs/2606.05922

08. LongHorizon-Harness: Advancing Long-Horizon Agents for Real-World Tasks
Ziyu Ma et al. (2026).
https://arxiv.org/abs/2608.01964

09. The Agent Operating System (AOS): A Reference Operating Architecture for Distributed Agentic Systems
Ankur Sharma, Deep Shah (2026).
https://arxiv.org/abs/2608.03214

10. AgentOS: From Application Silos to a Natural Language-Driven Data Ecosystem
Rui Liu et al. (2026).
https://arxiv.org/abs/2603.08938

11. UFO2: The Desktop AgentOS
Chaoyun Zhang et al. (2026).
https://arxiv.org/abs/2504.14603

12. Magentic-One: A Generalist Multi-Agent System for Solving Complex Tasks
Adam Fourney et al. (2024).
https://arxiv.org/abs/2411.04468

13. A Survey of Agent Interoperability Protocols: MCP, ACP, A2A, and ANP
Abul Ehtesham et al. (2025).
https://arxiv.org/abs/2505.02279

14. Agent S2: A Compositional Generalist-Specialist Framework for Computer Use Agents
Saaket Agashe et al. (2025).
https://arxiv.org/abs/2504.00906

15. The Unreasonable Effectiveness of Scaling Agents for Computer Use
Gonzalo Gonzalez-Pumariega et al. (2025/2026).
https://arxiv.org/abs/2510.02250

16. OpenCUA: Open Foundations for Computer-Use Agents
Xinyuan Wang et al. (2025).
https://arxiv.org/abs/2508.09123

17. UI-TARS-2 Technical Report: Advancing GUI Agent with Multi-Turn Reinforcement Learning
ByteDance Seed et al. (2025).
https://arxiv.org/abs/2509.02544

18. AgentGym: Evolving Large Language Model-based Agents across Diverse Environments
Zhiheng Xi et al. (2024).
https://arxiv.org/abs/2406.04151

19. The BrowserGym Ecosystem for Web Agent Research
Thibault Le Sellier De Chezelles et al. (2024/2025).
https://arxiv.org/abs/2412.05467

20. tau-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains
Shunyu Yao et al. (2024).
https://arxiv.org/abs/2406.12045

21. tau^2-Bench: Evaluating Conversational Agents in a Dual-Control Environment
Victor Barres et al. (2025).
https://arxiv.org/abs/2506.07982

22. AgentDojo: A Dynamic Environment to Evaluate Prompt Injection Attacks and Defenses for LLM Agents
Edoardo Debenedetti et al. (2024).
https://arxiv.org/abs/2406.13352

23. Agent-SafetyBench: Evaluating the Safety of LLM Agents
Zhexin Zhang et al. (2024).
https://arxiv.org/abs/2412.14470

24. Automated Design of Agentic Systems
Shengran Hu, Cong Lu, Jeff Clune (2024/2025).
https://arxiv.org/abs/2408.08435

25. Darwin Godel Machine: Open-Ended Evolution of Self-Improving Agents
Jenny Zhang et al. (2025).
https://arxiv.org/abs/2505.22954

26. AgentEvolver: Towards Efficient Self-Evolving Agent System
Yunpeng Zhai et al. (2025).
https://arxiv.org/abs/2511.10395

27. Reinforcement Learning for Self-Improving Agent with Skill Library (SAGE)
Jiongxiao Wang et al. (2025).
https://arxiv.org/abs/2512.17102

28. Self-Evolving Software Agents
Marco Robol, Paolo Giorgini (2026).
https://arxiv.org/abs/2604.27264

29. OpenHands: An Open Platform for AI Software Developers as Generalist Agents
Xingyao Wang et al. (2024/2025).
https://arxiv.org/abs/2407.16741

30. Qwen3-Coder-Next Technical Report
Qwen Team (2026).
https://arxiv.org/abs/2603.00729

31. Terminal-Bench 2.0: Benchmarking Agents on Hard, Realistic Tasks in Command Line Interfaces
Mike A. Merrill et al. (2026).
https://arxiv.org/abs/2601.11868

32. s1: Simple Test-Time Scaling
Niklas Muennighoff et al. (2025).
https://arxiv.org/abs/2501.19393

33. Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning
Bowen Jin et al. (2025).
https://arxiv.org/abs/2503.09516

34. Training Large Language Models to Reason in a Continuous Latent Space (Coconut)
Shibo Hao et al. (2024/2025).
https://arxiv.org/abs/2412.06769

35. Process Reward Models That Think (ThinkPRM)
Muhammad Khalifa et al. (2025).
https://arxiv.org/abs/2504.16828

36. Reasoning Models Don't Always Say What They Think
Yanda Chen et al. (2025).
https://arxiv.org/abs/2505.05410

37. Qwen3 Technical Report
Qwen Team (2025).
https://arxiv.org/abs/2505.09388

38. Kimi K2: Open Agentic Intelligence
Kimi Team (2025).
https://arxiv.org/abs/2507.20534

39. GLM-4.5: Agentic, Reasoning, and Coding (ARC) Foundation Models
GLM-4.5 Team (2025).
https://arxiv.org/abs/2508.06471

40. MiniMax-M1: Scaling Test-Time Compute Efficiently with Lightning Attention
MiniMax Team (2025).
https://arxiv.org/abs/2506.13585

Research cutoff: 31 Aug 2026. Preprints الحديثة وtechnical reports موسومة بحسب حالتها. لا يُعتبر ظهور paper في هذا atlas تصديقاً لكل claims أو أرقامها؛ المرجع الأولي هو الورقة نفسها، والاختبار المحلي هو شرط الترقية داخل MOUSSAIF.