Harness Engineering for Language Agents: The Harness Layer as Control, Agency, and Runtime

Chaoyue He *, Xin Zhou, Di Wang, Hong Xu, Wei Liu, Chunyan Miao

Posted Date: 23 April 2026

doi: 10.20944/preprints202603.1756.v2

Keywords: language agents; harness engineering; agent harness; control agency runtime (CAR); agent evaluation; prompt engineering; context engineering; HarnessCard; scaffolding; LLM agents; tool use

 doi:10.20944/preprints202603.1756.v2

Article

Harness Engineering for Language Agents: The Harness Layer as Control, Agency, and Runtime

Chaoyue He [1,*], Xin Zhou [1], Di Wang [1], Hong Xu [1], Wei Liu [2] and Chunyan Miao [1]

[1] Alibaba-NTU Global e-Sustainability CorpLab (ANGEL), Singapore
[2] Alibaba Group, China
* Correspondence: cyhe@ntu.edu.sg

Abstract

Language agents that act through tools, files, browsers, APIs, and persistent sessions are shaped by more than the base model or a single prompt. Their reliability depends on a *harness layer* that determines which instructions remain authoritative, what actions are available, how state is carried forward, and how failures are handled over time. This paper argues that this layer warrants explicit treatment in NLP. We propose and operationalize a working decomposition of the harness layer as **control, agency, and runtime (CAR)**; situate harness engineering in the arc from software engineering through prompt and context engineering; and audit **63** harness-relevant works, suggesting a meaningful visibility gap between academic papers and public engineering notes. We further argue that many reported agent gains may be partly **harness-sensitive** rather than purely model-driven, and propose HarnessCard as a lightweight reporting artifact. Grounded in papers, benchmarks, protocols, and engineering notes collected through **April 21, 2026**, we argue that progress in language agents should report not only the model, but also the harness layer that turns capability into governed action.

Keywords: language agents; harness engineering; agent harness; control agency runtime (CAR); agent evaluation; prompt engineering; context engineering; HarnessCard; scaffolding; LLM agents; tool use

1. Introduction

Reliable agency is designed, not inferred. Once language models act through tools, files, browsers, APIs, and persistent sessions, reliability depends on a *harness layer*: the extra-model layer that determines which instructions remain binding, what actions are available, how state is externalized, and how multi-step execution is kept bounded, recoverable, and inspectable. We argue that *harness engineering* is a useful name for work on that layer and that language-agent research should study it as an explicit object rather than leave it as hidden implementation residue. Figure 1 previews both the widening of the engineering object and our working CAR decomposition of the harness layer.

 doi:10.20944/preprints202603.1756.v2

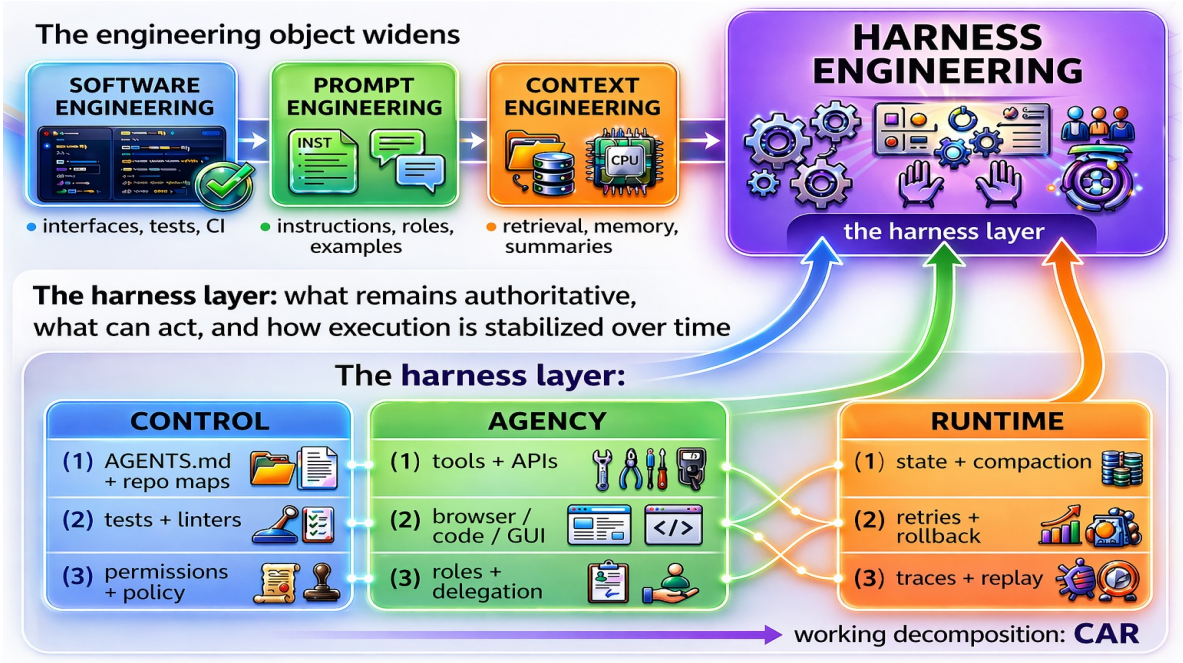

Figure 1. Top: the engineering object widens from software engineering [1] through prompt engineering [2], context engineering [3], to harness engineering [4,5]. **Bottom**: our decomposition of the *harness layer* into control, agency, and runtime (CAR).

Anthropic defines an *agent harness* (or scaffold) as the system that enables a model to act as an agent, and emphasizes that evaluating an "agent" means evaluating model and harness together rather than the model in isolation [4]. OpenAI uses the broader label *harness engineering* for long-horizon systems whose reliability depends on repository maps, `AGENTS.md`, architectural rules, cleanup loops, and runtime controls rather than on prompt wording alone [5–8]. Anthropic's note on long-running application development and its Managed Agents essay emphasize generator–evaluator structure, artifact handoffs, context resets, stable session interfaces, and meta-harness design for long-horizon work [9,10]. OpenAI's Agents SDK packages sandbox execution, file inspection, command running, approvals, and resume-capable workspaces as reusable harness primitives rather than one-off product logic [11]. Anthropic's earlier long-running harness work and research-system notes extend the same lesson beyond coding by emphasizing state externalization, progress files, orchestration, and recovery structure in multi-session or multi-agent work [12,13]. Meanwhile, academic work makes the layer explicit on its own terms: NLAH externalizes harness behavior as a portable executable natural-language artifact, Meta-Harness treats harness code as an optimization target, and a survey studies agent harnesses as a distinct research object [14–16]. Taken together, these materials motivate the working use in this paper: a *harness* is the extra-model system that couples durable control artifacts, mediated action interfaces, and runtime policies into a governed execution regime.

The underlying design problem, however, is older than the name. Work on tool use, browser agents, software agents, orchestration frameworks, interactive evaluation, runtime protocols, and deployment notes has long studied pieces of the same layer without naming it [17–25]. Our claim is therefore not that harness-like mechanisms are new. It is that practice makes the layer visible enough that the field now benefits from naming, decomposing, and reporting it directly.

We draw on a selective evidence base of harness-relevant papers, benchmarks, protocols, and engineering notes with a literature cut-off date of **April 21, 2026** (Appendix A). The contribution is fourfold. First, we situate harness engineering in the widening arc from software engineering through prompt and context engineering. Second, we propose and operationalize a working definition of the harness layer through a coupled *control–agency–runtime* (CAR) decomposition and use that decomposition to clarify the scientific object. Third, we add a descriptive audit of the in-scope evidence base to examine which parts of the harness are more visible in the literature and which are more

 doi:10.20944/preprints202603.1756.v2

often surfaced only in engineering notes. Fourth, we argue that language-agent progress should be interpreted and reported more *harness-sensitively*, and we propose HARNESSCARD as a lightweight disclosure artifact for that purpose. Our claim is not that scaffolds, runtimes, or orchestration are new, but that the harness layer has become a reportable scientific object in its own right, making claims more comparable, auditable, and reproducible.

Table 1. Public formulations motivating the harness layer as a distinct systems object; fuller version in Appendix A.2.

Public source	Year	Object	Why it matters for the harness layer
Anthropic evals note [4]	2026	agent harness / evaluation harness	defines the harness as the system that enables action and makes explicit that agent evaluation measures harness plus model
OpenAI harness note [5]	2026	harness engineering	names the broader practice and ties it to project guidance, constraints, cleanup loops, and long-horizon coding work
Anthropic application-harness note [9]	2026	long-running application harness	shows generator–evaluator structure, artifact handoffs, and context resets as first-class harness levers
Anthropic Managed Agents note [10]	2026	managed agents / meta-harness	argues for stable interfaces around evolving harnesses and decouples long-horizon sessions from any single sandbox implementation
OpenAI Agents SDK update [11]	2026	model-native harness / SDK	packages sandboxed execution, files, approvals, and resume bookkeeping as reusable harness primitives
OpenAI Codex notes [26,27]	2026	Codex harness	treats the harness as reusable runtime logic that can power multiple surfaces, not just a one-shot prompt wrapper
Anthropic long-running note [12]	2025	long-running harnesses	makes state externalization, progress tracking, and recovery first-class engineering levers
OpenAI developer guidance [6–8]	2026	`AGENTS.md`, long-horizon tasks	shows the harness as durable instruction plus executable verification and repair loops
MCP specification [28]	2025	protocol layer	moves tool interoperability and permission boundaries into an explicit systems interface

2. From Software Engineering to Prompt, Context, & Harness Engineering

Seen historically, harness engineering does not replace earlier forms of engineering around models; it nests them. Software engineering taught the field to care about interfaces, tests, modularity, and operational discipline [1]. Prompt engineering narrowed the immediate intervention surface to the design of instructions and examples [2]. Context engineering widened that surface to the evolving token state an agent carries from step to step, including retrieval, memory, tool context, and history [3]. Harness engineering widens it again: not only what enters context, but also how action is mediated, how runtime is controlled, how failures are repaired, and how governance and traces are maintained. Prompt design therefore remains inside control, context design spans control plus runtime, and harness engineering names the coupled layer that makes agency durable over time. Table 2 summarizes this widening by engineering question, typical artifacts, and what each frame still leaves under-described.

That widening matters because a system can improve even when the base model and prompt change little, simply because the system of record, retry policy, action surface, verifier loop, or recovery structure changed. Once agents act over time, the scientifically interesting question is no longer only what the model was told, but also which harness layer helped keep its behavior reliable.

2.1. The Label Is New, the Design Problem Is Not

The phrase *harness engineering* appears recently in public AI discourse, but the underlying design problem has been converging for years. ReAct made reasoning-and-acting trajectories explicit [17]. Toolformer and API-Bank made tool calls and evaluation concrete objects [18,19]. AgentBench, WebArena, GAIA, SWE-bench, BrowseComp, Terminal-Bench 2.0, and OSWorld moved interactive action, not static text generation, toward the center of evaluation [20–22,25,29–31]. SWE-agent, OpenHands, and CodeAct showed that interface design and action substrate choices can strongly affect software-agent performance [23,24,32]. A terminal-agent report explicitly distinguishes scaffolding, harness, and context engineering in deployed coding systems [33]. MCP made tool interoperability itself a protocol-level question rather than an ad hoc detail [28]. Recent work makes the abstraction itself harder to ignore: NLAH externalizes harness logic, Meta-Harness optimizes harness code directly,

 doi:10.20944/preprints202603.1756.v2

General Agent Evaluation treats protocol mediation as a first-class evaluation problem, MCP-Atlas operationalizes real MCP-server tool use, VeRO treats agent optimization as an evaluation-harness problem, and HAL turns cross-benchmark evaluation itself into a harness-managed infrastructure problem [14,15,34–37].

What was often missing was not the mechanism itself, but a commonly named abstraction for the layer that ties these elements together. Anthropic supplied one part of the vocabulary with *agent harness* and *evaluation harness* [4]. OpenAI then made the broader engineering practice explicit under the name *harness engineering* [5]. The significance of that naming move is methodological rather than merely terminological: it identifies a recurring systems object whose responsibilities cut across context specification, tool mediation, runtime control, verification, governance, and traces. Academic and technical work increasingly meets that abstraction halfway rather than merely inheriting it from product discourse [14–16,33,34].

2.2. Canonical Public Examples

Four families of public examples make the concept concrete. First, OpenAI's harness engineering note treats durable project instructions, architectural constraints, custom linters, cleanup loops, and automated review as the central levers for getting coding agents to produce maintainable code over long horizons [5]. Second, Anthropic's note on long-running application development shows how generator–evaluator splits, task lists, artifact handoffs, and context resets function as harness-level levers in both frontend design and autonomous coding [9]; its earlier long-running harness essay frames state externalization, initialization, progress files, and recovery discipline as prerequisites for useful multi-session agency [12]. Third, Anthropic's Managed Agents note and OpenAI's updated Agents SDK show a move toward reusable harness primitives: stable session interfaces, detached sandboxes, approvals, workspace manifests, and resume-capable tooling can materially change long-horizon behavior without changing the base model [10,11]. Fourth, Anthropic's multi-agent research-system notes make the same point outside a single repository loop: source hierarchy, tool access, parallel exploration, memory structure, and recovery policy can materially change behavior in long-horizon research and browsing settings [13,21,29,30]. Table 1 shows that these materials converge from different angles on the same point: if we want to understand language agents scientifically, we need to reason about the harness layer that turns a model into a system that can act, recover, and be governed.

This is also where the present paper diverges from adjacent surveys. Overviews of autonomous agents, multi-agent systems, and agent evaluation are valuable; a harness-focused survey explicitly catalogs harness components across representative systems, and an externalization-centered review places harness engineering inside a broader shift from weights to context to external cognitive infrastructure [16,38–42]. Our claim is not that those perspectives are wrong. It is that a CAR-centered, reporting-oriented harness perspective cuts across them and helps explain why systems that look similar at the model level often behave differently in practice.

Table 2. The engineering object widens from prompts to context and then to the harness around the model. The labels differ not only in scope, but in what they make scientifically visible.

Frame	Main engineering question	Typical artifacts	What remains under-described if this frame is treated as sufficient
Software engineering	How should the system stay correct and maintainable?	modules, interfaces, tests, CI, operational procedures	model-facing instructions, evolving context, and agent-specific control policies
Prompt engineering	What should the model be told?	system prompts, examples, roles, output schemas	retrieval, memory, runtime policy, permissions, and tool mediation
Context engineering	What should the model see right now?	retrieved snippets, message history, tool descriptions, summaries, notes	action interfaces, approval logic, recovery policy, and observability over time
Harness engineering	How should a language agent be governed over time?	durable instructions, tool contracts, checkpoints, graders, budgets, approvals, traces	the agent is no longer reduced to the model alone; the harness layer becomes the thing that must be reported

 doi:10.20944/preprints202603.1756.v2

3. A Working Definition: The Harness Layer

Building on these formulations, we propose a working definition for this paper. We define the *harness layer* as the extra-model layer that determines what an agent sees, what it can do, how its work unfolds over time, which feedback it receives, and how that behavior is constrained, observed, and evaluated. Harness engineering is the design and maintenance of that layer. Operationally, the harness layer mediates between model capability and situated action: it translates model outputs into governed execution and turns environmental feedback back into actionable state. Academic work converges on closely related moves: NLAH lifts harness behavior into editable natural-language contracts, Meta-Harness treats harness code as an optimization target, the harness survey argues that the harness is a unified research object rather than a loose bundle of implementation details, and an externalization-centered review places harness engineering alongside memory, skills, and protocols inside a broader account of offloaded agent capability [14–16,42].

We write the harness layer compactly as $H = \langle C, A, R \rangle$, where C is the *control* layer, A is the *agency* layer, and R is the *runtime* layer. We use the acronym *CAR* because it keeps the ordering explicit and highlights that the harness is not just runtime plumbing. The notation is not meant as a rigid ontology. Its purpose is explanatory: it marks three coupled functions that papers routinely bundle together while leaving under-described.

Control.

The control layer contains durable artifacts that shape behavior before a step is taken: repository maps, `AGENTS.md`, tool descriptions, system instructions, architecture rules, tests, linters, permission policies, and success criteria [5–7]. In other words, control is where human judgment becomes machine-readable constraint. NLAH makes this externalization explicit by lifting parts of harness behavior into inspectable natural-language modules rather than burying them entirely in controller code [14]. A key harness insight is that reliable agents are rarely bounded by prompt wording; they are often bounded by specifications.

Agency.

The agency layer determines how the model is allowed to act. It includes action substrates such as code execution or browser interaction, planner–verifier or orchestrator–worker structures, reviewer roles, and the concrete interfaces that define the action space [13,23,24,35,43]. We use *agency* here in a narrow systems sense: the mediated action surface and delegation structure that the harness permits, not a claim about unrestricted autonomy or general capability.

Runtime.

The runtime layer governs what happens as work unfolds over time: context assembly, memory and compaction, checkpointing, retries, backtracking, approval flows, budgets, trace collection, and replay support [3,9–12,26,44,45]. This is where long-horizon behavior succeeds or collapses. Runtime work spans both harness-managed state and model-assisted compaction: ClawVM focuses on durable external memory, while MEMENTO shows that context compression itself can become an explicit design variable rather than hidden prompt bookkeeping [44,45]. Many agent failures are runtime failures: stale state, brittle retry loops, overgrown context, or poor recovery from intermediate mistakes.

Two concrete mini-cases.

Consider first a repository coding agent. Two systems may share the same frontier model and nearly the same task prompt, yet behave very differently because one harness adds a repository map, root-level `AGENTS.md`, required tests, a linter, bounded shell access, a progress file, and manual approval for privileged actions. The CAR lens explains why these are not minor details: the map, tests, and approval policy live in control; the shell and file-edit surface live in agency; and the progress file, retries, and escalation logic live in runtime. The reported performance is already partly a property of the harness layer, not of the model alone.

 doi:10.20944/preprints202603.1756.v2

Now consider a browser or research agent. Two systems may share the same browsing-capable model and high-level task prompt, yet differ because one harness defines a source hierarchy, citation rules, note-taking format, uncertainty-triggered escalation, tool quotas, and replayable traces [13,21,29,30]. Here, control includes source authority and citation policy; agency includes the search, browser, and delegation surface; and runtime includes scratchpads, branching traces, and recovery when evidence conflicts. Again, what looks like "agent quality" is partly a property of the harness layer around the model.

This definition also clarifies what harness engineering is *not*. It is broader than prompt engineering because prompts are only one artifact inside a larger control structure. It is narrower than "agent systems" because not every property of the environment belongs to the harness. It overlaps with platform engineering and MLOps, but it is more specifically about the layer through which language-centered models become usable agents. The harness begins where the system actively shapes trajectories: curated context, mediated tools, retries, checkpoints, graders, permissions, traces, and similar control points.

Two boundary clarifications are especially important. First, harness engineering is not merely another name for prompting. A paper that reports only the prompt while omitting which files count as the system of record, which tools can be called, which tests are binding, what is remembered, and when humans or policies must intervene is often omitting the most consequential part of the system. Second, harness engineering is not equivalent to all software infrastructure around a model. A web front-end that forwards text to a model is not yet a harness in the sense we mean. The harness begins where the system actively shapes trajectories over time, mediates action, encodes constraints, manages recovery, and produces traces that let others inspect what happened.

Table 3. A descriptive audit of 63 in-scope works (Appendix A.3). Counts reflect whether a work primarily foregrounds a harness component; tags are non-exclusive, so rows exceed n. The literature preserves the basic asymmetry while making benchmark integrity, runtime reliability, and large-scale observability more explicit: papers and benchmarks emphasize interfaces, feedback, and observability, while engineering notes emphasize runtime, control, and governance.

Source type	Control	Agency	Interfaces	Runtime	Governance	Feedback	Observability
Papers / benchmarks ($n = 38$)	5	8	13	11	5	12	22
Notes / protocols / technical articles ($n = 25$)	10	2	7	12	9	3	6

4. A Descriptive Audit of the Evidence Base

We present a descriptive audit of the 63 in-scope works in Appendix A.3, reusing the primary harness-component tags already assigned in the inventory. While this audit is a visibility check over the selected evidence base rather than a field-wide prevalence estimate, the pattern is informative.

Table 3 indicates that papers and benchmarks in our evidence base most often foreground *interfaces*, *feedback*, and *observability*. Public engineering notes, protocol documents, and technical articles more often foreground *runtime*, *control*, and *governance*. In other words, the academic-facing literature more readily names the action surface, explicit diagnostics, and evaluation setting, whereas practitioner-facing documents more readily describe the durable instructions, recovery policies, and permission structures that make a system work day to day.

Specific implementations make harness functions more explicit in academic discourse through systems like NLAH, Meta-Harness, VeRO, General Agent Evaluation, MCP-Atlas, ProdCodeBench, AnalysisBench, Terminal-Bench 2.0, HAL, COMPOSITE-STEM, HORIZON, ATBench, Meerkat, agent-reliability profiling, and ClawVM [14,15,31,34–37,44,46–52]. The remaining asymmetry helps explain why the harness layer can still feel newer in academic NLP than it does in practice: the parts of the

 doi:10.20944/preprints202603.1756.v2

layer that matter most for deployment and transfer are still disproportionately surfaced in engineering notes rather than formal system papers. This gap is one reason a reporting artifact like HARNESSCARD is useful.

5. Why the Harness Layer Changes What Counts as Progress

Many reported agent gains can be partly harness-sensitive.

When two systems use similar frontier models yet behave very differently, the explanation often lies at least partly in the harness layer. Software agents can improve when their action substrate is redesigned, not only when their model is swapped [23,24]. Interactive benchmarks can become more tractable when traces, retries, hidden checks, and verifiers are better coordinated [20–22,31,53]. Long-running agents can improve when state is externalized and progress is resumable [12,54]. Research and browsing agents can improve when orchestration, source selection, and memory structure are better aligned with the task [13,30]. Even where models matter greatly, the last mile of reliability is frequently a harness question.

Specific implementations make this point more concrete. Meta-Harness shows that harness code itself can be optimized and that discovered harnesses can outperform strong hand-engineered baselines [15]. ProdCodeBench, AnalysisBench, and Terminal-Bench 2.0 show that verification structure, staged execution, and environment design materially affect coding, software-analysis, and terminal-agent solve rates [31,46,47]. ClawVM, MEMENTO, and HORIZON make runtime state handling, compaction, and long-horizon failure attribution explicit scientific variables rather than hidden implementation details [44,45,49].

Evaluation must become harness-sensitive.

Once an agent acts through tools and over time, evaluating the model in isolation misses the object. Anthropic makes this explicit: an agent evaluation is an evaluation of harness plus model [4]. That has four implications. First, evaluations should measure trajectories and outcomes, not only strings. Second, papers should report action budgets, retries, checkpoints, graders, and interventions because these can change results. Third, productivity or reliability claims become more interpretable when paired with explicit acceptance tests and operating envelopes rather than anecdotal claims. Fourth, variance should be expected: interactive systems are more sensitive to infrastructure noise and control policies than static benchmark runs. Work already moves in this direction: General Agent Evaluation proposes a Unified Protocol and Exgentic, MCP-Atlas evaluates real MCP-server workflows through a containerized harness, VeRO frames agent optimization through a dedicated evaluation harness, ATBench shifts safety assessment to trajectories, Meerkat shows that cross-trace audits can uncover benchmark gaming and sparse failures that per-trace judges miss, and HAL adds a standardized evaluation harness with shared logs across models, scaffolds, and benchmarks [34–37,50,51]. Reliability profiling work reinforces the same point by showing that benchmark accuracy alone obscures consistency, robustness, predictability, and safety across runs [52]. Anthropic argues that small leaderboard gaps on agentic coding evaluations deserve skepticism until the evaluation resource configuration is documented and matched [55].

Benchmark integrity is now itself a harness problem. HAL's standardized harness and shared logs reveal scaffold-level bugs and agent misbehavior at scale; Anthropic's BrowseComp eval-awareness note shows that web-enabled agents can recognize the test and retrieve benchmark answers; Meerkat surfaces sparse failures across large trace sets; and efficient subsetting results indicate that agent rankings shift under scaffold variation even when full-benchmark evaluation is expensive [37,51,56,57]. This shift is not confined to coding and web tasks: COMPOSITE-STEM uses an adapted agent harness to evaluate expert-written scientific workflows with files, tools, and criterion-based rubrics [48].

Harness sensitivity also changes what a fair comparison means. Matching model family while changing tool access, retry budgets, verifier strictness, or escalation policy is not a controlled comparison; nor is holding the prompt fixed while silently changing memory compaction or checkpointing.

 doi:10.20944/preprints202603.1756.v2

In agent settings, the experimental unit is the coupled execution regime. That is why papers should report not only success rate, but also the envelope within which success was obtained: budget ceilings, allowed side effects, approval requirements, and whether failures were recoverable or terminal. Without that envelope, benchmark numbers can look commensurable while actually reflecting different kinds of systems.

Reproducibility now depends on the harness layer.

A language-agent paper can appear more novel than it really is if the harness is under-described. Retry budgets, hidden human escalations, tool filters, repository instructions, or grader prompts often remain implicit even when they are load-bearing. This is not a minor reporting issue. It obscures which gains transfer across settings, which ones depend on domain-specific control logic, and which ones are artifacts of a particular runtime. Cross-trace auditing work now makes the stakes concrete by showing that benchmark gaming and sparse failures can remain invisible when trace structure and evaluation infrastructure are under-specified [51]. Shared evaluation harnesses and logs, as in Terminal-Bench 2.0 and HAL, help only when the coupled execution regime is itself reported clearly [31,37]. A harness-centered research practice would reverse that asymmetry: the extra-model layer would be described as carefully as the model.

Table 4. Recurring harness patterns and associated reliability profiles; broader inventory in Appendix A.3.

Pattern	Typical harness levers	Representative works	Common strengths and failure modes
Single-agent tool loop	prompt assembly, tool schemas, light memory, bounded retries	ReAct [17], Toolformer [18], API-Bank [19]	simple and efficient, but brittle when tasks are long, under-specified, or tool-heavy
Executable action substrate	code as action language, interpreter feedback, self-debugging	CodeAct [23], OpenHands [32]	flexible and compositional, but can amplify side effects without strong governance
Agent–computer interface	constrained command surface, file editing, search, browser or GUI actions	SWE-agent [24], MCP-Atlas [35], AnalysisBench [47], OSWorld [25]	large gains from interface design, but failures shift to navigation, grounding, and tool parameterization
Orchestrator–worker topology	decomposition, role specialization, routing, verifier or reviewer roles	AutoGen [43], MetaGPT [58], ChatDev [59], Anthropic research system [13]	better coverage and parallelism, but coordination overhead and cascading errors remain common
Long-running harness	state externalization, checkpoints, progress files, resumability, and compaction	Voyager [54], ClawVM [44], MEMENTO [45], HORIZON [49], Anthropic long-running harnesses [9,12]	supports hours-long work, but state drift, context decay, and brittle recovery remain central risks
Policy-aware deployment	permissions, sandboxing, escalation, audit logs, protocol mediation	τ-bench [60], ATBench [50], auto mode [61], monitoring notes [62], MCP [28]	improves safety and accountability, but can reduce autonomy or hide control logic if under-reported
Evaluation / trajectory harness	task adapters, versioned snapshots, multi-run aggregation, trace auditing	General Agent Evaluation [34], VeRO [36], HAL [37], Terminal-Bench 2.0 [31], Meerkat [51]	improves comparability and benchmark integrity, but protocol assumptions, suite design, and contamination controls become part of the result

6. Why NLP Should Treat the Harness Layer as an Explicit Object of Study

Some of the clearest descriptions of harness work came from product teams, but academic papers and technical reports are closing the gap [14,15,33–37,44]. That should be read as an opportunity rather than as a reason for the research community to stay away. NLP has repeatedly advanced by turning messy practice into explicit method: annotation, evaluation, retrieval, prompting, and human feedback all followed that path. Harness engineering is ready for the same move. The field can contribute formal task definitions, benchmark suites that isolate control-layer effects, reporting norms that travel across domains, and theories of when language is the right interface for state, tools, and oversight.

What makes harness engineering especially relevant to NLP is that the harness layer itself is often made of language-bearing artifacts. Repository maps, task decompositions, tool descriptions, approval prompts, error summaries, progress files, and policy messages are not just interface copy; they are operational control media. The harness therefore turns language from output modality into an

 doi:10.20944/preprints202603.1756.v2

instrument for specification, recovery, and oversight. That is exactly why NLP should care: questions of wording, authority, grounding, and state representation are no longer peripheral UX details, but part of system correctness.

A harness-centered program would also improve the quality of agent claims. The descriptive audit above suggests that academic papers more readily surface interfaces and evaluation than the durable control and runtime policies that often decide whether a system transfers. A stronger norm would reward the opposite behavior: state the harness clearly, vary it experimentally, and explain which parts are portable versus domain-specific. General Agent Evaluation, VeRO, ATBench, Meerkat, and HAL show that evaluation infrastructure, trace structure, and audit logic are themselves part of the scientific object, not just auxiliary benchmark plumbing [34,36,37,50,51].

Treating the harness as a layer also sharpens baselines. The most revealing comparison is often not only model A versus model B, but thinner versus richer control, narrower versus broader action surfaces, or stateless versus recoverable runtime on the same model. Layer-aware baselines would let the field estimate when progress comes from the model and when it comes from the layer around it. Without them, attribution and reproducibility remain entangled.

Adjacent literature on automated research systems, public agent ecosystems, and acceptance-test-centered productivity reporting reinforce neighboring parts of the same shift, even though they do not use the harness-layer framing directly [63–65]. A harness-focused survey and an externalization-centered review reach compatible high-level conclusions through different taxonomies, which strengthens the claim that the harness is becoming a stable unit of analysis rather than a transient product label [16,42].

Two likely objections.

A natural objection is that harness engineering is "just software engineering." That objection misses the object-specific nature of the problem. Harnesses for language agents are not generic infrastructure. Their control artifacts are often partly linguistic objects: instructions, repository maps, tool descriptions, summaries, approval prompts, grader criteria, and progress files. Likewise, their governance is often mediated through language, not only through low-level system calls. Another objection is that the term may be too new or too vendor-specific to anchor research. That concern is reasonable, but it cuts in favor of clarification rather than silence. When a recurring design problem becomes visible across multiple agent families, making the abstraction explicit helps the field compare systems more honestly.

7. Research Questions from the Harness-Layer Lens

The harness perspective makes several research questions easier to see. One concerns *authority in context*. Which artifacts should an agent trust most when repository documentation, retrieved snippets, tool outputs, policy text, and runtime observations disagree? OpenAI's emphasis on a repository "system of record," Anthropic's emphasis on progress files, Managed Agents' stable session interfaces, and the Agents SDK's workspace manifests are practical answers, but NLP can ask the deeper question of how language should represent state so that later decisions remain grounded [5,10–12]. This is not only a product question. It is a question about language as a control medium.

A second family of questions concerns *recovery*. Benchmarks often reward eventual completion, but long-running systems live or die by their ability to recover from bad intermediate steps. When should a harness retry, roll back, checkpoint, escalate, or terminate? How much memory of earlier failed attempts should be carried forward? Which summaries preserve the right information for future repair without contaminating later reasoning? These are not marginal implementation details. They determine whether an agent remains useful over repeated real-world runs. Trajectory-grounded diagnostic work and runtime-memory work are beginning to provide the right empirical objects for this question by localizing long-horizon breakdowns, delayed-trigger safety failures, sparse cross-trace violations, and the costs of carrying forward too much or too little intermediate state [45,49–51].

 doi:10.20944/preprints202603.1756.v2

A third family concerns *governance through language*. Policy checks, approval prompts, uncertainty statements, and provenance traces are often treated as UX residue. For language agents, however, they are part of the grammar through which the system and the human coordinate. What wording leads to calibrated escalation rather than either over-blocking or over-compliance? How should an agent explain a blocked action, a tool request, or a risky proposed next step? How should responsibility be assigned when a failure is partly model error and partly harness error? Work on auto-approved permissions, deployment monitoring, and eval awareness suggests that these questions are no longer peripheral [56,61,62]. They sit at the boundary of NLP, HCI, security, and software engineering, and they become sharper once the harness layer is treated as an explicit object of study.

A fourth family concerns *transfer*. Some improvements travel with the model, some with the harness, and some only with a task environment. A stronger repository map or verifier loop may transfer across coding domains yet fail to help browser tasks; a better browsing policy may transfer across websites yet not to shell-heavy workflows. Reporting model transfer, harness transfer, and task transfer separately would make agent papers more cumulative because readers could tell whether a claimed gain is about general capability, portable control logic, or local infrastructure tuning. Meta-Harness, General Agent Evaluation, MCP-Atlas, and efficient agent benchmarking make this especially visible: each isolates a different sense in which performance depends on the portability of harness logic, task adapters, protocol-level tool use, or evaluation suites [15,34,35,57].

Table 5. Compact main-paper HARNESSCARD; governance and observability are merged here only for space, while Appendix B separates them and Appendix B.2 gives a filled example.

Field	Minimum disclosure	Priority
Base model(s)	model name, version, and decoding or adaptation settings	Required
Control artifacts	instructions, repo maps, `AGENTS.md`, architecture rules, tests, linters, success criteria	Required
Runtime policy	memory or compaction strategy, checkpoints, retries, rollback or escalation policy, budgets	Required
Action substrate	tools, APIs, browser or GUI access, code execution, interface schemas, MCP usage	Required
Execution topology	single-agent vs multi-agent structure, verifier or reviewer roles, routing logic	Required
Feedback stack	tests, graders, hidden checks, reflection prompts, human interventions	Required
Governance / observability	permissions, sandboxing, provenance logs, replay support, failure categories	Required
Evaluation protocol	task set, number of runs, outcome criteria, variance treatment, budget limits	Required

8. Design Patterns & Failure Modes

A harness-centered view also makes recurring design patterns easier to compare. Some systems rely on a relatively thin single-agent tool loop; others give the model an executable action substrate such as code; others build an agent–computer interface for browsing or software editing; others use orchestrator–worker or planner–verifier structures; and others emphasize long-running state plus policy-aware deployment. These are not cosmetic packaging choices. They redistribute where error happens and which interventions are possible.

The same is true for failure. Context drift, action-schema mismatch, planning collapse, verifier overfitting, policy violations, and cost blow-up are often discussed as separate problems. Through the CAR lens, these failures are related: each type appears when control, agency, and runtime are poorly matched to the task. Table 4 summarizes representative patterns and the failures they tend to surface. The point is not that one pattern is best. It is that different harness choices produce different reliability profiles, and agent papers should say so explicitly.

This has methodological consequences. A harness-centered experiment should not only ask whether a system succeeds, but also which control artifacts, runtime policies, and interface choices were necessary for success. That means benchmarking patterns as patterns rather than treating them as unreported implementation residue. It also means that papers should report when a failure was recovered by the harness rather than by the model alone, because recovery behavior is often one of the most practically important parts of the system.

 doi:10.20944/preprints202603.1756.v2

9. A Research & Reporting Agenda

Harness engineering should not be treated as an implementation afterthought. It opens a distinct research agenda for NLP.

First, study control as executable specification.

Repository maps, `AGENTS.md`, architecture rules, and cleanup loops suggest that agentic "instruction following" is partly a problem of specification design, not only of compliance [5,6,14,15,33,42]. NLP can study how language, code, schemas, and tests should be composed so that guidance remains durable and authoritative.

Second, treat agency as an interface question.

The agent's action space is harness-mediated. Tool schemas, interface design, execution substrate, and benchmark environment choice can dominate both performance and safety [23–25,31,35,47,48,66]. On this view, agent capability is a property of the model inside a specific interface and control regime.

Third, treat runtime as a scientific variable.

Compaction, state persistence, recovery policy, and execution budgets are not residue. They determine whether a system remains coherent over long horizons [3,8–12,44,45,49]. Research should ask which runtime policies preserve the right information and when agents should backtrack, escalate, or recompute.

Fourth, normalize reporting of the harness.

We propose HarnessCard, a lightweight reporting artifact for language-agent systems. Table 5 gives the compact main-paper version. At minimum, it should disclose the base model, control artifacts, runtime policy, action substrate, feedback stack, governance layer, and evaluation protocol; Appendix B gives the template and Appendix B.2 a filled example. The goal is not bureaucracy but auditable, transferable results. Unlike model cards, HarnessCard documents the apparatus that makes an agent claim interpretable. This proposal is complementary to the unified-evaluation and evaluation-harness directions now emerging in General Agent Evaluation, VeRO, ATBench, and HAL [34,36,37,50]. The appendix mini-cases in Appendix C suggest that the template travels beyond coding, and authors can disclose proprietary harnesses faithfully at a higher level of abstraction, including what was withheld.

Fifth, build layer-aware baselines.

Controlled comparisons should vary one layer at a time: the same model with different control artifacts, the same action substrate with different runtime policy, or the same runtime with different verification and governance regimes. That is necessary to estimate when progress comes from the model and when it comes from the layer. Meta-Harness, ProdCodeBench, AnalysisBench, Terminal-Bench 2.0, and reliability profiling illustrate what this can look like in practice when harness search, verification structure, environment design, and multi-run behavior are treated as explicit experimental variables [15,31,46,47,52]. A concrete, falsifiable proposition follows from this literature: holding model family roughly fixed, changes to the control or runtime layer can materially change success rate, variance, and safety on long-horizon tasks. Reporting the harness clarifies both reproducibility and attribution.

10. Conclusions

Harness engineering makes explicit the extra-model layer that governs language agents once they act through tools and time. Through the lens of control, agency, and runtime as a working decomposition, many reported agent gains may be harness-sensitive rather than model-only, and many reproducibility failures are failures to disclose the layer that shapes instruction, action, and recovery. For NLP, the point is not to downplay models, but to study and report the harness layer

 doi:10.20944/preprints202603.1756.v2

clearly enough that model progress can be separated more credibly from harness progress, turning agent results into cumulative systems knowledge.

Acknowledgments

This research is supported by the RIE2025 Industry Alignment Fund (Award I2301E0026) and the Alibaba-NTU Global e-Sustainability CorpLab.

Limitations

This paper makes a selective argument rather than a complete field survey. Its aim is to sharpen an explicit systems layer, not to catalogue every agent paper. The evidence base is therefore intentionally selective and tilted toward work that illuminates the harness directly: interactive benchmarks, software agents, tool-use systems, protocols, and official engineering notes. That improves conceptual focus, but it also means that some adjacent literatures are under-represented, including robotics, embodied control, model-training methods, and general-purpose platform engineering.

A second limitation is conceptual. The surrounding vocabulary is still evolving. Different communities use overlapping labels such as *scaffold*, *agent loop*, *orchestrator*, *runtime*, or *evaluation harness*. Our control–agency–runtime decomposition is therefore an analytic proposal rather than a settled ontology. It is useful because it organizes recurring design decisions, but other decompositions are possible and may prove better as the literature matures. Externalization-oriented accounts that group memory, skills, protocols, and harnesses differently offer a compatible but not identical perspective [42].

A third limitation is evidentiary. We rely substantially on official engineering documents from a small number of frontier organizations because they currently provide the clearest descriptions of harness practice. Those sources are informative, but they are not a substitute for a mature peer-reviewed literature. They also reflect the priorities, products, and deployment settings of the organizations that publish them. Academic work should test how well those lessons transfer to other models, budgets, institutional settings, and languages.

A fourth limitation concerns the descriptive audit. Table 3 is a visibility check over this paper's in-scope evidence base, not a field-wide bibliometric estimate. The underlying tags are interpretive and non-exclusive. We use the audit to show an asymmetry in what different source types tend to foreground, not to claim a definitive prevalence distribution for the broader literature.

A fifth limitation concerns causal attribution. The paper argues that many reported agent improvements are harness-sensitive, but it does not itself provide controlled ablations that separate model effects from harness effects. In real systems the two are often entangled: a stronger model may need less scaffolding, while a stronger harness can make a weaker model look far more capable. A fuller empirical program still needs benchmark designs, reporting norms, and intervention studies that estimate how much each layer contributes.

A sixth limitation is scope. The argument is strongest for language agents that operate through tools, files, browsers, and persistent execution. Some language-model applications remain closer to structured prediction or short-form assistance, where the harness is thinner and the term may feel oversized. Likewise, most current public examples come from English-centric, high-resource environments. The research agenda should therefore be tested against multilingual, low-resource, and community-run settings rather than assumed to transfer unchanged.

Finally, HARNESSCARD is proposed here as a reporting artifact, not as a validated community standard. We give a concrete example, a cross-task coverage check, and a rationale, but we do not yet present author studies, reviewer trials, or benchmark evidence showing which fields are sufficient in practice, which are too burdensome, or how disclosure interacts with proprietary constraints. The proposal is intended to be revised in response to community use.

 doi:10.20944/preprints202603.1756.v2

Ethical Considerations

A harness-centered research agenda has clear benefits, but it also carries meaningful risks. Better harnesses can make agents more reliable, and reliability is dual-use. The same control, recovery, and tool-integration improvements that help with benign coding, browsing, writing, or enterprise workflows can also make high-risk automation easier to scale in domains where mistakes or misuse have serious consequences. For that reason, governance should be treated as internal to harness engineering rather than as an optional wrapper added after capability work is complete.

There are also risks of opacity and misrepresentation. Agent papers can overstate autonomy if manual escalation, privileged tool access, hidden review, or extensive hand-tuned control logic are omitted from the description. A harness-sensitive reporting norm is partly an ethical response to that problem: readers should be able to tell where judgment came from, how the system was constrained, which interventions were necessary to obtain the reported result, and which parts of the workflow remained human-dependent.

Privacy and security risks intensify at the harness layer because the harness is where traces, permissions, and connectors live. Runtime logs, memory files, approval records, tool transcripts, and retrieved artifacts may contain sensitive user, organizational, or proprietary information. Tool access expands the attack surface; poorly governed shells, browsers, databases, or retrieval backends can expose systems to leakage, unsafe actions, or policy bypass. The more capable the harness becomes, the more important it is to treat permissions, sandboxing, auditing, monitoring, and trace retention as first-class design decisions [28,50,51,56,60–62,67]. Trajectory-level safety evaluation and large-trace auditing also create their own retention and access-control burdens, because the very traces that make failures legible may themselves be sensitive artifacts.

A further concern is over-automation. A well-engineered harness can encourage organizations to delegate tasks whose failure modes are still poorly understood because the system feels disciplined and auditable. This can create a false sense of safety. The presence of retries, tests, checkpoints, approval prompts, or monitoring dashboards does not guarantee that the right thing is being optimized or that downstream stakeholders can contest a harmful action in time.

There is also a risk of concentration and uneven access. Durable project instructions, privileged observability stacks, and reusable runtime infrastructure may concentrate advantage in a small number of organizations that can afford intensive harness engineering. That can distort scientific comparison if public papers benchmark base models while keeping the decisive systems layer private. It can also produce an uneven research landscape in which access to the harness, not only access to the model, determines who can participate meaningfully in frontier agent work.

Finally, stronger harnesses can hide human cleanup work just as easily as they can amplify machine capability, and they can substantially increase compute usage through retries, persistent traces, and background execution. Honest reporting should therefore surface where humans remained in the loop and what budgets, retries, monitoring regimes, and persistent runtime costs were required to obtain the reported result.

Appendix roadmap. The appendix is organized into four grouped sections to reduce float fragmentation and empty space while keeping the same supplementary material: evidence base, public formulations, and exhaustive inventory (Appendix A); HARNESSCARD materials (Appendix B); search strings, glossary, and additional mini-cases (Appendix C); and expanded visual summaries (Appendix D).

Appendix A. Evidence Base, Public Formulations, and Exhaustive Inventory

Appendix A.1. Evidence Base and Selection Logic

This paper is not a survey of all agent research. It is a selective argument grounded in a structured cited evidence base assembled to support the harness claim. We searched ACL Anthology, arXiv, OpenReview, official engineering notes, standards documents, and public technical articles using query stems such as `harness`, `agent harness`, `natural-language harness`, `meta-harness`, `evaluation harness`, `scaffold`, `externalization`, `context engineering`, `tool use`, `agent-computer interface`,

 doi:10.20944/preprints202603.1756.v2

`interactive evals`, `trajectory safety`, `benchmark integrity`, `MCP servers`, `agent optimization`, `browser agents`, `terminal benchmarks`, `agent reliability`, `software engineering agents`, and `long-running agents`. We retained work if it materially illuminated at least one harness-relevant function: control/specification, runtime/state, mediated agency and interfaces, feedback or verification, governance, or observability/evaluation.

The cited evidence base contains 75 unique sources. We divide them into 63 in-scope harness-relevant works and 12 adjacent framing pieces used for historical context, boundary-setting, neighboring claims, or contemporaneous survey positioning. Of the 63 in-scope items, 38 are papers or benchmarks and 25 are official engineering notes, protocol documents, developer guides, or technical articles. The exhaustive list of the 63 in-scope items appears in Appendix A.3. Because facet labels such as control, agency, runtime, governance, or observability are partly interpretive and non-exclusive, we report them per work in the inventory rather than as a single aggregate facet count table. Table 3 reuses those same inventory tags to provide the main paper's lightweight visibility audit; it should be read as a structured view of this paper's evidence base rather than as a field-wide bibliometric estimate. Figure A1 visualizes the composition of the evidence base, and Table A1 gives the exact counts.

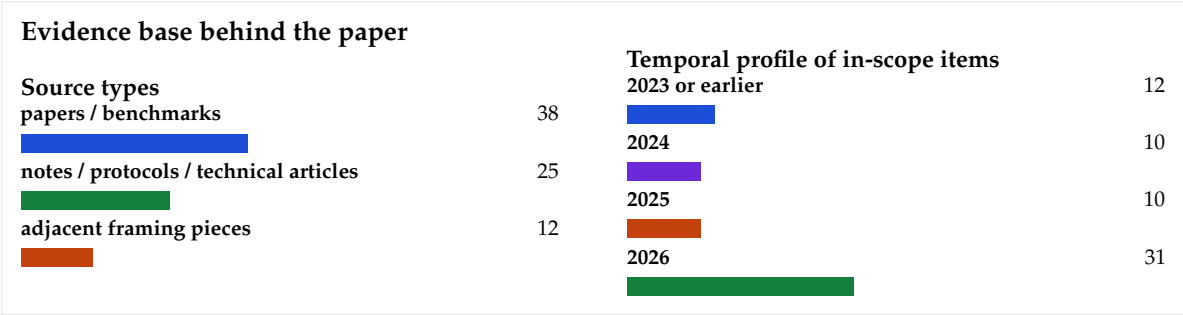

Figure A1. Exact counts for the cited evidence base. Source-type counts use all cited sources; the temporal profile uses only the 63 in-scope harness-relevant items.

Table A1. Exact profile of the cited evidence base.

View	Breakdown	Count
Total cited sources	unique cited sources	75
Scope split	in-scope harness-relevant works	63
Scope split	adjacent framing pieces	12
In-scope source type	papers or benchmarks	38
In-scope source type	engineering notes, protocol documents, developer guides, or technical articles	25
In-scope time band	2023 or earlier	12/63
In-scope time band	2024	10/63
In-scope time band	2025	10/63
In-scope time band	2026	31/63
In-scope time band	2024–2026 combined	51/63

Appendix A.2. Public Formulations and Official Examples

Table A2 expands the compact main-paper inventory of public formulations that anchor the working definition.

 doi:10.20944/preprints202603.1756.v2

Table A2. Public formulations and official examples that shape the paper's working definition.

Source	Year	Object or term	What it contributes to the concept
Anthropic evals note [4]	2026	agent harness / evaluation harness	defines the harness as the system that enables a model to act as an agent and stresses that agent evaluation measures harness plus model together
OpenAI harness note [5]	2026	harness engineering	names the broader practice and ties it to durable instructions, architectural constraints, cleanup loops, and long-horizon coding work
Anthropic application-harness note [9]	2026	long-running application harness	shows generator–evaluator structure, artifact handoffs, context resets, and task-list discipline as explicit harness levers
Anthropic Managed Agents note [10]	2026	managed agents / meta-harness	argues for stable interfaces that outlast any particular harness, session, or sandbox implementation
OpenAI Agents SDK update [11]	2026	model-native harness / SDK	packages sandbox execution, file inspection, approvals, and resume support as reusable harness primitives
OpenAI Codex harness note [26]	2026	Codex harness	treats the harness as reusable runtime logic and protocol support rather than a one-shot prompt surface
Anthropic long-running note [12]	2025	long-running harnesses	makes state externalization, progress tracking, clean-state discipline, and resumability central design levers
OpenAI developer guidance [6–8]	2026	`AGENTS.md`, long-horizon tasks	provides concrete examples of harness work as durable project guidance plus iterative plan–act–test–repair loops
MCP specification [28]	2025	protocol layer	shows that tool interoperability and permission boundaries can themselves be harness design objects

Appendix A.3. Exhaustive in-Scope Works

Tables A3–A6 provide the auditable list of the 63 in-scope harness-relevant works cited in this paper. Table A3 covers the earlier part of the inventory, Table A4 continues with further papers, notes, protocols, and developer materials, Table A5 records additional academic and official literature, and Table A6 covers benchmark-integrity, runtime, and deployment-governance literature.

 doi:10.20944/preprints202603.1756.v2

Table A3. Exhaustive in-scope works cited in the paper, Part I.

Work	Year	Type	Why it is in scope	Primary harness component(s)
Yao et al. [17]	2022	precursor paper	couples reasoning and acting in trajectories	agency, control
Schick et al. [18]	2023	paper	makes tool calls an explicit part of model behavior	interfaces, agency
Shinn et al. [68]	2023	paper	treats reflection as trajectory repair and memory support	runtime, feedback
Wang et al. [54]	2023	paper	shows skill libraries and resumable state for long tasks	runtime, feedback
Li et al. [69]	2023	paper	role-structured multi-agent execution	agency
Li et al. [19]	2023	benchmark	API tool use and tool evaluation	interfaces, observability
Liu et al. [20]	2023	benchmark	interactive agent evaluation	observability
Zhou et al. [21]	2023	benchmark	web interaction as environment-level action	interfaces, observability
Wu et al. [43]	2024	paper	orchestrated multi-agent conversations and control logic	agency, control
Hong et al. [58]	2023	paper	specialist-agent software workflows	agency
Mialon et al. [29]	2023	benchmark	general-assistant tasks that stress tool use and grounded action	observability, agency
Jimenez et al. [22]	2023	benchmark	executable software tasks and hidden tests	feedback, observability
Khattab et al. [70]	2023	paper	LM pipeline compilation and self-improvement	control, feedback
Qian et al. [59]	2024	paper	role-specialized software agents	agency
Wang et al. [23]	2024	paper	code as an executable action substrate	interfaces, agency
Xie et al. [25]	2024	benchmark	open computer-use environment	interfaces, observability
Yang et al. [24]	2024	paper	agent-computer interface for software engineering	interfaces, control
Yao et al. [60]	2024	benchmark	policy-aware tool-agent-user interaction	governance, observability
Wang et al. [32]	2024	paper	open software-agent platform with sandboxed runtime	interfaces, governance, runtime
Xu et al. [71]	2024	benchmark	consequential workplace tasks for agents	governance, observability

Table A4. Exhaustive in-scope works cited in the paper, Part II.

Work	Year	Type	Why it is in scope	Primary harness component(s)
Pan et al. [53]	2024	paper	trainable loops with verifiers and repair	feedback, observability
Wei et al. [30]	2025	benchmark	persistent browsing and recovery	interfaces, runtime
Schluntz and Zhang [72]	2024	note	codifies practical agent patterns and the workflow/agent distinction	control, feedback
Anthropic [73]	2025	note	explicit reflection insertion in tool loops	feedback
Hadfield et al. [13]	2025	note	orchestrator–worker research system	agency
Rajasekaran et al. [3]	2025	note	state curation and compaction as engineering	runtime
Aizawa et al. [66]	2025	note	interface and tool-surface design for agents	interfaces, governance
Dworken et al. [67]	2025	note	security and containment for coding agents	governance
Wu et al. [74]	2025	note	large tool surfaces and long-running tool access	interfaces, runtime
Young [12]	2025	note	state externalization, resumability, and recovery	runtime
Grace et al. [4]	2026	note	agent-harness and evaluation-harness distinction	observability, feedback
Segato [55]	2026	note	infrastructure variance in agentic evaluation	observability
Model Context Protocol [28]	2025	protocol	tool interoperability and permissions as protocol objects	interfaces, governance
Lopopolo [5]	2026	note	explicit naming of harness engineering and its practical levers	control, runtime, governance, observability
Chen [26]	2026	note	reusable runtime logic and protocol support	runtime, interfaces
Bolin [27]	2026	note	stepwise action loop and runtime structure	agency, runtime
OpenAI [6]	2026	developer guide	durable in-repository instruction surfaces	control
OpenAI [7]	2026	developer guide	operational guidance for agentic coding loops	control, runtime
Choi [8]	2026	developer blog	long-horizon execution discipline and resumability	runtime, control
Böckeler [75]	2026	technical article	software-architecture articulation of the harness concept	control, governance, observability

 doi:10.20944/preprints202603.1756.v2

Table A5. Exhaustive in-scope works cited in the paper, Part III.

Work	Year	Type	Why it is in scope	Primary harness component(s)
Pan et al. [14]	2026	paper	externalizes harness behavior as portable natural-language control plus shared runtime	control, runtime
Lee et al. [15]	2026	paper	treats harness code itself as an optimization target and searches over harness designs	runtime, feedback
Bandel et al. [34]	2026	paper	frames general-agent evaluation as a unified protocol and infrastructure problem	interfaces, observability
Bandi et al. [35]	2026	benchmark	evaluates real MCP-server tool use with a containerized harness and rich diagnostics	interfaces, observability
Ursekar et al. [36]	2026	paper	introduces a reproducible evaluation harness for agent optimization	feedback, observability
Rafique and Bindschaedler [44]	2026	paper	makes memory durability and prompt-state residency explicit harness responsibilities	runtime
Jha et al. [46]	2026	benchmark	brings production-derived coding tasks and verification signals into agent evaluation	feedback, observability
Pradel et al. [47]	2026	paper	studies agent architectures on tool-and-project setup plus validation-heavy analysis tasks	interfaces, feedback, observability
Li et al. [50]	2026	benchmark	provides trajectory-level safety evaluation with long-horizon delayed triggers and diagnosis	governance, observability
Wang et al. [49]	2026	benchmark	diagnoses cross-domain long-horizon failures with trajectory-grounded attribution	runtime, observability
Stein et al. [51]	2026	paper	audits large trace collections for sparse safety violations and benchmark gaming	governance, observability
Rajasekaran [9]	2026	note	shows long-running application harness design with generator–evaluator structure and artifact handoffs	control, runtime
Martin et al. [10]	2026	note	introduces a meta-harness that decouples harness logic from sessions and sandboxes	interfaces, runtime, governance

Table A6. Exhaustive in-scope works cited in the paper, Part IV.

Work	Year	Type	Why it is in scope	Primary harness component(s)
OpenAI [11]	2026	product note	packages sandbox execution, files, approvals, and resume bookkeeping as reusable harness primitives	interfaces, runtime, control
Anthropic [61]	2026	note	studies classifier-mediated auto-approval as a harness governance mechanism	governance, control
Bui [33]	2026	technical article	explicitly distinguishes scaffolding, harness, and context engineering in terminal coding agents	control, runtime, interfaces
Merrill et al. [31]	2026	benchmark	hard terminal benchmark with published evaluation harness and multi-run agent analysis	interfaces, runtime, observability
Kapoor et al. [37]	2025	paper	standardizes large-scale agent evaluation harnesses and releases logs for cross-scaffold analysis	observability, runtime, feedback
Waters et al. [48]	2026	benchmark	adapts an agent harness to expert-written STEM workflows with mixed rubric and exact-match grading	feedback, observability, interfaces
Anthropic [56]	2026	note	documents benchmark-aware agent behavior in a web-enabled evaluation setting	observability, governance
OpenAI [62]	2026	note	makes monitoring infrastructure and trace review part of the deployment harness	governance, observability
Rabanser et al. [52]	2026	paper	decomposes agent performance into consistency, robustness, predictability, and safety metrics	observability, feedback
Ndzomga [57]	2026	paper	studies scaffold-driven distribution shift in cost-aware agent benchmarking	observability, runtime

Appendix B. HarnessCard Materials

Appendix B.1. Expanded HarnessCard

The template below expands the compact main-paper version of HarnessCard into a fuller disclosure schema. Table A7 separates governance and observability and adds recommended release and risk fields.

 doi:10.20944/preprints202603.1756.v2

Table A7. HARNESSCARD: a lightweight reporting artifact for language-agent systems.

Field	What should be disclosed	Priority
Base model(s)	model name, version, decoding settings, and any finetuning or adapters	Required
Control artifacts	system instructions, `AGENTS.md`, repo maps, architecture rules, schemas, tests, linters, done-when criteria	Required
Runtime policy	memory type, compaction or summarization policy, checkpointing, retry or rollback policy, budget limits	Required
Action substrate	tools, APIs, browser or GUI access, code execution, interface schemas, MCP usage	Required
Execution topology	single-agent vs multi-agent structure, planner/verifier roles, reviewer loops, routing logic	Required
Feedback stack	tests, graders, reflection prompts, hidden checks, human interventions, or repair loops	Required
Governance layer	permissions, sandboxing, escalation rules, policy checks, provenance logging, audit support	Required
Observability	stored traces, replay support, latency and cost logging, failure categories	Required
Evaluation protocol	task set, number of runs, success criteria, variance treatment, held-out checks or budget limits	Required
Release artifacts	prompts or programs, tool specs, traces, configs, environment setup, reproducibility notes	Recommended
Known limitations and risks	unresolved failure modes, portability caveats, safety concerns, or red-team findings	Recommended

Appendix B.2. Illustrative HARNESSCARD*: Repository Coding Agent*

The filled example below shows how the fields in HARNESSCARD can be instantiated for a repository coding agent. Table A8 is illustrative rather than product-specific and is meant to show the expected granularity of disclosure.

 doi:10.20944/preprints202603.1756.v2

Table A8. A filled illustrative HARNESSCARD for a repository coding agent. The example is a synthesis of recurring patterns in public materials rather than a reverse-engineered product specification.

Field	Illustrative disclosure	Why it matters
Base model(s)	frontier coding model configured through repo or user profiles; effort tuned for long tasks	keeps model choice distinct from harness choice
Control artifacts	root-level `AGENTS.md`; repository map; build/test/lint commands; architecture rules; done-when criteria	reveals the durable instructions and constraints the agent actually reads
Runtime policy	repository treated as system of record; thread history; progress file; compaction near context limits; bounded retries	makes long-horizon state handling explicit
Action substrate	file edits, shell commands, test runs, diff generation, PR review, optional MCP tools	discloses what the model can actually do in the environment
Execution topology	plan → edit → run tools → observe → repair → update status → repeat; optional reviewer loop	captures the control structure rather than only the model
Feedback stack	failing tests, custom linter messages, self-review, grader checks, occasional human review	surfaces the verification signals that shape behavior
Governance layer	sandbox mode, approval policy for privileged actions, least-privilege connectors, audit trail	keeps permissions and safety visible rather than implicit
Observability	persisted thread events, replay support, latency and cost logs, categorized failures	makes debugging and comparison scientifically possible
Success criteria	merged change passes required checks, stays within budget, and leaves updated status artifacts	completion becomes operationally verifiable, not merely verbal
Known risks	stale docs, state drift, verifier overfitting, hidden human intervention, over-trusting automated review	shows why limitation disclosure belongs inside the reporting standard

Appendix C. Search Strings, Glossary, and Additional Mini-Cases

The evidence base was assembled using query stems such as `harness`, `agent harness`, `natural-language harness`, `meta-harness`, `evaluation harness`, `scaffold`, `externalization`, `context engineering`, `prompt engineering`, `tool use`, `interactive evals`, `trajectory safety`, `benchmark integrity`, `MCP servers`, `agent optimization`, `browser agents`, `terminal benchmarks`, `agent reliability`, `agent-computer interface`, `long-running agents`, and `software engineering agents`. We excluded work that used the word *agent* only in a loose product-marketing sense without giving enough detail about control artifacts, runtime structure, action interfaces, evaluation logic, or governance. We also excluded papers whose contribution was primarily model training or alignment unless the paper materially illuminated the extra-model layer around agent execution. The 12 adjacent framing pieces used outside the in-scope inventory are historical, boundary-setting, survey-positioning, or runtime-framing citations rather than direct evidence for the harness taxonomy. Table A9 provides a compact glossary for neighboring terms, and Table A10 extends the cross-task illustration with additional mini-cases.

 doi:10.20944/preprints202603.1756.v2

Table A9. A compact glossary for neighboring terms that are often conflated in agent papers.

Term	Working meaning in this paper
Prompt engineering	writing and organizing instructions, examples, and role structure for desired model behavior
Context engineering	curating the evolving token state supplied to the model, including retrieval, memory, and tool context [3]
Agent harness / scaffold	the extra-model system that enables a model to act as an agent [4]
Harness layer	the extra-model layer that, in this paper's working definition, couples control artifacts, mediated action interfaces, and runtime policies into governed execution
Harness engineering	the design and maintenance of the control, agency, and runtime layer around the model [5]
Meta-harness	a more stable interface layer around evolving task-specific harnesses, sessions, and sandboxes [10]
Evaluation harness	the system that turns tasks, metrics, graders, and infrastructure into an executable evaluation regime [4]
Action substrate	the interface through which the agent can act: code, shell, browser, GUI, APIs, or role-structured delegation

Table A10. Mini-cases illustrating how the same control–agency–runtime decomposition appears across task families.

Task family	Control levers	Runtime levers	Agency levers and likely risks
Repository coding agent	repo map, `AGENTS.md`, tests, linters, architectural constraints	compaction, checkpoints, retries, cleanup passes, cost budgets	shell/file edit/PR review; risks include stale docs, verifier overfitting, and hidden human repair
Browser or research agent	source hierarchy, citation rules, task decomposition, grading rubric	search history, scratchpads, branching traces, escalation on uncertainty	browse, fetch, cite, summarize; risks include source drift, unsupported synthesis, and provenance loss
Enterprise support agent	policy text, workflow scripts, escalation rules, approval thresholds	queue state, customer history, retry and timeout policy, audit logs	tool/API access, human handoff, permissions; risks include privacy leakage, over-escalation, and inconsistent policy application
Agent optimizer / evaluation harness	target-agent spec, reference evaluation procedure, budget policy	versioned snapshots, multi-run aggregation, trace capture, replay	code edits plus edit–execute–evaluate loops; risks include grader gaming, reward misspecification, and infrastructure variance

Appendix D. Expanded Timeline and Framework

Figures A2 and A3 restate the main paper's timeline and CAR decomposition in larger visual forms for quick reference. The first expands the historical widening view, and the second enlarges the control–agency–runtime breakdown.

 doi:10.20944/preprints202603.1756.v2

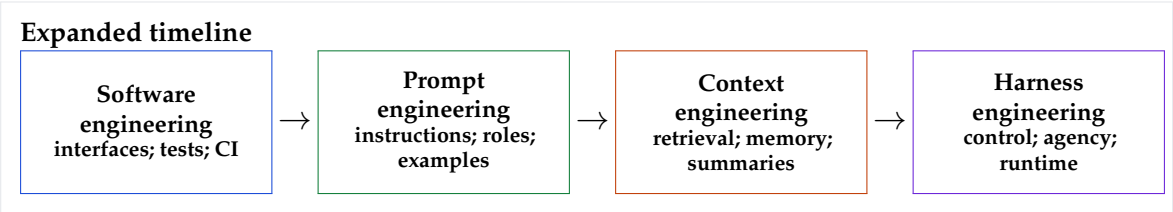

Figure A2. Expanded timeline view. The label *harness engineering* is recent, but the design problem had already appeared in trajectories, tool use, interactive evaluation, action interfaces, context management, and long-running execution.

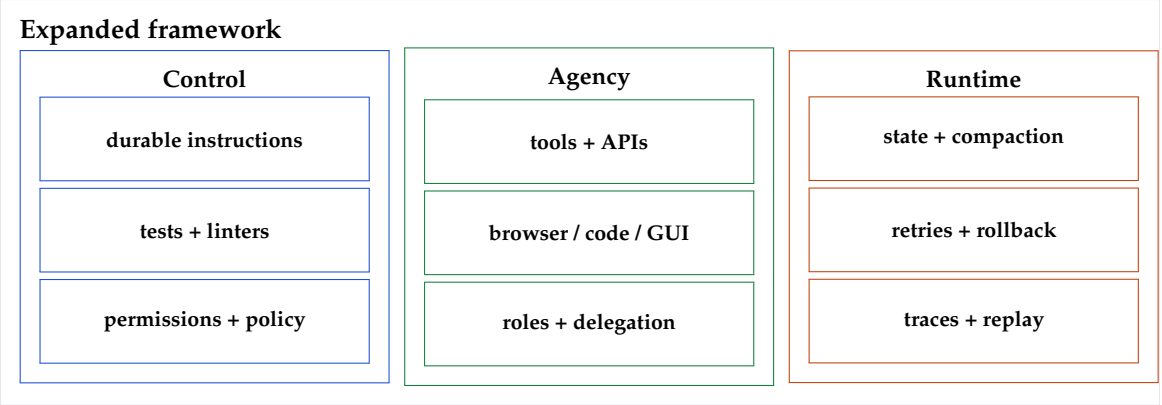

Figure A3. Expanded framework view. The harness layer becomes explicit when the extra-model layer is decomposed into control, agency, and runtime.

References

1. Naur, P.; Randell, B. *Software Engineering: Report of a conference sponsored by the NATO Science Committee, Garmisch, Germany, 7-11 Oct. 1968, Brussels, Scientific Affairs Division, NATO*; 1969.
2. Liu, X.; Wang, J.; Yuan, X.; Sun, J.; Dong, G.; Di, P.; Wang, W.; Wang, D. Prompting frameworks for large language models: A survey. *ACM Computing Surveys* **2023**.
3. Rajasekaran, P.; Dixon, E.; Ryan, C.; Hadfield, J.; Ayub, R.; Moran, H.; Rueb, C.; Jennings, C. Effective context engineering for AI agents, 2025. Anthropic engineering note.
4. Grace, M.; Hadfield, J.; Olivares, R.; Jonghe, J.D. Demystifying evals for AI agents, 2026. Anthropic engineering note.
5. Lopopolo, R. Harness engineering: leveraging Codex in an agent-first world, 2026. OpenAI engineering note.
6. OpenAI. Custom instructions with AGENTS.md, 2026. OpenAI Codex developer guide.
7. OpenAI. Best practices, 2026. OpenAI Codex developer guide.
8. Choi, D. Run long horizon tasks with Codex, 2026. OpenAI developer blog.
9. Rajasekaran, P. Harness design for long-running application development, 2026. Anthropic engineering note.
10. Martin, L.; Cemaj, G.; Cohen, M. Scaling Managed Agents: Decoupling the brain from the hands, 2026. Anthropic engineering note.
11. OpenAI. The next evolution of the Agents SDK, 2026. OpenAI product note.
12. Young, J. Effective harnesses for long-running agents, 2025. Anthropic engineering note.
13. Hadfield, J.; Zhang, B.; Lien, K.; Scholz, F.; Fox, J.; Ford, D. How we built our multi-agent research system, 2025. Anthropic engineering note.
14. Pan, L.; Zou, L.; Guo, S.; Ni, J.; Zheng, H.T. Natural-Language Agent Harnesses. *arXiv preprint arXiv:2603.25723* **2026**.
15. Lee, Y.; Nair, R.; Zhang, Q.; Lee, K.; Khattab, O.; Finn, C. Meta-Harness: End-to-End Optimization of Model Harnesses. *arXiv preprint arXiv:2603.28052* **2026**.
16. Meng, Q.; Wang, Y.; Chen, L.; Wang, Q.; Lu, C.; Wu, W.; Gao, Y.; Wu, Y.; Hu, Y. Agent Harness for Large Language Model Agents: A Survey. *Preprints* **2026**. https://doi.org/10.20944/preprints202604.0428.v2.

 doi:10.20944/preprints202603.1756.v2

17. Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.R.; Cao, Y. ReAct: Synergizing reasoning and acting in language models. In Proceedings of the The eleventh international conference on learning representations, 2022.
18. Schick, T.; Dwivedi-Yu, J.; Dessì, R.; Raileanu, R.; Lomeli, M.; Hambro, E.; Zettlemoyer, L.; Cancedda, N.; Scialom, T. Toolformer: Language models can teach themselves to use tools. *Advances in neural information processing systems* **2023**, *36*, 68539–68551.
19. Li, M.; Zhao, Y.; Yu, B.; Song, F.; Li, H.; Yu, H.; Li, Z.; Huang, F.; Li, Y. API-Bank: A comprehensive benchmark for tool-augmented LLMs. In Proceedings of the Proceedings of the 2023 conference on empirical methods in natural language processing, 2023, pp. 3102–3116.
20. Liu, X.; Yu, H.; Zhang, H.; Xu, Y.; Lei, X.; Lai, H.; Gu, Y.; Ding, H.; Men, K.; Yang, K.; et al. AgentBench: Evaluating LLMs as agents. *arXiv preprint arXiv:2308.03688* **2023**.
21. Zhou, S.; Xu, F.F.; Zhu, H.; Zhou, X.; Lo, R.; Sridhar, A.; Cheng, X.; Ou, T.; Bisk, Y.; Fried, D.; et al. WebArena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854* **2023**.
22. Jimenez, C.E.; Yang, J.; Wettig, A.; Yao, S.; Pei, K.; Press, O.; Narasimhan, K. SWE-bench: Can language models resolve real-world GitHub issues? *arXiv preprint arXiv:2310.06770* **2023**.
23. Wang, X.; Chen, Y.; Yuan, L.; Zhang, Y.; Li, Y.; Peng, H.; Ji, H. Executable code actions elicit better LLM agents. In Proceedings of the Forty-first International Conference on Machine Learning, 2024.
24. Yang, J.; Jimenez, C.E.; Wettig, A.; Lieret, K.; Yao, S.; Narasimhan, K.; Press, O. SWE-agent: Agent-computer interfaces enable automated software engineering. *Advances in Neural Information Processing Systems* **2024**, *37*, 50528–50652.
25. Xie, T.; Zhang, D.; Chen, J.; Li, X.; Zhao, S.; Cao, R.; Hua, T.J.; Cheng, Z.; Shin, D.; Lei, F.; et al. OSWorld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *Advances in Neural Information Processing Systems* **2024**, *37*, 52040–52094.
26. Chen, C. Unlocking the Codex harness: how we built the App Server, 2026. OpenAI engineering note.
27. Bolin, M. Unrolling the Codex agent loop, 2026. OpenAI engineering note.
28. Model Context Protocol. Model Context Protocol specification, 2025. Model Context Protocol specification, version 2025-11-25.
29. Mialon, G.; Fourrier, C.; Wolf, T.; LeCun, Y.; Scialom, T. GAIA: a benchmark for general AI assistants. In Proceedings of the The Twelfth International Conference on Learning Representations, 2023.
30. Wei, J.; Sun, Z.; Papay, S.; McKinney, S.; Han, J.; Fulford, I.; Chung, H.W.; Passos, A.T.; Fedus, W.; Glaese, A. BrowseComp: A simple yet challenging benchmark for browsing agents. *arXiv preprint arXiv:2504.12516* **2025**.
31. Merrill, M.A.; Shaw, A.G.; Carlini, N.; Li, B.; Raj, H.; Bercovich, I.; Shi, L.; Shin, J.Y.; Walshe, T.; Buchanan, E.K.; et al. Terminal-Bench: Benchmarking agents on hard, realistic tasks in command line interfaces. *arXiv preprint arXiv:2601.11868* **2026**.
32. Wang, X.; Li, B.; Song, Y.; Xu, F.F.; Tang, X.; Zhuge, M.; Pan, J.; Song, Y.; Li, B.; Singh, J.; et al. OpenHands: An open platform for AI software developers as generalist agents. *arXiv preprint arXiv:2407.16741* **2024**.
33. Bui, N.D. Building AI Coding Agents for the Terminal: Scaffolding, Harness, Context Engineering, and Lessons Learned. *arXiv preprint arXiv:2603.05344* **2026**.
34. Bandel, E.; Yehudai, A.; Eden, L.; Sagron, Y.; Perlitz, Y.; Venezian, E.; Razinkov, N.; Ergas, N.; Shachor Ifergan, S.; Shlomov, S.; et al. General Agent Evaluation. *arXiv preprint arXiv:2602.22953* **2026**.
35. Bandi, C.; Hertzberg, B.; Boo, G.; Polakam, T.; Da, J.; Hassaan, S.; Sharma, M.; Park, A.; Hernandez, E.; Rambado, D.; et al. MCP-Atlas: A Large-Scale Benchmark for Tool-Use Competency with Real MCP Servers. *arXiv preprint arXiv:2602.00933* **2026**.
36. Ursekar, V.; Shanker, A.; Chatrath, V.; Xue, Y.E.; Denton, S. VeRO: An Evaluation Harness for Agents to Optimize Agents. *arXiv preprint arXiv:2602.22480* **2026**.
37. Kapoor, S.; Stroebl, B.; Kirgis, P.; Nadgir, N.; Siegel, Z.S.; Wei, B.; Xue, T.; Chen, Z.; Chen, F.; Utpala, S.; et al. Holistic agent leaderboard: The missing infrastructure for AI agent evaluation. *arXiv preprint arXiv:2510.11977* **2025**.
38. Wang, L.; Ma, C.; Feng, X.; Zhang, Z.; Yang, H.; Zhang, J.; Chen, Z.; Tang, J.; Chen, X.; Lin, Y.; et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science* **2024**, *18*, 186345.
39. Piccialli, F.; Chiaro, D.; Sarwar, S.; Cerciello, D.; Qi, P.; Mele, V. AgentAI: A comprehensive survey on autonomous agents in distributed AI for industry 4.0. *Expert Systems with Applications* **2025**, *291*, 128404.

 doi:10.20944/preprints202603.1756.v2

40. Luo, J.; Zhang, W.; Yuan, Y.; Zhao, Y.; Yang, J.; Gu, Y.; Wu, B.; Chen, B.; Qiao, Z.; Long, Q.; et al. Large language model agent: A survey on methodology, applications and challenges. *arXiv preprint arXiv:2503.21460* **2025**.
41. Yehudai, A.; Eden, L.; Li, A.; Uziel, G.; Zhao, Y.; Bar-Haim, R.; Cohan, A.; Shmueli-Scheuer, M. Survey on evaluation of LLM-based agents. *arXiv preprint arXiv:2503.16416* **2025**.
42. Zhou, C.; Chai, H.; Chen, W.; Guo, Z.; Shan, R.; Song, Y.; Xu, T.; Yang, Y.; Yu, A.; Zhang, W.; et al. Externalization in LLM Agents: A Unified Review of Memory, Skills, Protocols and Harness Engineering. *arXiv preprint arXiv:2604.08224* **2026**.
43. Wu, Q.; Bansal, G.; Zhang, J.; Wu, Y.; Li, B.; Zhu, E.; Jiang, L.; Zhang, X.; Zhang, S.; Liu, J.; et al. AutoGen: Enabling next-gen LLM applications via multi-agent conversations. In Proceedings of the First conference on language modeling, 2024.
44. Rafique, M.; Bindschaedler, L. ClawVM: Harness-Managed Virtual Memory for Stateful Tool-Using LLM Agents. In Proceedings of the Proceedings of the 6th European Workshop on Machine Learning and Systems (EuroMLSys '26), 2026. Also available as arXiv:2604.10352.
45. Kontonis, V.; Zeng, Y.; Garg, S.; Chen, L.; Tang, H.; Wang, Z.; Awadallah, A.; Horvitz, E.; Langford, J.; Papailiopoulos, D. MEMENTO: Teaching LLMs to Manage Their Own Context. *arXiv preprint arXiv:2604.09852* **2026**.
46. Jha, S.; Paltenghi, M.; Maddila, C.; Murali, V.; Ugare, S.; Chandra, S. ProdCodeBench: A Production-Derived Benchmark for Evaluating AI Coding Agents. *arXiv preprint arXiv:2604.01527* **2026**.
47. Pradel, M.; Cadar, C.; Bouzenia, I. Evaluating LLM Agents on Automated Software Analysis Tasks. *arXiv preprint arXiv:2604.11270* **2026**.
48. Waters, K.; Nuzzi, L.; Looram, T.; Tomasiello, A.; Kamdoum, A.G.K.; Li, B.; Sileo, D.; Kretov, E.; Fournier-Facio, F.; Soloupis, G.; et al. COMPOSITE-STEM. *arXiv preprint arXiv:2604.09836* **2026**.
49. Wang, X.J.; Bai, H.; Sun, Y.; Wang, H.; Zhang, S.; Hu, W.; Schroder, M.; Mutlu, B.; Song, D.; Nowak, R.D. The Long-Horizon Task Mirage? Diagnosing Where and Why Agentic Systems Break. *arXiv preprint arXiv:2604.11978* **2026**.
50. Li, Y.; Luo, H.; Xie, Y.; Fu, Y.; Yang, Z.; Shao, S.; Ren, Q.; Qu, W.; Fu, Y.; Yang, Y.; et al. ATBench: A Diverse and Realistic Trajectory Benchmark for Long-Horizon Agent Safety. *arXiv preprint arXiv:2604.02022* **2026**.
51. Stein, A.; Brown, D.; Hassani, H.; Naik, M.; Wong, E. Detecting Safety Violations Across Many Agent Traces. *arXiv preprint arXiv:2604.11806* **2026**.
52. Rabanser, S.; Kapoor, S.; Kirgis, P.; Liu, K.; Utpala, S.; Narayanan, A. Towards a science of AI agent reliability. *arXiv preprint arXiv:2602.16666* **2026**.
53. Pan, J.; Wang, X.; Neubig, G.; Jaitly, N.; Ji, H.; Suhr, A.; Zhang, Y. Training software engineering agents and verifiers with SWE-Gym. *arXiv preprint arXiv:2412.21139* **2024**.
54. Wang, G.; Xie, Y.; Jiang, Y.; Mandlekar, A.; Xiao, C.; Zhu, Y.; Fan, L.; Anandkumar, A. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291* **2023**.
55. Segato, G. Quantifying infrastructure noise in agentic coding evals, 2026. Anthropic engineering note.
56. Anthropic. Eval awareness in Claude Opus 4.6's BrowseComp performance, 2026. Anthropic engineering note.
57. Ndzomga, F. Efficient Benchmarking of AI Agents. *arXiv preprint arXiv:2603.23749* **2026**.
58. Hong, S.; Zhuge, M.; Chen, J.; Zheng, X.; Cheng, Y.; Wang, J.; Zhang, C.; Wang, Z.; Yau, S.K.S.; Lin, Z.; et al. MetaGPT: Meta programming for a multi-agent collaborative framework. In Proceedings of the The twelfth international conference on learning representations, 2023.
59. Qian, C.; Liu, W.; Liu, H.; Chen, N.; Dang, Y.; Li, J.; Yang, C.; Chen, W.; Su, Y.; Cong, X.; et al. ChatDev: Communicative agents for software development. In Proceedings of the Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers), 2024, pp. 15174–15186.
60. Yao, S.; Shinn, N.; Razavi, P.; Narasimhan, K. τ-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains. *arXiv preprint arXiv:2406.12045* **2024**.
61. Anthropic. Claude Code auto mode: a safer way to skip permissions, 2026. Anthropic engineering note.
62. OpenAI. How we monitor internal coding agents for misalignment, 2026. OpenAI safety note.
63. He, C.; Zhou, X.; Wang, D.; Xu, H.; Liu, W.; Miao, C. The AutoResearch Moment: From Experimenter to Research Director **2026**.
64. He, C.; Zhou, X.; Wang, D.; Xu, H.; Liu, W.; Miao, C. OpenClaw as Language Infrastructure: A Case-Centered Survey of a Public Agent Ecosystem in the Wild **2026**.

 doi:10.20944/preprints202603.1756.v2

65. He, C.; Zhou, X.; Wang, D.; Xu, H.; Liu, W.; Miao, C. Human-AI productivity claims should be reported as time-to-acceptance under explicit acceptance tests **2026**.
66. Aizawa, K.; Zhang, B.; Witten, Z.; Jiang, D.; Al-Sheikh, S.; Bell, M.; Vo, M.; Chu, T.; Welsh, J.; Parra, D.S.; et al. Writing effective tools for agents — with agents, 2025. Anthropic engineering note.
67. Dworken, D.; Weller-Davies, O.; Choi, M.; Wu, C.; Vorwerck, M.; Isken, A.; Bradwell, K.; Garcia, K. Beyond permission prompts: making Claude Code more secure and autonomous, 2025. Anthropic engineering note.
68. Shinn, N.; Cassano, F.; Gopinath, A.; Narasimhan, K.; Yao, S. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems* **2023**, *36*, 8634–8652.
69. Li, G.; Hammoud, H.; Itani, H.; Khizbullin, D.; Ghanem, B. CAMEL: Communicative agents for "mind" exploration of large language model society. *Advances in neural information processing systems* **2023**, *36*, 51991–52008.
70. Khattab, O.; Singhvi, A.; Maheshwari, P.; Zhang, Z.; Santhanam, K.; Vardhamanan, S.; Haq, S.; Sharma, A.; Joshi, T.T.; Moazam, H.; et al. DSPy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714* **2023**.
71. Xu, F.F.; Song, Y.; Li, B.; Tang, Y.; Jain, K.; Bao, M.; Wang, Z.Z.; Zhou, X.; Guo, Z.; Cao, M.; et al. TheAgentCompany: benchmarking LLM agents on consequential real world tasks. *arXiv preprint arXiv:2412.14161* **2024**.
72. Schluntz, E.; Zhang, B. Building effective agents, 2024. Anthropic engineering note.
73. Anthropic. The "think" tool: Enabling Claude to stop and think in complex tool use situations, 2025. Anthropic engineering note.
74. Wu, B.; Jones, A.; Renault, A.; Tay, H.; Noble, J.; Picard, N.; Jiang, S.; et al. Introducing advanced tool use on the Claude Developer Platform, 2025. Anthropic engineering note.
75. Böckeler, B. Harness engineering, 2026. Thoughtworks article.