MOUSSAIF FRONTIER REASONING & AGENTIC AGI 2026

الجزء الثالث - حالة الفن القصوى في نماذج الاستدلال والوكلاء

الكتاب التقني المكثف لحالة الفن: من Ultra Reasoning Models إلى Frontier Agentic Systems

STATE OF THE ART • REASONING • TEST-TIME COMPUTE • RLVR • VERIFIERS • AGENTS • MEMORY • COMPUTER USE • MCP • A2A

Research cutoff: 18 August 2026

تركيب علمي جديد مكمل لكتاب MOUSSAIF Unified AGI وAll-vs-All Agentic OS Atlas. لا يكرر النظرية الداخلية، بل يضعها أمام 45 مصدراً خارجياً حديثاً ويستخرج الفروق الهندسية القابلة للاختبار.

الميثاق الإبستيمي وحدود الادعاء

Evidence-first / Claim-ranked

هذا المجلد لا يعامل عبارة **near-AGI** كقياس علمي مثبت. المقصود بها اتجاه هندسي: ارتفاع القدرة على حل مهام جديدة، الاستدلال تحت ميزانية متغيرة، استعمال الأدوات، الاستمرار عبر أفق زمني طويل، وإغلاق الحلقة بين القرار والأثر. لذلك نفصل بين أربع رتب: **Established** لما تدعمه ورقة أو قياس موثق؛ **Reported** لما تبلغه جهة مطورة عن نموذجها؛ **Inference** لما نستنتجه عبر عدة مصادر؛ و**MOUSSAIF Hypothesis** لما نقترحه كاختبار داخلي.

قاعدة الهدم

- أي مصطلح لا يغيّر قراراً أو تجربة يُخفض إلى وصف.
- أي تفوق بلا baseline ومكرر لا يرقى إلى E5.
- أي benchmark مشبع أو ملوث لا يعامل كدليل عمومية.
- القدرة لا تساوي الاستقلالية، والاستقلالية لا تساوي السلامة.

قاعدة المقارنة

- لا نقارن أرقام vendors مباشرة ما لم يكن harness متطابقاً.
- نفرّق بين model capability وagent harness capability.
- نذكر test-time compute كميزانية، لا كسحر.
- نعتبر tool success أثراً يحتاج post-action verification.

```
EvidenceScore = SourceQuality × Reproducibility × HarnessParity × ExternalValidity
```

Source ledger / دفتر المصدر

S01 · Official · 2026-07-09
GPT-5.6 System Card
Official system card; Sol/Terra/Luna, reasoning effort, agentic safety, prompt-injection and self-improvement evaluations.

S02 · Official · 2026-06-09
Claude Fable 5 & Mythos 5 System Card
Official system card for Anthropic's most capable widely released Fable 5 class; always-on adaptive thinking, long-running agents, 1M context and safety evaluations.

S03 · Official · 2026-05-19
Gemini 3.5: frontier intelligence with action
Official model report; action-oriented workflows, agentic coding, subagents and benchmark reporting.

S06 · Research · 2026
DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence
arXiv:2606.19348; hybrid attention, million-token context, effort scaling, agentic benchmarks.

S29 · Research · 2026
The 2025 AI Agent Index: Documenting Technical and Safety Features of Deployed Agentic AI Systems
arXiv:2602.17753; deployed-agent capability/safety documentation.

S15 · Research · 2026
Inference Scaling, Learning to Reason, and Agentic Systems
arXiv:2504.09037v4; synthesis of inference scaling, verifiers, reasoning learning and agentic systems.

S18 · Research · 2024
Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters
arXiv:2408.03314; adaptive compute allocation and verifier-guided search.

S20 · Research · 2025
100 Days After DeepSeek-R1: A Survey on Replication Studies and More Directions for Reasoning Language Models
arXiv:2505.00551; replication lessons for SFT and RLVR.

S21 · Research · 2025
DeepSeek-R1 Thoughtology: Let's <think> about LLM Reasoning
arXiv:2504.07128; analysis of reasoning length, rumination, control and safety.

S23 · Official · 2025-12-18
Evaluating Chain-of-Thought Monitorability
Official framework with intervention/process/outcome-property evaluations of CoT monitorability.

Deep technical extraction / الاستخراج الهندسي

Operational discipline. أي ادعاء frontier يجب أن يحمل خمسة حقول: model، provenance/version، harness، budget، evaluation date. غياب أحدها يجعل المقارنة غير قابلة لإعادة البناء. نعامل vendor system cards كدليل أولي ممتاز على design intent والسلامة والاختبارات المنشورة، لكن claims المقارنة تبقى *reported* إلى أن تتكرر على harness مستقل.

- **Measurement unit:** task trajectory لا الجواب النهائي فقط.
- **Control:** نفس الأدوات والوقت وعدد المحاولات عند المقارنة.
- **Disproof:** ablation يزيل طبقة واحدة ويبحث عن تغير موثوق.
- **Near-AGI signal:** transfer + recovery + world grounding، لا score مفرد.

Research specification / مواصفة الاختبار

Variable / layer	Measure	Failure signal
Statistics	CI/effect size	leaderboard noise
Replication	independent run	single-lab claim
Claim rank	established/reported/inference	marketing conflation
Snapshot control	exact version/date	moving target
Budget parity	tokens/tools/retries equalized	unfair ranking
Pass@1 vs best-of-N	both reported	hidden search budget
Contamination	private/dynamic tests	memorization

الخلاصة التنفيذية: أين تتحرك الجبهة؟

Four coupled frontiers

تتحرك الجبهة في 2026 على أربعة محاور مترابطة. الأول هو **inference-time intelligence**: النموذج لا ينتج جواباً واحداً، بل يوزع compute، يولد مسارات، يقيّمها ويعيد التخطيط. الثاني هو **reasoning-trained post-training** عبر RLVR وreward models وcurricula توسع أنماط الحل. الثالث هو **tool-grounded reasoning** حيث يصبح الكود، البحث، الواجهات والبيئة جزءاً من التفكير نفسه. الرابع هو **persistent agency**: ذاكرة، state، مهام ممتدة، استئناف، تحقق، وتنسيق عدة وكلاء.

الاقتراب من AGI لا يظهر هنا كقفزة أحادية في parameter count، بل كتحول من دالة تقريب نصي إلى **systemic cognition**: نموذج + tools + memory + verifier + budget controller + execution harness + safety policy + environment. هذه هي المنطقة التي يلتقي فيها corpus الخارجي مع MOUSSAIF: الفرق الحاكم، bounded HyperSkills، Vascular Compute، MCD/QMCD، Falsification Court وHyperLoops، World-Effect Loop.

الاستنتاج المركزي: أكثر تحسنات 2025-2026 إثارة ليست "thinking longer" وحدها، بل القدرة على تخصيص التفكير، مراجعته، تحويله إلى فعل، ثم الاحتفاظ بحالة تنفيذ يمكن استعادتها والتحقق منها.

Source ledger / دفتر المصدر

S01 · Official · 2026-07-09
GPT-5.6 System Card
Official system card; Sol/Terra/Luna, reasoning effort, agentic safety, prompt-injection and self-improvement evaluations.

S03 · Official · 2026-05-19
Gemini 3.5: frontier intelligence with action
Official model report; action-oriented workflows, agentic coding, subagents and benchmark reporting.

S06 · Research · 2026
DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence
arXiv:2606.19348; hybrid attention, million-token context, effort scaling, agentic benchmarks.

S08 · Research · 2025
DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning
arXiv:2501.12948; RLVR-style reasoning emergence, cold start, distillation.

S12 · Research · 2026
Kimi K3: Open Frontier Intelligence
arXiv:2607.24653; 2.8T MoE / 104B active, native vision, 1M context, reasoning-effort levels, long-horizon agentic RL and open weights.

S30 · Research · 2026
Benchmark Test-Time Scaling of General LLM Agents
arXiv:2602.18998; sequential vs parallel scaling across search, coding, reasoning and tools.

S31 · Research · 2026
The Long-Horizon Task Mirage? Diagnosing Where and Why Agents Fail as Task Horizon Increases
arXiv:2604.11978; cross-domain degradation with horizon length.

S15 · Research · 2026
Inference Scaling, Learning to Reason, and Agentic Systems
arXiv:2504.09037v4; synthesis of inference scaling, verifiers, reasoning learning and agentic systems.

S18 · Research · 2024
Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters
arXiv:2408.03314; adaptive compute allocation and verifier-guided search.

S20 · Research · 2025
100 Days After DeepSeek-R1: A Survey on Replication Studies and More Directions for Reasoning Language Models
arXiv:2505.00551; replication lessons for SFT and RLVR.

Deep technical extraction / الاستخراج الهندسي

Operational discipline. أي ادعاء frontier يجب أن يحمل خمسة حقول: model، provenance/version، harness، budget، evaluation date. غياب أحدها يجعل المقارنة غير قابلة لإعادة البناء. نعامل vendor system cards كدليل أولي ممتاز على design intent والسلامة والاختبارات المنشورة، لكن claims المقارنة تبقى *reported* إلى أن تتكرر على harness مستقل.

- **Measurement unit:** task trajectory لا الجواب النهائي فقط.
- **Control:** نفس الأدوات والوقت وعدد المحاولات عند المقارنة.
- **Disproof:** ablation يزيل طبقة واحدة ويبحث عن تغير موثوق.
- **Near-AGI signal:** transfer + recovery + world grounding، لا score مفرد.

Research specification / مواصفة الاختبار

Variable / layer	Measure	Failure signal
Safety	impact/reversibility tests	unsafe success
Generalization	cross-domain held-outs	benchmark specialization
Statistics	CI/effect size	leaderboard noise
Replication	independent run	single-lab claim
Claim rank	established/reported/inference	marketing conflation
Snapshot control	exact version/date	moving target
Budget parity	tokens/tools/retries equalized	unfair ranking

خريطة corpus: 45 جذراً جديداً

24 reasoning + 21 agentic roots

Part I — Frontier Reasoning

24 مصدراً تغطي النماذج الأعلى، process reward، test-time scaling، RL post-training، CoT monitoring، long-context، multimodal reasoning، reasoning budgets، open-weight frontier.

GPT-5.6, Claude Fable 5/Opus 5, Gemini 3.5/3.1 Deep Think, Grok 4.6, DeepSeek V4/V4-Pro/R1, Qwen3.8-Max/Coder/Omni, Kimi K3, gpt-oss، ثم أوراق inference scaling وPRM والmonitorability.

Part II — Frontier Agentic Systems

21 مصدراً تغطي computer use، science agents، deep research، deployed agents، coding evolution، web agents، long-horizon memory، monitoring، benchmark standardization، MCP، وA2A.

ChatGPT agent, Deep Research, Co-Scientist, Agent Index, General AgentBench, HORIZON, UltraHorizon, OSWorld 2.0, Odysseys, SWE-EVO, AMA-Bench, SAM SentinelBench, CUBE, MCP, A2A.

لا تدخل أوراق corpus السابق ذي الـ30 جذراً في عداد هذه الـ45؛ هذا مجلد توسعة مستقل يركز على frontier 2025-2026.

Source ledger / دفتر المصدر

S01 · Official · 2026-07-09
GPT-5.6 System Card
Official system card; Sol/Terra/Luna, reasoning effort, agentic safety, prompt-injection and self-improvement evaluations.

S45 · Research · 2026
Agents in the Wild: Where Research Meets Deployment
arXiv:2607.19336; deployment-oriented synthesis of reliability, memory, tools and evaluation.

S15 · Research · 2026
Inference Scaling, Learning to Reason, and Agentic Systems
arXiv:2504.09037v4; synthesis of inference scaling, verifiers, reasoning learning and agentic systems.

S18 · Research · 2024
Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters
arXiv:2408.03314; adaptive compute allocation and verifier-guided search.

S20 · Research · 2025
100 Days After DeepSeek-R1: A Survey on Replication Studies and More Directions for Reasoning Language Models
arXiv:2505.00551; replication lessons for SFT and RLVR.

S21 · Research · 2025
DeepSeek-R1 Thoughtology: Let's <think> about LLM Reasoning
arXiv:2504.07128; analysis of reasoning length, rumination, control and safety.

S23 · Official · 2025-12-18
Evaluating Chain-of-Thought Monitorability
Official framework with intervention/process/outcome-property evaluations of CoT monitorability.

S24 · Official · 2025-03-10
Detecting Misbehavior in Frontier Reasoning Models
Official research on monitoring reasoning traces for reward hacking and exploit behavior.

S29 · Research · 2026
The 2025 AI Agent Index: Documenting Technical and Safety Features of Deployed Agentic AI Systems
arXiv:2602.17753; deployed-agent capability/safety documentation.

S30 · Research · 2026
Benchmark Test-Time Scaling of General LLM Agents
arXiv:2602.18998; sequential vs parallel scaling across search, coding, reasoning and tools.

Deep technical extraction / الاستخراج الهندسي

Operational discipline. أي ادعاء frontier يجب أن يحمل خمسة حقول: model، provenance/version، harness، budget، وevaluation date. غياب أحدها يجعل المقارنة غير قابلة لإعادة البناء. نعامل vendor system cards كدليل أولي ممتاز على design intent والسلامة والاختبارات المنشورة، لكن claims المقارنة تبقى *reported* إلى أن تتكرر على harness مستقل.

- **Measurement unit:** task trajectory لا الجواب النهائي فقط.
- **Control:** نفس الأدوات والوقت وعدد المحاولات عند المقارنة.
- **Disproof:** ablation يزيل طبقة واحدة ويبحث عن تغير موثوق.
- **Near-AGI signal:** transfer + recovery + world grounding، لا score مفرد.

Research specification / مواصفة الاختبار

Variable / layer	Measure	Failure signal
Process traces	failure-stage labels	outcome-only blindness
Cost	latency/$/tool calls	capability without economics
Safety	impact/reversibility tests	unsafe success
Generalization	cross-domain held-outs	benchmark specialization
Statistics	CI/effect size	leaderboard noise
Replication	independent run	single-lab claim
Claim rank	established/reported/inference	marketing conflation

مشهد النماذج العليا في أغسطس 2026
Frontier model landscape

العائلة	الفكرة الهندسية الأبرز	ما يجب قياسه
GPT-5.6	reasoning effort curves + tool/computer-use safeguards	effort-efficiency, action correctness, monitorability
Claude Fable 5 / Opus 5	adaptive thinking, 1M context, long-horizon agentic coding; Opus adds effort ladder through max	sustained reasoning, delegation/verification behavior, cost-quality frontier
Gemini 3.5 / Deep Think	frontier intelligence with action + science high-compute mode	agentic throughput, multimodal reasoning, subagent utility
Grok 4.6	long-running agentic RL + coding/interactive workflows + xhigh effort	long-trajectory reliability, self-testing, coding/knowledge-work agent performance
DeepSeek V4 / V4-Pro	1M context, hybrid/sparse attention efficiency, reasoning effort, agent-oriented deployment	context fidelity, open-model efficiency, long-horizon/tool generalization
Qwen3.8-Max family	frontier multimodal MoE + coding/cowork; Coder-Next and Omni specialization	agentic coding, multimodal grounding, long-context efficiency and budget calibration
Kimi K3	2.8T open MoE + native vision + 1M context + long-horizon agentic RL	open frontier scaling efficiency, persistent agent execution, multimodal coding/reasoning

الأرقام المعلنة عبر جهات مختلفة ليست leaderboard موحداً، نعالجها كـreported evidence ونركز على الأنماط المعمارية.

Source ledger / دفتر المصدر

S01 · Official · 2026-07-09
GPT-5.6 System Card
Official system card; Sol/Terra/Luna, reasoning effort, agentic safety, prompt-injection and self-improvement evaluations.

S02 · Official · 2026-06-09
Claude Fable 5 & Mythos 5 System Card
Official system card for Anthropic's most capable widely released Fable 5 class; always-on adaptive thinking, long-running agents, 1M context and safety evaluations.

S03 · Official · 2026-05-19
Gemini 3.5: frontier intelligence with action
Official model report; action-oriented workflows, agentic coding, subagents and benchmark reporting.

S04 · Official · 2026-02
Gemini 3.1 Deep Think / Accelerating Mathematical and Scientific Discovery
Official research/model documentation for specialized high-compute scientific reasoning.

S05 · Official · 2026-08-12
Introducing Grok 4.6
Official current flagship release; long-running agents, coding, interactive/visual work, agentic RL, configurable reasoning effort and 500k context.

S06 · Research · 2026
DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence
arXiv:2606.19348; hybrid attention, million-token context, effort scaling, agentic benchmarks.

S07 · Research · 2026-08-13
DeepSeek V4-Pro-0813 launch and agent capability update
Current-release reporting on V4 Pro availability and improved AI-agent capabilities; used as deployment-state evidence, not as an independent technical paper.

S09 · Official · 2026-08-02
Qwen3.8-Max: A New Bar for Coding and Cowork
Official current flagship report; 2.4T-class MoE, multimodal/long-context frontier, coding and cowork/agentic work. Vendor benchmark claims are treated as reported.

S10 · Research · 2026
Qwen3-Coder-Next Technical Report
arXiv:2603.00729; repository reasoning, agent interaction patterns, executable training/evaluation environments.

S11 · Research · 2026
Qwen3.5-Omni Technical Report
arXiv:2604.15804; hybrid-attention MoE omnimodal reasoning across text, vision and audio, long-sequence interaction and grounding.

Deep technical extraction / الاستخراج الهندسي

قراءة المشهد. لا توجد جبهة واحدة؛ هناك خمسة محاور متداخلة: inference، post-training-reasoning، multimodal grounding، context/memory efficiency، time scaling، وagentic execution. النموذج قد يتصدر محوراً ويتراجع في آخر. لذلك الأفضل تمثيل كل عائلة بمتجه قدرة لا بترتيب كلي.

```
CapabilityVector = [R, TTS, Tool, Memory, CUA, SWE, Safety, Cost]
```

- افصل **native model capability** عن gains الناتجة من harness.
- قِس hard-tail بدلاً من المتوسط فقط.
- سجِل latency وtokens وtool calls إلى جانب accuracy.
- راقب degradation مع horizon؛ كثير من الفروق لا تظهر في المهام القصيرة.

Research specification / مواصفة الاختبار

Variable / layer	Measure	Failure signal
Safety envelope	injection/destructive-action tests	capability without control
Throughput	tokens/s under target context	benchmark-only speed
Open/closed control	ablation/debug access	hidden harness effects
External replication	independent harness delta	vendor-only comparison
Activated compute	active params / token	capacity-cost confusion
Context fidelity	state retention under noise	attention dilution
KV efficiency	memory/throughput at long context	bandwidth bottleneck

Ultra Reasoning Model تشريح

From token predictor to compute-adaptive solver

نموذج الاستدلال المتقدم يمكن تمثيله بخمس طبقات لا تظهر كلها في weights: **Base model** يوفر المعرفة والتمثيل؛ **post-training** يعلّم سياسات حل؛ **budget controller** يحدد مقدار compute؛ **/search verifier layer** يوسع أو يرشح المسارات؛ و**tool layer** يربط الاستدلال ببرهان خارجي أو تنفيذ.

```
y* = argmax_y V(y | x, τ, tools) subject to C_reason + C_tool ≤ B(x)
```

حيث B(x) ليست ثابتة. التقدم الحقيقي يظهر عندما يتعلم النظام أن المسألة السهلة لا تستحق 20k thought tokens، وأن المسألة الصعبة تحتاج فرعاً جديداً أو أداة لا تكراراً لفظياً. هذه النقطة تفسر لماذا بدأت الواجهات الحديثة تعرض درجات **effort** أو budget modes، ولماذا صار البحث في adaptive test-time compute محوراً مستقلاً.

- **Failure 1:** overthinking - compute إضافي يخفض الدقة.
- **Failure 2:** verifier bias - المسار الذي يبدو منطقياً يفوز رغم خطئه.
- **Failure 3:** tool myopia - الأداة تستخدم لأن لها مكافأة لا لأنها تقلل عدم اليقين.
- **Failure 4:** trace gaming - النموذج يتعلم إرضاء monitor بدلاً من حل المهمة.

Deep technical extraction / الاستخراج الهندسي

الآلية العامة. Ultra reasoning model ليس "CoT أطول"؛ هو policy قادرة على اختيار مقدار deliberation، توليد بدائل، استخدام verifier أو reward signal، والتوقف حين تنخفض القيمة الحدية للحوسبة. التحسن الحقيقي يظهر عندما يرتفع الأداء على hard tail أسرع من ارتفاع التكلفة.

```
Stop when ΔExpectedUtility / ΔCompute ≤ λ_cost
```

- **State:** problem representation + partial proof + uncertainty.
- **Action:** continue / branch / verify / tool-call / terminate.
- **Metric:** accuracy-cost frontier لا accuracy منفردة.
- **Failure:** verifier gaming، rumination، أو branch explosion.

Source ledger / دفتر المصدر

S15 · Research · 2026
Inference Scaling, Learning to Reason, and Agentic Systems
arXiv:2504.09037v4; synthesis of inference scaling, verifiers, reasoning learning and agentic systems.

S18 · Research · 2024
Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters
arXiv:2408.03314; adaptive compute allocation and verifier-guided search.

S19 · Research · 2024
Simple and Provable Scaling Laws for the Test-Time Compute of Large Language Models
arXiv:2411.19477; tournament/league aggregation and scaling guarantees under assumptions.

S21 · Research · 2025
DeepSeek-R1 Thoughtology: Let's <think> about LLM Reasoning
arXiv:2504.07128; analysis of reasoning length, rumination, control and safety.

S22 · Research · 2026
Mid-Think: Training-Free Intermediate-Budget Reasoning
arXiv:2601.07036; inference budget control between short and long reasoning modes.

S23 · Official · 2025-12-18
Evaluating Chain-of-Thought Monitorability
Official framework with intervention/process/outcome-property evaluations of CoT monitorability.

S24 · Official · 2025-03-10
Detecting Misbehavior in Frontier Reasoning Models
Official research on monitoring reasoning traces for reward hacking and exploit behavior.

S01 · Official · 2026-07-09
GPT-5.6 System Card
Official system card; Sol/Terra/Luna, reasoning effort, agentic safety, prompt-injection and self-improvement evaluations.

S04 · Official · 2026-02
Gemini 3.1 Deep Think / Accelerating Mathematical and Scientific Discovery
Official research/model documentation for specialized high-compute scientific reasoning.

S08 · Research · 2025
DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning
arXiv:2501.12948; RLVR-style reasoning emergence, cold start, distillation.

Research specification / مواصفة الاختبار

Variable / layer	Measure	Failure signal
Parallelism	wall-clock speedup	merge overhead
Self-consistency	agreement vs correctness	correlated error
Tool escalation	gain after external evidence	unnecessary calls
Uncertainty	calibration / abstention	false certainty
Latency	p50/p95 end-to-end	unusable deep mode
Trace quality	monitorability probes	unfaithful trace
Hard-tail gain	top difficulty subset	average-score illusion

RLVR: لماذا غيّر post-training قواعد اللعبة؟

Reinforcement Learning with Verifiable Rewards

قوة RLVR تأتي من أن reward يمكن حسابه آلياً في مجالات ذات جواب قابل للتحقق: رياضيات، كود، theorem proving، بعض مهام الأدوات. هذا يسمح بتوليد عدد كبير من trajectories ثم دفع السياسة نحو السلاسل التي تنتهي بنتيجة صحيحة. DeepSeek-R1 جعل هذا المسار مركزياً في النقاش العام، بينما أظهرت عائلات لاحقة أن الفكرة يمكن دمجها مع tool-use data، curricula، SFT وRL multimodal.

```
J(θ)=E_{x,y~πθ}[ r_verifiable(x,y) - β·KL(πθ || π_ref) ]
```

لكن reward verifiability لا يعني *reasoning validity*. قد تصل السياسة إلى الجواب الصحيح بمسار هش، أو تستغل checker، أو تتعلم أنماط benchmark. لذلك الجبهة التالية ليست فقط RL أكبر، بل richer rewards: process signals، environment-grounded success، constraints، cost penalties، وadversarial evaluation.

صلة MOUSSAIF: RLVR يقابل "محكمة نتيجة" واحدة. Falsification Court المقترحة تحتاج طبقات أوسع: صحة المصدر، صحة الأدوات، أثر العالم، اتساق الذاكرة، وقابلية الرجوع.

Source ledger / دفتر المصدر

S08 · Research · 2025
DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning
arXiv:2501.12948; RLVR-style reasoning emergence, cold start, distillation.

S16 · Research · 2026
A Survey of Process Reward Models: From Outcome Supervision to Process Supervision
arXiv:2510.08049; PRM data, training, inference-time use and RL.

S17 · Research · 2026
Reward Modeling for Reinforcement Learning-Based LLM Reasoning
arXiv:2602.09305; model-based, rule-based and self-reward designs.

S20 · Research · 2025
100 Days After DeepSeek-R1: A Survey on Replication Studies and More Directions for Reasoning Language Models
arXiv:2505.00551; replication lessons for SFT and RLVR.

S24 · Official · 2025-03-10
Detecting Misbehavior in Frontier Reasoning Models
Official research on monitoring reasoning traces for reward hacking and exploit behavior.

S15 · Research · 2026
Inference Scaling, Learning to Reason, and Agentic Systems
arXiv:2504.09037v4; synthesis of inference scaling, verifiers, reasoning learning and agentic systems.

S21 · Research · 2025
DeepSeek-R1 Thoughtology: Let's <think> about LLM Reasoning
arXiv:2504.07128; analysis of reasoning length, rumination, control and safety.

S22 · Research · 2026
Mid-Think: Training-Free Intermediate-Budget Reasoning
arXiv:2601.07036; inference budget control between short and long reasoning modes.

S23 · Official · 2025-12-18
Evaluating Chain-of-Thought Monitorability
Official framework with intervention/process/outcome-property evaluations of CoT monitorability.

S18 · Research · 2024
Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters
arXiv:2408.03314; adaptive compute allocation and verifier-guided search.

Deep technical extraction / الاستخراج الهندسي

RLVR core. عندما تكون النتيجة قابلة للتحقق آلياً - كود يمر بالاختبارات، جواب رياضي، أو constraint واضح - يمكن تحويل reasoning إلى policy optimization بدلاً من imitation فقط. الجبهة البحثية هي تصميم rewards لا تكافئ shortcuts، وتوزيع difficulty بحيث لا ينهار التعلم إلى مسائل سهلة أو trajectories متشابهة.

- افصل reward على **correctness** عن reward على **format/style**.
- اختبر robustness بتغيير verifier أو test set.
- قِس entropy/diversity كي لا يصبح policy أحادي المسار.
- استخدم hold-outs تمنع memorized proofs أو test leakage.

```
R_total = R_verifiable - β·HackRisk - γ·ComputeWaste
```

Research specification / مواصفة الاختبار

Variable / layer	Measure	Failure signal
Reasoning length	quality per token	rumination
Data leakage	private/novel tests	memorized trajectories
Distillation	student vs teacher under equal budget	trace imitation without transfer
Tool feedback	environment pass rate	false positive tool signal
Safety reward	adversarial exploit rate	gaming safer-looking traces
Calibration	predicted vs empirical success	overconfidence
Reproducibility	seed variance / run-to-run stability	unstable post-training

Outcome Reward vs Process Reward

When should a model be rewarded?

Outcome Reward Model (ORM) يحكم على النهاية. **Process Reward Model (PRM)** يحكم على خطوات أو مقاطع trajectory. النتيجة العملية ليست أن PRM أفضل دائماً، بل أن resolution الأعلى يعطي إشارة أكثر غنى عندما تكون الخطوات قابلة للتقييم، مقابل كلفة annotation واحتمال تعليم "شكل التفكير المقبول" بدلاً من التفكير الفعال.

$$R_{process}(\tau)=\Sigma_t w_t \cdot r(s_t,a_t) \; ; \; R_{outcome}(\tau)=r(s_T)$$

ORM يفيد عندما

- التحقق النهائي قوي ورخيص.
- الخطوات الداخلية غير موثوقة أو مخفية.
- الهدف هو action-level success.
- process labels قد تفرض أسلوباً لا ضرورة له.

PRM يفيد عندما

- الأخطاء المحلية قابلة للكشف.
- المسار طويل والجواب النهائي sparse.
- نحتاج verifier أثناء search.
- نريد attribution لسبب الفشل.

الاتجاه الحالي هو hybrid reward stack: outcome verifier + selective process checks + tool/environment assertions + cost budget. هذا أقرب إلى نظام تشغيل استدلالي من "reward model" مفرد.

Deep technical extraction / الاستخراج الهندسي

Outcome مقابل process. ORM يجيب: هل انتهت المسألة صحيحاً؟ PRM يسأل: أين بدأ الانحراف؟ PRM مفيد للبحث والتشخيص، لكنه قد يعلّم النموذج "شكل" reasoning المقبول بدلاً من صحة الآلية. لذلك frontier stack يحتاج heterogeneous verification: rules/tests عند الإمكان، process judge عند الحاجة، وexecution evidence للمهام الأداتية.

- قارن step labels مع final outcome تحت distribution shift.
- افحص calibration: هل score العالي يعني احتمال نجاح أعلى فعلاً؟
- استخدم disagreement بين judges كإشارة uncertainty.
- لا تجعل نفس model يولد ويحكم دون control baseline.

Source ledger / دفتر المصدر

S16 · Research · 2026
A Survey of Process Reward Models: From Outcome Supervision to Process Supervision
arXiv:2510.08049; PRM data, training, inference-time use and RL.

S17 · Research · 2026
Reward Modeling for Reinforcement Learning-Based LLM Reasoning
arXiv:2602.09305; model-based, rule-based and self-reward designs.

S24 · Official · 2025-03-10
Detecting Misbehavior in Frontier Reasoning Models
Official research on monitoring reasoning traces for reward hacking and exploit behavior.

S08 · Research · 2025
DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning
arXiv:2501.12948; RLVR-style reasoning emergence, cold start, distillation.

S15 · Research · 2026
Inference Scaling, Learning to Reason, and Agentic Systems
arXiv:2504.09037v4; synthesis of inference scaling, verifiers, reasoning learning and agentic systems.

S20 · Research · 2025
100 Days After DeepSeek-R1: A Survey on Replication Studies and More Directions for Reasoning Language Models
arXiv:2505.00551; replication lessons for SFT and RLVR.

S21 · Research · 2025
DeepSeek-R1 Thoughtology: Let's <think> about LLM Reasoning
arXiv:2504.07128; analysis of reasoning length, rumination, control and safety.

S22 · Research · 2026
Mid-Think: Training-Free Intermediate-Budget Reasoning
arXiv:2601.07036; inference budget control between short and long reasoning modes.

S23 · Official · 2025-12-18
Evaluating Chain-of-Thought Monitorability
Official framework with intervention/process/outcome-property evaluations of CoT monitorability.

S18 · Research · 2024
Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters
arXiv:2408.03314; adaptive compute allocation and verifier-guided search.

Research specification / مواصفة الاختبار

Variable / layer	Measure	Failure signal
Data leakage	private/novel tests	memorized trajectories
Distillation	student vs teacher under equal budget	trace imitation without transfer
Tool feedback	environment pass rate	false positive tool signal
Safety reward	adversarial exploit rate	gaming safer-looking traces
Calibration	predicted vs empirical success	overconfidence
Reproducibility	seed variance / run-to-run stability	unstable post-training
Reward validity	cross-verifier agreement; exact tests	shortcut/reward hacking

GRPO وما بعده: تحسين السياسة بدون ناقد ضخم

Group-relative optimization intuition

تُفهم GRPO عملياً كطريقة تجعل مجموعة completions لنفس السؤال مرجعاً نسبياً للمكافأة، فتقل الحاجة إلى value model منفصل وتصبح الإشارة مرتبطة بأداء المجموعة. أهميتها ليست في المعادلة وحدها، بل في أنها جعلت توسيع reasoning RL أكثر قابلية للتنفيذ في مختبرات لا تملك pipeline RLHF تقليدي ثقيل.

```
A_i ≈ (r_i - mean(r_1…r_G)) / (std(r_1…r_G)+ε)
```

المشكلة تظهر عند ضيق توزيع المجموعة: إذا كانت كل الحلول سيئة أو متشابهة، يصبح signal ضعيفاً. كذلك قد يزيد reward hacking مع checkers سطحية. لذلك frontier engineering يتجه نحو **diverse sampling**، tool-grounded rewards، automatic difficulty estimation، curricula، cross-checking، verifiers بين.

- استعمل group diversity كمورد، لا مجرد batch trick.
- افصل reward على correctness عن reward على brevity أو style.
- راقب collapse إلى reasoning template واحد.
- اختبر policy على out-of-distribution tasks لا على training-family فقط.

Deep technical extraction / الاستخراج الهندسي

Policy optimization. عائلة GRPO-style methods تقلل الاعتماد على critic ضخم عبر مقارنة samples داخل group واستخدام relative advantages. القيمة الهندسية هي جعل RL على نماذج كبيرة أقل كلفة، لكن الاستقرار يعتمد على KL control، reward scaling، sampling diversity، وطبيعة المسألة.

```
A_i ≈ (r_i - mean(r_group)) / (std(r_group)+ε)
```

- راقب collapse نحو templates قصيرة.
- اختبر حساسية النتيجة لحجم group.
- افصل gain الناتج عن المزيد من samples عن gain الناتج عن RL.
- سجل compute الكامل: rollout generation غالباً هو العنصر الأغلى.

Source ledger / دفتر المصدر

S08 · Research · 2025
DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning
arXiv:2501.12948; RLVR-style reasoning emergence, cold start, distillation.

S20 · Research · 2025
100 Days After DeepSeek-R1: A Survey on Replication Studies and More Directions for Reasoning Language Models
arXiv:2505.00551; replication lessons for SFT and RLVR.

S17 · Research · 2026
Reward Modeling for Reinforcement Learning-Based LLM Reasoning
arXiv:2602.09305; model-based, rule-based and self-reward designs.

S15 · Research · 2026
Inference Scaling, Learning to Reason, and Agentic Systems
arXiv:2504.09037v4; synthesis of inference scaling, verifiers, reasoning learning and agentic systems.

S16 · Research · 2026
A Survey of Process Reward Models: From Outcome Supervision to Process Supervision
arXiv:2510.08049; PRM data, training, inference-time use and RL.

S21 · Research · 2025
DeepSeek-R1 Thoughtology: Let's <think> about LLM Reasoning
arXiv:2504.07128; analysis of reasoning length, rumination, control and safety.

S22 · Research · 2026
Mid-Think: Training-Free Intermediate-Budget Reasoning
arXiv:2601.07036; inference budget control between short and long reasoning modes.

S23 · Official · 2025-12-18
Evaluating Chain-of-Thought Monitorability
Official framework with intervention/process/outcome-property evaluations of CoT monitorability.

S24 · Official · 2025-03-10
Detecting Misbehavior in Frontier Reasoning Models
Official research on monitoring reasoning traces for reward hacking and exploit behavior.

S18 · Research · 2024
Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters
arXiv:2408.03314; adaptive compute allocation and verifier-guided search.

Research specification / مواصفة الاختبار

Variable / layer	Measure	Failure signal
Calibration	predicted vs empirical success	overconfidence
Reproducibility	seed variance / run-to-run stability	unstable post-training
Reward validity	cross-verifier agreement; exact tests	shortcut/reward hacking
Rollout diversity	unique paths; entropy; pass@N	mode collapse
KL / policy drift	KL and behavior delta	over-regularization or drift
Difficulty curriculum	success by difficulty bin	easy-task saturation
Credit assignment	step ablation / contribution	spurious intermediate reward

قانون Test-Time Compute

The new scaling axis

بعد أعوام من scaling في parameters وtokens، صار inference نفسه محور توسع. ورقة Snell et al. بيّنت أن adaptive allocation يمكن أن يجعل model أصغر يتفوق على model أكبر في أجزاء من توزيع الصعوبة، بينما تقدم أعمال لاحقة تحليلات نظرية لطرق aggregation/search. الفكرة الأساسية: ليس كل prompt يستحق budget متساوياً.

```
B*(x)=argmax_B [ P(correct | x,B) - λ·Cost(B) ]
```

عملياً توجد أربع طرق رئيسية: **longer single trajectory**، **parallel sampling**، **search with verifier**، و**iterative revise/reflect**. لا توجد واحدة مهيمنة مطلقاً؛ المسائل التي لديها checker قوي تستفيد من sampling/search، المسائل المفتوحة قد تستفيد أكثر من tool use أو decomposition.

> الحد الحاسم نحو AGI ليس "كم token يمكن أن يفكر؟" بل "هل يعرف أين يصرف token، ومتى يتوقف، ومتى يستدعي بيئة خارجية؟".

Source ledger / دفتر المصدر

S18 · Research · 2024
Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters
arXiv:2408.03314; adaptive compute allocation and verifier-guided search.

S19 · Research · 2024
Simple and Provable Scaling Laws for the Test-Time Compute of Large Language Models
arXiv:2411.19477; tournament/league aggregation and scaling guarantees under assumptions.

S15 · Research · 2026
Inference Scaling, Learning to Reason, and Agentic Systems
arXiv:2504.09037v4; synthesis of inference scaling, verifiers, reasoning learning and agentic systems.

S22 · Research · 2026
Mid-Think: Training-Free Intermediate-Budget Reasoning
arXiv:2601.07036; inference budget control between short and long reasoning modes.

S21 · Research · 2025
DeepSeek-R1 Thoughtology: Let's <think> about LLM Reasoning
arXiv:2504.07128; analysis of reasoning length, rumination, control and safety.

S23 · Official · 2025-12-18
Evaluating Chain-of-Thought Monitorability
Official framework with intervention/process/outcome-property evaluations of CoT monitorability.

S24 · Official · 2025-03-10
Detecting Misbehavior in Frontier Reasoning Models
Official research on monitoring reasoning traces for reward hacking and exploit behavior.

S01 · Official · 2026-07-09
GPT-5.6 System Card
Official system card; Sol/Terra/Luna, reasoning effort, agentic safety, prompt-injection and self-improvement evaluations.

S04 · Official · 2026-02
Gemini 3.1 Deep Think / Accelerating Mathematical and Scientific Discovery
Official research/model documentation for specialized high-compute scientific reasoning.

S08 · Research · 2025
DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning
arXiv:2501.12948; RLVR-style reasoning emergence, cold start, distillation.

Deep technical extraction / الاستخراج الهندسي

Test-time compute. scaling عند inference يمكن أن يأخذ شكل tokens أطول، best-of-N، verifier-guided selection، branching/search، أو parallel agents. لا توجد سياسة مثلى ثابتة لكل الأسئلة؛ frontier الحقيقي هو **adaptive allocation** حسب difficulty وuncertainty.

- بناء difficulty estimator قبل الإنفاق المكلف.
- ميزانية مشتركة عادلة عند مقارنة methods.
- حساب diminishing returns لكل additional sample/branch.
- قياس wall-clock، لا token count فقط، عند parallelism.

```
B*(x)=argmax_B E[Quality(x,B)]-λ·Cost(B)
```

Research specification / مواصفة الاختبار

Variable / layer	Measure	Failure signal
Self-consistency	agreement vs correctness	correlated error
Tool escalation	gain after external evidence	unnecessary calls
Uncertainty	calibration / abstention	false certainty
Latency	p50/p95 end-to-end	unusable deep mode
Trace quality	monitorability probes	unfaithful trace
Hard-tail gain	top difficulty subset	average-score illusion
Difficulty estimator	error by difficulty bucket	misallocated compute

Search over reasoning trajectories

Best-of-N is only the beginning

عند توليد N مسارات، قيمة النظام تعتمد على **verifier quality × diversity**. إذا كانت المسارات correlated فإن زيادة N تهدر compute. وإذا كان verifier ضعيفاً، فقد يزيد N احتمال اختيار إجابة مقنعة خاطئة. لذلك بدأت frontier systems تمزج execution checks، debate-like comparison، tournaments، tree search، وadaptive branching.

```
Gain(N) ≈ Diversity(τ_1…τ_N) × VerifierAccuracy - CorrelationPenalty
```

- **Knockout:** مقارنة زوجية تدريجية، رخيص نسبياً لكنه حساس لخطأ مبكر.
- **League:** كل مرشح يواجه عدة منافسين؛ أكثر كلفة لكن أكثر robust.
- **Tool-verified:** الكود/المحاكي/solver يخرج المقارنة من اللغة.
- **Self-revision:** نفس trajectory يعاد بناؤها بعد evidence جديد.

في MOUSSAIF يمكن ترجمة ذلك إلى QMCD: لا نفتح branch إضافياً إلا إذا كان *Expected Information Gain* أو risk أعلى من كلفته. وهذا يمنع "شجرة أفكار" غير محدودة لا تفعل سوى توسيع النص.

Source ledger / دفتر المصدر

S19 · Research · 2024
Simple and Provable Scaling Laws for the Test-Time Compute of Large Language Models
arXiv:2411.19477; tournament/league aggregation and scaling guarantees under assumptions.

S15 · Research · 2026
Inference Scaling, Learning to Reason, and Agentic Systems
arXiv:2504.09037v4; synthesis of inference scaling, verifiers, reasoning learning and agentic systems.

S18 · Research · 2024
Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters
arXiv:2408.03314; adaptive compute allocation and verifier-guided search.

S21 · Research · 2025
DeepSeek-R1 Thoughtology: Let's <think> about LLM Reasoning
arXiv:2504.07128; analysis of reasoning length, rumination, control and safety.

S22 · Research · 2026
Mid-Think: Training-Free Intermediate-Budget Reasoning
arXiv:2601.07036; inference budget control between short and long reasoning modes.

S23 · Official · 2025-12-18
Evaluating Chain-of-Thought Monitorability
Official framework with intervention/process/outcome-property evaluations of CoT monitorability.

S24 · Official · 2025-03-10
Detecting Misbehavior in Frontier Reasoning Models
Official research on monitoring reasoning traces for reward hacking and exploit behavior.

S01 · Official · 2026-07-09
GPT-5.6 System Card
Official system card; Sol/Terra/Luna, reasoning effort, agentic safety, prompt-injection and self-improvement evaluations.

S04 · Official · 2026-02
Gemini 3.1 Deep Think / Accelerating Mathematical and Scientific Discovery
Official research/model documentation for specialized high-compute scientific reasoning.

S08 · Research · 2025
DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning
arXiv:2501.12948; RLVR-style reasoning emergence, cold start, distillation.

Deep technical extraction / الاستخراج الهندسي

Reasoning search. branching يحول inference إلى مسألة search في فضاء partial solutions. أفضل استراتيجية تعتمد على branching factor، جودة heuristic/verifier، وإمكانية الرجوع. Tree-like search يفيد عندما توجد بدائل يمكن تقييمها مبكراً؛ أما إذا كان verifier ضعيفاً فقد يضاعف الضجيج فقط.

- حدد state canonicalization لمنع الفروع المكررة.
- استخدم pruning مبني على evidence لا confidence لغوي فقط.
- قارن beam/tree/MCTS/best-of-N تحت نفس budget.
- راقب search overhead مقابل gain على hard problems.

Research specification / مواصفة الاختبار

Variable / layer	Measure	Failure signal
Verifier precision	selection accuracy	bad pruning
Stopping rule	marginal gain/token	overthinking
Parallelism	wall-clock speedup	merge overhead
Self-consistency	agreement vs correctness	correlated error
Tool escalation	gain after external evidence	unnecessary calls
Uncertainty	calibration / abstention	false certainty
Latency	p50/p95 end-to-end	unusable deep mode

Verifiers كطبقة مستقلة

Verifier = a model of error, not truth itself

الـverifier الحديث قد يكون LLM judge، PRM، ORM، execution sandbox، theorem prover، unit tests، rule-based، أو ensemble. كل نوع يملك blind spots مختلفة. أهم تحول هو فصل **generator** عن **evaluator** منطقياً حتى لو شاركا weights؛ لأن الجهة التي تولد الحل ليست مرجعاً كافياً لتصديقه.

Verifier	قوته	فشله النموذجي
Rule / exact check	حاسم ورخيص	ضيق المجال
Execution	يربط الكود بالواقع	اختبارات ناقصة
PRM	إشارة خطوة بخطوة	style bias / annotation cost
LLM Judge	مرن واسع المجال	shared blind spots / persuasion
Ensemble	خفض خطأ فردي	كلفة وتواطؤ أخطاء مترابطة

المعيار القوي: verifier يجب أن يخضع هو أيضاً لـadversarial examples، calibration، وablation.

Source ledger / دفتر المصدر

S16 · Research · 2026
A Survey of Process Reward Models: From Outcome Supervision to Process Supervision
arXiv:2510.08049; PRM data, training, inference-time use and RL.

S17 · Research · 2026
Reward Modeling for Reinforcement Learning-Based LLM Reasoning
arXiv:2602.09305; model-based, rule-based and self-reward designs.

S19 · Research · 2024
Simple and Provable Scaling Laws for the Test-Time Compute of Large Language Models
arXiv:2411.19477; tournament/league aggregation and scaling guarantees under assumptions.

S24 · Official · 2025-03-10
Detecting Misbehavior in Frontier Reasoning Models
Official research on monitoring reasoning traces for reward hacking and exploit behavior.

S15 · Research · 2026
Inference Scaling, Learning to Reason, and Agentic Systems
arXiv:2504.09037v4; synthesis of inference scaling, verifiers, reasoning learning and agentic systems.

S18 · Research · 2024
Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters
arXiv:2408.03314; adaptive compute allocation and verifier-guided search.

S21 · Research · 2025
DeepSeek-R1 Thoughtology: Let's <think> about LLM Reasoning
arXiv:2504.07128; analysis of reasoning length, rumination, control and safety.

S22 · Research · 2026
Mid-Think: Training-Free Intermediate-Budget Reasoning
arXiv:2601.07036; inference budget control between short and long reasoning modes.

S23 · Official · 2025-12-18
Evaluating Chain-of-Thought Monitorability
Official framework with intervention/process/outcome-property evaluations of CoT monitorability.

S01 · Official · 2026-07-09
GPT-5.6 System Card
Official system card; Sol/Terra/Luna, reasoning effort, agentic safety, prompt-injection and self-improvement evaluations.

Deep technical extraction / الاستخراج الهندسي

Verifier stack. الـverifier الجيد ليس judge واحداً. رتب الأدلة: deterministic tests > formal checks > environment postconditions > specialized reward model > general self-judge. كلما هبطنا في السلم زادت الحاجة إلى uncertainty وcross-check.

```
DecisionConfidence = Σ w_k·Evidence_k - ConflictPenalty
```

- اختبر independence بين proposer وverifier.
- أعد تقييم samples adversarially لتقدير reward hacking.
- حافظ على raw evidence لا verdict فقط.
- استخدم verifier disagreement لتشغيل branch إضافي أو escalation.

Research specification / مواصفة الاختبار

Variable / layer	Measure	Failure signal
Parallelism	wall-clock speedup	merge overhead
Self-consistency	agreement vs correctness	correlated error
Tool escalation	gain after external evidence	unnecessary calls
Uncertainty	calibration / abstention	false certainty
Latency	p50/p95 end-to-end	unusable deep mode
Trace quality	monitorability probes	unfaithful trace
Hard-tail gain	top difficulty subset	average-score illusion

Reasoning Budget Control

Thinking is a resource allocation problem

Qwen3 جعل budget mechanism جزءاً واضحاً من واجهة reasoning، بينما أصبحت effort levels شائعة في نماذج مغلقة. هذا يعني أن inference policy نفسها أصبحت API surface. النظام الأفضل لا يختار budget يدوياً لكل prompt، بل يتنبأ بالصعوبة والمخاطر وقيمة الأداة.

```
BudgetScore = α·Difficulty + β·Risk + γ·Novelty + δ·ToolUncertainty - η·LatencyCost
```

Mid-Think يركز على المنطقة الوسطى: كثير من المهام لا تحتاج "no-think" ولا maximum deliberation. هذه المنطقة مهمة اقتصادياً لأنها حيث يتحول reasoning من benchmark feature إلى production primitive.

Vascular Compute bridge: يمكن لـMCDJ أن يصبح scheduler: الفروع ذات Active Difference أعلى تحصل على compute، والفروع التي لا تغير prediction أو action تُضغط.

Source ledger / دفتر المصدر

S09 · Official · 2026-08-02
Qwen3.8-Max: A New Bar for Coding and Cowork
Official current flagship report; 2.4T-class MoE, multimodal/long-context frontier, coding and cowork/agentic work. Vendor benchmark claims are treated as reported.

S22 · Research · 2026
Mid-Think: Training-Free Intermediate-Budget Reasoning
arXiv:2601.07036; inference budget control between short and long reasoning modes.

S01 · Official · 2026-07-09
GPT-5.6 System Card
Official system card; Sol/Terra/Luna, reasoning effort, agentic safety, prompt-injection and self-improvement evaluations.

S02 · Official · 2026-06-09
Claude Fable 5 & Mythos 5 System Card
Official system card for Anthropic's most capable widely released Fable 5 class; always-on adaptive thinking, long-running agents, 1M context and safety evaluations.

S15 · Research · 2026
Inference Scaling, Learning to Reason, and Agentic Systems
arXiv:2504.09037v4; synthesis of inference scaling, verifiers, reasoning learning and agentic systems.

S18 · Research · 2024
Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters
arXiv:2408.03314; adaptive compute allocation and verifier-guided search.

S19 · Research · 2024
Simple and Provable Scaling Laws for the Test-Time Compute of Large Language Models
arXiv:2411.19477; tournament/league aggregation and scaling guarantees under assumptions.

S21 · Research · 2025
DeepSeek-R1 Thoughtology: Let's <think> about LLM Reasoning
arXiv:2504.07128; analysis of reasoning length, rumination, control and safety.

S23 · Official · 2025-12-18
Evaluating Chain-of-Thought Monitorability
Official framework with intervention/process/outcome-property evaluations of CoT monitorability.

S24 · Official · 2025-03-10
Detecting Misbehavior in Frontier Reasoning Models
Official research on monitoring reasoning traces for reward hacking and exploit behavior.

Deep technical extraction / الاستخراج الهندسي

Budget control. التفكير الأقوى لا يعني دائماً التفكير الأطول. هناك مسائل تستفيد من fast mode، وأخرى تحتاج deep deliberation أو tool use. router جيد يتنبأ بالقيمة الحدية للcompute ويختار effort level قبل وأثناء trajectory.

```
Effort_{t+1}=f(Uncertainty_t, Stakes, Progress, VerifierSignal)
```

- اختبر underthinking وoverthinking كفشلين متناظرين.
- أوقف trajectory إذا توقف progress رغم زيادة tokens.
- اسمح بالتصعيد من fast إلى deep بعد failed verification.
- قارن fixed budget مع adaptive budget على difficulty bins.

Research specification / مواصفة الاختبار

Variable / layer	Measure	Failure signal
Trace quality	monitorability probes	unfaithful trace
Hard-tail gain	top difficulty subset	average-score illusion
Difficulty estimator	error by difficulty bucket	misallocated compute
Budget	quality-cost frontier	overspend
Branch factor	gain per additional branch	branch explosion
Verifier precision	selection accuracy	bad pruning
Stopping rule	marginal gain/token	overthinking

Long-Context Reasoning ≠ Long Memory

Million tokens are not a memory architecture

DeepSeek-V4 وQwen3.8-Max/Kimi K3 يدفعان سياقاً طويلاً جداً، لكن نافذة 1M token لا تحل تلقائياً causal memory، state compaction، selective forgetting أو cross-session persistence. long-context هو **address space**؛ الذاكرة الوكيلية هي policy تكتب وتسترجع وتنسى.

التحديات الأساسية: cost quadratic/linear، retrieval degradation، position dilution حسب interference، attention، ومشكلة أن history الخام تحتوي actions قديمة لم تعد صالحة بعد تغير البيئة. لذلك frontier architectures تجمع sparse/hybrid attention مع vector/graph retrieval، episodic stores، summaries، وstate objects.

```
UsefulContext = RelevantState / (TotalTokens × RetrievalNoise)
```

من منظور Self-Reconstruction Memory، MOUSSAIF ليست "أرشيف محادثة"، بل قدرة على إعادة بناء الحالة التشغيلية: الهدف، tool outputs، assumptions، فروع سقطت، وسبب آخر قرار.

Source ledger / دفتر المصدر

S06 · Research · 2026
DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence
arXiv:2606.19348; hybrid attention, million-token context, effort scaling, agentic benchmarks.

S11 · Research · 2026
Qwen3.5-Omni Technical Report
arXiv:2604.15804; hybrid-attention MoE omnimodal reasoning across text, vision and audio, long-sequence interaction and grounding.

S36 · Research · 2026
AMA-Bench: Evaluating Long-Horizon Memory for Agentic Applications
arXiv:2602.22769; causal/objective memory gaps and AMA-Agent.

S37 · Research · 2026
Agent Memory: Characterization and System Implications of Stateful Long-Horizon Workloads
arXiv:2606.06448; system taxonomy, cost profiling and fleet implications.

S38 · Research · 2026
SAM: State-Adaptive Memory for Long-Horizon Reasoning Agent
arXiv:2605.24468; evolving-state-conditioned recall and compact cues.

S01 · Official · 2026-07-09
GPT-5.6 System Card
Official system card; Sol/Terra/Luna, reasoning effort, agentic safety, prompt-injection and self-improvement evaluations.

S02 · Official · 2026-06-09
Claude Fable 5 & Mythos 5 System Card
Official system card for Anthropic's most capable widely released Fable 5 class; always-on adaptive thinking, long-running agents, 1M context and safety evaluations.

S03 · Official · 2026-05-19
Gemini 3.5: frontier intelligence with action
Official model report; action-oriented workflows, agentic coding, subagents and benchmark reporting.

S04 · Official · 2026-02
Gemini 3.1 Deep Think / Accelerating Mathematical and Scientific Discovery
Official research/model documentation for specialized high-compute scientific reasoning.

S05 · Official · 2026-08-12
Introducing Grok 4.6
Official current flagship release; long-running agents, coding, interactive/visual work, agentic RL, configurable reasoning effort and 500k context.

Deep technical extraction / الاستخراج الهندسي

Context vs memory. نافذة مليون token لا تعني أن كل الماضي متاح وظيفياً. long context يعاني attention dilution، stale actions، contradictory state، وكلفة إعادة القراءة. الذاكرة الوكيلية تحتاج policy للكتابة والضغط والاسترجاع، لا مجرد window أكبر.

- **Hot state:** الهدف الحالي، constraints، آخر observations.
- **Episodic:** أحداث قابلة للاسترجاع مع timestamps/outcomes.
- **Semantic:** facts/procedures مستقرة نسبياً.
- **Reconstruction:** summary قابل لإعادة بناء لماذا تغيرت الخطة.

```
UsefulMemory = RelevantState × Freshness × CausalLinkage
```

Research specification / مواصفة الاختبار

Variable / layer	Measure	Failure signal
Throughput	tokens/s under target context	benchmark-only speed
Open/closed control	ablation/debug access	hidden harness effects
External replication	independent harness delta	vendor-only comparison
Activated compute	active params / token	capacity-cost confusion
Context fidelity	state retention under noise	attention dilution
KV efficiency	memory/throughput at long context	bandwidth bottleneck
Effort knob	quality across effort levels	uncalibrated mode

Efficiency Frontier وSparse Attention

Reasoning is bounded by memory bandwidth too

القدرة على التفكير أطول تصطدم بهندسة الذاكرة قبل أن تصطدم بالمعادلات. DeepSeek-V4 يبرز hybrid attention وmillion-token efficiency؛ Qwen3 وKimi وDeepSeek تستخدم MoE أو attention optimizations مختلفة لتخفيض activated compute. النتيجة هي تحويل cost curve بحيث يصبح long-horizon reasoning قابلاً للتشغيل.

```
EffectiveReasoningBudget ≈ FLOPs_budget × KV_efficiency × ActiveParamsEfficiency
```

- **MoE:** يرفع total capacity دون تفعيل كل المعلمات لكل token.
- **Sparse/Hybrid Attention:** يخفض كلفة السياق الطويل.
- **KV compression/caching:** حاسم عند agent loops المتكررة.
- **Parallel agents:** ينقل bottleneck إلى orchestration والتزامن.

الـfrontier لم يعد نموذجاً فقط؛ إنه co-design بين scheduler، cache policy، kernels، architecture، وagent harness.

Source ledger / دفتر المصدر

S06 · Research · 2026
DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence
arXiv:2606.19348; hybrid attention, million-token context, effort scaling, agentic benchmarks.

S07 · Research · 2026-08-13
DeepSeek V4-Pro-0813 launch and agent capability update
Current-release reporting on V4 Pro availability and improved AI-agent capabilities; used as deployment-state evidence, not as an independent technical paper.

S09 · Official · 2026-08-02
Qwen3.8-Max: A New Bar for Coding and Cowork
Official current flagship report; 2.4T-class MoE, multimodal/long-context frontier, coding and cowork/agentic work. Vendor benchmark claims are treated as reported.

S12 · Research · 2026
Kimi K3: Open Frontier Intelligence
arXiv:2607.24653; 2.8T MoE / 104B active, native vision, 1M context, reasoning-effort levels, long-horizon agentic RL and open weights.

S01 · Official · 2026-07-09
GPT-5.6 System Card
Official system card; Sol/Terra/Luna, reasoning effort, agentic safety, prompt-injection and self-improvement evaluations.

S02 · Official · 2026-06-09
Claude Fable 5 & Mythos 5 System Card
Official system card for Anthropic's most capable widely released Fable 5 class; always-on adaptive thinking, long-running agents, 1M context and safety evaluations.

S03 · Official · 2026-05-19
Gemini 3.5: frontier intelligence with action
Official model report; action-oriented workflows, agentic coding, subagents and benchmark reporting.

S04 · Official · 2026-02
Gemini 3.1 Deep Think / Accelerating Mathematical and Scientific Discovery
Official research/model documentation for specialized high-compute scientific reasoning.

S05 · Official · 2026-08-12
Introducing Grok 4.6
Official current flagship release; long-running agents, coding, interactive/visual work, agentic RL, configurable reasoning effort and 500k context.

S10 · Research · 2026
Qwen3-Coder-Next Technical Report
arXiv:2603.00729; repository reasoning, agent interaction patterns, executable training/evaluation environments.

Deep technical extraction / الاستخراج الهندسي

Efficiency frontier. sparse/hybrid attention، MoE، KV-cache engineering، quantization وspeculative execution تغيّر اقتصاد reasoning. هذه ليست تفاصيل inference ثانوية؛ هي التي تسمح بتوسيع horizon أو تشغيل عدة agents ضمن نفس budget.

- سجل total parameters لا activated parameters فقط.
- قِس throughput مع long context لا short prompt.
- افصل memory bandwidth bottleneck عن FLOPs bottleneck.
- للـswarm، أضف network/merge overhead إلى الحساب.

```
SystemUtility ∝ Capability / (Latency × $Cost × MemoryFootprint)
```

Research specification / مواصفة الاختبار

Variable / layer	Measure	Failure signal
Coding	repo-level tests / evolution	snippet overfit
Memory	reconstruction/recovery	history != memory
Safety envelope	injection/destructive-action tests	capability without control
Throughput	tokens/s under target context	benchmark-only speed
Open/closed control	ablation/debug access	hidden harness effects
External replication	independent harness delta	vendor-only comparison
Activated compute	active params / token	capacity-cost confusion

Multimodal Reasoning

Vision is entering the reasoning loop, not just the prompt

image-grounded planning: قراءة شاشة، اختيار عنصر، توليد كود من صورة، أو استخدام vision كدليل في خطوة استدلال. الفرق المهم هو الانتقال من "extract visual features ثم أجب" إلى GPT reasoning families وGemini وKimi K3 وQwen3.5-Omni/Qwen3.8 تجعل الصورة والفيديو والواجهة عناصر في trajectory.

Kimi K3 يجمع native vision مع 1M context وagentic RL طويل الأفق، بينما Qwen3.5-Omni يركز على omnimodal long-sequence interaction والgrounding عبر الصوت والصورة، وتدفع Qwen3.8-Max multimodality إلى flagship coding/cowork. هذه الاتجاهات تقرب agent من بيئة فعلية، لكنها تضيف failure modes: OCR hallucination، coordinate drift، stale screenshot، وعدم استقرار perception تحت UI changes.

```
ActionConfidence = ReasoningConfidence × PerceptionConfidence × StateFreshness
```

Deep technical extraction / الاستخراج الهندسي

Multimodal loop. عندما تدخل الصورة أو واجهة الكمبيوتر في reasoning، يصبح perception جزءاً من state estimator. الخطأ البصري قد ينتقل إلى planning ثم يتضخم عبر أفعال لا رجعة فيها. لذلك يجب حفظ timestamp/screenshot region وconfidence مع كل observation.

- فرّق OCR confidence عن semantic inference.
- أعد الالتقاط قبل destructive click إذا كانت state قديمة.
- اختبر robustness لتغير layout والحجم واللغة.
- قِس action grounding، لا VQA accuracy فقط.

```
P(correct action)≈P(perception)·P(plan|perception)·P(execution)
```

Source ledger / دفتر المصدر

S11 · Research · 2026
Qwen3.5-Omni Technical Report
arXiv:2604.15804; hybrid-attention MoE omnimodal reasoning across text, vision and audio, long-sequence interaction and grounding.

S12 · Research · 2026
Kimi K3: Open Frontier Intelligence
arXiv:2607.24653; 2.8T MoE / 104B active, native vision, 1M context, reasoning-effort levels, long-horizon agentic RL and open weights.

S03 · Official · 2026-05-19
Gemini 3.5: frontier intelligence with action
Official model report; action-oriented workflows, agentic coding, subagents and benchmark reporting.

S01 · Official · 2026-07-09
GPT-5.6 System Card
Official system card; Sol/Terra/Luna, reasoning effort, agentic safety, prompt-injection and self-improvement evaluations.

S02 · Official · 2026-06-09
Claude Fable 5 & Mythos 5 System Card
Official system card for Anthropic's most capable widely released Fable 5 class; always-on adaptive thinking, long-running agents, 1M context and safety evaluations.

S04 · Official · 2026-02
Gemini 3.1 Deep Think / Accelerating Mathematical and Scientific Discovery
Official research/model documentation for specialized high-compute scientific reasoning.

S05 · Official · 2026-08-12
Introducing Grok 4.6
Official current flagship release; long-running agents, coding, interactive/visual work, agentic RL, configurable reasoning effort and 500k context.

S06 · Research · 2026
DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence
arXiv:2606.19348; hybrid attention, million-token context, effort scaling, agentic benchmarks.

S07 · Research · 2026-08-13
DeepSeek V4-Pro-0813 launch and agent capability update
Current-release reporting on V4 Pro availability and improved AI-agent capabilities; used as deployment-state evidence, not as an independent technical paper.

S09 · Official · 2026-08-02
Qwen3.8-Max: A New Bar for Coding and Cowork
Official current flagship report; 2.4T-class MoE, multimodal/long-context frontier, coding and cowork/agentic work. Vendor benchmark claims are treated as reported.

Research specification / مواصفة الاختبار

Variable / layer	Measure	Failure signal
Open/closed control	ablation/debug access	hidden harness effects
External replication	independent harness delta	vendor-only comparison
Activated compute	active params / token	capacity-cost confusion
Context fidelity	state retention under noise	attention dilution
KV efficiency	memory/throughput at long context	bandwidth bottleneck
Effort knob	quality across effort levels	uncalibrated mode
Tool use	tool success + postcondition	function-call theater

Tool-Integrated Reasoning

Thinking and acting are being co-trained

DeepSeek-V4 يبني على خط دمج reasoning مع tool/agent workloads؛ Qwen3-Coder-Next يدرب على agent-style interaction داخل environments؛ GPT-5.6 وClaude/Gemini/Grok تُطلق مع computer/tool workflows. هذه ليست مجرد function calling API. في النظام الناضج، tool observation تغير belief state وتعيد ترتيب الخطة.

```
b_{t+1} = Update(b_t, action_t, observation_t, provenance_t)
```

- Tool choice يجب أن يعتمد على expected information gain.
- النتيجة تحتاج schema + provenance + freshness.
- الفشل الصامت أخطر من exception؛ لذلك نحتاج post-condition checks.
- tool call لا يساوي world effect حتى نفحص الحالة الجديدة.

هذا هو الجسر المباشر إلى World-Effect Loop: Execute → Observe → Verify → Attribute → Persist.

Source ledger / دفتر المصدر

S07 · Research · 2026-08-13
DeepSeek V4-Pro-0813 launch and agent capability update
Current-release reporting on V4 Pro availability and improved AI-agent capabilities; used as deployment-state evidence, not as an independent technical paper.

S10 · Research · 2026
Qwen3-Coder-Next Technical Report
arXiv:2603.00729; repository reasoning, agent interaction patterns, executable training/evaluation environments.

S01 · Official · 2026-07-09
GPT-5.6 System Card
Official system card; Sol/Terra/Luna, reasoning effort, agentic safety, prompt-injection and self-improvement evaluations.

S02 · Official · 2026-06-09
Claude Fable 5 & Mythos 5 System Card
Official system card for Anthropic's most capable widely released Fable 5 class; always-on adaptive thinking, long-running agents, 1M context and safety evaluations.

S03 · Official · 2026-05-19
Gemini 3.5: frontier intelligence with action
Official model report; action-oriented workflows, agentic coding, subagents and benchmark reporting.

S05 · Official · 2026-08-12
Introducing Grok 4.6
Official current flagship release; long-running agents, coding, interactive/visual work, agentic RL, configurable reasoning effort and 500k context.

S04 · Official · 2026-02
Gemini 3.1 Deep Think / Accelerating Mathematical and Scientific Discovery
Official research/model documentation for specialized high-compute scientific reasoning.

S06 · Research · 2026
DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence
arXiv:2606.19348; hybrid attention, million-token context, effort scaling, agentic benchmarks.

S09 · Official · 2026-08-02
Qwen3.8-Max: A New Bar for Coding and Cowork
Official current flagship report; 2.4T-class MoE, multimodal/long-context frontier, coding and cowork/agentic work. Vendor benchmark claims are treated as reported.

S11 · Research · 2026
Qwen3.5-Omni Technical Report
arXiv:2604.15804; hybrid-attention MoE omnimodal reasoning across text, vision and audio, long-sequence interaction and grounding.

Deep technical extraction / الاستخراج الهندسي

Tool-integrated cognition. الاستدلال الناضج يتعامل مع tool result كتحديث belief لا كملاحظة هامشية. كل call يحتاج provenance، postcondition، expected observation، precondition، وfailure code. إذا لم تتغير الخطة بعد observation مهم فهناك احتمال أن النموذج "يتكلم حول الأداة" بدل أن يستخدمها معرفيًا.

- اختبر counterfactual tool outputs لكشف anchoring.
- استعمل typed schemas للحد من ambiguity.
- ميّز partial result، empty result، timeout، وsilent corruption.
- نفّذ post-action state check قبل إعلان النجاح.

Research specification / مواصفة الاختبار

Variable / layer	Measure	Failure signal
Safety envelope	injection/destructive-action tests	capability without control
Throughput	tokens/s under target context	benchmark-only speed
Open/closed control	ablation/debug access	hidden harness effects
External replication	independent harness delta	vendor-only comparison
Activated compute	active params / token	capacity-cost confusion
Context fidelity	state retention under noise	attention dilution
KV efficiency	memory/throughput at long context	bandwidth bottleneck

Case Study: GPT-5.6

Reasoning effort + deployment safety as one system

المهم في GPT-5.6 ليس اسم العائلة فقط، بل طريقة توصيف القدرة. النظام الرسمي يعرض Sol/Terra/Luna، ويعامل **reasoning effort** كمنحنى بدلاً من score واحد. كما يربط agentic capability بتقييمات AI self-improvement، destructive-action avoidance، computer-use confirmations، prompt injection وevals external.

هذه الصياغة مفيدة هندسياً لأنها تكسر عادة benchmark leaderboard: capability يجب أن تقاس مع safeguards وaction policy. كما أن system card يذكر أن النماذج reasoning models مدربة عبر reinforcement learning للتفكير قبل الإجابة، وأن تقييم self-improvement ما زال منفصلاً عن الادعاء العام بالقدرة.

درس معماري: frontier model لا يكفي دون deployment envelope يحدد ما يستطيع فعله، متى يحتاج confirmation، وكيف تُرصد prompt injection والآثار المدمرة.

Source ledger / دفتر المصدر

S01 · Official · 2026-07-09
GPT-5.6 System Card
Official system card; Sol/Terra/Luna, reasoning effort, agentic safety, prompt-injection and self-improvement evaluations.

S23 · Official · 2025-12-18
Evaluating Chain-of-Thought Monitorability
Official framework with intervention/process/outcome-property evaluations of CoT monitorability.

S24 · Official · 2025-03-10
Detecting Misbehavior in Frontier Reasoning Models
Official research on monitoring reasoning traces for reward hacking and exploit behavior.

S02 · Official · 2026-06-09
Claude Fable 5 & Mythos 5 System Card
Official system card for Anthropic's most capable widely released Fable 5 class; always-on adaptive thinking, long-running agents, 1M context and safety evaluations.

S03 · Official · 2026-05-19
Gemini 3.5: frontier intelligence with action
Official model report; action-oriented workflows, agentic coding, subagents and benchmark reporting.

S04 · Official · 2026-02
Gemini 3.1 Deep Think / Accelerating Mathematical and Scientific Discovery
Official research/model documentation for specialized high-compute scientific reasoning.

S05 · Official · 2026-08-12
Introducing Grok 4.6
Official current flagship release; long-running agents, coding, interactive/visual work, agentic RL, configurable reasoning effort and 500k context.

S06 · Research · 2026
DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence
arXiv:2606.19348; hybrid attention, million-token context, effort scaling, agentic benchmarks.

S07 · Research · 2026-08-13
DeepSeek V4-Pro-0813 launch and agent capability update
Current-release reporting on V4 Pro availability and improved AI-agent capabilities; used as deployment-state evidence, not as an independent technical paper.

S09 · Official · 2026-08-02
Qwen3.8-Max: A New Bar for Coding and Cowork
Official current flagship report; 2.4T-class MoE, multimodal/long-context frontier, coding and cowork/agentic work. Vendor benchmark claims are treated as reported.

Deep technical extraction / الاستخراج الهندسي

كيف تقرأ model report. نحتاج خمسة مستويات: post-training، pretraining architecture/ safety/evaluation harness، tool/agent context/memory، reasoning. لا ننقل أرقام benchmark بلا شروطها؛ نبحث عن design choices القابلة لإعادة الاختبار: memory، effort knob، recovery protocol، tool mechanism، أو multimodal grounding.

- Baseline: نفس task set مع model أبسط أو effort أقل.
- Ablation: إزالة memory/tool/parallelism عند الإمكان.
- Cost: tokens، latency، tool calls، retries.
- Generalization: هل الميزة تفيد خارج benchmark الذي صُممت حوله؟

Research specification / مواصفة الاختبار

Variable / layer	Measure	Failure signal
Safety envelope	injection/destructive-action tests	capability without control
Throughput	tokens/s under target context	benchmark-only speed
Open/closed control	ablation/debug access	hidden harness effects
External replication	independent harness delta	vendor-only comparison
Activated compute	active params / token	capacity-cost confusion
Context fidelity	state retention under noise	attention dilution
KV efficiency	memory/throughput at long context	bandwidth bottleneck

Case Study: Claude Fable 5 + Opus 5

Adaptive thinking and long-horizon agency

في الحالة الحالية لدى Anthropic، **Claude Fable 5** هو أعلى نموذج واسع الإتاحة من حيث capability العامة، مع adaptive thinking دائم ونافذة 1M، بينما **Claude Opus 5** يقدم مساراً أقل كلفة مع deep reasoning، agentic coding، long-horizon work، effort ladder يصل إلى **max**. الوثائق تصف Opus 5 كقفزة على Opus 4.8 في test-time compute scaling، وتذكر ميلاً أكبر للتفويض إلى subagents والتحقق الذاتي.

الدلالة البحثية ليست من "يفوز" في benchmark واحد، بل أن Anthropic أصبح يعامل verification، delegation، context، thinking mode، وcost كمنتج واحد. هذا يقرب evaluation من sustained coherence: هل يحافظ النموذج على intent ويكتشف الفشل ويغلق العمل دون stubs عبر trajectory طويلة؟

```
LongRunReliability(H)=GoalFidelity × Recovery × Verification × ToolCorrectness
```

Source ledger / دفتر المصدر

S02 · Official · 2026-06-09
Claude Fable 5 & Mythos 5 System Card
Official system card for Anthropic's most capable widely released Fable 5 class; always-on adaptive thinking, long-running agents, 1M context and safety evaluations.

S13 · Official · 2026-07-24
Claude Opus 5 — model documentation and system card
Official current Opus release; deep reasoning, agentic coding, long-horizon work, 1M context and test-time compute scaling up to max effort.

S28 · Official · 2026-02-18
Measuring AI Agent Autonomy in Practice
Official empirical analysis of user oversight and autonomous execution in Claude Code.

S01 · Official · 2026-07-09
GPT-5.6 System Card
Official system card; Sol/Terra/Luna, reasoning effort, agentic safety, prompt-injection and self-improvement evaluations.

S03 · Official · 2026-05-19
Gemini 3.5: frontier intelligence with action
Official model report; action-oriented workflows, agentic coding, subagents and benchmark reporting.

S04 · Official · 2026-02
Gemini 3.1 Deep Think / Accelerating Mathematical and Scientific Discovery
Official research/model documentation for specialized high-compute scientific reasoning.

S05 · Official · 2026-08-12
Introducing Grok 4.6
Official current flagship release; long-running agents, coding, interactive/visual work, agentic RL, configurable reasoning effort and 500k context.

S06 · Research · 2026
DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence
arXiv:2606.19348; hybrid attention, million-token context, effort scaling, agentic benchmarks.

S07 · Research · 2026-08-13
DeepSeek V4-Pro-0813 launch and agent capability update
Current-release reporting on V4 Pro availability and improved AI-agent capabilities; used as deployment-state evidence, not as an independent technical paper.

S09 · Official · 2026-08-02
Qwen3.8-Max: A New Bar for Coding and Cowork
Official current flagship report; 2.4T-class MoE, multimodal/long-context frontier, coding and cowork/agentic work. Vendor benchmark claims are treated as reported.

Deep technical extraction / الاستخراج الهندسي

كيف تقرأ model report. نحتاج خمسة مستويات: post-training، pretraining architecture/، safety/evaluation harness، tool/agent، context/memory، reasoning. لا ننقل أرقام benchmark بلا شروطها؛ نبحث عن design choices القابلة لإعادة الاختبار: memory، effort knob، recovery، tool protocol، mechanism، أو multimodal grounding.

- Baseline: نفس task set مع model أبسط أو effort أقل.
- Ablation: إزالة memory/tool/parallelism عند الإمكان.
- Cost: tokens، latency، tool calls، retries.
- Generalization: هل الميزة تفيد خارج benchmark الذي صُممت حوله؟

Research specification / مواصفة الاختبار

Variable / layer	Measure	Failure signal
Effort knob	quality across effort levels	uncalibrated mode
Tool use	tool success + postcondition	function-call theater
Multimodal grounding	action/VQA with perturbations	visual misgrounding
Coding	repo-level tests / evolution	snippet overfit
Memory	reconstruction/recovery	history != memory
Safety envelope	injection/destructive-action tests	capability without control
Throughput	tokens/s under target context	benchmark-only speed

Case Study: Gemini 3.1 Deep Think + 3.5

From specialized reasoning to intelligence-with-action

Gemini 3.1 Deep Think يمثل high-compute mode مخصصاً لمسائل العلم والهندسة، بينما 3.5 يعيد صياغة المنتج حول **frontier intelligence with action**. هذا التحول مهم: النموذج السريع لا يكتفي بجودة reasoning، بل يستهدف agent throughput، coding، subagents وmulti-step workflows.

الدرس هو أن architecture frontier تنقسم إلى مسارين متكاملين: **Deep Mode** للمسائل التي تبرر deliberation كثيف، و**Action Mode** الذي يوازن latency مع reliability عبر خطوات كثيرة. نظام AGI عملي سيحتاج الاثنين مع router لاختيار mode.

$$Mode^*(x)=\arg\max_{m\in\{fast,deep,agentic\}} Utility(x,m)-Cost(x,m)$$

Source ledger / دفتر المصدر

S03 · Official · 2026-05-19
Gemini 3.5: frontier intelligence with action
Official model report; action-oriented workflows, agentic coding, subagents and benchmark reporting.

S04 · Official · 2026-02
Gemini 3.1 Deep Think / Accelerating Mathematical and Scientific Discovery
Official research/model documentation for specialized high-compute scientific reasoning.

S01 · Official · 2026-07-09
GPT-5.6 System Card
Official system card; Sol/Terra/Luna, reasoning effort, agentic safety, prompt-injection and self-improvement evaluations.

S02 · Official · 2026-06-09
Claude Fable 5 & Mythos 5 System Card
Official system card for Anthropic's most capable widely released Fable 5 class; always-on adaptive thinking, long-running agents, 1M context and safety evaluations.

S05 · Official · 2026-08-12
Introducing Grok 4.6
Official current flagship release; long-running agents, coding, interactive/visual work, agentic RL, configurable reasoning effort and 500k context.

S06 · Research · 2026
DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence
arXiv:2606.19348; hybrid attention, million-token context, effort scaling, agentic benchmarks.

S07 · Research · 2026-08-13
DeepSeek V4-Pro-0813 launch and agent capability update
Current-release reporting on V4 Pro availability and improved AI-agent capabilities; used as deployment-state evidence, not as an independent technical paper.

S09 · Official · 2026-08-02
Qwen3.8-Max: A New Bar for Coding and Cowork
Official current flagship report; 2.4T-class MoE, multimodal/long-context frontier, coding and cowork/agentic work. Vendor benchmark claims are treated as reported.

S10 · Research · 2026
Qwen3-Coder-Next Technical Report
arXiv:2603.00729; repository reasoning, agent interaction patterns, executable training/evaluation environments.

S11 · Research · 2026
Qwen3.5-Omni Technical Report
arXiv:2604.15804; hybrid-attention MoE omnimodal reasoning across text, vision and audio, long-sequence interaction and grounding.

Deep technical extraction / الاستخراج الهندسي

كيف تقرأ model report. نحتاج خمسة مستويات: post-training/pretraining architecture، safety/evaluation harness، tool/agent context/memory، reasoning. لا ننقل أرقام benchmark بلا شروطها؛ نبحث عن design choices القابلة لإعادة الاختبار: memory effort knob، recovery protocol، tool mechanism، أو multimodal grounding.

- Baseline: نفس task set مع model أبسط أو effort أقل.
- Ablation: إزالة memory/tool/parallelism عند الإمكان.
- Cost: tokens، latency، tool calls، retries.
- Generalization: هل الميزة تفيد خارج benchmark الذي صُممت حوله؟

Research specification / مواصفة الاختبار

Variable / layer	Measure	Failure signal
Memory	reconstruction/recovery	history != memory
Safety envelope	injection/destructive-action tests	capability without control
Throughput	tokens/s under target context	benchmark-only speed
Open/closed control	ablation/debug access	hidden harness effects
External replication	independent harness delta	vendor-only comparison
Activated compute	active params / token	capacity-cost confusion
Context fidelity	state retention under noise	attention dilution

Case Study: DeepSeek V4 / V4-Pro / R1

Open frontier: RL, efficient million-token context and agent deployment

سلالة DeepSeek تلخص مسارين متصلين: R1 جعل reasoning-through-RL محوراً، ثم V4 دفع native million-token context وeffort scaling عبر هندسة هجينة عالية الكفاءة. وفي 13 أغسطس ظهر V4-Pro-0813 كحالة نشر أحدث مع تركيز معلن على تحسين قدرات الوكلاء. المهم هنا هو co-design بين long context، post-training، attention/KV efficiency، وserving economics.

- **R1:** دراسة emergence والاستدلال تحت verifiable reward.
- **V4:** جعل 1M context وtest-time scaling أكثر قابلية للتشغيل.
- **V4-Pro deployment:** نقل architecture إلى agentic workloads فعلية مع تحديث release-state.

مقارنات الورقة والجهة المطورة تبقى reported إلى أن تتكرر على harness مستقل؛ أما وجود model release وسياق مليون token فهو دليل deployment منفصل عن ادعاء التفوق.

Source ledger / دفتر المصدر

S06 · Research · 2026
DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence
arXiv:2606.19348; hybrid attention, million-token context, effort scaling, agentic benchmarks.

S07 · Research · 2026-08-13
DeepSeek V4-Pro-0813 launch and agent capability update
Current-release reporting on V4 Pro availability and improved AI-agent capabilities; used as deployment-state evidence, not as an independent technical paper.

S08 · Research · 2025
DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning
arXiv:2501.12948; RLVR-style reasoning emergence, cold start, distillation.

S01 · Official · 2026-07-09
GPT-5.6 System Card
Official system card; Sol/Terra/Luna, reasoning effort, agentic safety, prompt-injection and self-improvement evaluations.

S02 · Official · 2026-06-09
Claude Fable 5 & Mythos 5 System Card
Official system card for Anthropic's most capable widely released Fable 5 class; always-on adaptive thinking, long-running agents, 1M context and safety evaluations.

S03 · Official · 2026-05-19
Gemini 3.5: frontier intelligence with action
Official model report; action-oriented workflows, agentic coding, subagents and benchmark reporting.

S04 · Official · 2026-02
Gemini 3.1 Deep Think / Accelerating Mathematical and Scientific Discovery
Official research/model documentation for specialized high-compute scientific reasoning.

S05 · Official · 2026-08-12
Introducing Grok 4.6
Official current flagship release; long-running agents, coding, interactive/visual work, agentic RL, configurable reasoning effort and 500k context.

S09 · Official · 2026-08-02
Qwen3.8-Max: A New Bar for Coding and Cowork
Official current flagship report; 2.4T-class MoE, multimodal/long-context frontier, coding and cowork/agentic work. Vendor benchmark claims are treated as reported.

S10 · Research · 2026
Qwen3-Coder-Next Technical Report
arXiv:2603.00729; repository reasoning, agent interaction patterns, executable training/evaluation environments.

Deep technical extraction / الاستخراج الهندسي

كيف تقرأ model report. نحتاج خمسة مستويات: post-training، pretraining architecture/، safety/evaluation harness، tool/agent، context/memory، reasoning. لا ننقل أرقام benchmark بلا شروطها؛ نبحث عن design choices القابلة لإعادة الاختبار: memory، effort knob، recovery، tool protocol، mechanism، أو multimodal grounding.

- Baseline: نفس task set مع model أبسط أو effort أقل.
- Ablation: إزالة memory/tool/parallelism عند الإمكان.
- Cost: tokens، latency، tool calls، retries.
- Generalization: هل الميزة تفيد خارج benchmark الذي صُممت حوله؟

Research specification / مواصفة الاختبار

Variable / layer	Measure	Failure signal
Tool use	tool success + postcondition	function-call theater
Multimodal grounding	action/VQA with perturbations	visual misgrounding
Coding	repo-level tests / evolution	snippet overfit
Memory	reconstruction/recovery	history != memory
Safety envelope	injection/destructive-action tests	capability without control
Throughput	tokens/s under target context	benchmark-only speed
Open/closed control	ablation/debug access	hidden harness effects

Case Study: Qwen3.8-Max / Coder-Next / Omni

Frontier MoE, code environments and omnimodal grounding

Qwen3.8-Max يدفع عائلة Qwen إلى frontier MoE multimodal موجهة بقوة إلى coding وcowork، بينما Qwen3-Coder-Next يحتفظ بقيمة بحثية في repository reasoning والتفاعل داخل بيئات تنفيذ، وQwen3.5-Omni يضيف omnimodal reasoning/grounding عبر النص والصورة والصوت. هذا يظهر كيف تتحول عائلة model إلى substrate لعدة runtimes.

الدرس الهندسي هو separation of concerns: base general intelligence، ثم domain mid-training، ثم environment/harness. هذا يسمح بقياس أين جاءت الزيادة: weights العامة أم coder data أم tool scaffolding.

اختبار مطلوب: نفس backbone + نفس budget + ثلاثة harnesses مختلفة، ثم ablation للذاكرة والأدوات. من دون ذلك لا نعرف هل التحسن "model" أم "system".

Source ledger / دفتر المصدر

S09 · Official · 2026-08-02
Qwen3.8-Max: A New Bar for Coding and Cowork
Official current flagship report; 2.4T-class MoE, multimodal/long-context frontier, coding and cowork/agentic work. Vendor benchmark claims are treated as reported.

S10 · Research · 2026
Qwen3-Coder-Next Technical Report
arXiv:2603.00729; repository reasoning, agent interaction patterns, executable training/evaluation environments.

S11 · Research · 2026
Qwen3.5-Omni Technical Report
arXiv:2604.15804; hybrid-attention MoE omnimodal reasoning across text, vision and audio, long-sequence interaction and grounding.

S01 · Official · 2026-07-09
GPT-5.6 System Card
Official system card; Sol/Terra/Luna, reasoning effort, agentic safety, prompt-injection and self-improvement evaluations.

S02 · Official · 2026-06-09
Claude Fable 5 & Mythos 5 System Card
Official system card for Anthropic's most capable widely released Fable 5 class; always-on adaptive thinking, long-running agents, 1M context and safety evaluations.

S03 · Official · 2026-05-19
Gemini 3.5: frontier intelligence with action
Official model report; action-oriented workflows, agentic coding, subagents and benchmark reporting.

S04 · Official · 2026-02
Gemini 3.1 Deep Think / Accelerating Mathematical and Scientific Discovery
Official research/model documentation for specialized high-compute scientific reasoning.

S05 · Official · 2026-08-12
Introducing Grok 4.6
Official current flagship release; long-running agents, coding, interactive/visual work, agentic RL, configurable reasoning effort and 500k context.

S06 · Research · 2026
DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence
arXiv:2606.19348; hybrid attention, million-token context, effort scaling, agentic benchmarks.

S07 · Research · 2026-08-13
DeepSeek V4-Pro-0813 launch and agent capability update
Current-release reporting on V4 Pro availability and improved AI-agent capabilities; used as deployment-state evidence, not as an independent technical paper.

Deep technical extraction / الاستخراج الهندسي

كيف تقرأ model report. نحتاج خمسة مستويات: pretraining architecture/ post-training، safety/evaluation harness، tool/agent context/memory، reasoning. لا ننقل أرقام benchmark بلا شروطها؛ نبحث عن design choices القابلة لإعادة الاختبار: memory effort knob، recovery protocol، tool mechanism، أو multimodal grounding.

- Baseline: نفس task set مع model أبسط أو effort أقل.
- Ablation: إزالة memory/tool/parallelism عند الإمكان.
- Cost: tokens، latency، tool calls، retries.
- Generalization: هل الميزة تفيد خارج benchmark الذي صُممت حوله؟

Research specification / مواصفة الاختبار

Variable / layer	Measure	Failure signal
Tool use	tool success + postcondition	function-call theater
Multimodal grounding	action/VQA with perturbations	visual misgrounding
Coding	repo-level tests / evolution	snippet overfit
Memory	reconstruction/recovery	history != memory
Safety envelope	injection/destructive-action tests	capability without control
Throughput	tokens/s under target context	benchmark-only speed
Open/closed control	ablation/debug access	hidden harness effects

Case Study: Kimi K3

Open 3T-class long-horizon agentic intelligence

Kimi K3 يمثل قفزة من "agent feature" إلى co-design كامل: 2.8T MoE مع reasoning-effort levels، 1M context، native vision، 104B activated، وRL agentic million-token مع persistent rollout/sandbox state. التقرير يضع reasoning، knowledge work، long-horizon coding، وvision ضمن مسار post-training واحد، مع open weights.

أهم درس هندسي هو أن long-horizon agentic capability تحتاج model architecture + training curriculum + persistent environment state معاً. الحجم وحده لا يفسر ذلك. open weights تجعل K3 substrate مهماً للـablation، لكن ادعاءات المقارنة في التقرير تبقى self-reported حتى تكرر مستقلاً.

```
AgenticScale = ModelCapacity × PersistentRL × EnvironmentState × Verification
```

صلة MOUSSAIF: هذا يضع bounded HyperLoops، HyperRouting، وSelf-Reconstruction أمام baseline مفتوح قوي جداً بدلاً من مقارنة scaffold مع نموذج أقدم.

Source ledger / دفتر المصدر

S12 · Research · 2026
Kimi K3: Open Frontier Intelligence
arXiv:2607.24653; 2.8T MoE / 104B active, native vision, 1M context, reasoning-effort levels, long-horizon agentic RL and open weights.

S01 · Official · 2026-07-09
GPT-5.6 System Card
Official system card; Sol/Terra/Luna, reasoning effort, agentic safety, prompt-injection and self-improvement evaluations.

S02 · Official · 2026-06-09
Claude Fable 5 & Mythos 5 System Card
Official system card for Anthropic's most capable widely released Fable 5 class; always-on adaptive thinking, long-running agents, 1M context and safety evaluations.

S03 · Official · 2026-05-19
Gemini 3.5: frontier intelligence with action
Official model report; action-oriented workflows, agentic coding, subagents and benchmark reporting.

S04 · Official · 2026-02
Gemini 3.1 Deep Think / Accelerating Mathematical and Scientific Discovery
Official research/model documentation for specialized high-compute scientific reasoning.

S05 · Official · 2026-08-12
Introducing Grok 4.6
Official current flagship release; long-running agents, coding, interactive/visual work, agentic RL, configurable reasoning effort and 500k context.

S06 · Research · 2026
DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence
arXiv:2606.19348; hybrid attention, million-token context, effort scaling, agentic benchmarks.

S07 · Research · 2026-08-13
DeepSeek V4-Pro-0813 launch and agent capability update
Current-release reporting on V4 Pro availability and improved AI-agent capabilities; used as deployment-state evidence, not as an independent technical paper.

S09 · Official · 2026-08-02
Qwen3.8-Max: A New Bar for Coding and Cowork
Official current flagship report; 2.4T-class MoE, multimodal/long-context frontier, coding and cowork/agentic work. Vendor benchmark claims are treated as reported.

S10 · Research · 2026
Qwen3-Coder-Next Technical Report
arXiv:2603.00729; repository reasoning, agent interaction patterns, executable training/evaluation environments.

Deep technical extraction / الاستخراج الهندسي

كيف تقرأ model report. نحتاج خمسة مستويات: post-training/pretraining architecture، safety/evaluation harness، tool/agent، context/memory، reasoning. لا ننقل أرقام benchmark بلا شروطها؛ نبحث عن design choices القابلة لإعادة الاختبار: memory، effort knob، recovery، tool protocol، mechanism، أو multimodal grounding.

- Baseline: نفس task set مع model أبسط أو effort أقل.
- Ablation: إزالة memory/tool/parallelism عند الإمكان.
- Cost: tokens، latency، tool calls، retries.
- Generalization: هل الميزة تفيد خارج benchmark الذي صُممت حوله؟

Research specification / مواصفة الاختبار

Variable / layer	Measure	Failure signal
Memory	reconstruction/recovery	history != memory
Safety envelope	injection/destructive-action tests	capability without control
Throughput	tokens/s under target context	benchmark-only speed
Open/closed control	ablation/debug access	hidden harness effects
External replication	independent harness delta	vendor-only comparison
Activated compute	active params / token	capacity-cost confusion
Context fidelity	state retention under noise	attention dilution

Case Study: Grok 4.6 + gpt-oss

Current proprietary agent frontier vs open reasoning substrate

Grok 4.6، الصادر في 12 أغسطس 2026، يركز على interactive/visual work، coding، long-running agents، وagentic RL، مع reasoning effort يصل إلى xhigh و500k context. في المقابل gpt-oss يمثل open-weight reasoning substrate يمكن التحكم في effort وفحصه محلياً. المساران يكشفان trade-off مختلفاً بين integrated frontier harness وexperiment control.

الفرق البحثي: closed frontier يعطيك integrated system وtelemetry واسعة، بينما open weights تعطيك ablations، fine-tuning، experiment control، وinspection. برنامج AGI جاد يحتاج الاثنين: frontier external baselines + open experimental substrate.

> لا نساوي "open" بالشفافية الكاملة؛ weights وحدها لا تكشف data، training curriculum أو كل post-training details.

Source ledger / دفتر المصدر

S05 · Official · 2026-08-12
Introducing Grok 4.6
Official current flagship release; long-running agents, coding, interactive/visual work, agentic RL, configurable reasoning effort and 500k context.

S14 · Official · 2025-08-05
Introducing gpt-oss
Official open-weight reasoning release; controllable reasoning effort and open debugging/training affordances.

S01 · Official · 2026-07-09
GPT-5.6 System Card
Official system card; Sol/Terra/Luna, reasoning effort, agentic safety, prompt-injection and self-improvement evaluations.

S02 · Official · 2026-06-09
Claude Fable 5 & Mythos 5 System Card
Official system card for Anthropic's most capable widely released Fable 5 class; always-on adaptive thinking, long-running agents, 1M context and safety evaluations.

S03 · Official · 2026-05-19
Gemini 3.5: frontier intelligence with action
Official model report; action-oriented workflows, agentic coding, subagents and benchmark reporting.

S04 · Official · 2026-02
Gemini 3.1 Deep Think / Accelerating Mathematical and Scientific Discovery
Official research/model documentation for specialized high-compute scientific reasoning.

S06 · Research · 2026
DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence
arXiv:2606.19348; hybrid attention, million-token context, effort scaling, agentic benchmarks.

S07 · Research · 2026-08-13
DeepSeek V4-Pro-0813 launch and agent capability update
Current-release reporting on V4 Pro availability and improved AI-agent capabilities; used as deployment-state evidence, not as an independent technical paper.

S09 · Official · 2026-08-02
Qwen3.8-Max: A New Bar for Coding and Cowork
Official current flagship report; 2.4T-class MoE, multimodal/long-context frontier, coding and cowork/agentic work. Vendor benchmark claims are treated as reported.

S10 · Research · 2026
Qwen3-Coder-Next Technical Report
arXiv:2603.00729; repository reasoning, agent interaction patterns, executable training/evaluation environments.

Deep technical extraction / الاستخراج الهندسي

كيف تقرأ model report. نحتاج خمسة مستويات: post-training، pretraining architecture/، safety/evaluation harness، tool/agent، context/memory، reasoning. لا ننقل أرقام benchmark بلا شروطها؛ نبحث عن design choices القابلة لإعادة الاختبار: memory، effort knob، recovery، tool protocol، mechanism، أو multimodal grounding.

- Baseline: نفس task set مع model أبسط أو effort أقل.
- Ablation: إزالة memory/tool/parallelism عند الإمكان.
- Cost: tokens، latency، tool calls، retries.
- Generalization: هل الميزة تفيد خارج benchmark الذي صُممت حوله؟

Research specification / مواصفة الاختبار

Variable / layer	Measure	Failure signal
KV efficiency	memory/throughput at long context	bandwidth bottleneck
Effort knob	quality across effort levels	uncalibrated mode
Tool use	tool success + postcondition	function-call theater
Multimodal grounding	action/VQA with perturbations	visual misgrounding
Coding	repo-level tests / evolution	snippet overfit
Memory	reconstruction/recovery	history != memory
Safety envelope	injection/destructive-action tests	capability without control

مصفوفة المقارنة: ما الذي يتنافس فعلاً؟

Capabilities are vectors, not a single rank

المحور	سؤال القياس	فشل leaderboard التقليدي
Reasoning depth	هل يرفع effort الأداء على hard tail؟	score متوسط يخفي difficulty curve
Token efficiency	كم compute للوصول إلى نفس الجودة؟	لا يحسب latency/cost
Tool grounding	هل evidence الخارجي يغير الخطة؟	function-call success فقط
Long horizon	هل يحافظ على state ويستعيد الفشل؟	مهام قصيرة مستقلة
Computer use	هل perception/action loop مستقر؟	vision QA منفصل
Memory	هل يعرف ماذا يكتب/ينسى/يسترجع؟	long context كبديل زائف
Safety	هل autonomy محكومة بالتفويض؟	رفض نصي دون effect-level tests

لذلك "أفضل نموذج" عبارة ناقصة. نحتاج frontier vector وfrontier Pareto deployment-specific.

Source ledger / دفتر المصدر

S01 · Official · 2026-07-09
GPT-5.6 System Card
Official system card; Sol/Terra/Luna, reasoning effort, agentic safety, prompt-injection and self-improvement evaluations.

S02 · Official · 2026-06-09
Claude Fable 5 & Mythos 5 System Card
Official system card for Anthropic's most capable widely released Fable 5 class; always-on adaptive thinking, long-running agents, 1M context and safety evaluations.

S03 · Official · 2026-05-19
Gemini 3.5: frontier intelligence with action
Official model report; action-oriented workflows, agentic coding, subagents and benchmark reporting.

S06 · Research · 2026
DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence
arXiv:2606.19348; hybrid attention, million-token context, effort scaling, agentic benchmarks.

S12 · Research · 2026
Kimi K3: Open Frontier Intelligence
arXiv:2607.24653; 2.8T MoE / 104B active, native vision, 1M context, reasoning-effort levels, long-horizon agentic RL and open weights.

S04 · Official · 2026-02
Gemini 3.1 Deep Think / Accelerating Mathematical and Scientific Discovery
Official research/model documentation for specialized high-compute scientific reasoning.

S05 · Official · 2026-08-12
Introducing Grok 4.6
Official current flagship release; long-running agents, coding, interactive/visual work, agentic RL, configurable reasoning effort and 500k context.

S07 · Research · 2026-08-13
DeepSeek V4-Pro-0813 launch and agent capability update
Current-release reporting on V4 Pro availability and improved AI-agent capabilities; used as deployment-state evidence, not as an independent technical paper.

S09 · Official · 2026-08-02
Qwen3.8-Max: A New Bar for Coding and Cowork
Official current flagship report; 2.4T-class MoE, multimodal/long-context frontier, coding and cowork/agentic work. Vendor benchmark claims are treated as reported.

S10 · Research · 2026
Qwen3-Coder-Next Technical Report
arXiv:2603.00729; repository reasoning, agent interaction patterns, executable training/evaluation environments.

Deep technical extraction / الاستخراج الهندسي

قراءة المشهد. لا توجد جبهة واحدة؛ هناك خمسة محاور متداخلة: post-training، inference-reasoning، multimodal grounding، context/memory efficiency، time scaling، وagentic execution. النموذج قد يتصدر محوراً ويتراجع في آخر. لذلك الأفضل تمثيل كل عائلة بمتجه قدرة لا بترتيب كلي.

```
CapabilityVector = [R, TTS, Tool, Memory, CUA, SWE, Safety, Cost]
```

- افصل **native model capability** عن gains الناتجة من harness.
- قِس hard-tail بدلاً من المتوسط فقط.
- سجّل latency وtokens وtool calls إلى جانب accuracy.
- راقب degradation مع horizon؛ كثير من الفروق لا تظهر في المهام القصيرة.

Research specification / مواصفة الاختبار

Variable / layer	Measure	Failure signal
Throughput	tokens/s under target context	benchmark-only speed
Open/closed control	ablation/debug access	hidden harness effects
External replication	independent harness delta	vendor-only comparison
Activated compute	active params / token	capacity-cost confusion
Context fidelity	state retention under noise	attention dilution
KV efficiency	memory/throughput at long context	bandwidth bottleneck
Effort knob	quality across effort levels	uncalibrated mode

Contamination، Benchmark Saturation

When a score stops measuring progress

كلما صار benchmark مشهوراً، زادت احتمالات contamination، tuning غير مباشر، harness engineering، frontier research. ceiling effects. بدأ يتحرك إلى hidden tests، live benchmarks، realistic environments، economically valuable tasks. لكن حتى هذه يمكن أن تنحاز إلى tool configuration أو budget.

- افصل **model** عن **scaffold** في التقرير.
- ثبت tool budget، timeouts عند المقارنة.
- استخدم fresh/held-out tasks.
- راقب memorization indicators.
- انشر failure taxonomy لا score فقط.

```
BenchmarkValue ≈ Novelty × ExternalValidity × HarnessControl × FailureObservability
```

Source ledger / دفتر المصدر

S02 · Official · 2026-06-09
Claude Fable 5 & Mythos 5 System Card
Official system card for Anthropic's most capable widely released Fable 5 class; always-on adaptive thinking, long-running agents, 1M context and safety evaluations.

S06 · Research · 2026
DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence
arXiv:2606.19348; hybrid attention, million-token context, effort scaling, agentic benchmarks.

S10 · Research · 2026
Qwen3-Coder-Next Technical Report
arXiv:2603.00729; repository reasoning, agent interaction patterns, executable training/evaluation environments.

S29 · Research · 2026
The 2025 AI Agent Index: Documenting Technical and Safety Features of Deployed Agentic AI Systems
arXiv:2602.17753; deployed-agent capability/safety documentation.

S40 · Research · 2026
AgentAtlas: Beyond Outcome Leaderboards for LLM Agents
arXiv:2605.20530; behavioral/process evaluation beyond success rate.

S41 · Research · 2026
CUBE: A Standard for Unifying Agent Benchmarks
arXiv:2603.15798; common benchmark interface and environment standardization.

S01 · Official · 2026-07-09
GPT-5.6 System Card
Official system card; Sol/Terra/Luna, reasoning effort, agentic safety, prompt-injection and self-improvement evaluations.

S15 · Research · 2026
Inference Scaling, Learning to Reason, and Agentic Systems
arXiv:2504.09037v4; synthesis of inference scaling, verifiers, reasoning learning and agentic systems.

S18 · Research · 2024
Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters
arXiv:2408.03314; adaptive compute allocation and verifier-guided search.

S20 · Research · 2025
100 Days After DeepSeek-R1: A Survey on Replication Studies and More Directions for Reasoning Language Models
arXiv:2505.00551; replication lessons for SFT and RLVR.

Deep technical extraction / الاستخراج الهندسي

Benchmark hygiene. saturation تعني أن المقياس فقد التمييز، لا أن AGI تحقق. contamination، hidden retries، prompt overfitting، أو unequal tool budgets يمكن أن تعطي صورة مضللة. لذلك ينتقل frontier evaluation إلى long horizon، private tests، dynamic tasks، وprocess traces.

- سجل exact model snapshot والتاريخ.
- افصل pass@1 عن best-of-N.
- استخدم confidence intervals ومهام متعددة العائلات.
- احتفظ failure taxonomy بجانب success rate.

Research specification / مواصفة الاختبار

Variable / layer	Measure	Failure signal
Process traces	failure-stage labels	outcome-only blindness
Cost	latency/$/tool calls	capability without economics
Safety	impact/reversibility tests	unsafe success
Generalization	cross-domain held-outs	benchmark specialization
Statistics	CI/effect size	leaderboard noise
Replication	independent run	single-lab claim
Claim rank	established/reported/inference	marketing conflation

Overthinking, Rumination, Reward Hacking

More thought can make a model worse

تحليل Thoughtology وغيره يبين أن reasoning length ليست monotonic quality. بعض النماذج تعيد صياغة نفس الفرضية، تتمسك بمسار قديم أو تنتج "reflection theater". في RL، قد يتعلم model استغلال reward proxy. في agents، قد يتحول ذلك إلى loops فعلية تكلف أموالاً أو تعدّل ملفات.

مضادات

- marginal-value stopping.
- novelty-aware branch scoring.
- external evidence checkpoints.
- loop detector + budget hard cap.

علامات overthinking

- تكرار semantic مع tokens أكثر.
- انخفاض entropy دون evidence جديد.
- رفض التوقف رغم verifier قوي.
- إعادة فتح فروع محسومة.

```
Continue iff E[ΔUtility_next] > MarginalComputeCost
```

Deep technical extraction / الاستخراج الهندسي

Pathologies of reasoning. overthinking يظهر عندما ينفق النموذج compute دون information gain، يعيد صياغة الفرضِ نفسه، أو يبدل جواباً صحيحاً إلى خاطئ. reward hacking يظهر حين يتعلم policy مؤشرات verifier بدلاً من الحقيقة.

```
Progress_t = NewConstraint + ErrorRemoved + EvidenceGain
```

- إذا بقي Progress≈0 لعدة خطوات، جرّب verify/branch/stop.
- راقب token length-conditioned accuracy.
- اختبر adversarial verifier perturbations.
- لا تكافئ "التفصيل" بذاته؛ كافئ resolution القابل للتحقق.

Source ledger / دفتر المصدر

S21 · Research · 2025
DeepSeek-R1 Thoughtology: Let's <think> about LLM Reasoning
arXiv:2504.07128; analysis of reasoning length, rumination, control and safety.

S24 · Official · 2025-03-10
Detecting Misbehavior in Frontier Reasoning Models
Official research on monitoring reasoning traces for reward hacking and exploit behavior.

S22 · Research · 2026
Mid-Think: Training-Free Intermediate-Budget Reasoning
arXiv:2601.07036; inference budget control between short and long reasoning modes.

S08 · Research · 2025
DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning
arXiv:2501.12948; RLVR-style reasoning emergence, cold start, distillation.

S15 · Research · 2026
Inference Scaling, Learning to Reason, and Agentic Systems
arXiv:2504.09037v4; synthesis of inference scaling, verifiers, reasoning learning and agentic systems.

S16 · Research · 2026
A Survey of Process Reward Models: From Outcome Supervision to Process Supervision
arXiv:2510.08049; PRM data, training, inference-time use and RL.

S17 · Research · 2026
Reward Modeling for Reinforcement Learning-Based LLM Reasoning
arXiv:2602.09305; model-based, rule-based and self-reward designs.

S20 · Research · 2025
100 Days After DeepSeek-R1: A Survey on Replication Studies and More Directions for Reasoning Language Models
arXiv:2505.00551; replication lessons for SFT and RLVR.

S23 · Official · 2025-12-18
Evaluating Chain-of-Thought Monitorability
Official framework with intervention/process/outcome-property evaluations of CoT monitorability.

S18 · Research · 2024
Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters
arXiv:2408.03314; adaptive compute allocation and verifier-guided search.

Research specification / مواصفة الاختبار

Variable / layer	Measure	Failure signal
Tool feedback	environment pass rate	false positive tool signal
Safety reward	adversarial exploit rate	gaming safer-looking traces
Calibration	predicted vs empirical success	overconfidence
Reproducibility	seed variance / run-to-run stability	unstable post-training
Reward validity	cross-verifier agreement; exact tests	shortcut/reward hacking
Rollout diversity	unique paths; entropy; pass@N	mode collapse
KL / policy drift	KL and behavior delta	over-regularization or drift

Chain-of-Thought Monitorability

Reasoning traces as a safety sensor

OpenAI خصص عملاً لقياس monitorability بدلاً من افتراض أن CoT مرآة للنوايا. الإطار يميز process، intervention، وoutcome-property evaluations. الهدف هو سؤال قابل للقياس: هل monitor يستطيع التقاط misbehavior قبل النتيجة؟ وهل training ضد monitor يجعل traces أقل صدقاً؟

هذا يفرض فصلين: **reasoning quality** و**reasoning observability**. قد يتحسن الأول بينما يسوء الثاني. ومع انتقال النماذج إلى hidden/internal reasoning، يصبح الاعتماد على trace وحدها أضعف؛ لذلك نحتاج telemetry من state changes، tool calls، actions، وmemory writes.

> في MOUSSAIF، Courts يجب أن تقرأ trace كدليل واحد ضمن Evidence Graph، لا كحقيقة سيادية.

Source ledger / دفتر المصدر

S23 · Official · 2025-12-18
Evaluating Chain-of-Thought Monitorability
Official framework with intervention/process/outcome-property evaluations of CoT monitorability.

S24 · Official · 2025-03-10
Detecting Misbehavior in Frontier Reasoning Models
Official research on monitoring reasoning traces for reward hacking and exploit behavior.

S01 · Official · 2026-07-09
GPT-5.6 System Card
Official system card; Sol/Terra/Luna, reasoning effort, agentic safety, prompt-injection and self-improvement evaluations.

S15 · Research · 2026
Inference Scaling, Learning to Reason, and Agentic Systems
arXiv:2504.09037v4; synthesis of inference scaling, verifiers, reasoning learning and agentic systems.

S18 · Research · 2024
Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters
arXiv:2408.03314; adaptive compute allocation and verifier-guided search.

S19 · Research · 2024
Simple and Provable Scaling Laws for the Test-Time Compute of Large Language Models
arXiv:2411.19477; tournament/league aggregation and scaling guarantees under assumptions.

S21 · Research · 2025
DeepSeek-R1 Thoughtology: Let's <think> about LLM Reasoning
arXiv:2504.07128; analysis of reasoning length, rumination, control and safety.

S22 · Research · 2026
Mid-Think: Training-Free Intermediate-Budget Reasoning
arXiv:2601.07036; inference budget control between short and long reasoning modes.

S04 · Official · 2026-02
Gemini 3.1 Deep Think / Accelerating Mathematical and Scientific Discovery
Official research/model documentation for specialized high-compute scientific reasoning.

S08 · Research · 2025
DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning
arXiv:2501.12948; RLVR-style reasoning emergence, cold start, distillation.

Deep technical extraction / الاستخراج الهندسي

Monitorability. reasoning traces مفيدة للرصد إذا كانت مرتبطة بالسياسة الفعلية، لكنها ليست مرآة مضمونة للحساب الداخلي. لذلك نختبر monitorability تجريبياً: هل التلاعب بالCoT يغير outcome؟ هل يمكن monitor اكتشاف misbehavior قبل ظهور الضرر؟ وهل التدريب القوي على إخفاء الأثر يقلل الإشارة؟

- process-property tests.
- intervention tests على أجزاء من trace.
- outcome correlation تحت tasks adversarial.
- comparison بين hidden/visible scratchpad policies.

Research specification / مواصفة الاختبار

Variable / layer	Measure	Failure signal
Verifier precision	selection accuracy	bad pruning
Stopping rule	marginal gain/token	overthinking
Parallelism	wall-clock speedup	merge overhead
Self-consistency	agreement vs correctness	correlated error
Tool escalation	gain after external evidence	unnecessary calls
Uncertainty	calibration / abstention	false certainty
Latency	p50/p95 end-to-end	unusable deep mode

إلى أي حد اقتربنا من AGI؟

What improved, what remains structurally open

ما تحسن بوضوح: reasoning تحت compute أعلى، برمجة long-context، computer interaction، web research، tool use، multimodal perception، repo-level، وبعض multi-agent orchestration. ما لم يُحسم: robust transfer إلى مهام غير ممثلة، continual learning، long-horizon reliability الآمن، self-modeling الصادق، causal world models العامة، وbounded autonomy تحت أهداف متغيرة.

لذلك أقرب تعريف هندسي لـ near-AGI هنا هو **verification** × **adaptivity** × **persistence** × **coverage**، لا benchmark IQ. يمكن لنظام أن يملك reasoning قوي لكنه ينهار بعد 200 خطوة؛ أو agent قوي لكنه بلا calibration.

```
GeneralAgentCapability ≈ Breadth × Adaptation × Horizon × Grounding × Recoverability
```

Source ledger / دفتر المصدر

S29 · Research · 2026
The 2025 AI Agent Index: Documenting Technical and Safety Features of Deployed Agentic AI Systems
arXiv:2602.17753; deployed-agent capability/safety documentation.

S31 · Research · 2026
The Long-Horizon Task Mirage? Diagnosing Where and Why Agents Fail as Task Horizon Increases
arXiv:2604.11978; cross-domain degradation with horizon length.

S32 · Research · 2025
UltraHorizon: Benchmarking Agent Capabilities in Ultra Long-Horizon Scenarios
arXiv:2509.21766; 200k+ token / 400+ tool-call trajectories and long-horizon failure taxonomy.

S33 · Research · 2026
OSWorld 2.0: Benchmarking Computer Use Agents on Long-Horizon Real-World Tasks
arXiv:2606.29537; realistic long-horizon computer-use evaluation.

S45 · Research · 2026
Agents in the Wild: Where Research Meets Deployment
arXiv:2607.19336; deployment-oriented synthesis of reliability, memory, tools and evaluation.

S01 · Official · 2026-07-09
GPT-5.6 System Card
Official system card; Sol/Terra/Luna, reasoning effort, agentic safety, prompt-injection and self-improvement evaluations.

S15 · Research · 2026
Inference Scaling, Learning to Reason, and Agentic Systems
arXiv:2504.09037v4; synthesis of inference scaling, verifiers, reasoning learning and agentic systems.

S18 · Research · 2024
Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters
arXiv:2408.03314; adaptive compute allocation and verifier-guided search.

S20 · Research · 2025
100 Days After DeepSeek-R1: A Survey on Replication Studies and More Directions for Reasoning Language Models
arXiv:2505.00551; replication lessons for SFT and RLVR.

S21 · Research · 2025
DeepSeek-R1 Thoughtology: Let's <think> about LLM Reasoning
arXiv:2504.07128; analysis of reasoning length, rumination, control and safety.

Deep technical extraction / الاستخراج الهندسي

قرب AGI كمنتجه. الدليل الأقوى ليس لقباً تسويقياً بل انخفاض الحاجة إلى task-specific scaffolding مع بقاء reliability. نبحث عن transfer إلى tools جديدة، open objectives، persistent self-recovery، world evidence acquisition، وability to operate under novel constraints.

- **Breadth:** عدد عائلات المهام غير المعلّبة مسبقاً.
- **Depth:** hard-tail reasoning under bounded compute.
- **Agency:** closed-loop action with verification.
- **Adaptation:** تعديل strategy دون weight update مباشر.

```
NearAGIIndex := Breadth×Depth×Adaptation×Reliability (research heuristic)
```

Research specification / مواصفة الاختبار

Variable / layer	Measure	Failure signal
Contamination	private/dynamic tests	memorization
Horizon	success vs task length	short-task optimism
Process traces	failure-stage labels	outcome-only blindness
Cost	latency/$/tool calls	capability without economics
Safety	impact/reversibility tests	unsafe success
Generalization	cross-domain held-outs	benchmark specialization
Statistics	CI/effect size	leaderboard noise

Bridge I → MOUSSAIF

Reasoning frontier mapped to the internal OS

Frontier concept	MOUSSAIF layer	Experiment
reasoning effort	Vascular Compute / MCD	allocate tokens by active difference
parallel sampling	QMCD branching	branch only when expected gain positive
PRM / verifier	Falsification Court	heterogeneous verifier ensemble
tool-integrated reasoning	HyperSkills + Tool Contracts	post-condition-verified tool use
long context	HKOS + Self-Reconstruction	raw context vs stateful memory ablation
reasoning monitorability	Trace / Audit Plane	predict failures before world effect

هذه المطابقة لا تثبت تفوق MOUSSAIF؛ إنها تحول النظرية إلى matrix تجارب. إذا لم يتغير الأداء بعد إضافة طبقة، تُخفض قيمتها.

Source ledger / دفتر المصدر

S15 · Research · 2026
Inference Scaling, Learning to Reason, and Agentic Systems
arXiv:2504.09037v4; synthesis of inference scaling, verifiers, reasoning learning and agentic systems.

S16 · Research · 2026
A Survey of Process Reward Models: From Outcome Supervision to Process Supervision
arXiv:2510.08049; PRM data, training, inference-time use and RL.

S18 · Research · 2024
Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters
arXiv:2408.03314; adaptive compute allocation and verifier-guided search.

S23 · Official · 2025-12-18
Evaluating Chain-of-Thought Monitorability
Official framework with intervention/process/outcome-property evaluations of CoT monitorability.

S01 · Official · 2026-07-09
GPT-5.6 System Card
Official system card; Sol/Terra/Luna, reasoning effort, agentic safety, prompt-injection and self-improvement evaluations.

S20 · Research · 2025
100 Days After DeepSeek-R1: A Survey on Replication Studies and More Directions for Reasoning Language Models
arXiv:2505.00551; replication lessons for SFT and RLVR.

S21 · Research · 2025
DeepSeek-R1 Thoughtology: Let's <think> about LLM Reasoning
arXiv:2504.07128; analysis of reasoning length, rumination, control and safety.

S24 · Official · 2025-03-10
Detecting Misbehavior in Frontier Reasoning Models
Official research on monitoring reasoning traces for reward hacking and exploit behavior.

S29 · Research · 2026
The 2025 AI Agent Index: Documenting Technical and Safety Features of Deployed Agentic AI Systems
arXiv:2602.17753; deployed-agent capability/safety documentation.

S30 · Research · 2026
Benchmark Test-Time Scaling of General LLM Agents
arXiv:2602.18998; sequential vs parallel scaling across search, coding, reasoning and tools.

Deep technical extraction / الاستخراج الهندسي

الربط المنهجي. لا نبحث عن تشابه أسماء بين MOUSSAIF والأدبيات، بل عن equivalence تشغيلية يمكن هدمها. مثال: MCD لا يصبح test-time scaling إلا إذا تنبأ فعلاً بأي tasks تستحق compute إضافيا. Self-Reconstruction Memory لا تصبح أفضل من context الطويل إلا إذا حسنت recovery أو خفضت cost تحت نفس model.

- لكل bridge: mechanism، baseline، metric، falsifier.
- إذا لم يغير القرار، يخفض المفهوم إلى وصف.
- إذا كرر feature قياسياً، يعاد تسميته كـadapter لا "نظرية جديدة".
- القيمة في differential gain لا في التوافق اللغوي.

Research specification / مواصفة الاختبار

Variable / layer	Measure	Failure signal
Safety	impact/reversibility tests	unsafe success
Generalization	cross-domain held-outs	benchmark specialization
Statistics	CI/effect size	leaderboard noise
Replication	independent run	single-lab claim
Claim rank	established/reported/inference	marketing conflation
Snapshot control	exact version/date	moving target
Budget parity	tokens/tools/retries equalized	unfair ranking

PART II — FRONTIER AGENTIC SYSTEMS

من الإجابة إلى تنفيذ مهمة ممتدة

الوكيل في 2026 لم يعد prompt loop بسيطاً. الأنظمة الأقوى تجمع browser/computer control، tool runtime، reasoning model، verification، subagents، background execution، permissions، environment state، memory، وrecovery. هذا الجزء يركز على **system-level frontier** لا على model benchmark فقط.

السؤال الحاكم: ما الذي يجعل agent قادراً على الاستمرار في مهمة حقيقية لعشرات أو مئات الخطوات دون أن يفقد الهدف، يخترع حالة، يعيد العمل أو يتجاوز التفويض؟

Source ledger / دفتر المصدر

S25 · Official · 2025-07-17
Introducing ChatGPT Agent: Bridging Research and Action
Official deployed agent combining research, browsing, computer interaction and action.

S26 · Official · 2025-02-02
Introducing Deep Research
Official reasoning agent for multi-step web research and source synthesis.

S27 · Official · 2026-05-19
Co-Scientist: A Multi-Agent AI Partner to Accelerate Research
Official multi-agent scientific discovery system for hypothesis generation and evolution.

S28 · Official · 2026-02-18
Measuring AI Agent Autonomy in Practice
Official empirical analysis of user oversight and autonomous execution in Claude Code.

S29 · Research · 2026
The 2025 AI Agent Index: Documenting Technical and Safety Features of Deployed Agentic AI Systems
arXiv:2602.17753; deployed-agent capability/safety documentation.

S30 · Research · 2026
Benchmark Test-Time Scaling of General LLM Agents
arXiv:2602.18998; sequential vs parallel scaling across search, coding, reasoning and tools.

S31 · Research · 2026
The Long-Horizon Task Mirage? Diagnosing Where and Why Agents Fail as Task Horizon Increases
arXiv:2604.11978; cross-domain degradation with horizon length.

S32 · Research · 2025
UltraHorizon: Benchmarking Agent Capabilities in Ultra Long-Horizon Scenarios
arXiv:2509.21766; 200k+ token / 400+ tool-call trajectories and long-horizon failure taxonomy.

S40 · Research · 2026
AgentAtlas: Beyond Outcome Leaderboards for LLM Agents
arXiv:2605.20530; behavioral/process evaluation beyond success rate.

S41 · Research · 2026
CUBE: A Standard for Unifying Agent Benchmarks
arXiv:2603.15798; common benchmark interface and environment standardization.

Deep technical extraction / الاستخراج الهندسي

Agent architecture الحلقة الدنيا هي Sense → Update State → Select Goal/Action → Execute → Observe → Verify. frontier يضيف subagents، tool registry، planner، memory، وpolicy gates وtrace store. التحدي هو حفظ invariantات النظام عبر كل دورة.

```
State_{t+1}=Transition(State_t, Obs_t, Evidence_t, Policy)
```

- لا تسمح لوكيل فرعي بتغيير objective الأعلى بلا authorization.
- كل action خارجي يحمل idempotency/rollback policy إن أمكن.
- completion يحتاج evidence وليس self-report.
- persistent agent يحتاج explicit stop/sleep/resume semantics.

Research specification / مواصفة الاختبار

Variable / layer	Measure	Failure signal
Checkpoint	resume success	catastrophic restart
World effect	expected vs observed delta	phantom completion
Stop policy	loop rate / runaway cost	non-termination
Goal state	goal checksum / drift rate	subgoal takeover
Mission DAG	dependency violations	linear-plan brittleness
Action schema	invalid-call rate	ambiguous tool use
Postconditions	verified completion rate	self-reported success

معمارية General Agent

A POMDP with tools, memory and policy gates

أفضل abstraction هو POMDP: agent لا يرى العالم كاملاً؛ يملك belief state، يختار action، يحصل على observation، ويحدث الذاكرة. LLM يشارك في policy لكنه ليس النظام كله.

```
Agent = (Policy, BeliefState, Memory, Tools, Environment, Constraints, Verifiers)
```

```
a_t ~ π(a | b_t, goal, budget, policy); b_{t+1}=F(b_t,o_t,m_t)
```

الـharness يحدد checkpointing، concurrency، retries، tool schemas، action space، وpermissions. لذلك نفس model قد يظهر agentic capability مختلفة جذرياً عبر harnessين.

- Goal decomposition
- State management
- Tool routing
- Memory write/read policy
- Verification and rollback
- Human escalation

Source ledger / دفتر المصدر

S29 · Research · 2026
The 2025 AI Agent Index: Documenting Technical and Safety Features of Deployed Agentic AI Systems
arXiv:2602.17753; deployed-agent capability/safety documentation.

S30 · Research · 2026
Benchmark Test-Time Scaling of General LLM Agents
arXiv:2602.18998; sequential vs parallel scaling across search, coding, reasoning and tools.

S45 · Research · 2026
Agents in the Wild: Where Research Meets Deployment
arXiv:2607.19336; deployment-oriented synthesis of reliability, memory, tools and evaluation.

S25 · Official · 2025-07-17
Introducing ChatGPT Agent: Bridging Research and Action
Official deployed agent combining research, browsing, computer interaction and action.

S28 · Official · 2026-02-18
Measuring AI Agent Autonomy in Practice
Official empirical analysis of user oversight and autonomous execution in Claude Code.

S31 · Research · 2026
The Long-Horizon Task Mirage? Diagnosing Where and Why Agents Fail as Task Horizon Increases
arXiv:2604.11978; cross-domain degradation with horizon length.

S32 · Research · 2025
UltraHorizon: Benchmarking Agent Capabilities in Ultra Long-Horizon Scenarios
arXiv:2509.21766; 200k+ token / 400+ tool-call trajectories and long-horizon failure taxonomy.

S40 · Research · 2026
AgentAtlas: Beyond Outcome Leaderboards for LLM Agents
arXiv:2605.20530; behavioral/process evaluation beyond success rate.

S41 · Research · 2026
CUBE: A Standard for Unifying Agent Benchmarks
arXiv:2603.15798; common benchmark interface and environment standardization.

S43 · Protocol · 2026-07-28
MCP Specification 2026-07-28
Stable open protocol revision for tool/context integration, tasks, subscriptions, caching and multi-round trips.

Deep technical extraction / الاستخراج الهندسي

Agent architecture. الحلقة الدنيا هي Sense → Update State → Select Goal/Action → Execute → Observe → Verify. frontier يضيف tool registry، subagents، planner، memory، وpolicy gates. التحدي هو حفظ invariantات النظام عبر كل دورة. trace store

```
State_{t+1}=Transition(State_t, Obs_t, Evidence_t, Policy)
```

- لا تسمح لوكيل فرعي بتغيير objective الأعلى بلا authorization.
- كل action خارجي يحمل idempotency/rollback policy إن أمكن.
- completion يحتاج evidence وليس self-report.
- persistent agent يحتاج explicit stop/sleep/resume semantics.

Research specification / مواصفة الاختبار

Variable / layer	Measure	Failure signal
Postconditions	verified completion rate	self-reported success
Permissions	least-privilege violations	overbroad autonomy
Memory	reconstruction fidelity	state loss
Retries	recovery gain / duplicate side effects	blind retry
Human gates	necessary escalation precision	confirmation fatigue
Trace	coverage of decisions/actions	undebuggable failures
Checkpoint	resume success	catastrophic restart

Agent Harness vs Foundation Model

Where does capability actually live?

في 2026، جزء متزايد من performance يأتي من harness engineering: prompt/state formatting، parallel tool calls، code execution، caching، browser abstraction، memory and retry policies. لذلك يجب أن يسجل كل benchmark: model version، system prompt، tools، timeouts، token budget، number of attempts، computer image size، network policy.

```
ObservedPerformance = f(Model, Harness, Tools, Budget, Environment, Grader)
```

هذه المعادلة البسيطة تمنع خطأً شائعاً: نسب كل تحسين إلى weights. CUBE وAgentAtlas يدفعان نحو standardization/process analysis لأن outcome leaderboard وحده لا يكفي.

Deep technical extraction / الاستخراج الهندسي

Harness contribution. system prompt، tool API، browser abstraction، retry policy، memory، caching، parallelization قد تغير النتيجة أكثر من فرق model صغير. لذلك أي تقييم علمي للagents يجب أن يثبت harness أو يعامله كمتغير مستقل.

- سجل prompt and tool versions.
- ثبّت timeout/network/sandbox conditions.
- قِس tool errors separately from reasoning errors.
- قَارن base model داخل أكثر من harness والعكس.

```
ΔPerformance = ΔModel + ΔHarness + Interaction(Model,Harness)
```

Source ledger / دفتر المصدر

S40 · Research · 2026
AgentAtlas: Beyond Outcome Leaderboards for LLM Agents
arXiv:2605.20530; behavioral/process evaluation beyond success rate.

S41 · Research · 2026
CUBE: A Standard for Unifying Agent Benchmarks
arXiv:2603.15798; common benchmark interface and environment standardization.

S29 · Research · 2026
The 2025 AI Agent Index: Documenting Technical and Safety Features of Deployed Agentic AI Systems
arXiv:2602.17753; deployed-agent capability/safety documentation.

S25 · Official · 2025-07-17
Introducing ChatGPT Agent: Bridging Research and Action
Official deployed agent combining research, browsing, computer interaction and action.

S28 · Official · 2026-02-18
Measuring AI Agent Autonomy in Practice
Official empirical analysis of user oversight and autonomous execution in Claude Code.

S30 · Research · 2026
Benchmark Test-Time Scaling of General LLM Agents
arXiv:2602.18998; sequential vs parallel scaling across search, coding, reasoning and tools.

S31 · Research · 2026
The Long-Horizon Task Mirage? Diagnosing Where and Why Agents Fail as Task Horizon Increases
arXiv:2604.11978; cross-domain degradation with horizon length.

S32 · Research · 2025
UltraHorizon: Benchmarking Agent Capabilities in Ultra Long-Horizon Scenarios
arXiv:2509.21766; 200k+ token / 400+ tool-call trajectories and long-horizon failure taxonomy.

S43 · Protocol · 2026-07-28
MCP Specification 2026-07-28
Stable open protocol revision for tool/context integration, tasks, subscriptions, caching and multi-round trips.

S44 · Protocol · 2025
Agent2Agent (A2A) Protocol
Open protocol for interoperable agent-to-agent communication across vendors/ frameworks.

Research specification / مواصفة الاختبار

Variable / layer	Measure	Failure signal
Permissions	least-privilege violations	overbroad autonomy
Memory	reconstruction fidelity	state loss
Retries	recovery gain / duplicate side effects	blind retry
Human gates	necessary escalation precision	confirmation fatigue
Trace	coverage of decisions/actions	undebuggable failures
Checkpoint	resume success	catastrophic restart
World effect	expected vs observed delta	phantom completion

Planning، Mission Graphs

Plan once is not enough

long-horizon agent يحتاج plan قابل للتعديل، لا chain ثابتة. الأفضل تمثيل المهمة كـDAG من subgoals مع rollback points، evidence requirements، dependencies، وstop conditions. كل node قد ينفذ skill أو subagent.

```
MissionDAG = (V_tasks, E_dep, Preconditions, Postconditions, Evidence)
```

HORIZON وUltraHorizon يبرزان أن الفشل يزيد مع horizon، غالباً بسبب compounding errors، state drift، وin-context locking. الحل ليس فقط "خطة أطول"، بل plan repair عند observation جديد.

- Replan on invalidated assumption.
- Checkpoint after irreversible side effects.
- Separate discovery from execution.
- Attach evidence to each completed node.

Source ledger / دفتر المصدر

S31 · Research · 2026
The Long-Horizon Task Mirage? Diagnosing Where and Why Agents Fail as Task Horizon Increases
arXiv:2604.11978; cross-domain degradation with horizon length.

S32 · Research · 2025
UltraHorizon: Benchmarking Agent Capabilities in Ultra Long-Horizon Scenarios
arXiv:2509.21766; 200k+ token / 400+ tool-call trajectories and long-horizon failure taxonomy.

S45 · Research · 2026
Agents in the Wild: Where Research Meets Deployment
arXiv:2607.19336; deployment-oriented synthesis of reliability, memory, tools and evaluation.

S25 · Official · 2025-07-17
Introducing ChatGPT Agent: Bridging Research and Action
Official deployed agent combining research, browsing, computer interaction and action.

S28 · Official · 2026-02-18
Measuring AI Agent Autonomy in Practice
Official empirical analysis of user oversight and autonomous execution in Claude Code.

S29 · Research · 2026
The 2025 AI Agent Index: Documenting Technical and Safety Features of Deployed Agentic AI Systems
arXiv:2602.17753; deployed-agent capability/safety documentation.

S30 · Research · 2026
Benchmark Test-Time Scaling of General LLM Agents
arXiv:2602.18998; sequential vs parallel scaling across search, coding, reasoning and tools.

S40 · Research · 2026
AgentAtlas: Beyond Outcome Leaderboards for LLM Agents
arXiv:2605.20530; behavioral/process evaluation beyond success rate.

S41 · Research · 2026
CUBE: A Standard for Unifying Agent Benchmarks
arXiv:2603.15798; common benchmark interface and environment standardization.

S43 · Protocol · 2026-07-28
MCP Specification 2026-07-28
Stable open protocol revision for tool/context integration, tasks, subscriptions, caching and multi-round trips.

Deep technical extraction / الاستخراج الهندسي

Planning under changing state. mission DAG أفضل من plan نصي ثابت لأنه يربط dependencies وpre/postconditions ويتيح repair موضعي. plan node المكتمل يحتاج evidence، وإلا قد تبنى بقية السلسلة على إنجاز وهمي.

- ضع checkpoints قبل side effects غير القابلة للعكس.
- أعد فتح node إذا ظهر invalidated assumption لاحقاً.
- فرّق discovery nodes عن execution nodes.
- استخدم critical path لتخصيص compute والparallelism.

Research specification / مواصفة الاختبار

Variable / layer	Measure	Failure signal
Retries	recovery gain / duplicate side effects	blind retry
Human gates	necessary escalation precision	confirmation fatigue
Trace	coverage of decisions/actions	undebuggable failures
Checkpoint	resume success	catastrophic restart
World effect	expected vs observed delta	phantom completion
Stop policy	loop rate / runaway cost	non-termination
Goal state	goal checksum / drift rate	subgoal takeover

Tool Use كفضاء فعل

Tools are part of cognition and risk

كل أداة توسع action space وتضيف failure surface. search يمكن أن يجلب prompt injection؛ shell يمكن أن يدمر ملفات؛ database tool يمكن أن يكتب state خاطئة. لذلك tool contract يحتاج postconditions، idempotency، authorization، schema، وaudit record.

```
ToolRisk = Impact × Irreversibility × Uncertainty × Privilege
```

frontier systems تضيف confirmations أو policy gates للأفعال الحساسة. GPT-5.6 مثلاً يوثق computer-use confirmation policy، بينما MCP يحاول توحيد الاتصال بالأدوات والموارد. هندسياً، الأداة ليست "function"؛ هي capability تحت governance.

Source ledger / دفتر المصدر

S01 · Official · 2026-07-09
GPT-5.6 System Card
Official system card; Sol/Terra/Luna, reasoning effort, agentic safety, prompt-injection and self-improvement evaluations.

S43 · Protocol · 2026-07-28
MCP Specification 2026-07-28
Stable open protocol revision for tool/context integration, tasks, subscriptions, caching and multi-round trips.

S25 · Official · 2025-07-17
Introducing ChatGPT Agent: Bridging Research and Action
Official deployed agent combining research, browsing, computer interaction and action.

S28 · Official · 2026-02-18
Measuring AI Agent Autonomy in Practice
Official empirical analysis of user oversight and autonomous execution in Claude Code.

S29 · Research · 2026
The 2025 AI Agent Index: Documenting Technical and Safety Features of Deployed Agentic AI Systems
arXiv:2602.17753; deployed-agent capability/safety documentation.

S30 · Research · 2026
Benchmark Test-Time Scaling of General LLM Agents
arXiv:2602.18998; sequential vs parallel scaling across search, coding, reasoning and tools.

S31 · Research · 2026
The Long-Horizon Task Mirage? Diagnosing Where and Why Agents Fail as Task Horizon Increases
arXiv:2604.11978; cross-domain degradation with horizon length.

S32 · Research · 2025
UltraHorizon: Benchmarking Agent Capabilities in Ultra Long-Horizon Scenarios
arXiv:2509.21766; 200k+ token / 400+ tool-call trajectories and long-horizon failure taxonomy.

S40 · Research · 2026
AgentAtlas: Beyond Outcome Leaderboards for LLM Agents
arXiv:2605.20530; behavioral/process evaluation beyond success rate.

S41 · Research · 2026
CUBE: A Standard for Unifying Agent Benchmarks
arXiv:2603.15798; common benchmark interface and environment standardization.

Deep technical extraction / الاستخراج الهندسي

Capability governance. العقد الجيد يصف tool يوسع action space ويزيد blast radius. audit evidence. least وarguments، privileges، side effects، retry semantics، postconditions أهم من confidence العالي.

```
ActionGate = Authorized ∧ Preconditions ∧ Risk≤Threshold
```

- read-only افتراضياً عند discovery.
- confirmation للـfinancial/destructive/identity-sensitive actions.
- sanitize untrusted tool/web outputs.
- post-condition inspection بعد write/delete/send.

Research specification / مواصفة الاختبار

Variable / layer	Measure	Failure signal
Stop policy	loop rate / runaway cost	non-termination
Goal state	goal checksum / drift rate	subgoal takeover
Mission DAG	dependency violations	linear-plan brittleness
Action schema	invalid-call rate	ambiguous tool use
Postconditions	verified completion rate	self-reported success
Permissions	least-privilege violations	overbroad autonomy
Memory	reconstruction fidelity	state loss

Computer Use Agents

Vision-action loops in real operating systems

Computer-use agent يواجه بيئة مرئية جزئياً يمكن أن تتغير بين screenshot وآخر. النجاح يتطلب perception، grounding، action planning، timing، error recovery وstate freshness. OSWorld 2.0 يدفع التقييم إلى long-horizon real-world tasks بدلاً من click puzzles.

```
CUA Success = Perception × Grounding × Action × Recovery × StateFreshness
```

- خطأ pixel صغير قد يغير action effect بالكامل.
- UI latency يجعل observation قديمة.
- modals/permissions تعطل trajectory.
- dangerous actions تحتاج confirmation لا مجرد confidence.

هذه بيئة ممتازة لاختبار World-Effect Loop لأن كل خطوة قابلة للملاحظة وإعادة الفحص.

Source ledger / دفتر المصدر

S33 · Research · 2026
OSWorld 2.0: Benchmarking Computer Use Agents on Long-Horizon Real-World Tasks
arXiv:2606.29537; realistic long-horizon computer-use evaluation.

S12 · Research · 2026
Kimi K3: Open Frontier Intelligence
arXiv:2607.24653; 2.8T MoE / 104B active, native vision, 1M context, reasoning-effort levels, long-horizon agentic RL and open weights.

S01 · Official · 2026-07-09
GPT-5.6 System Card
Official system card; Sol/Terra/Luna, reasoning effort, agentic safety, prompt-injection and self-improvement evaluations.

S02 · Official · 2026-06-09
Claude Fable 5 & Mythos 5 System Card
Official system card for Anthropic's most capable widely released Fable 5 class; always-on adaptive thinking, long-running agents, 1M context and safety evaluations.

S25 · Official · 2025-07-17
Introducing ChatGPT Agent: Bridging Research and Action
Official deployed agent combining research, browsing, computer interaction and action.

S26 · Official · 2025-02-02
Introducing Deep Research
Official reasoning agent for multi-step web research and source synthesis.

S27 · Official · 2026-05-19
Co-Scientist: A Multi-Agent AI Partner to Accelerate Research
Official multi-agent scientific discovery system for hypothesis generation and evolution.

S34 · Research · 2026
Odysseys: Benchmarking Web Agents on Realistic Long Horizon Tasks
arXiv:2604.24964; realistic web task horizon and interaction complexity.

S35 · Research · 2026
SWE-EVO: Benchmarking Coding Agents in Long-Horizon Software Evolution
arXiv:2512.18470; multi-step repository evolution beyond issue-level patching.

S39 · Research · 2026
SentinelBench: A Benchmark for Long-Running Monitoring Agents
arXiv:2606.05342; persistent monitoring as a distinct long-running agent workload.

Deep technical extraction / الاستخراج الهندسي

GUI agent واجهة الكمبيوتر partially observable وغير مستقرة. screenshot ليس state كامل؛ hidden tabs وmodals، focus، network delay تؤثر في النتيجة. benchmark قوي يجب أن يقيس recovery بعد misclick أو state change، لا success في path مثالي فقط.

- ground element by semantics + geometry.
- refresh observation بعد transitions.
- validate destructive targets قبل click.
- store action→screenshot diff للaudit.

Research specification / مواصفة الاختبار

Variable / layer	Measure	Failure signal
Artifact diff	expected vs produced artifact	silent corruption
Citation fidelity	claim-source match	hallucinated provenance
Observation freshness	age of state before action	stale UI/web state
Grounding	element/source grounding accuracy	misclick/misattribution
Navigation	steps / backtrack rate	local loop
Sandbox	escape/side-effect containment	unsafe execution
Testing	tests or factual cross-checks	unchecked output

Web Agents

Research, navigation and adversarial content

Web agent يتعامل مع DOM أو screenshots أو كلاهما، ومع صفحات متغيرة ومحتوى غير موثوق. Odysseys يركز على realistic long-horizon tasks، بينما Deep Research وChatGPT agent يعكسان الانتقال إلى deployed systems تقوم بالبحث والتنفيذ.

أكبر خطر ليس فقد رابط، بل **instruction contamination**: الويب نفسه قد يحتوي أوامر مضللة. لذلك retrieval يجب أن يفصل content authority عن instruction authority.

```
WebEvidenceWeight = SourceTrust × Relevance × Freshness × Independence
```

> قاعدة: retrieved text is data, not authority.

Source ledger / دفتر المصدر

S34 · Research · 2026
Odysseys: Benchmarking Web Agents on Realistic Long Horizon Tasks
arXiv:2604.24964; realistic web task horizon and interaction complexity.

S25 · Official · 2025-07-17
Introducing ChatGPT Agent: Bridging Research and Action
Official deployed agent combining research, browsing, computer interaction and action.

S26 · Official · 2025-02-02
Introducing Deep Research
Official reasoning agent for multi-step web research and source synthesis.

S01 · Official · 2026-07-09
GPT-5.6 System Card
Official system card; Sol/Terra/Luna, reasoning effort, agentic safety, prompt-injection and self-improvement evaluations.

S27 · Official · 2026-05-19
Co-Scientist: A Multi-Agent AI Partner to Accelerate Research
Official multi-agent scientific discovery system for hypothesis generation and evolution.

S33 · Research · 2026
OSWorld 2.0: Benchmarking Computer Use Agents on Long-Horizon Real-World Tasks
arXiv:2606.29537; realistic long-horizon computer-use evaluation.

S35 · Research · 2026
SWE-EVO: Benchmarking Coding Agents in Long-Horizon Software Evolution
arXiv:2512.18470; multi-step repository evolution beyond issue-level patching.

S39 · Research · 2026
SentinelBench: A Benchmark for Long-Running Monitoring Agents
arXiv:2606.05342; persistent monitoring as a distinct long-running agent workload.

S40 · Research · 2026
AgentAtlas: Beyond Outcome Leaderboards for LLM Agents
arXiv:2605.20530; behavioral/process evaluation beyond success rate.

S42 · Research · 2026
CocoaBench: Evaluating Unified Digital Agents in the Wild
arXiv:2604.11201; cross-application digital agent evaluation.

Deep technical extraction / الاستخراج الهندسي

Open-web agency. الويب يجمع noisy retrieval وuntrusted instructions وتغيراً زمنياً سريعاً. query planning يجب أن يوازن coverage مع source independence. prompt injection defense تبدأ بفصل "المحتوى الذي نقرأه" عن "السلطة التي تحدد ما نفعله".

- trusted-domain preference للحقائق الحساسة.
- cross-source confirmation لل claims المتغيرة.
- freshness timestamps لكل evidence.
- never execute instructions embedded in retrieved content دون policy authorization.

Research specification / مواصفة الاختبار

Variable / layer	Measure	Failure signal
Testing	tests or factual cross-checks	unchecked output
Source trust	primary-source share	low-quality evidence
Prompt injection	attack success rate	instruction contamination
State recovery	recovery after perturbation	fragile trajectory
Latency	wall-clock / wait handling	timing failures
Async events	event handling correctness	missed updates
Artifact diff	expected vs produced artifact	silent corruption

Coding Agents

From patch generation to software evolution

SWE-bench نقل التقييم من snippets إلى issues، لكن SWE-EVO يدفع أبعد: software evolution طويلة الأفق، refactors، tests، dependencies، وQwen3-Coder-Next وClaude/Grok/DeepSeek frontier models صارت مدربة أو مهيأة لrepository reasoning وterminal work. multi-step repository changes.

القدرة الحاسمة ليست كتابة patch؛ بل **close the engineering loop**: reproduce → inspect → change → test → analyze failure → iterate → verify no regression.

```
SWE Agent Quality = Resolution × TestQuality × RegressionSafety × Maintainability
```

Source ledger / دفتر المصدر

S35 · Research · 2026
SWE-EVO: Benchmarking Coding Agents in Long-Horizon Software Evolution
arXiv:2512.18470; multi-step repository evolution beyond issue-level patching.

S10 · Research · 2026
Qwen3-Coder-Next Technical Report
arXiv:2603.00729; repository reasoning, agent interaction patterns, executable training/evaluation environments.

S02 · Official · 2026-06-09
Claude Fable 5 & Mythos 5 System Card
Official system card for Anthropic's most capable widely released Fable 5 class; always-on adaptive thinking, long-running agents, 1M context and safety evaluations.

S05 · Official · 2026-08-12
Introducing Grok 4.6
Official current flagship release; long-running agents, coding, interactive/visual work, agentic RL, configurable reasoning effort and 500k context.

S06 · Research · 2026
DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence
arXiv:2606.19348; hybrid attention, million-token context, effort scaling, agentic benchmarks.

S25 · Official · 2025-07-17
Introducing ChatGPT Agent: Bridging Research and Action
Official deployed agent combining research, browsing, computer interaction and action.

S26 · Official · 2025-02-02
Introducing Deep Research
Official reasoning agent for multi-step web research and source synthesis.

S27 · Official · 2026-05-19
Co-Scientist: A Multi-Agent AI Partner to Accelerate Research
Official multi-agent scientific discovery system for hypothesis generation and evolution.

S33 · Research · 2026
OSWorld 2.0: Benchmarking Computer Use Agents on Long-Horizon Real-World Tasks
arXiv:2606.29537; realistic long-horizon computer-use evaluation.

S34 · Research · 2026
Odysseys: Benchmarking Web Agents on Realistic Long Horizon Tasks
arXiv:2604.24964; realistic web task horizon and interaction complexity.

Deep technical extraction / الاستخراج الهندسي

SWE loop. coding agent قوي يحتاج reproduction قبل static tests، repository map، patch/ runtime feedback، وlong-horizon software evolution يضيف regression checks. architectural consistency، dependency drift، وmulti-commit reasoning.

- metric: issue resolution + tests + no regression.
- measure unnecessary diff size.
- track compile/test failures as observations لا كـdead ends.
- checkpoint commits تمنع فقد مسار ناجح.

Research specification / مواصفة الاختبار

Variable / layer	Measure	Failure signal
Navigation	steps / backtrack rate	local loop
Sandbox	escape/side-effect containment	unsafe execution
Testing	tests or factual cross-checks	unchecked output
Source trust	primary-source share	low-quality evidence
Prompt injection	attack success rate	instruction contamination
State recovery	recovery after perturbation	fragile trajectory
Latency	wall-clock / wait handling	timing failures

Deep Research Agents

Search is now an agentic workload

Deep research يختلف عن RAG لأنه يقرر ماذا يبحث، متى يغير query، كيف يربط مصادر متعارضة، ومتى يتوقف. هذا يخلق trajectory طويلة من information acquisition. النجاح يقاس ليس فقط بصحة الخلاصة بل contradiction handling، coverage، provenance، وcost.

```
ResearchUtility = Coverage × SourceQuality × SynthesisAccuracy × CitationFidelity / Cost
```

في MOUSSAIF، هذا يطابق Governing Question Engine + Evidence Graph + Falsification Court. السؤال الحاكم ليس "ماذا وجدت؟" بل "ما الفرق الذي ما زال يمنع الحسم؟".

Source ledger / دفتر المصدر

S26 · Official · 2025-02-02
Introducing Deep Research
Official reasoning agent for multi-step web research and source synthesis.

S12 · Research · 2026
Kimi K3: Open Frontier Intelligence
arXiv:2607.24653; 2.8T MoE / 104B active, native vision, 1M context, reasoning-effort levels, long-horizon agentic RL and open weights.

S45 · Research · 2026
Agents in the Wild: Where Research Meets Deployment
arXiv:2607.19336; deployment-oriented synthesis of reliability, memory, tools and evaluation.

S25 · Official · 2025-07-17
Introducing ChatGPT Agent: Bridging Research and Action
Official deployed agent combining research, browsing, computer interaction and action.

S27 · Official · 2026-05-19
Co-Scientist: A Multi-Agent AI Partner to Accelerate Research
Official multi-agent scientific discovery system for hypothesis generation and evolution.

S33 · Research · 2026
OSWorld 2.0: Benchmarking Computer Use Agents on Long-Horizon Real-World Tasks
arXiv:2606.29537; realistic long-horizon computer-use evaluation.

S34 · Research · 2026
Odysseys: Benchmarking Web Agents on Realistic Long Horizon Tasks
arXiv:2604.24964; realistic web task horizon and interaction complexity.

S35 · Research · 2026
SWE-EVO: Benchmarking Coding Agents in Long-Horizon Software Evolution
arXiv:2512.18470; multi-step repository evolution beyond issue-level patching.

S39 · Research · 2026
SentinelBench: A Benchmark for Long-Running Monitoring Agents
arXiv:2606.05342; persistent monitoring as a distinct long-running agent workload.

S40 · Research · 2026
AgentAtlas: Beyond Outcome Leaderboards for LLM Agents
arXiv:2605.20530; behavioral/process evaluation beyond success rate.

Deep technical extraction / الاستخراج الهندسي

Research agent. البحث الوكيلي هو active information acquisition: اختيار query، فحص المصادر، اكتشاف gap، ثم إعادة توجيه البحث. جودة التقرير النهائية لا تكفي؛ يجب قياس citation fidelity، source diversity، contradiction resolution، وcoverage.

```
NextQuery*=argmax ExpectedInformationGain(gap, query)
```

- فصل primary source عن secondary commentary.
- تسجيل لماذا استُبعد مصدر.
- تحديد stop condition عندما لا تضيف queries فرقاً.
- إبقاء unresolved disagreements صريحة.

Research specification / مواصفة الاختبار

Variable / layer	Measure	Failure signal
Async events	event handling correctness	missed updates
Artifact diff	expected vs produced artifact	silent corruption
Citation fidelity	claim-source match	hallucinated provenance
Observation freshness	age of state before action	stale UI/web state
Grounding	element/source grounding accuracy	misclick/misattribution
Navigation	steps / backtrack rate	local loop
Sandbox	escape/side-effect containment	unsafe execution

Co-Scientist وScientific Agents

From answer generation to hypothesis evolution

Co-Scientist يمثل نمطاً مختلفاً: multi-agent system لتوليد hypotheses، نقدها وتطويرها. في البحث العلمي، agency ليست مجرد تنفيذ أدوات؛ هي إدارة مساحة فرضيات تحت evidence ناقص.

الخطأ الكبير هو conflation بين creative hypothesis generation وscientific validation. نظام جيد يجب أن يحفظ lineage: من اقترح الفرضية، ما evidence، ما counterexample، وما experiment القادر على إسقاطها.

```
HypothesisScore = Novelty × Plausibility × Testability × ExpectedInformationGain
```

Source ledger / دفتر المصدر

S27 · Official · 2026-05-19
Co-Scientist: A Multi-Agent AI Partner to Accelerate Research
Official multi-agent scientific discovery system for hypothesis generation and evolution.

S40 · Research · 2026
AgentAtlas: Beyond Outcome Leaderboards for LLM Agents
arXiv:2605.20530; behavioral/process evaluation beyond success rate.

S25 · Official · 2025-07-17
Introducing ChatGPT Agent: Bridging Research and Action
Official deployed agent combining research, browsing, computer interaction and action.

S26 · Official · 2025-02-02
Introducing Deep Research
Official reasoning agent for multi-step web research and source synthesis.

S33 · Research · 2026
OSWorld 2.0: Benchmarking Computer Use Agents on Long-Horizon Real-World Tasks
arXiv:2606.29537; realistic long-horizon computer-use evaluation.

S34 · Research · 2026
Odysseys: Benchmarking Web Agents on Realistic Long Horizon Tasks
arXiv:2604.24964; realistic web task horizon and interaction complexity.

S35 · Research · 2026
SWE-EVO: Benchmarking Coding Agents in Long-Horizon Software Evolution
arXiv:2512.18470; multi-step repository evolution beyond issue-level patching.

S39 · Research · 2026
SentinelBench: A Benchmark for Long-Running Monitoring Agents
arXiv:2606.05342; persistent monitoring as a distinct long-running agent workload.

S42 · Research · 2026
CocoaBench: Evaluating Unified Digital Agents in the Wild
arXiv:2604.11201; cross-application digital agent evaluation.

S43 · Protocol · 2026-07-28
MCP Specification 2026-07-28
Stable open protocol revision for tool/context integration, tasks, subscriptions, caching and multi-round trips.

Deep technical extraction / الاستخراج الهندسي

Scientific agency. hypothesis generation يجب أن تقترن بـcritic. multi-agent falsifiability. مفيد إذا كانت أدواره مستقلة فعلاً: statistician، methodologist، literature checker، proposer، adversarial reviewer. consensus غير المتنوع لا يرفع الدليل.

- لكل hypothesis prediction مميز.
- experiment يفرق بينها وبين baseline theory.
- حفظ lineage للتعديلات.
- عدم تحويل plausibility score إلى claim of discovery.

Research specification / مواصفة الاختبار

Variable / layer	Measure	Failure signal
Testing	tests or factual cross-checks	unchecked output
Source trust	primary-source share	low-quality evidence
Prompt injection	attack success rate	instruction contamination
State recovery	recovery after perturbation	fragile trajectory
Latency	wall-clock / wait handling	timing failures
Async events	event handling correctness	missed updates
Artifact diff	expected vs produced artifact	silent corruption

Multi-Agent Orchestration

Parallelism is not intelligence for free

Kimi K3 long-horizon/parallel agent training وCo-Scientist وGrok workflows تعكس انتقالاً من single-agent إلى multi-agent teams. يفيد عندما subproblems مستقلة أو تحتاج expertise متعارض. لكنه قد يكرر العمل، يضاعف hallucinations أو ينتج consensus زائف.

```
TeamGain = Σ SpecialistGain - CoordinationCost - CorrelatedError
```

- role specialization يجب أن يكون measurable.
- merge يحتاج evidence-aware aggregator.
- parallelism يناسب breadth أكثر من dependency chains.
- agent spawning نفسه قرار budgeted.

Source ledger / دفتر المصدر

S12 · Research · 2026
Kimi K3: Open Frontier Intelligence
arXiv:2607.24653; 2.8T MoE / 104B active, native vision, 1M context, reasoning-effort levels, long-horizon agentic RL and open weights.

S27 · Official · 2026-05-19
Co-Scientist: A Multi-Agent AI Partner to Accelerate Research
Official multi-agent scientific discovery system for hypothesis generation and evolution.

S05 · Official · 2026-08-12
Introducing Grok 4.6
Official current flagship release; long-running agents, coding, interactive/visual work, agentic RL, configurable reasoning effort and 500k context.

S30 · Research · 2026
Benchmark Test-Time Scaling of General LLM Agents
arXiv:2602.18998; sequential vs parallel scaling across search, coding, reasoning and tools.

S31 · Research · 2026
The Long-Horizon Task Mirage? Diagnosing Where and Why Agents Fail as Task Horizon Increases
arXiv:2604.11978; cross-domain degradation with horizon length.

S32 · Research · 2025
UltraHorizon: Benchmarking Agent Capabilities in Ultra Long-Horizon Scenarios
arXiv:2509.21766; 200k+ token / 400+ tool-call trajectories and long-horizon failure taxonomy.

S36 · Research · 2026
AMA-Bench: Evaluating Long-Horizon Memory for Agentic Applications
arXiv:2602.22769; causal/objective memory gaps and AMA-Agent.

S37 · Research · 2026
Agent Memory: Characterization and System Implications of Stateful Long-Horizon Workloads
arXiv:2606.06448; system taxonomy, cost profiling and fleet implications.

S38 · Research · 2026
SAM: State-Adaptive Memory for Long-Horizon Reasoning Agent
arXiv:2605.24468; evolving-state-conditioned recall and compact cues.

S39 · Research · 2026
SentinelBench: A Benchmark for Long-Running Monitoring Agents
arXiv:2606.05342; persistent monitoring as a distinct long-running agent workload.

Deep technical extraction / الاستخراج الهندسي

Multi-agent economics. parallelism يرفع breadth لكنه يستهلك budget ويزيد merge error. نحتاج spawn policy: متى تبرر قيمة المعلومات إنشاء subagent مستقل؟ correlated agents قد يعيدون نفس الخطأ بألفاظ مختلفة.

```
Spawn if E[MarginalGain] > Compute + Coordination + CorrelationRisk
```

- تنويع roles/prompts/tools عند الحاجة.
- merge على evidence لا majority vote فقط.
- قياس duplicate work.
- استخدام sequential handoff للمهام ذات dependencies القوية.

Research specification / مواصفة الاختبار

Variable / layer	Measure	Failure signal
Forgetting	stale item suppression	obsolete state
Objective memory	goal-relevance recall	attention hijack
Horizon	success by action count	long-run collapse
State drift	invariant violation rate	corrupted summaries
Checkpoint	rollback/resume success	irrecoverable error
Coordination	duplicate work / merge cost	swarm overhead
Loop detection	repeat-pattern recall	infinite action cycles

Agent Memory: من النص إلى state

Memory is execution state management

AMA-Bench وAgent Memory characterization وSAM توضح أن ذاكرة الوكيل ليست chat history. هي stream من objectives، observations، actions، وsimilarity. machine-generated state. retrieval وحده قد يفشل لأنه لا يمثل causality أو objective relevance.

```
MemoryItem = (state, cause, action, outcome, confidence, validity_interval)
```

الذاكرة الجيدة تحتاج intent-driven recall، causal links، selective forgetting، consolidation، write policy، وcost profiling. وهذا يطابق Self-Reconstruction Memory: إعادة بناء ما يجب أن يعرفه agent الآن، لا إعادة تشغيل كل الماضي.

Source ledger / دفتر المصدر

S36 · Research · 2026
AMA-Bench: Evaluating Long-Horizon Memory for Agentic Applications
arXiv:2602.22769; causal/objective memory gaps and AMA-Agent.

S37 · Research · 2026
Agent Memory: Characterization and System Implications of Stateful Long-Horizon Workloads
arXiv:2606.06448; system taxonomy, cost profiling and fleet implications.

S38 · Research · 2026
SAM: State-Adaptive Memory for Long-Horizon Reasoning Agent
arXiv:2605.24468; evolving-state-conditioned recall and compact cues.

S30 · Research · 2026
Benchmark Test-Time Scaling of General LLM Agents
arXiv:2602.18998; sequential vs parallel scaling across search, coding, reasoning and tools.

S31 · Research · 2026
The Long-Horizon Task Mirage? Diagnosing Where and Why Agents Fail as Task Horizon Increases
arXiv:2604.11978; cross-domain degradation with horizon length.

S32 · Research · 2025
UltraHorizon: Benchmarking Agent Capabilities in Ultra Long-Horizon Scenarios
arXiv:2509.21766; 200k+ token / 400+ tool-call trajectories and long-horizon failure taxonomy.

S39 · Research · 2026
SentinelBench: A Benchmark for Long-Running Monitoring Agents
arXiv:2606.05342; persistent monitoring as a distinct long-running agent workload.

S40 · Research · 2026
AgentAtlas: Beyond Outcome Leaderboards for LLM Agents
arXiv:2605.20530; behavioral/process evaluation beyond success rate.

S45 · Research · 2026
Agents in the Wild: Where Research Meets Deployment
arXiv:2607.19336; deployment-oriented synthesis of reliability, memory, tools and evaluation.

S12 · Research · 2026
Kimi K3: Open Frontier Intelligence
arXiv:2607.24653; 2.8T MoE / 104B active, native vision, 1M context, reasoning-effort levels, long-horizon agentic RL and open weights.

Deep technical extraction / الاستخراج الهندسي

Stateful memory. agent memory يجب أن تسجل causality وvalidity، لا embedding فقط. العنصر المفيد يجيب: ماذا حدث؟ لماذا؟ ما effect؟ هل ما زال صالحاً؟ هذا يسمح بتمييز knowledge مستقرة من temporary state.

- write policy تمنع تخزين كل token.
- consolidation يدمج الأحداث المتكررة.
- forget/expire rules لل state القديمة.
- retrieval يجمع semantic relevance مع goal/causal relevance.

```
Recall = SemRel + GoalRel + CausalRel + Freshness - Conflict
```

Research specification / مواصفة الاختبار

Variable / layer	Measure	Failure signal
Merge	evidence-weighted conflict resolution	majority hallucination
Causal memory	cause-to-effect retrieval	semantic-only recall
Forgetting	stale item suppression	obsolete state
Objective memory	goal-relevance recall	attention hijack
Horizon	success by action count	long-run collapse
State drift	invariant violation rate	corrupted summaries
Checkpoint	rollback/resume success	irrecoverable error

Long-Horizon Memory Benchmarks

What do we actually test?

المقاييس الحديثة تبتعد عن "هل يتذكر fact بعد 100k token؟" إلى: هل يستفيد من نتيجة action قديم لتغيير action جديد؟ هل ينسى القاعدة التي لم تعد صالحة؟ هل يحافظ على objective وسط آلاف observations؟

Benchmark	محور	الدلالة
AMA-Bench	agent trajectories	causal/objective memory
MemoryAgentBench family	incremental streams	retrieval + learning + forgetting
UltraHorizon	very long tasks	memory/planning/tool endurance
HORIZON	horizon diagnostics	where failure accelerates

Source ledger / دفتر المصدر

S36 · Research · 2026
AMA-Bench: Evaluating Long-Horizon Memory for Agentic Applications
arXiv:2602.22769; causal/objective memory gaps and AMA-Agent.

S37 · Research · 2026
Agent Memory: Characterization and System Implications of Stateful Long-Horizon Workloads
arXiv:2606.06448; system taxonomy, cost profiling and fleet implications.

S31 · Research · 2026
The Long-Horizon Task Mirage? Diagnosing Where and Why Agents Fail as Task Horizon Increases
arXiv:2604.11978; cross-domain degradation with horizon length.

S32 · Research · 2025
UltraHorizon: Benchmarking Agent Capabilities in Ultra Long-Horizon Scenarios
arXiv:2509.21766; 200k+ token / 400+ tool-call trajectories and long-horizon failure taxonomy.

S30 · Research · 2026
Benchmark Test-Time Scaling of General LLM Agents
arXiv:2602.18998; sequential vs parallel scaling across search, coding, reasoning and tools.

S38 · Research · 2026
SAM: State-Adaptive Memory for Long-Horizon Reasoning Agent
arXiv:2605.24468; evolving-state-conditioned recall and compact cues.

S39 · Research · 2026
SentinelBench: A Benchmark for Long-Running Monitoring Agents
arXiv:2606.05342; persistent monitoring as a distinct long-running agent workload.

S40 · Research · 2026
AgentAtlas: Beyond Outcome Leaderboards for LLM Agents
arXiv:2605.20530; behavioral/process evaluation beyond success rate.

S45 · Research · 2026
Agents in the Wild: Where Research Meets Deployment
arXiv:2607.19336; deployment-oriented synthesis of reliability, memory, tools and evaluation.

S12 · Research · 2026
Kimi K3: Open Frontier Intelligence
arXiv:2607.24653; 2.8T MoE / 104B active, native vision, 1M context, reasoning-effort levels, long-horizon agentic RL and open weights.

Deep technical extraction / الاستخراج الهندسي

Memory evaluation. benchmark الجيد لا يسأل فقط عن استرجاع fact، بل عن action selection بعد سلسلة من الأحداث. يجب أن يحتوي distractors، تغير state invalidation، long objectives، وgaps بين cause/effect.

- measure recall precision/utility، لا hit rate فقط.
- اختبر selective forgetting.
- قِس cost لكل useful retrieval.
- run same trajectory مع causal memory، vector memory، raw context.

Research specification / مواصفة الاختبار

Variable / layer	Measure	Failure signal
Objective memory	goal-relevance recall	attention hijack
Horizon	success by action count	long-run collapse
State drift	invariant violation rate	corrupted summaries
Checkpoint	rollback/resume success	irrecoverable error
Coordination	duplicate work / merge cost	swarm overhead
Loop detection	repeat-pattern recall	infinite action cycles
Recovery	MTTR / restored correctness	failure accumulation

Test-Time Scaling for Agents

Sequential vs parallel trajectories

General AgentBench يميز scaling عبر **sequential interaction** و**parallel trajectories**. في agent setting، compute إضافي لا يعني tokens فقط؛ يعني tool calls، environment resets، subagents، containers، browsers، وربما real-world cost.

```
AgentBudget = Tokens + ToolCost + EnvTime + ParallelWorkers + SideEffectRisk
```

التخصيص الأمثل يختلف حسب phase: discovery قد يستفيد من parallel search، بينما execution ذو dependencies يحتاج sequential verification. هذا بالضبط ما يفترض أن يفعله cognitive router.

Source ledger / دفتر المصدر

S30 · Research · 2026
Benchmark Test-Time Scaling of General LLM Agents
arXiv:2602.18998; sequential vs parallel scaling across search, coding, reasoning and tools.

S12 · Research · 2026
Kimi K3: Open Frontier Intelligence
arXiv:2607.24653; 2.8T MoE / 104B active, native vision, 1M context, reasoning-effort levels, long-horizon agentic RL and open weights.

S31 · Research · 2026
The Long-Horizon Task Mirage? Diagnosing Where and Why Agents Fail as Task Horizon Increases
arXiv:2604.11978; cross-domain degradation with horizon length.

S32 · Research · 2025
UltraHorizon: Benchmarking Agent Capabilities in Ultra Long-Horizon Scenarios
arXiv:2509.21766; 200k+ token / 400+ tool-call trajectories and long-horizon failure taxonomy.

S36 · Research · 2026
AMA-Bench: Evaluating Long-Horizon Memory for Agentic Applications
arXiv:2602.22769; causal/objective memory gaps and AMA-Agent.

S37 · Research · 2026
Agent Memory: Characterization and System Implications of Stateful Long-Horizon Workloads
arXiv:2606.06448; system taxonomy, cost profiling and fleet implications.

S38 · Research · 2026
SAM: State-Adaptive Memory for Long-Horizon Reasoning Agent
arXiv:2605.24468; evolving-state-conditioned recall and compact cues.

S39 · Research · 2026
SentinelBench: A Benchmark for Long-Running Monitoring Agents
arXiv:2606.05342; persistent monitoring as a distinct long-running agent workload.

S40 · Research · 2026
AgentAtlas: Beyond Outcome Leaderboards for LLM Agents
arXiv:2605.20530; behavioral/process evaluation beyond success rate.

S45 · Research · 2026
Agents in the Wild: Where Research Meets Deployment
arXiv:2607.19336; deployment-oriented synthesis of reliability, memory, tools and evaluation.

Deep technical extraction / الاستخراج الهندسي

Compute for agents. ميزانية agent تشمل tokens وtool latency وenvironment time وparallel workers. بعض الفروع يمكن تشغيلها بالتوازي، لكن execution الذي يحمل side effects يحتاج ترتيباً وحواجز تحقق.

- parallelize evidence gathering.
- serialize irreversible writes.
- allocate extra reasoning عند tool disagreement.
- قِس end-to-end latency وdollar cost وsuccess.

Research specification / مواصفة الاختبار

Variable / layer	Measure	Failure signal
Horizon	success by action count	long-run collapse
State drift	invariant violation rate	corrupted summaries
Checkpoint	rollback/resume success	irrecoverable error
Coordination	duplicate work / merge cost	swarm overhead
Loop detection	repeat-pattern recall	infinite action cycles
Recovery	MTTR / restored correctness	failure accumulation
Spawn policy	marginal gain/subagent	compute amplification

Horizon Tax: لماذا تنهار الوكالة مع الزمن؟

Compounding error is only part of the story

HORIZON وUltraHorizon يظهران degradation منهجياً مع طول المهمة. نموذج naïve يفترض أن النجاح الكلي يساوي حاصل ضرب دقات الخطوات، لكن agents تملك recovery؛ لذلك المشكلة الأهم هي **state corruption** و**capability gaps** التي لا يصلحها retry.

```
P_success(H) = Π_t P(step_t | state_t) × RecoveryFactor - StateDriftPenalty
```

- in-context locking: التمسك بإطار قديم.
- goal drift: يتحول subgoal إلى الهدف الحقيقي.
- tool fatigue: تكرار نفس أداة رغم فشلها.
- memory overwrite: summary خاطئ يصبح "حقيقة".

Source ledger / دفتر المصدر

S31 · Research · 2026
The Long-Horizon Task Mirage? Diagnosing Where and Why Agents Fail as Task Horizon Increases
arXiv:2604.11978; cross-domain degradation with horizon length.

S32 · Research · 2025
UltraHorizon: Benchmarking Agent Capabilities in Ultra Long-Horizon Scenarios
arXiv:2509.21766; 200k+ token / 400+ tool-call trajectories and long-horizon failure taxonomy.

S30 · Research · 2026
Benchmark Test-Time Scaling of General LLM Agents
arXiv:2602.18998; sequential vs parallel scaling across search, coding, reasoning and tools.

S36 · Research · 2026
AMA-Bench: Evaluating Long-Horizon Memory for Agentic Applications
arXiv:2602.22769; causal/objective memory gaps and AMA-Agent.

S37 · Research · 2026
Agent Memory: Characterization and System Implications of Stateful Long-Horizon Workloads
arXiv:2606.06448; system taxonomy, cost profiling and fleet implications.

S38 · Research · 2026
SAM: State-Adaptive Memory for Long-Horizon Reasoning Agent
arXiv:2605.24468; evolving-state-conditioned recall and compact cues.

S39 · Research · 2026
SentinelBench: A Benchmark for Long-Running Monitoring Agents
arXiv:2606.05342; persistent monitoring as a distinct long-running agent workload.

S40 · Research · 2026
AgentAtlas: Beyond Outcome Leaderboards for LLM Agents
arXiv:2605.20530; behavioral/process evaluation beyond success rate.

S45 · Research · 2026
Agents in the Wild: Where Research Meets Deployment
arXiv:2607.19336; deployment-oriented synthesis of reliability, memory, tools and evaluation.

S12 · Research · 2026
Kimi K3: Open Frontier Intelligence
arXiv:2607.24653; 2.8T MoE / 104B active, native vision, 1M context, reasoning-effort levels, long-horizon agentic RL and open weights.

Deep technical extraction / الاستخراج الهندسي

Horizon degradation. طول المهمة يخلق state drift وretry. compounding uncertainty. وحده لا يصلح هدفاً انحرف أو summary فاسداً. لذلك نحتاج periodic invariants: ما الهدف؟ ما constraints؟ ما completed evidence؟ وما next critical node؟

- goal checksum كل N actions.
- detect repeated tool/action loops.
- reconstruct state من evidence عند anomaly.
- measure success as function of horizon bins.

Research specification / مواصفة الاختبار

Variable / layer	Measure	Failure signal
Role diversity	error correlation	fake diversity
Merge	evidence-weighted conflict resolution	majority hallucination
Causal memory	cause-to-effect retrieval	semantic-only recall
Forgetting	stale item suppression	obsolete state
Objective memory	goal-relevance recall	attention hijack
Horizon	success by action count	long-run collapse
State drift	invariant violation rate	corrupted summaries

Loop Resistance وFailure Recovery

The agent must know when it is stuck

frontier deployments بدأت تتكلم صراحة عن loop resistance وgraceful error recovery. هذا يعكس انتقال معيار الجودة من answer correctness إلى **operational resilience**. agent الذي يفشل ثم يكتشف فشله ويعيد المسار قد يكون أفضل من agent أعلى دقة خطوة بخطوة لكنه ينهار بصمت.

```
Resilience = DetectFailure × Diagnose × Replan × Recover × Verify
```

Loop detector جيد يراقب unchanged world state، identical tool signatures، semantic repetition، وdeclining information gain. عند threshold يجب أن يصعد إلى meta-level: غيّر strategy، اطلب help، أو توقف.

Deep technical extraction / الاستخراج الهندسي

Recovery policy. resilience قدرة مستقلة عن first-pass accuracy. النظام يحتاج classify failure ثم اختيار rollback، replan، alternate tool، retry أو blind retry. قد يضاعف human escalation الضرر.

```
Recovery = Detect × Diagnose × AlternativeQuality × Verification
```

- حفظ last-known-good checkpoint.
- حد أقصى للتكرار المتطابق.
- raise uncertainty بعد repeated failures.
- post-recovery regression check.

Source ledger / دفتر المصدر

S02 · Official · 2026-06-09
Claude Fable 5 & Mythos 5 System Card
Official system card for Anthropic's most capable widely released Fable 5 class; always-on adaptive thinking, long-running agents, 1M context and safety evaluations.

S31 · Research · 2026
The Long-Horizon Task Mirage? Diagnosing Where and Why Agents Fail as Task Horizon Increases
arXiv:2604.11978; cross-domain degradation with horizon length.

S45 · Research · 2026
Agents in the Wild: Where Research Meets Deployment
arXiv:2607.19336; deployment-oriented synthesis of reliability, memory, tools and evaluation.

S30 · Research · 2026
Benchmark Test-Time Scaling of General LLM Agents
arXiv:2602.18998; sequential vs parallel scaling across search, coding, reasoning and tools.

S32 · Research · 2025
UltraHorizon: Benchmarking Agent Capabilities in Ultra Long-Horizon Scenarios
arXiv:2509.21766; 200k+ token / 400+ tool-call trajectories and long-horizon failure taxonomy.

S36 · Research · 2026
AMA-Bench: Evaluating Long-Horizon Memory for Agentic Applications
arXiv:2602.22769; causal/objective memory gaps and AMA-Agent.

S37 · Research · 2026
Agent Memory: Characterization and System Implications of Stateful Long-Horizon Workloads
arXiv:2606.06448; system taxonomy, cost profiling and fleet implications.

S38 · Research · 2026
SAM: State-Adaptive Memory for Long-Horizon Reasoning Agent
arXiv:2605.24468; evolving-state-conditioned recall and compact cues.

S39 · Research · 2026
SentinelBench: A Benchmark for Long-Running Monitoring Agents
arXiv:2606.05342; persistent monitoring as a distinct long-running agent workload.

S40 · Research · 2026
AgentAtlas: Beyond Outcome Leaderboards for LLM Agents
arXiv:2605.20530; behavioral/process evaluation beyond success rate.

Research specification / مواصفة الاختبار

Variable / layer	Measure	Failure signal
Loop detection	repeat-pattern recall	infinite action cycles
Recovery	MTTR / restored correctness	failure accumulation
Spawn policy	marginal gain/subagent	compute amplification
Role diversity	error correlation	fake diversity
Merge	evidence-weighted conflict resolution	majority hallucination
Causal memory	cause-to-effect retrieval	semantic-only recall
Forgetting	stale item suppression	obsolete state

World-Effect وVerification

Action is not complete until the world is re-read

أكبر فرق بين chatbot وagent هو side effects. لذلك نحتاج closed loop: **intent → authorize → act → observe → verify → record**. نجاح API call لا يثبت أن المطلوب تحقق؛ ربما أنشأ record في wrong account أو فشل downstream silently.

```
EffectVerified = ExpectedStateDelta ≈ ObservedStateDelta
```

هذه المعادلة هي قلب World-Effect Loop. كل action ذات impact مرتفع يجب أن تحمل source of truth، expected delta، وrollback path إن أمكن.

> عند عدم القدرة على فحص world state، يجب خفض claim من "تم" إلى "تم إرسال الأمر/الطلب".

Source ledger / دفتر المصدر

S25 · Official · 2025-07-17
Introducing ChatGPT Agent: Bridging Research and Action
Official deployed agent combining research, browsing, computer interaction and action.

S33 · Research · 2026
OSWorld 2.0: Benchmarking Computer Use Agents on Long-Horizon Real-World Tasks
arXiv:2606.29537; realistic long-horizon computer-use evaluation.

S45 · Research · 2026
Agents in the Wild: Where Research Meets Deployment
arXiv:2607.19336; deployment-oriented synthesis of reliability, memory, tools and evaluation.

S30 · Research · 2026
Benchmark Test-Time Scaling of General LLM Agents
arXiv:2602.18998; sequential vs parallel scaling across search, coding, reasoning and tools.

S31 · Research · 2026
The Long-Horizon Task Mirage? Diagnosing Where and Why Agents Fail as Task Horizon Increases
arXiv:2604.11978; cross-domain degradation with horizon length.

S32 · Research · 2025
UltraHorizon: Benchmarking Agent Capabilities in Ultra Long-Horizon Scenarios
arXiv:2509.21766; 200k+ token / 400+ tool-call trajectories and long-horizon failure taxonomy.

S36 · Research · 2026
AMA-Bench: Evaluating Long-Horizon Memory for Agentic Applications
arXiv:2602.22769; causal/objective memory gaps and AMA-Agent.

S37 · Research · 2026
Agent Memory: Characterization and System Implications of Stateful Long-Horizon Workloads
arXiv:2606.06448; system taxonomy, cost profiling and fleet implications.

S38 · Research · 2026
SAM: State-Adaptive Memory for Long-Horizon Reasoning Agent
arXiv:2605.24468; evolving-state-conditioned recall and compact cues.

S39 · Research · 2026
SentinelBench: A Benchmark for Long-Running Monitoring Agents
arXiv:2606.05342; persistent monitoring as a distinct long-running agent workload.

Deep technical extraction / الاستخراج الهندسي

World-effect closure. نجاح tool call لا يثبت أن العالم تغير كما أردنا. نحتاج expected state delta، observation بعد التنفيذ، attribution، rollback status. هذه الطبقة تحول agent من "مولد أوامر" إلى نظام effect-aware.

```
VerifiedEffect = Observe(World_after) - Observe(World_before) ≈
                 ExpectedDelta
```

- تحديد tolerances للنتيجة المقبولة.
- فحص partial success.
- ربط effect بالaction id.
- عدم إعلان completion إذا verification غير متاحة.

Research specification / مواصفة الاختبار

Variable / layer	Measure	Failure signal
Role diversity	error correlation	fake diversity
Merge	evidence-weighted conflict resolution	majority hallucination
Causal memory	cause-to-effect retrieval	semantic-only recall
Forgetting	stale item suppression	obsolete state
Objective memory	goal-relevance recall	attention hijack
Horizon	success by action count	long-run collapse
State drift	invariant violation rate	corrupted summaries

Agentic Safety: التفويض قبل الذكاء

Permissions, confirmations and prompt injection

مع alignment، agents ينتقل من content policy إلى capability governance. المخاطر الرئيسية: prompt injection من privilege، accidental destructive actions، credential abuse، tools، data exfiltration، escalation وgoal hijacking.

Control	وظيفته
Least privilege	حصر capability لكل task
Confirmation gates	أفعال عالية الأثر أو غير قابلة للعكس
Sandboxing	تقليل blast radius
Output validation	منع tool injection/data poisoning
Audit trail	إعادة بناء القرار والأثر

GPT-5.6 system card يقدم أمثلة صريحة على prompt-injection وcomputer-use confirmations؛ هذه ليست إضافات UX بل safety architecture.

Source ledger / دفتر المصدر

S01 · Official · 2026-07-09
GPT-5.6 System Card
Official system card; Sol/Terra/Luna, reasoning effort, agentic safety, prompt-injection and self-improvement evaluations.

S29 · Research · 2026
The 2025 AI Agent Index: Documenting Technical and Safety Features of Deployed Agentic AI Systems
arXiv:2602.17753; deployed-agent capability/safety documentation.

S43 · Protocol · 2026-07-28
MCP Specification 2026-07-28
Stable open protocol revision for tool/context integration, tasks, subscriptions, caching and multi-round trips.

S28 · Official · 2026-02-18
Measuring AI Agent Autonomy in Practice
Official empirical analysis of user oversight and autonomous execution in Claude Code.

S39 · Research · 2026
SentinelBench: A Benchmark for Long-Running Monitoring Agents
arXiv:2606.05342; persistent monitoring as a distinct long-running agent workload.

S41 · Research · 2026
CUBE: A Standard for Unifying Agent Benchmarks
arXiv:2603.15798; common benchmark interface and environment standardization.

S44 · Protocol · 2025
Agent2Agent (A2A) Protocol
Open protocol for interoperable agent-to-agent communication across vendors/frameworks.

S45 · Research · 2026
Agents in the Wild: Where Research Meets Deployment
arXiv:2607.19336; deployment-oriented synthesis of reliability, memory, tools and evaluation.

S25 · Official · 2025-07-17
Introducing ChatGPT Agent: Bridging Research and Action
Official deployed agent combining research, browsing, computer interaction and action.

S33 · Research · 2026
OSWorld 2.0: Benchmarking Computer Use Agents on Long-Horizon Real-World Tasks
arXiv:2606.29537; realistic long-horizon computer-use evaluation.

Deep technical extraction / الاستخراج الهندسي

Safety envelope. الاستقلالية ليست binary؛ هي permissions متدرجة حسب impact وreversibility. frontier agent يجب أن يعرف ما يمكنه فعله مباشرة، ما يحتاج confirmation، وما يجب رفضه أو تصعيده.

- least privilege credentials.
- untrusted-input isolation.
- rate/volume limits للأفعال المتكررة.
- audit + post-action inspection.

```
AutonomyLevel = f(TaskRisk, Trust, Reversibility, Observability)
```

Research specification / مواصفة الاختبار

Variable / layer	Measure	Failure signal
Audit	evidence completeness	unattributable action
Identity	agent/tool principal binding	confused deputy
Authorization	capability-specific grants	ambient authority
Discovery	correct capability metadata	wrong-tool selection
Schema	validation failure rate	malformed actions
Versioning	compatibility across upgrades	protocol drift
Cancellation	bounded stop latency	zombie tasks

MCP: طبقة توصيل الأدوات والسياق

Model Context Protocol 2026-07-28

MCP يحول tool/context integration إلى protocol بدلاً من adapters مخصصة لكل تطبيق. revision 2026-07-28 يضيف/يثبت مفاهيم مثل long-running، subscriptions، discovery/negotiation، multi-round trips، caching، tasks، وstandard routing semantics وفق SDKs الرسمية.

قيمته لـAGI OS ليست أنه "يخلق agent"، بل أنه يفصل **capability exposure** عن model logic. هذا يسمح لـMOUSSAIXT/Hermes runtime أن يكتشف الأدوات ويتعاقد معها دون hardcoding.

```
Agent ↔ MCP Client ↔ MCP Server ↔ Resource/Tool/Task
```

لكن MCP لا يحل authorization أو trust وحده؛ يجب أن يضاف policy layer وprovenance validation.

Source ledger / دفتر المصدر

S43 · Protocol · 2026-07-28
MCP Specification 2026-07-28
Stable open protocol revision for tool/context integration, tasks, subscriptions, caching and multi-round trips.

S01 · Official · 2026-07-09
GPT-5.6 System Card
Official system card; Sol/Terra/Luna, reasoning effort, agentic safety, prompt-injection and self-improvement evaluations.

S28 · Official · 2026-02-18
Measuring AI Agent Autonomy in Practice
Official empirical analysis of user oversight and autonomous execution in Claude Code.

S29 · Research · 2026
The 2025 AI Agent Index: Documenting Technical and Safety Features of Deployed Agentic AI Systems
arXiv:2602.17753; deployed-agent capability/safety documentation.

S39 · Research · 2026
SentinelBench: A Benchmark for Long-Running Monitoring Agents
arXiv:2606.05342; persistent monitoring as a distinct long-running agent workload.

S41 · Research · 2026
CUBE: A Standard for Unifying Agent Benchmarks
arXiv:2603.15798; common benchmark interface and environment standardization.

S44 · Protocol · 2025
Agent2Agent (A2A) Protocol
Open protocol for interoperable agent-to-agent communication across vendors/frameworks.

S45 · Research · 2026
Agents in the Wild: Where Research Meets Deployment
arXiv:2607.19336; deployment-oriented synthesis of reliability, memory, tools and evaluation.

S25 · Official · 2025-07-17
Introducing ChatGPT Agent: Bridging Research and Action
Official deployed agent combining research, browsing, computer interaction and action.

S33 · Research · 2026
OSWorld 2.0: Benchmarking Computer Use Agents on Long-Horizon Real-World Tasks
arXiv:2606.29537; realistic long-horizon computer-use evaluation.

Deep technical extraction / الاستخراج الهندسي

MCP as interoperability layer. protocol يوحّد discovery/access إلى /tools/resources/prompts، لكنه لا يحل governance تلقائياً. secure deployment يحتاج authorization، identity، cancellation، timeouts، schema validation، وaudit around protocol messages.

- capability discovery لا يعني capability permission.
- version negotiation مهم لاستقرار long-running agents.
- tasks/subscriptions تحتاج lifecycle state واضح.
- cache/freshness يجب ألا يخفي تغير resource.

Research specification / مواصفة الاختبار

Variable / layer	Measure	Failure signal
Versioning	compatibility across upgrades	protocol drift
Cancellation	bounded stop latency	zombie tasks
Durability	resume after crash	lost long task
Observability	trace completeness	opaque orchestration
Sandbox	resource/FS/network limits	blast radius
Secrets	exposure incidents	credential leakage
Idempotency	duplicate-effect rate	retry damage

A2A: طبقة تواصل agent-to-agent

Interoperability above the tool layer

A2A صُمم لتمكين agents من vendors/frameworks مختلفة من التعاون. موقعه مختلف عن MCP: MCP يركز على context/tools، بينما A2A يركز على agent-to-agent tasks, capabilities and collaboration. هذا يخلق stack ناشئاً: tools ↔ MCP ↔ agent runtime ↔ model، وremote agent ↔ A2A ↔ agent.

```
Interoperability = SharedProtocol × CapabilityDiscovery × Identity × Policy × Traceability
```

البروتوكولات المفتوحة open protocols تقلل lock-in لكنها توسع attack surface. لذلك federated agents تحتاج signed artifacts، scope، identity، وresult verification.

Source ledger / دفتر المصدر

S44 · Protocol · 2025
Agent2Agent (A2A) Protocol
Open protocol for interoperable agent-to-agent communication across vendors/frameworks.

S43 · Protocol · 2026-07-28
MCP Specification 2026-07-28
Stable open protocol revision for tool/context integration, tasks, subscriptions, caching and multi-round trips.

S01 · Official · 2026-07-09
GPT-5.6 System Card
Official system card; Sol/Terra/Luna, reasoning effort, agentic safety, prompt-injection and self-improvement evaluations.

S28 · Official · 2026-02-18
Measuring AI Agent Autonomy in Practice
Official empirical analysis of user oversight and autonomous execution in Claude Code.

S29 · Research · 2026
The 2025 AI Agent Index: Documenting Technical and Safety Features of Deployed Agentic AI Systems
arXiv:2602.17753; deployed-agent capability/safety documentation.

S39 · Research · 2026
SentinelBench: A Benchmark for Long-Running Monitoring Agents
arXiv:2606.05342; persistent monitoring as a distinct long-running agent workload.

S41 · Research · 2026
CUBE: A Standard for Unifying Agent Benchmarks
arXiv:2603.15798; common benchmark interface and environment standardization.

S45 · Research · 2026
Agents in the Wild: Where Research Meets Deployment
arXiv:2607.19336; deployment-oriented synthesis of reliability, memory, tools and evaluation.

S25 · Official · 2025-07-17
Introducing ChatGPT Agent: Bridging Research and Action
Official deployed agent combining research, browsing, computer interaction and action.

S33 · Research · 2026
OSWorld 2.0: Benchmarking Computer Use Agents on Long-Horizon Real-World Tasks
arXiv:2606.29537; realistic long-horizon computer-use evaluation.

Deep technical extraction / الاستخراج الهندسي

A2A semantics. agent-to-agent communication يحتاج أكثر من chat: capability discovery، status، artifacts، cancellation، task ownership، وtrust boundary. عند cross-vendor delegation يجب أن يبقى evidence قابلاً للتحقق لدى orchestrator.

- signed/attributed results عند الحساسية.
- deadline وretry semantics.
- no implicit privilege inheritance.
- merge conflicting remote results عبر verifier مستقل.

Research specification / مواصفة الاختبار

Variable / layer	Measure	Failure signal
Secrets	exposure incidents	credential leakage
Idempotency	duplicate-effect rate	retry damage
Audit	evidence completeness	unattributable action
Identity	agent/tool principal binding	confused deputy
Authorization	capability-specific grants	ambient authority
Discovery	correct capability metadata	wrong-tool selection
Schema	validation failure rate	malformed actions

Trace Plane وObservability

You cannot govern what you cannot reconstruct

AgentAtlas وAgent Index يدفعان إلى ما وراء success rate: نحتاج process telemetry. الحد الأدنى: memory writes، observations، tool calls، model version، prompts/system policies، final state delta وretries، token/tool cost، timestamps، errors، authorization events.

```
Trace = {Decision, Evidence, Action, Result, Cost, Policy, StateBefore, StateAfter}
```

هذا يجعل debugging وscience ممكنين. بدون trace لا يمكن التمييز بين model failure وtool failure وenvironment drift وbad prompt.

Source ledger / دفتر المصدر

S29 · Research · 2026
The 2025 AI Agent Index: Documenting Technical and Safety Features of Deployed Agentic AI Systems
arXiv:2602.17753; deployed-agent capability/safety documentation.

S40 · Research · 2026
AgentAtlas: Beyond Outcome Leaderboards for LLM Agents
arXiv:2605.20530; behavioral/process evaluation beyond success rate.

S45 · Research · 2026
Agents in the Wild: Where Research Meets Deployment
arXiv:2607.19336; deployment-oriented synthesis of reliability, memory, tools and evaluation.

S01 · Official · 2026-07-09
GPT-5.6 System Card
Official system card; Sol/Terra/Luna, reasoning effort, agentic safety, prompt-injection and self-improvement evaluations.

S28 · Official · 2026-02-18
Measuring AI Agent Autonomy in Practice
Official empirical analysis of user oversight and autonomous execution in Claude Code.

S39 · Research · 2026
SentinelBench: A Benchmark for Long-Running Monitoring Agents
arXiv:2606.05342; persistent monitoring as a distinct long-running agent workload.

S41 · Research · 2026
CUBE: A Standard for Unifying Agent Benchmarks
arXiv:2603.15798; common benchmark interface and environment standardization.

S43 · Protocol · 2026-07-28
MCP Specification 2026-07-28
Stable open protocol revision for tool/context integration, tasks, subscriptions, caching and multi-round trips.

S44 · Protocol · 2025
Agent2Agent (A2A) Protocol
Open protocol for interoperable agent-to-agent communication across vendors/ frameworks.

S25 · Official · 2025-07-17
Introducing ChatGPT Agent: Bridging Research and Action
Official deployed agent combining research, browsing, computer interaction and action.

Deep technical extraction / الاستخراج الهندسي

Trace plane. بدون traces لا نستطيع تمييز model failure من tool failure أو policy gate. الحد الأدنى: decision reason، error، tool input/output digest، state hash، action، timestamp، verification result، class، وcost.

- redact secrets مع بقاء audit utility.
- ربط parent/child spans للـsubagents.
- metrics للـlooping/retries/state drift.
- retention policy حسب الحساسية.

Research specification / مواصفة الاختبار

Variable / layer	Measure	Failure signal
Durability	resume after crash	lost long task
Observability	trace completeness	opaque orchestration
Sandbox	resource/FS/network limits	blast radius
Secrets	exposure incidents	credential leakage
Idempotency	duplicate-effect rate	retry damage
Audit	evidence completeness	unattributable action
Identity	agent/tool principal binding	confused deputy

Execution Environments وSandboxes

Agents need worlds they can safely break

Coding agents وcomputer-use benchmarks تعتمد containers/VMs لأن agent يجب أن يجرب ويفشل. sandbox ليس فقط security boundary؛ هو **scientific instrument** يجعل state resettable، reproducible، وgradeable.

- snapshot قبل task.
- deterministic seeds حيث أمكن.
- network/credential policy صريحة.
- resource ceilings.
- post-run diff.

Qwen3-Coder-Next يصف executable environments في التدريب/التقييم، وCUBE يحاول توحيد benchmark interfaces. هذا يمهد لـagentic CI/CD.

Source ledger / دفتر المصدر

S10 · Research · 2026
Qwen3-Coder-Next Technical Report
arXiv:2603.00729; repository reasoning, agent interaction patterns, executable training/evaluation environments.

S41 · Research · 2026
CUBE: A Standard for Unifying Agent Benchmarks
arXiv:2603.15798; common benchmark interface and environment standardization.

S35 · Research · 2026
SWE-EVO: Benchmarking Coding Agents in Long-Horizon Software Evolution
arXiv:2512.18470; multi-step repository evolution beyond issue-level patching.

S01 · Official · 2026-07-09
GPT-5.6 System Card
Official system card; Sol/Terra/Luna, reasoning effort, agentic safety, prompt-injection and self-improvement evaluations.

S28 · Official · 2026-02-18
Measuring AI Agent Autonomy in Practice
Official empirical analysis of user oversight and autonomous execution in Claude Code.

S29 · Research · 2026
The 2025 AI Agent Index: Documenting Technical and Safety Features of Deployed Agentic AI Systems
arXiv:2602.17753; deployed-agent capability/safety documentation.

S39 · Research · 2026
SentinelBench: A Benchmark for Long-Running Monitoring Agents
arXiv:2606.05342; persistent monitoring as a distinct long-running agent workload.

S43 · Protocol · 2026-07-28
MCP Specification 2026-07-28
Stable open protocol revision for tool/context integration, tasks, subscriptions, caching and multi-round trips.

S44 · Protocol · 2025
Agent2Agent (A2A) Protocol
Open protocol for interoperable agent-to-agent communication across vendors/frameworks.

S45 · Research · 2026
Agents in the Wild: Where Research Meets Deployment
arXiv:2607.19336; deployment-oriented synthesis of reliability, memory, tools and evaluation.

Deep technical extraction / الاستخراج الهندسي

Execution isolation. code agents تحتاج بيئة قابلة لإعادة الضبط، limits لـCPU/RAM/network، filesystem boundaries، وsandbox. artifact capture ليس فقط security feature؛ هو شرط reproducibility لأن كل run يبدأ من state معروف.

- immutable base image.
- ephemeral workspace + explicit artifact export.
- network allowlist للمهام الحساسة.
- capture stdout/stderr/tests as evidence.

Research specification / مواصفة الاختبار

Variable / layer	Measure	Failure signal
Observability	trace completeness	opaque orchestration
Sandbox	resource/FS/network limits	blast radius
Secrets	exposure incidents	credential leakage
Idempotency	duplicate-effect rate	retry damage
Audit	evidence completeness	unattributable action
Identity	agent/tool principal binding	confused deputy
Authorization	capability-specific grants	ambient authority

Deployment Reliability

Async, resumability, checkpoints and budgets

الوكيل المنتج لا يعمل فقط في session قصيرة. يحتاج background execution، resume بعد crash وhuman، idempotent tasks، lease/lock semantics، checkpointing، budget enforcement وhandoff. هذه خصائص distributed systems بقدر ما هي AI.

```
DurableAgent = CognitivePolicy + WorkflowEngine + PersistentState + Idempotency + Audit
```

هنا تلتقي Hermes-style runtimes مع research agents: cron/schedules وحدها ليست كافية؛ يجب حفظ causal state وrecovery plan. long-running monitoring كما في SentinelBench يختبر نمطاً لا تغطيه benchmarks التقليدية.

Deep technical extraction / الاستخراج الهندسي

Production reliability. agent production يحتاج heartbeat، durable task state، queues، retries، idempotency، cancellation، وLLM reasoning. dead-letter handling قد يكون ممتازاً لكن runtime هشاً؛ لذلك SLO يجب أن يقيس completion evidence وrecovery time لا API uptime فقط.

- exactly-once side effects عبر idempotency keys حيث أمكن.
- resume بعد crash من checkpoint.
- budget guards تمنع runaway loops.
- version pinning للموديل والأدوات.

Source ledger / دفتر المصدر

S39 · Research · 2026
SentinelBench: A Benchmark for Long-Running Monitoring Agents
arXiv:2606.05342; persistent monitoring as a distinct long-running agent workload.

S45 · Research · 2026
Agents in the Wild: Where Research Meets Deployment
arXiv:2607.19336; deployment-oriented synthesis of reliability, memory, tools and evaluation.

S28 · Official · 2026-02-18
Measuring AI Agent Autonomy in Practice
Official empirical analysis of user oversight and autonomous execution in Claude Code.

S01 · Official · 2026-07-09
GPT-5.6 System Card
Official system card; Sol/Terra/Luna, reasoning effort, agentic safety, prompt-injection and self-improvement evaluations.

S29 · Research · 2026
The 2025 AI Agent Index: Documenting Technical and Safety Features of Deployed Agentic AI Systems
arXiv:2602.17753; deployed-agent capability/safety documentation.

S41 · Research · 2026
CUBE: A Standard for Unifying Agent Benchmarks
arXiv:2603.15798; common benchmark interface and environment standardization.

S43 · Protocol · 2026-07-28
MCP Specification 2026-07-28
Stable open protocol revision for tool/context integration, tasks, subscriptions, caching and multi-round trips.

S44 · Protocol · 2025
Agent2Agent (A2A) Protocol
Open protocol for interoperable agent-to-agent communication across vendors/ frameworks.

S25 · Official · 2025-07-17
Introducing ChatGPT Agent: Bridging Research and Action
Official deployed agent combining research, browsing, computer interaction and action.

S33 · Research · 2026
OSWorld 2.0: Benchmarking Computer Use Agents on Long-Horizon Real-World Tasks
arXiv:2606.29537; realistic long-horizon computer-use evaluation.

Research specification / مواصفة الاختبار

Variable / layer	Measure	Failure signal
Schema	validation failure rate	malformed actions
Versioning	compatibility across upgrades	protocol drift
Cancellation	bounded stop latency	zombie tasks
Durability	resume after crash	lost long task
Observability	trace completeness	opaque orchestration
Sandbox	resource/FS/network limits	blast radius
Secrets	exposure incidents	credential leakage

Frontier Benchmark Atlas

What each benchmark reveals

Benchmark	Environment	Primary stress
General AgentBench	search/code/reason/tools	test-time scaling
HORIZON	cross-domain	horizon-dependent failure
UltraHorizon	very long discovery	200k+ token endurance
OSWorld 2.0	desktop apps	real computer use
Odysseys	web	realistic long-horizon browsing
SWE-EVO	repositories	software evolution
AMA-Bench	agent trajectories	causal memory
SentinelBench	long-running monitor	persistence over time

لا يوجد benchmark واحد يقيس "agentic AGI". القوة في portfolio يغطي task types وfailure modes مختلفة.

Source ledger / دفتر المصدر

S30 · Research · 2026
Benchmark Test-Time Scaling of General LLM Agents
arXiv:2602.18998; sequential vs parallel scaling across search, coding, reasoning and tools.

S31 · Research · 2026
The Long-Horizon Task Mirage? Diagnosing Where and Why Agents Fail as Task Horizon Increases
arXiv:2604.11978; cross-domain degradation with horizon length.

S32 · Research · 2025
UltraHorizon: Benchmarking Agent Capabilities in Ultra Long-Horizon Scenarios
arXiv:2509.21766; 200k+ token / 400+ tool-call trajectories and long-horizon failure taxonomy.

S33 · Research · 2026
OSWorld 2.0: Benchmarking Computer Use Agents on Long-Horizon Real-World Tasks
arXiv:2606.29537; realistic long-horizon computer-use evaluation.

S34 · Research · 2026
Odysseys: Benchmarking Web Agents on Realistic Long Horizon Tasks
arXiv:2604.24964; realistic web task horizon and interaction complexity.

S35 · Research · 2026
SWE-EVO: Benchmarking Coding Agents in Long-Horizon Software Evolution
arXiv:2512.18470; multi-step repository evolution beyond issue-level patching.

S36 · Research · 2026
AMA-Bench: Evaluating Long-Horizon Memory for Agentic Applications
arXiv:2602.22769; causal/objective memory gaps and AMA-Agent.

S39 · Research · 2026
SentinelBench: A Benchmark for Long-Running Monitoring Agents
arXiv:2606.05342; persistent monitoring as a distinct long-running agent workload.

S15 · Research · 2026
Inference Scaling, Learning to Reason, and Agentic Systems
arXiv:2504.09037v4; synthesis of inference scaling, verifiers, reasoning learning and agentic systems.

S18 · Research · 2024
Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters
arXiv:2408.03314; adaptive compute allocation and verifier-guided search.

Deep technical extraction / الاستخراج الهندسي

Portfolio evaluation. كل benchmark يكشف طبقة مختلفة؛ لا نحولها إلى average وحيد يخفي trade-offs. نرسم capability profile: web، desktop، SWE، memory، monitoring، test-time scaling، safety. ثم نبحث عن failure clusters المشتركة.

- normalize budgets قبل aggregation.
- احتفظ بـ per-task traces للـdiagnosis.
- اختبر transfer بين benchmark families.
- استخدم rolling/dynamic sets لتقليل contamination.

Research specification / مواصفة الاختبار

Variable / layer	Measure	Failure signal
Replication	repeat/independent test	one-off success
Negative result	registry downgrade/patch	theory immune to failure
World effect	external-state verification	simulation-only success
Concept definition	operational fields complete	name without mechanism
External analogue	closest published mechanism	reinvention
Differential	unique predicted gain	no novelty
Baseline	strongest simple comparator	straw baseline

Deployed Frontier Systems

Research systems vs production envelopes

تمثل أنماطاً مختلفة من deployment. Kimi Agent Swarm وChatGPT Agent، Deep Research، Claude Code/agentic workflows، Gemini 3.5 agent-first tooling، Grok Build/workflows بعضها general digital work، بعضها coding، بعضها scientific/research، وبعضها parallel orchestration.

المقارنة الصحيحة ليست "أي agent أذكى؟" بل: ما action space؟ ما permissions؟ هل لديه persistent memory؟ ما maximum horizon؟ هل يدعم background tasks؟ كيف يثبت completion؟

```
DeploymentEnvelope = Capabilities ∩ Permissions ∩ Reliability ∩ Safety
```

Source ledger / دفتر المصدر

S25 · Official · 2025-07-17
Introducing ChatGPT Agent: Bridging Research and Action
Official deployed agent combining research, browsing, computer interaction and action.

S26 · Official · 2025-02-02
Introducing Deep Research
Official reasoning agent for multi-step web research and source synthesis.

S27 · Official · 2026-05-19
Co-Scientist: A Multi-Agent AI Partner to Accelerate Research
Official multi-agent scientific discovery system for hypothesis generation and evolution.

S28 · Official · 2026-02-18
Measuring AI Agent Autonomy in Practice
Official empirical analysis of user oversight and autonomous execution in Claude Code.

S03 · Official · 2026-05-19
Gemini 3.5: frontier intelligence with action
Official model report; action-oriented workflows, agentic coding, subagents and benchmark reporting.

S05 · Official · 2026-08-12
Introducing Grok 4.6
Official current flagship release; long-running agents, coding, interactive/visual work, agentic RL, configurable reasoning effort and 500k context.

S12 · Research · 2026
Kimi K3: Open Frontier Intelligence
arXiv:2607.24653; 2.8T MoE / 104B active, native vision, 1M context, reasoning-effort levels, long-horizon agentic RL and open weights.

S01 · Official · 2026-07-09
GPT-5.6 System Card
Official system card; Sol/Terra/Luna, reasoning effort, agentic safety, prompt-injection and self-improvement evaluations.

S29 · Research · 2026
The 2025 AI Agent Index: Documenting Technical and Safety Features of Deployed Agentic AI Systems
arXiv:2602.17753; deployed-agent capability/safety documentation.

S39 · Research · 2026
SentinelBench: A Benchmark for Long-Running Monitoring Agents
arXiv:2606.05342; persistent monitoring as a distinct long-running agent workload.

Deep technical extraction / الاستخراج الهندسي

Deployment comparison. الأنظمة المنشورة تختلف في action space والقيود أكثر مما تختلف في UI. عند المقارنة نسجل: persistent/background execution؟ computer use؟ coding terminal؟ memory؟ subagents؟ confirmation gates؟ provenance؟ هذه هي dimensions التي تحدد agentic envelope.

- لا نستنتج capability غير موثقة من marketing label.
- قارن task class نفسها.
- افصل availability عن research preview.
- سجل تاريخ snapshot لأن المنتجات تتغير سريعاً.

Research specification / مواصفة الاختبار

Variable / layer	Measure	Failure signal
Schema	validation failure rate	malformed actions
Versioning	compatibility across upgrades	protocol drift
Cancellation	bounded stop latency	zombie tasks
Durability	resume after crash	lost long task
Observability	trace completeness	opaque orchestration
Sandbox	resource/FS/network limits	blast radius
Secrets	exposure incidents	credential leakage

ما الذي ينقص Frontier Agentic AI؟

Open gaps after the 2026 wave

رغم القفزة، بقيت ست فجوات بنيوية: **reliable long horizon، continual learning without catastrophic drift، causal world models، robust self-evaluation، transfer under novel tools/environments، وgovernance of autonomous side effects.**

المشكلة ليست نقص features. كثير من الأنظمة تملك memory/tool/browser/subagents، لكن coupling بينها هش. AGI-oriented OS يجب أن يضمن أن permissions، evidence، state، وobjectives تنتقل بين الطبقات دون فقد أو تناقض.

```
AGI Gap ≠ MissingModule ; AGI Gap ≈ MissingReliableIntegration
```

Source ledger / دفتر المصدر

S29 · Research · 2026
The 2025 AI Agent Index: Documenting Technical and Safety Features of Deployed Agentic AI Systems
arXiv:2602.17753; deployed-agent capability/safety documentation.

S31 · Research · 2026
The Long-Horizon Task Mirage? Diagnosing Where and Why Agents Fail as Task Horizon Increases
arXiv:2604.11978; cross-domain degradation with horizon length.

S32 · Research · 2025
UltraHorizon: Benchmarking Agent Capabilities in Ultra Long-Horizon Scenarios
arXiv:2509.21766; 200k+ token / 400+ tool-call trajectories and long-horizon failure taxonomy.

S36 · Research · 2026
AMA-Bench: Evaluating Long-Horizon Memory for Agentic Applications
arXiv:2602.22769; causal/objective memory gaps and AMA-Agent.

S45 · Research · 2026
Agents in the Wild: Where Research Meets Deployment
arXiv:2607.19336; deployment-oriented synthesis of reliability, memory, tools and evaluation.

S15 · Research · 2026
Inference Scaling, Learning to Reason, and Agentic Systems
arXiv:2504.09037v4; synthesis of inference scaling, verifiers, reasoning learning and agentic systems.

S18 · Research · 2024
Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters
arXiv:2408.03314; adaptive compute allocation and verifier-guided search.

S23 · Official · 2025-12-18
Evaluating Chain-of-Thought Monitorability
Official framework with intervention/process/outcome-property evaluations of CoT monitorability.

S30 · Research · 2026
Benchmark Test-Time Scaling of General LLM Agents
arXiv:2602.18998; sequential vs parallel scaling across search, coding, reasoning and tools.

S33 · Research · 2026
OSWorld 2.0: Benchmarking Computer Use Agents on Long-Horizon Real-World Tasks
arXiv:2606.29537; realistic long-horizon computer-use evaluation.

Deep technical extraction / الاستخراج الهندسي

Frontier gaps. ما يزال integration هو عنق الزجاجة: reasoning قوي قد يعمل مع memory ضعيفة، tool use جيد مع verification ضعيف، أو long context بلا continual adaptation. الاقتراب من AGI يتطلب أن تبقى هذه القدرات متماسكة تحت novelty والزمن.

- continual learning دون drift.
- causal world models قابلة للتحديث.
- calibrated self-evaluation.
- transfer إلى tools/envs غير مألوفة.
- governance لل side effects طويلة المدى.

Research specification / مواصفة الاختبار

Variable / layer	Measure	Failure signal
Maturity	E0-E5 evidence rank	artifact=deployment confusion
Replication	repeat/independent test	one-off success
Negative result	registry downgrade/patch	theory immune to failure
World effect	external-state verification	simulation-only success
Concept definition	operational fields complete	name without mechanism
External analogue	closest published mechanism	reinvention
Differential	unique predicted gain	no novelty

All-vs-All Mapping إلى MOUSSAIF OS

12-layer integration view

MOUSSAIF layer	Frontier counterpart
L0 Constitution/Safety	confirmation, least privilege, prompt-injection defense
L1 Goal/Question	deep research query reformulation, planning
L2 MCD/QMCD/MCAI	adaptive effort, branch selection, test-time scaling
L3 HKOS/World Model	long-context + causal memory + retrieval
L4 Retrieval	web/deep research evidence acquisition
L5 HyperSkills	tool schemas / coding actions / reusable procedures
L6 HyperAbilities	composed capabilities across tools/models
L7 HyperLoops	durable workflows / agent trajectories
L8 Router/Agents	subagents, swarm, orchestration
L9 Tools	MCP, computer use, terminal/browser
L10 Verification	verifiers, tests, world-state diff
L11 Memory/Evolution	SAM/AMA causal memory, agent traces, learning

Source ledger / دفتر المصدر

S01 · Official · 2026-07-09
GPT-5.6 System Card
Official system card; Sol/Terra/Luna, reasoning effort, agentic safety, prompt-injection and self-improvement evaluations.

S12 · Research · 2026
Kimi K3: Open Frontier Intelligence
arXiv:2607.24653; 2.8T MoE / 104B active, native vision, 1M context, reasoning-effort levels, long-horizon agentic RL and open weights.

S30 · Research · 2026
Benchmark Test-Time Scaling of General LLM Agents
arXiv:2602.18998; sequential vs parallel scaling across search, coding, reasoning and tools.

S36 · Research · 2026
AMA-Bench: Evaluating Long-Horizon Memory for Agentic Applications
arXiv:2602.22769; causal/objective memory gaps and AMA-Agent.

S43 · Protocol · 2026-07-28
MCP Specification 2026-07-28
Stable open protocol revision for tool/context integration, tasks, subscriptions, caching and multi-round trips.

S44 · Protocol · 2025
Agent2Agent (A2A) Protocol
Open protocol for interoperable agent-to-agent communication across vendors/frameworks.

S15 · Research · 2026
Inference Scaling, Learning to Reason, and Agentic Systems
arXiv:2504.09037v4; synthesis of inference scaling, verifiers, reasoning learning and agentic systems.

S18 · Research · 2024
Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters
arXiv:2408.03314; adaptive compute allocation and verifier-guided search.

S23 · Official · 2025-12-18
Evaluating Chain-of-Thought Monitorability
Official framework with intervention/process/outcome-property evaluations of CoT monitorability.

S29 · Research · 2026
The 2025 AI Agent Index: Documenting Technical and Safety Features of Deployed Agentic AI Systems
arXiv:2602.17753; deployed-agent capability/safety documentation.

Deep technical extraction / الاستخراج الهندسي

All-vs-all rule. كل طبقة MOUSSAIF يجب أن تربط بآلية frontier قابلة للقياس لا باسم مشابه. نبني matrix من ثلاث قيم: overlap، support، وnovel differential. أعلى أولوية للتجربة عندما يكون الادعاء الداخلي قوياً لكن الدعم الخارجي غير مباشر.

```
BridgePriority = ClaimImportance × Uncertainty × Testability
```

- support ≠ validation.
- overlap يمنع إعادة اختراع المفاهيم.
- novel differential يحتاج baseline.
- negative result يجب أن يحدث patch في theory registry.

Research specification / مواصفة الاختبار

Variable / layer	Measure	Failure signal
Maturity	E0-E5 evidence rank	artifact=deployment confusion
Replication	repeat/independent test	one-off success
Negative result	registry downgrade/patch	theory immune to failure
World effect	external-state verification	simulation-only success
Concept definition	operational fields complete	name without mechanism
External analogue	closest published mechanism	reinvention
Differential	unique predicted gain	no novelty

برنامج الاختبار 2026-2028

20 experiments that can falsify the architecture

1. MCD-based token budget vs fixed high effort.
2. QMCD branch scheduler vs best-of-N.
3. PRM + execution verifier vs single judge.
4. raw 1M context vs reconstructed state memory.
5. causal memory graph vs vector similarity only.
6. HyperSkill contracts with/without postconditions.
7. World-Effect verification vs tool-call success.
8. single agent vs 4-way swarm under equal compute.
9. parallel search then sequential execution hybrid.
10. loop detector on UltraHorizon trajectories.
11. goal-drift detector on HORIZON tasks.
12. MCP tool discovery with policy layer ablation.
13. A2A remote-agent delegation with signed evidence.
14. computer-use confirmation policy vs no policy.
15. recovery checkpointing on SWE-EVO.
16. trace-based failure classifier before outcome.
17. self-judge vs heterogeneous verifier ensemble.
18. Deep Research governing-question controller.
19. multimodal state freshness detector for GUI agents.
20. end-to-end 12-layer ablation under fixed model/budget.

نجاح البرنامج: ليس رفع score فقط، بل تحديد أي طبقة تغيّر reliability curve وأين تتوقف فائدتها.

Source ledger / دفتر المصدر

S18 · Research · 2024
Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters
arXiv:2408.03314; adaptive compute allocation and verifier-guided search.

S23 · Official · 2025-12-18
Evaluating Chain-of-Thought Monitorability
Official framework with intervention/process/outcome-property evaluations of CoT monitorability.

S30 · Research · 2026
Benchmark Test-Time Scaling of General LLM Agents
arXiv:2602.18998; sequential vs parallel scaling across search, coding, reasoning and tools.

S31 · Research · 2026
The Long-Horizon Task Mirage? Diagnosing Where and Why Agents Fail as Task Horizon Increases
arXiv:2604.11978; cross-domain degradation with horizon length.

S32 · Research · 2025
UltraHorizon: Benchmarking Agent Capabilities in Ultra Long-Horizon Scenarios
arXiv:2509.21766; 200k+ token / 400+ tool-call trajectories and long-horizon failure taxonomy.

S33 · Research · 2026
OSWorld 2.0: Benchmarking Computer Use Agents on Long-Horizon Real-World Tasks
arXiv:2606.29537; realistic long-horizon computer-use evaluation.

S35 · Research · 2026
SWE-EVO: Benchmarking Coding Agents in Long-Horizon Software Evolution
arXiv:2512.18470; multi-step repository evolution beyond issue-level patching.

S36 · Research · 2026
AMA-Bench: Evaluating Long-Horizon Memory for Agentic Applications
arXiv:2602.22769; causal/objective memory gaps and AMA-Agent.

S43 · Protocol · 2026-07-28
MCP Specification 2026-07-28
Stable open protocol revision for tool/context integration, tasks, subscriptions, caching and multi-round trips.

S15 · Research · 2026
Inference Scaling, Learning to Reason, and Agentic Systems
arXiv:2504.09037v4; synthesis of inference scaling, verifiers, reasoning learning and agentic systems.

Deep technical extraction / الاستخراج الهندسي

Experimental discipline. التجارب المقترحة يجب أن تستخدم frozen task sets، equalized budgets، preregistered primary metrics، وfailure taxonomy. لا يكفي عرض أفضل run. نحتاج seeds/repeats حيث stochasticity مهمة، وحفظ raw traces لإعادة التحليل.

- Primary: task success under fixed budget.
- Secondary: cost/latency/recovery/safety.
- Ablation: one layer removed at a time.
- Decision rule: confidence interval + practical effect size.

Research specification / مواصفة الاختبار

Variable / layer	Measure	Failure signal
Concept definition	operational fields complete	name without mechanism
External analogue	closest published mechanism	reinvention
Differential	unique predicted gain	no novelty
Baseline	strongest simple comparator	straw baseline
Metric	direct measurable outcome	proxy-only claim
Falsifier	condition that lowers claim	unfalsifiable theory
Effect size	practical + statistical delta	tiny decorative gain

المراجع المشروحة I

S01–S24 — Frontier Reasoning Corpus

S01 — OpenAI. *GPT-5.6 System Card*. (2026-07-09). Official system card; Sol/Terra/Luna reasoning effort, agentic safety, prompt-injection and self-improvement evaluations.

S02 — Anthropic. *Claude Fable 5 & Mythos 5 System Card*. (2026-06-09). Official system card for Anthropic's most capable widely released Fable 5 class; always-on adaptive thinking, long-running agents, 1M context and safety evaluations.

S03 — Google DeepMind. *Gemini 3.5: frontier intelligence with action*. (2026-05-19). Official model report; action-oriented workflows, agentic coding, subagents and benchmark reporting.

S04 — Google DeepMind. *Gemini 3.1 Deep Think / Accelerating Mathematical and Scientific Discovery*. (2026-02). Official research/model documentation for specialized high-compute scientific reasoning.

S05 — SpaceXAI. *Introducing Grok 4.6*. (2026-08-12). Official current flagship release; long-running agents, coding, interactive/visual work, agentic RL, configurable reasoning effort and 500k context.

S06 — DeepSeek-AI. *DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence*. (2026). arXiv:2606.19348; hybrid attention, million-token context, effort scaling, agentic benchmarks.

S07 — Reuters / DeepSeek release coverage. *DeepSeek V4-Pro-0813 launch and agent capability update*. (2026-08-13). Current-release reporting on V4 Pro availability and improved AI-agent capabilities; used as deployment-state evidence, not as an independent technical paper.

S08 — DeepSeek-AI. *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*. (2025). arXiv:2501.12948; RLVR-style reasoning emergence, cold start, distillation.

S09 — Qwen Team. *Qwen3.8-Max: A New Bar for Coding and Cowork*. (2026-08-02). Official /current flagship report; 2.4T-class MoE, multimodal/long-context frontier, coding and cowork agentic work. Vendor benchmark claims are treated as reported.

S10 — Qwen Team. *Qwen3-Coder-Next Technical Report*. (2026). arXiv:2603.00729; repository reasoning, agent interaction patterns, executable training/evaluation environments.

S11 — Qwen Team. *Qwen3.5-Omni Technical Report*. (2026). arXiv:2604.15804; hybrid-attention MoE omnimodal reasoning across text, vision and audio, long-sequence interaction and grounding.

S12 — Kimi Team. *Kimi K3: Open Frontier Intelligence*. (2026) / arXiv:2607.24653; 2.8T MoE 104B active, native vision, 1M context, reasoning-effort levels, long-horizon agentic RL and open weights.

S13 — Anthropic. *Claude Opus 5 — model documentation and system card*. (2026-07-24). Official -current Opus release; deep reasoning, agentic coding, long-horizon work, 1M context and test time compute scaling up to max effort.

S14 — OpenAI. *Introducing gpt-oss*. (2025-08-05). Official open-weight reasoning release; controllable reasoning effort and open debugging/training affordances.

S15 — Survey authors. *Inference Scaling, Learning to Reason, and Agentic Systems*. (2026). arXiv:2504.09037v4; synthesis of inference scaling, verifiers, reasoning learning and agentic systems.

S16 — Zheng et al.. *A Survey of Process Reward Models: From Outcome Supervision to Process Supervision*. (2026). arXiv:2510.08049; PRM data, training, inference-time use and RL.

S17 — Survey authors. *Reward Modeling for Reinforcement Learning-Based LLM Reasoning*. (2026). arXiv:2602.09305; model-based, rule-based and self-reward designs.

S18 — Snell et al.. *Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters*. (2024). arXiv:2408.03314; adaptive compute allocation and verifier-guided search.

S19 — Chen et al.. *Simple and Provable Scaling Laws for the Test-Time Compute of Large Language Models*. (2024). arXiv:2411.19477; tournament/league aggregation and scaling guarantees under assumptions.

S20 — Zhang et al.. *Days After DeepSeek-R1: A Survey on Replication Studies and More Directions for Reasoning Language Models*. (2025). arXiv:2505.00551; replication lessons for SFT and RLVR.

S21 — Marjanovic et al.. *DeepSeek-R1 Thoughtology: Let's <think> about LLM Reasoning*. (2025). arXiv:2504.07128; analysis of reasoning length, rumination, control and safety.

S22 — Mid-Think authors. *Mid-Think: Training-Free Intermediate-Budget Reasoning*. (2026). arXiv:2601.07036; inference budget control between short and long reasoning modes.

S23 — OpenAI. *Evaluating Chain-of-Thought Monitorability*. (2025-12-18). Official framework with intervention/process/outcome-property evaluations of CoT monitorability.

S24 — OpenAI. *Detecting Misbehavior in Frontier Reasoning Models*. (2025-03-10). Official research on monitoring reasoning traces for reward hacking and exploit behavior.

Corpus taxonomy / خريطة الـ45 مصدراً

Cluster	Sources	Research use	Main caution
Frontier model/system reports	S01-S14	architecture, post-training, effort, context, multimodality, tool/agent envelopes	vendor benchmark claims require independent harness parity
Reasoning science	S15-S24	test-time scaling, PRMs, reward design, RLVR replication, monitorability	mechanism-specific gains may not transfer to open-world agency
Deployed agent systems	S25-S29	research/action workflows, autonomy, scientific multi-agent systems	product envelopes change quickly; snapshot date matters
Long-horizon and process benchmarks	S30-S42	horizon failure, GUI/web/SWE, memory, monitoring, process evaluation	benchmark success is domain-bounded, not proof of general agency
Interoperability protocols	S43-S44	MCP tool/context integration and A2A remote-agent coordination	protocol interoperability does not automatically provide authorization or safety
Deployment synthesis	S45	bridge from research architectures to reliability constraints in the wild	survey/synthesis evidence must be traced back to primary results

Reading order. ابدأ بـS15-S24 لفهم كيف يتغير reasoning مع compute وreward؛ انتقل إلى S01-S14 لفحص كيف تُجسّد العائلات frontier هذه الآليات في systems؛ ثم S30-S42 لاختبار ما ينهار عند horizon والبيئات الحقيقية؛ وأخيراً S43-S45 لفهم طبقة التشغيل البيني والنشر. لا تُستخدم هذه المراجع كتصويت على "AGI"، بل كشبكة قيود تصميم وتجارب.

Evidence policy. Official/system card يثبت ما أعلنت الجهة أنها صممته أو اختبرته في ذلك الإصدار؛ arXiv/benchmark يثبت ادعاء الورقة ضمن إعدادها؛ الاستنتاجات العابرة للمصادر موسومة في الكتاب كـsynthesis. أي رقم يتغير مع harness أو model snapshot لا يعامل كثابت معماري.

المراجع المشروحة II + الخاتمة

S25-S45 — Frontier Agentic Systems

S25 — OpenAI. *Introducing ChatGPT Agent: Bridging Research and Action* (2025-07-17). Official deployed agent combining research, browsing, computer interaction and action.

S26 — OpenAI. *Introducing Deep Research* (2025-02-02). Official reasoning agent for multi-step web research and source synthesis.

S27 — Google DeepMind. *Co-Scientist: A Multi-Agent AI Partner to Accelerate Research* (2026-05-19). Official multi-agent scientific discovery system for hypothesis generation and evolution.

S28 — Anthropic. *Measuring AI Agent Autonomy in Practice* (2026-02-18). Official empirical analysis of user oversight and autonomous execution in Claude Code.

S29 — AI Agent Index authors. *The 2025 AI Agent Index: Documenting Technical and Safety Features of Deployed Agentic AI Systems* (2026). arXiv:2602.17753; deployed-agent capability/safety documentation.

S30 — General AgentBench authors. *Benchmark Test-Time Scaling of General LLM Agents* (2026). arXiv:2602.18998; sequential vs parallel scaling across search, coding, reasoning and tools.

S31 — HORIZON authors. *The Long-Horizon Task Mirage? Diagnosing Where and Why Agents Fail as Task Horizon Increases* (2026). arXiv:2604.11978; cross-domain degradation with horizon length.

S32 — Luo et al. *UltraHorizon: Benchmarking Agent Capabilities in Ultra Long-Horizon Scenarios* (2025). arXiv:2509.21766; 200k+ token / 400+ tool-call trajectories and long-horizon failure taxonomy.

S33 — XLANG et al. *OSWorld 2.0: Benchmarking Computer Use Agents on Long-Horizon Real-World Tasks* (2026). arXiv:2606.29537; realistic long-horizon computer-use evaluation.

S34 — Jang et al. *Odysseys: Benchmarking Web Agents on Realistic Long Horizon Tasks* (2026). arXiv:2604.24964; realistic web task horizon and interaction complexity.

S35 — SWE-EVO authors. *SWE-EVO: Benchmarking Coding Agents in Long-Horizon Software Evolution* (2026). arXiv:2512.18470; multi-step repository evolution beyond issue-level patching.

S36 — Zhao et al. *AMA-Bench: Evaluating Long-Horizon Memory for Agentic Applications* (2026). arXiv:2602.22769; causal/objective memory gaps and AMA-Agent.

S37 — Omri et al. *Agent Memory: Characterization and System Implications of Stateful Long-Horizon Workloads* (2026). arXiv:2606.06448; system taxonomy, cost profiling and fleet implications.

S38 — Hu et al. *SAM: State-Adaptive Memory for Long-Horizon Reasoning Agent* (2026). arXiv:2605.24468; evolving-state-conditioned recall and compact cues.

S39 — Microsoft Research et al. *SentinelBench: A Benchmark for Long-Running Monitoring Agents* (2026). arXiv:2606.05342; persistent monitoring as a distinct long-running agent workload.

S40 — Mazaheri et al. *AgentAtlas: Beyond Outcome Leaderboards for LLM Agents* (2026). arXiv:2605.20530; behavioral/process evaluation beyond success rate.

S41 — Lacoste et al. *CUBE: A Standard for Unifying Agent Benchmarks* (2026). arXiv:2603.15798; common benchmark interface and environment standardization.

S42 — CocoaBench Team. *CocoaBench: Evaluating Unified Digital Agents in the Wild* (2026). arXiv:2604.11201; cross-application digital agent evaluation.

S43 — Model Context Protocol. *MCP Specification 2026-07-28* (2026-07-28). Stable open protocol revision for tool/context integration, tasks, subscriptions, caching and multi-round trips.

S44 — Google / Linux Foundation. *Agent2Agent (A2A) Protocol* (2025). Open protocol for interoperable agent-to-agent communication across vendors/frameworks.

S45 — Agents in the Wild authors. *Agents in the Wild: Where Research Meets Deployment* (2026). arXiv:2607.19336; deployment-oriented synthesis of reliability, memory, tools and evaluation.

الخلاصة الحاكمة: الجبهة الأقرب إلى AGI في 2026 ليست model أكبر وحده، بل نظام يخصص compute، يستدعي evidence وأدوات، يحافظ على state، ينسق وكلاء، يكتشف فشله، يتحقق من world effect، ثم يكتب ذاكرة يمكن إعادة بنائها. هذا هو الموضع الذي يجب أن يُختبر فيه MOUSSAIF: لا كادعاء اسم، بل كطبقة تكامل قابلة للإزالة والمقارنة والهدم.

Frontier synthesis / الخلاصة الهندسية للحزأين

Frontier axis	state 2026	AGI-relevant question
Reasoning	RL-trained deliberation + adaptive effort + verifier/search stacks	هل يخصص compute بكفاءة لمسائل جديدة دون rumination؟
Context & memory	very-long context + external/state-adaptive memories	هل يحافظ على causal/objective state بدلاً من history خام؟
Tools & action	co-trained tool use, computer/web/code environments	هل يغلق الحلقة من belief إلى verified world effect؟
Multi-agent	subagents, swarms, scientific councils, parallel RL	هل gain أكبر من coordination cost وcorrelated error؟
Evaluation	long-horizon, process, GUI/web/SWE/memory benchmark suites	هل ينتقل الأداء خارج benchmark family؟
Safety	prompt-injection defenses, confirmations, least-privilege patterns	هل autonomy تبقى bounded مع ارتفاع capability؟
Protocols	MCP + A2A style interoperability	هل يمكن نقل state/evidence/authority بين الأنظمة بلا loss؟
MOUSSAIF test	layer OS hypothesis mapped to these frontier mechanisms-12	أي طبقة ترفع reliability curve تحت equal model/budget؟

Minimum frontier baseline for future experiments. Model snapshot مثبت؛ adaptive reasoning؛ tool schema + sandbox؛ budget؛ explicit mission state؛ memory with validity/provenance؛ trace plane؛ independent verification؛ post-action world-state check؛ benchmark portfolio يغطي web/GUI، code، long horizon، reasoning، وmemory. أي نظرية AGI داخلية يجب أن تتفوق على هذا baseline أو تبرهن أنها تخفض الكلفة/الخطر عند جودة مساوية.

Freshness rule. هذا الكتاب يحمل cutoff صريحاً في 18 أغسطس 2026. بعد ذلك، أي model/system جديد يضاف كسجل delta: ما capability الجديد؟ ما benchmark أو protocol الذي تغير؟ وما bridge الذي يصبح أقوى أو أضعف؟ بهذه الطريقة يبقى corpus قابلاً للتحديث دون إعادة كتابة النظرية من الصفر.