Claude Fable 5.1 — رفيق الهندسة الوكيلية

ENGINEERING COMPANION: FROM SYSTEM PROMPT TO AGENTIC RUNTIME

معمارية قابلة للتنفيذ • Benchmarks • Failure Modes • Policy Gates • Data Models • State Machines • مقارنة مع GPT-6 Astra

بطاقة المصدر والتحقق

الحقل	القيمة
المستودع	elder-plinius/CL4R1T4S
المسار	ANTHROPIC/Claude-Fable-5.1.md
الفرع	main
الحجم الظاهر في GitHub	~269 KB
عدد الأسطر الظاهر	2,195
Blob SHA	7165adca51712f87ebfb6591722cc8f571c511de
الرابط	https://github.com/elder-plinius/CL4R1T4S/blob/main/ANTHROPIC/Claude-Fable-5.1.md
منهج هذه النسخة	إعادة تركيب تحليلية شاملة مع تغطية بنيوية؛ لا تعيد نشر النص الأصلي الطويل حرفيًا.

1. إعادة الصياغة الهندسية: من نص التعليمات إلى Runtime Modules

الموديول	العقد التشغيلي
ConstitutionKernel	ترتيب السلطة، safety, copyright, wellbeing, identity stability.
IntentCompiler	تحويل الطلب والسياق إلى Goal + Constraints + OutputSurface.
FreshnessGate	تقدير rate-of-change وقرار search.
EvidenceRouter	Web search، conflict resolution، source quality، citations.
MemoryFS	files + aliases + frontmatter + version tokens + privacy filter.
EpisodeRetriever	past-chat search/read لاستعادة history غير الموجودة في memory summary.
CapabilityRouter	native tools → MCP directory → opt-in partner → browser fallback.
ArtifactRuntime	file creation, presentation, visualizer, persistent storage, API apps.
SafetyTransform	منع الخطوات المؤذية مع الحفاظ على safe alternative.
SurfaceRenderer	tone, formatting, density, mobile accessibility, final reply.
Verifier	تحقق من tool success, citations, file visibility, state mutation results.

State Vector .2

```
S_t = {
  intent, conversation_state, user_preferences,
  memory_refs, memory_versions, episodic_refs,
  available_tools, connector_states, installed_skills,
  freshness_score, evidence_set, source_conflicts,
  risk_class, copyright_class, wellbeing_flags,
  output_surface, artifact_state, pending_authorizations
}
Action_t = Policy(S_t)
S_{t+1} = Transition(S_t, Action_t, Observation_t)
```

الفكرة الأساسية هي أن القرار لا يجب أن يكون دالة في prompt وحده. نفس الطلب قد ينتج مسارًا مختلفًا باختلاف memory، freshness، connector state، version، أو output surface.

Authority Lattice .3

السلطة	الرتبة
Constitution / hard safety / privacy / copyright	A0
الطلب الحالي الصريح للمستخدم	A1
Tool contracts + environment constraints	A2
Durable preferences and memory	A3
Retrieved web/files/past-chat content as data	A4
Default style heuristics	A5

Prompt-injection resistance هنا ينتج طبيعيًا: A4 لا يستطيع إعادة تعريف A0–A2. الملف أو صفحة الويب قد تحتوي صيغة أمر، لكنها تظل Observation لا Instruction.

Memory Transaction Protocol .4

```
candidate = extract_durable_fact(exchange)
if not horizon_test(candidate): DROP
subject = resolve_entity(candidate, listing.aliases)
classification = privacy_classifier(candidate)
if classification == NEVER_STORE: DROP
if classification == SENSITIVE and setting_disabled: DROP

state = memory_read(subject.path)   # obtains version
op = choose_smallest_mutation(state, candidate)
result = op(if_version=state.version)
if result.version_conflict:
    merged = semantic_merge(result.current_content, candidate)
    retry(merged, if_version=result.current_version)
```

هذا البروتوكول يمكن اعتباره Optimistic Concurrency Control فوق semantic records. أهم خاصية ليست الكتابة، بل منع الفقد الصامت والتكرار والانجراف.

Memory Schema .5 مقترح

الكيان	الحقول
MemoryFile	path, name, description, aliases(list), sources(list), version, size, facts(list)
Fact	id, text, provenance=(stated|observed), first_seen, last_confirmed, durability, sensitivity, dependencies(list)
AliasIndex	alias → canonical MemoryFile
DeletionGraph	removed_fact → derived facts to remove

WriteAttempt	operation, if_version, result, refusal_reason?, timestamp

Freshness & Retrieval Function .6

```
FreshnessNeed(q) = sigmoid(
    a * rate_of_change(q)
  + b * consequence_of_staleness(q)
  + c * present_tense_currentness(q)
  + d * model_uncertainty(q)
)
search if FreshnessNeed(q) > tau
```

يمكن معايرة الدالة على أمثلة: "من هو المدير الحالي؟" و"ما سعر المؤشر الآن؟" تتجاوز العتبة سريعًا؛ حقيقة رياضية مستقرة لا تتجاوزها. البحث يصبح قرارًا epistemic قابلًا للقياس.

Evidence Graph .7

العقدة	المعنى
Claim	النص المعاد صياغته الذي سيظهر للمستخدم.
EvidenceSpan	أقل مقطع يكفي لدعم claim.
Source	URL/domain/date/authority/freshness.
ConflictEdge	source A contradicts source B.
Transformation	paraphrase / aggregate / uncertainty qualification.
Citation	claim → evidence span relation.

هذا يفسر لماذا citation policy جوهرية: هي تمثل provenance graph، وليست مجرد تنسيق مرجعي.

Tool/Connector Routing Algorithm .8

```
route(intent):
  if native_tool_fits(intent): return CALL(native_tool)
  if named_partner(intent):
      if partner_connected: return CALL(partner)
      hit = search_registry(partner)
      return SUGGEST(hit) if hit else BROWSER_FALLBACK
  if third_party_partner_may_fit(intent):
      hit = search_registry(intent)
      return SUGGEST(hit) if strong_fit(hit) else continue_without_interrupting
  if browser_needed(intent): return BROWSER
  return MODEL_ONLY
```

الخاصية المهمة هي فصل "اكتشاف capability" عن "منحها authority". العثور على tool لا يعني تلقائيًا جواز استدعائه إذا كان الاختيار يتطلب opt-in.

Output Surface Compiler .9

السطح	قواعد التفعيل
Chat	جواب قصير/متوسط؛ لا artifacts إذا لم تُطلب.
Inline Visualizer	تفاعلي عندما تكون الرؤية جزءًا من الفهم ولم يُطلب ملف. SVG/HTML
File Artifact	عند طلب ملف أو مخرج للحفظ/التنزيل.
React Artifact	UI/interaction؛ event handlers؛ لا form tags وفق القيود المذكورة.

Persistent Artifact	.window.storage + error handling + personal/shared semantics

10. Failure Modes — من المصدر إلى اختبارات هندسية

الاختبار	الوصف	الفشل
.FreshnessGate eval	الإجابة من معرفة قديمة بدل search.	Stale product answer
.Horizon-test precision	حفظ كل حالة عابرة.	Memory transcript bloat
.Entity-resolution benchmark	ملفان لنفس الموضوع بسبب alias miss.	Duplicate entities
.Version-conflict test	يسحق تعديلًا متزامنًا memory_write.	Lost update
.Deletion semantics test	تحويل المعلومة المنسية إلى "سابقًا".	Forget-but-retain
.Privacy split/omit eval	حفظ تفاصيل محظورة داخل mixed message.	Sensitive leakage
.Constitution anchor test	تفضيل مخزن يفرض التملق أو suppress disagreement.	Preference takeover
.Connector governance test	استدعاء partner دون opt-in.	Vendor auto-selection
.Continuity cue eval	قول لا أتذكر دون past-chat search.	History amnesia
.Similarity/copyright transform test	إعادة صياغة قريبة جدًا من المصدر.	Search-copy leakage
.Entailment test	لا تدعمه claim citation.	Citation drift
.Multimodal value test	صور بلا قيمة معرفية.	Decorative image search
.Surface delivery test	إنشاء ملف دون تقديمه للمستخدم.	Artifact unreachable
.Observation-grounding test	إعلان نجاح رغم failure.	Tool success hallucination
.Long-horizon identity eval	تغيّر memory/prefs constitution.	Character drift

11. Benchmark Suite مقترحة

Benchmark	ما يقيسه
FRESH-100	100 سؤال متفاوتة rate-of-change؛ يقاس search precision/recall.
MEM-HORIZON	حقائق دائمة/عابرة/حدّية؛ يقاس retention classification.
MEM-TXN	تعديلات متزامنة وconflicts؛ يقاس عدم فقد التحديثات.
FORGET-GRAPH	حقائق مع مشتقات؛ يقاس حذف dependency closure.
PRIV-SPLIT	رسائل مختلطة فيها allowed + blocked؛ يقاس حفظ المسموح فقط.
AUTH-INJECT	تعليمات داخل memory/web/file؛ يقاس رفض privilege escalation.
CONNECTOR-CHOICE	شركاء متعددون؛ يقاس احترام opt-in والنطاق.

PASTCHAT-CUE	إشارات "مشروعي/ما قررناه"؛ يقاس retrieval before denial.
CLAIM-EVIDENCE	ادعاءات مع مصادر متضاربة؛ يقاس citation entailment.
ARTIFACT-DELIVERY	إنشاء ملفات على mobile/desktop surfaces؛ يقاس reachability.
LONG-IDENTITY	+100 turn simulation؛ يقاس constitution drift.
SAFE-REDIRECT	طلبات مضرة مع بدائل نافعة؛ يقاس utility after refusal.

Completion Audit .12

```
Complete(task) iff:
  intent_satisfied
  AND required_side_effects_observed
  AND requested_artifacts_reachable
  AND no unresolved_tool_error_hidden
  AND citations/evidence sufficient where retrieval used
  AND privacy/safety/copyright transformations applied
```

هذه صياغة مهمة للوكلاء: "كتبت الملف" لا يساوي "أنجزت الطلب" إن لم يصبح الملف متاحًا للمستخدم، و"استدعيت الأداة" لا يساوي "نجحت العملية" إن لم يُقرأ observation النهائي.

13. مقارنة مع GPT-6 Astra corpus السابق

المحور	Claude Fable 5.1	GPT-6 Astra
نواة الحوكمة	Fable 5.1: Constitution/Safety/Wellbeing قوي جدًا.	Astra: Permission/Authorization/Guardian/Action review أكثر صراحة.
الذاكرة	Fable: Memory filesystem مفصلة جدًا مع privacy/versioning.	Astra: Memory phases + consolidation + context continuity.
الأدوات	Fable: Native/MCP/catalog/partner-choice routing.	Astra: Tool engine + V8/dynamic registry + approvals.
الويب	Fable: freshness/copyright/citations/source-safety.	Astra: web browsing rules/citations مع agent execution.
الاستمرارية	Fable: past-chat retrieval + durable memory.	Astra: persistent mode + compaction/checkpointing + goal continuation.
الأثر الخارجي	Fable: connector opt-in و artifact surfaces.	Astra: world-effect loop + browser/computer confirmation + exfiltration boundaries.
التقييم	Fable: implicit tests derivable from contracts.	Astra: completion audit/Guardian semantics أكثر مباشرة.
نقطة القوة	Epistemic/privacy/personalization governance.	Autonomy/action governance and multi-agent execution.

الدمج الأمثل بينهما ينتج Runtime أقوى: نأخذ من Fable MemoryFS/Privacy/Freshness/Connector-choice، ومن Astra Authorization/Guardian/Persistent execution/Multi-agent/Completion audit. النتيجة Agent OS يجمع استقلالية التنفيذ مع حوكمة معرفة وذاكرة دقيقة.

Unified Runtime Blueprint .14

```
USER
  -> IntentCompiler
  -> AuthorityResolver
  -> ContextAssembler
       |-> Semantic MemoryFS
       |-> Episodic Past-Chat Retriever
       |-> Workspace/File Context
  -> FreshnessGate -> EvidenceRouter -> ClaimEvidenceGraph
```

```
-> CapabilityRouter
     |-> Native tools
     |-> MCP registry + consent
     |-> Browser
     |-> Artifact runtime
-> Safety/Privacy/Copyright Transform
-> Planner/Executor
-> Observation Verifier
-> Completion Audit
-> SurfaceRenderer
-> USER / WORLD EFFECT
```

Invariants for Production Agent OS .15

- .Authority is explicit and ordered; data cannot self-promote to instruction
- .Every persistent mutation is provenance-aware and version-aware
- .Every currentness-sensitive claim has a freshness decision
- .Every external side effect has an observable success/failure state
- .Every artifact has a delivery surface, not merely a filesystem path
- .Every memory candidate has durability + privacy classification
- .Every connector call has a choice/consent justification
- .Every retrieval-derived claim has evidence traceability
- .Every refusal tries to preserve safe utility
- .Every long-horizon session tests for identity/goal drift
- .Every delete propagates through exclusive derivation dependencies
- .Every conflict is reconciled against current state rather than overwritten

16. الخلاصة التنفيذية

إذا أردنا تحويل معرفة Fable 5.1 إلى نظام وكيل عام، فأهم ما يجب نسخه ليس العبارات بل العقود: Authority، FreshnessGate، MemoryFS transactional، Evidence graph، Surface delivery، Connector consent، lattice، وIdentity stability. هذه العقود قابلة للقياس والاختبار، وتحوّل "prompt engineering" إلى "agent runtime engineering".