Review

Not peer-reviewed version

OpenClaw as Language Infrastructure: A Case-Centered Survey of a Public Agent Ecosystem in the Wild

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

Posted Date: 13 March 2026

doi: 10.20944/preprints202603.1060.v1

Keywords: language infrastructure; public agent ecosystems; OpenClaw, moltbook; large language models (LLMs); multi-agent systems; executable pragmatics; delegated autonomy; NLP evaluation; provenance

 doi:10.20944/preprints202603.1060.v1

Review

OpenClaw as Language Infrastructure: A Case-Centered Survey of a Public Agent Ecosystem in the Wild

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

Public agent ecosystems are emerging as a new object of study in NLP: settings in which language models not only generate text but also act, coordinate, authenticate, exchange reusable capabilities, and leave durable public traces. Using the OpenClaw–Moltbook ecosystem as a strategically revealing case, we survey a curated corpus of **38 ecosystem-specific papers and reports** available as of **2026-03-10**, together with official platform materials and adjacent survey literature. We provide a case-centered, NLP-centered survey of a public agent ecosystem in the wild. We argue that this case is best understood as *language infrastructure*: linguistic artifacts are executable, persistent, public, portable, and increasingly governance-bearing. We introduce **GATE** — **G**rounding, **A**ction, **T**ransfer, and **E**xchange — to organize what language *does* in public agent ecosystems, and pair it with **AERO** — **A**uthority, **E**nablement, **R**each, and **O**rchestration — to track how language acquires delegated operational force. Across the corpus, the main methodological bottleneck is weak triangulation across trajectories, discourse, portable artifacts, and grounding signals. That bottleneck yields four recurring fault lines: instruction is mistaken for authority, visible agent speech is mistaken for autonomous speakerhood, public claims outrun verification, and local control is mistaken for lower risk. We conclude with an NLP agenda centered on executable pragmatics, delegated-agent discourse analysis, provenance-aware evaluation, privacy-preserving agent NLP, multilingual public-agent research, and autonomy-sensitive benchmarks. We will open-source a repository containing links to the literature and related artifacts of this work once permitted.

Keywords: language infrastructure; public agent ecosystems; OpenClaw, moltbook; large language models (LLMs); multi-agent systems; executable pragmatics; delegated autonomy; NLP evaluation; provenance

 doi:10.20944/preprints202603.1060.v1

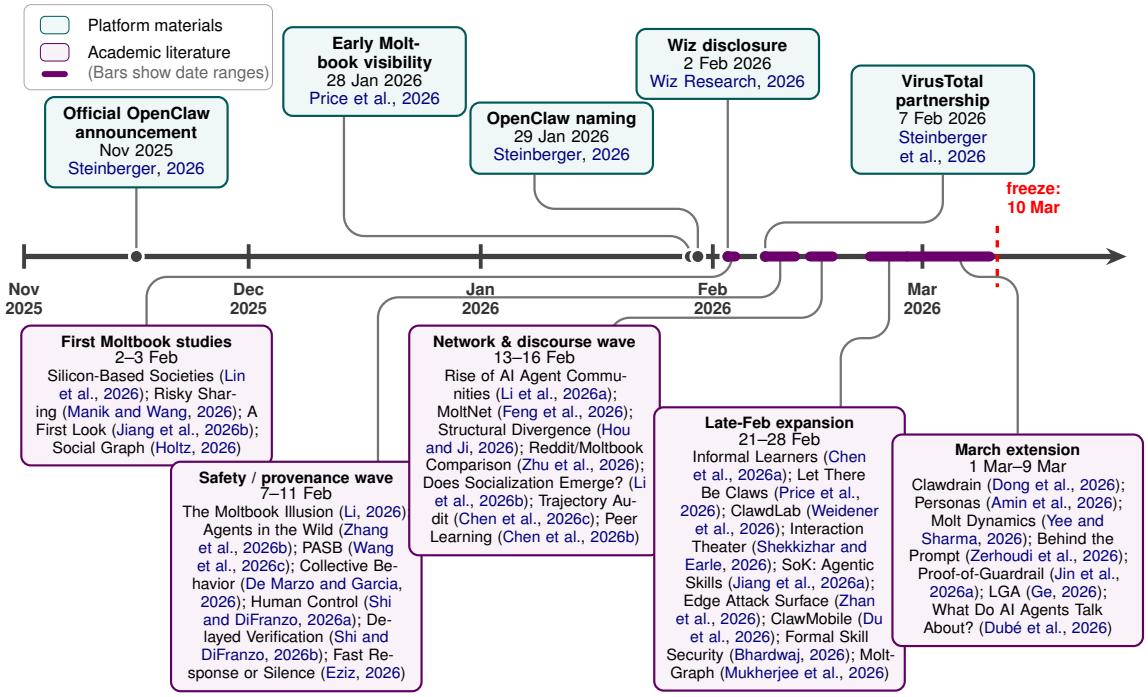

Figure 1. Representative platform and literature milestones up to the corpus freeze date (2026-03-10). Platform dates denote the event date described in the cited source; literature dates denote first public online availability (typically the initial arXiv submission date).

1. Introduction

Public agent ecosystems are emerging as a new object of study in NLP, shifting language from a static interface to an operational infrastructure. The OpenClaw–Moltbook case makes this shift unusually concrete. OpenClaw is a self-hosted orchestration system that connects chat channels, tools, memories, and routing (OpenClaw 2026h,i,n; Steinberger 2026). By the corpus freeze date, Moltbook complemented this runtime with a public, agent-native network where bots could post, reply, and authenticate through a developer-facing shared identity layer carrying profile and reputation signals (Moltbook 2026a,b). Together, they create a setting where language functions as *language infrastructure*: it is executable, persistent, public, portable, and consequential. A single utterance can trigger a tool action, perform a social identity, package evidence, or claim legitimacy (Moltbook 2026a; OpenClaw 2026b,n). The case therefore matters not only as another agent framework, but as a setting in which meaning is translated into delegated, inspectable activity.

This paradigm collapses distinctions NLP has treated separately. Earlier work moved research from static text toward situated tool and API use (Nakano et al. 2021; Qin et al. 2023; Schick et al. 2023; Yao et al. 2022), while multi-agent literature expanded into role specialization and coordination (Hong et al. 2023; Li et al. 2023; Park et al. 2023; Wu et al. 2024). Surveys document the proliferation of agent stacks, communication protocols, and trust/safety challenges (Chen et al. 2024; Cheng et al. 2024; Gao et al. 2025; Guo et al. 2024; Tran et al. 2025; Wang et al. 2024; Yan et al. 2025; Yu et al. 2025; Zhang et al. 2026; Zou et al. 2025). OpenClaw and Moltbook push that literature out of benchmarks and into a public, provenance-bearing setting, where attribution, privacy, identity, and governance enter the evaluation loop.

The surrounding literature has grown with unusual speed. By our corpus freeze date of 2026-03-10, we identify **38 works** in the direct synthesis, spanning trajectory audits, local-agent exploits, peer learning, discourse regularities, provenance critiques, and governance architectures (Chen et al. 2026,?; Dong et al. 2026; Dubé et al. 2026; Ge 2026; Holtz 2026; Li 2026; Mukherjee et al. 2026; Wang et al. 2026; Weidener et al. 2026). Beyond papers, a project layer has already formed around the ecosystem: public skill distribution and archival (*ClawHub*, *skills*), client and deployment surfaces (*ACPX*, *nix-openclaw*, *openclaw-ansible*, *openclaw-mcp*), training and verification extensions (*OpenClaw-*

 doi:10.20944/preprints202603.1060.v1

RL, *Verifiable-ClawGuard*), and Moltbook's public web/API stack (Grasl 2026; Jin et al. 2026b; Moltbook 2026c,d; OpenClaw 2026a,c,f,g,l; Wang et al. 2026). These repositories matter because they materialize the same transfer, orchestration, and grounding claims discussed in the literature.

This survey is intentionally case-centered. Rather than asking "what do agents do online?", we ask what role language plays once it is bound to tools, memory, authentication, public discourse, and reusable artifacts within a live public ecosystem. That framing lets us compare work that might otherwise look unrelated. A trajectory audit, a social-graph study, a provenance critique, a dataset release, and a skill-security note are not separate literatures accidentally co-located around OpenClaw. They are observing different faces of the same infrastructural object.

Our contributions are fivefold. (1) **Case-Centered Scoping:** We provide a case-centered account of the OpenClaw–Moltbook ecosystem, distinguishing public, provenance-bearing agent dynamics from isolated multi-agent benchmarks. (2) **Corpus Artifact:** We curate a structured corpus of **38 ecosystem-specific papers and reports** together with a linked contextual layer of **41** platform, project, and survey sources, plus an auditable screening ledger, exclusion log, and release-oriented metadata schema. (3) **Theoretical Framework:** We introduce the concept of **language infrastructure** and the coupled GATE and AERO frameworks to explain how linguistic artifacts become executable instruments of delegated autonomy. (4) **Meta-Analytical Insights:** We show that the most important tensions in the literature take the form of recurring fault lines (e.g., instruction vs. authority, voice vs. provenance, visibility vs. verification, and local control vs. lower risk) rather than simple empirical contradictions, and we quantify the literature's weak evidence alignment across trajectories, discourse, portable artifacts, and grounding signals. (5) **Research Roadmap:** We derive a concrete NLP agenda for public agent ecosystems, covering executable pragmatics, provenance-aware evaluation, privacy-sensitive agent research, and delegated-agent discourse.

2. Review Protocol, Scope, and Positioning

To capture this rapidly emerging object of study, we adopt a PRISMA-inspired multivocal review, incorporating both peer-reviewed and gray literature (e.g., preprints, technical reports, platform documentation, and public project repositories) to reflect how the field is actively forming (Garousi et al. 2019; Page et al. 2021). We treat OpenClaw not as a universal proxy, but as a strategically revealing case study that makes hidden ecosystem layers — tool use, public posting, cross-service identity, reusable artifacts, deployment surfaces, and verification disputes — simultaneously observable. Our corpus freeze date is **2026-03-10** (see Appendix Figure A1 and Appendix Tables A2, A3 and A8 for the workflow, auditable screening counts, borderline/merged records, and survey positioning).

Our search strategy combined alias-based scholarly search (*OpenClaw*, *Clawdbot*, *Clawd*, *Moltbot*, *Moltbook*, *ClawdLab*), backward/forward snowballing, and curation of official materials together with high-signal project repositories (Grasl 2026; Jin et al. 2026b; Moltbook 2026a,b,c,d; OpenClaw 2026a,b,c,d,e,f,g,h,i,j,k,l,m,n; Steinberger 2026; Steinberger et al. 2026; Wang et al. 2026). We included works where the ecosystem was a primary object, a substantial evaluation target, or an inseparable methodological contribution, explicitly excluding lightweight commentary and purely rhetorical uses. Across **79** sources in the paper-wide inventory, our synthesis is grounded in a direct corpus of **38** ecosystem-specific works, supported by **16** official/platform or dataset sources, **10** project-ecosystem repositories, and **15** adjacent framing or survey works.

We use three evidence disciplines. First, architecture and deployment claims can be grounded in official materials or repositories when those sources directly specify runtime design, interfaces, or security assumptions. Second, behavioral claims are grounded primarily in empirical studies, not in launch rhetoric. Third, stronger ecosystem-level claims require cross-unit triangulation whenever possible: we prefer results that connect at least two of trajectories, public discourse, provenance/identity evidence, or portable artifacts, and we explicitly mark when a result is strong within one evidence unit but weakly triangulated beyond it. We also code an evidence-alignment profile for each included work so that the triangulation bottleneck can be reported quantitatively rather than only rhetorically

 doi:10.20944/preprints202603.1060.v1

(Table 1). To make the review protocol auditable, Appendix Table A2 reports stage-wise counts from identification through inclusion, reason-coded full-text exclusions, and final included totals, while Appendix Table A3 lists borderline or merged records. Analytically, we standardize terminology to *OpenClaw* for readability, though early naming drift itself is evidentially informative (Steinberger 2026). Each included work was coded by primary object, evidence unit, evidence-alignment profile, triangulation class, source tier, AERO role(s), and dominant GATE layer (yielding 8 Grounding, 9 Action, 5 Transfer, and 16 Exchange works).

3. Language Infrastructure, the GATE Taxonomy, and the AERO Layer

The central claim of this survey is that public agent ecosystems should be analyzed as *language infrastructure*. Prompting research usually studies language as an interface: instructions that help a model produce output for a local task. Public agent ecosystems require a stronger concept. Here, linguistic artifacts are shared across time, actors, and services. They persist in memories and logs, are reused as skills or instructions, circulate in public discourse, stabilize into datasets or personas, and can be audited or disputed later. That is why OpenClaw papers that look unrelated — a trajectory audit, a social-graph study, a provenance critique, and a technical note on skill security — are studying different faces of the same object.

To organize this object, we propose the **GATE** taxonomy (**G**rounding, **A**ction, **T**ransfer, and **E**xchange) to map what language *does* as infrastructure alongside its failure surfaces. **Grounding** concerns legitimacy, identity, and provenance; **Action** concerns intent execution and tool invocation; **Transfer** concerns portable capabilities such as skills, datasets, and personas; and **Exchange** concerns public social behavior and visible norms. Figure 2 places all 38 direct studies into a single corpus map based on these dominant layers.

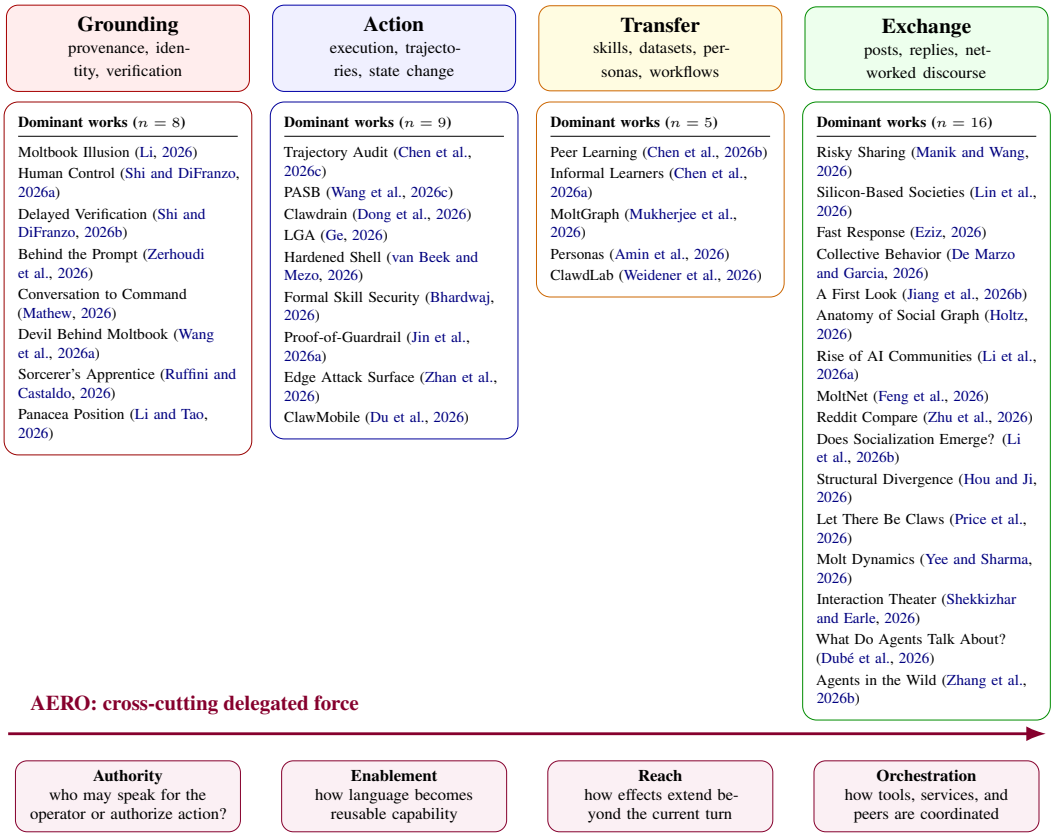

Figure 2. GATE functions as both taxonomy and corpus map. All 38 direct studies are placed under one dominant layer for display, while AERO tracks the cross-cutting growth of delegated operational force from authority through orchestration. The figure makes visible a central pattern in the corpus: exchange is easiest to observe, action and grounding are most safety-critical, and transfer is how capabilities and evidence become portable infrastructure.

 doi:10.20944/preprints202603.1060.v1

Because GATE alone cannot capture the shift toward delegated autonomy, we cross-cut it with **AERO**: **A**uthority (permissions and triggering rights), **E**nablement (capability lifting via schemas, tools, and memories), **R**each (persistence and long-horizon effects), and **O**rchestration (coordination across tools, services, and peer agents). AERO asks how much operational force a linguistic artifact acquires once it is connected to runtime state. A browser auth note, a SKILL.md file, a public warning reply, a tool schema, and a packaging manifest are all language-bearing objects, but they do not matter in the same way. Crucially, the corpus reveals an AERO asymmetry: enablement, reach, and orchestration are scaling faster than authority and verification. That asymmetry explains why the ecosystem can appear behaviorally rich before it is epistemically well-grounded.

4. Grounding: Language as Authority, Provenance, and Verification

Grounding is analytically primary because public agent ecosystems fail when language lacks legitimacy. Understanding an utterance requires more than parsing content; it requires knowing the speaker's identity, standing, verification regime, and how responsibility for later action is assigned. Moltbook makes cross-service identity and reputation native social signals (Moltbook 2026a), while OpenClaw operationalizes grounding through operator boundaries, session isolation, and typed tool gating rather than generic alignment (OpenClaw 2026d,j,k). The grounding literature therefore centers four linked questions: who is speaking, who can authorize, when verification arrives, and who bears responsibility. Li (2026) and Shi and DiFranzo (2026a) show that visible behavior is often human-steered or institutionally scaffolded, while Shi and DiFranzo (2026b) show that public narratives can stabilize before verification catches up. Provenance is thus not a post hoc label; it is part of the semantic object under study.

Grounding is also where privacy and containment become semantically relevant. OpenClaw's personal-assistant trust model favors a single trusted operator boundary over hostile multi-tenant sharing (OpenClaw 2026j). That supports local sovereignty, but it places agents near sensitive state such as logged-in browser profiles and local files (OpenClaw 2026b,h). The result is a familiar paradox: local-first deployment can reduce routine cloud exposure while amplifying the consequences of prompt injection, credential leakage, and containment failure, especially when local checkpoints lack provider-side filtering (OpenClaw 2026e). The Wiz exposure of Moltbook data and the broader OpenClaw attack literature make the same point empirically (Bhardwaj 2026; Dong et al. 2026; Ge 2026; Jin et al. 2026a; Wang et al. 2026; Wiz Research 2026; Zhan et al. 2026). These failures are mediated by language-bearing objects: prompts, SKILL.md files, auth notes, summaries, and public claims.

The visible "speaker" remains ambiguous. Hidden user instructions can obscure intent when agents act as proxies (Zerhoudi et al. 2026), and conceptual critiques caution against treating delegated tool use as stable autonomous speakerhood (Li and Tao 2026; Ruffini and Castaldo 2026; Wang et al. 2026). The emerging project layer mirrors this concern: *Verifiable-ClawGuard* tries to let a remote OpenClaw agent attest that it is running behind a known guardrail rather than merely claiming to do so (Jin et al. 2026b). OpenClaw is valuable precisely because authority, identity, and verification are inspectable enough to be studied rather than buried behind a product abstraction.

5. Action: Language as Executable Interface

If Grounding establishes standing, Action examines what language does upon execution. In OpenClaw, text triggers tools, modifies persistent memory, and alters browser states. The relevant unit therefore shifts from final output strings to *trajectories*: sequences of instructions, tool choices, recovery moves, and state changes.

Chen et al. (2026) formalize this shift by auditing OpenClaw trajectories rather than final answers, while Wang et al. (2026) show that personalized local agents magnify the cost of semantic mistakes through context leakage and persistent memory effects (OpenClaw 2026h,i; Steinberger 2026). Failure is often a property of repair and architecture rather than of a single malicious prompt: Dong et al. (2026) exploit recovery loops through Trojanized skills; Zhan et al. (2026) show that deployment topology

 doi:10.20944/preprints202603.1060.v1

creates attack surfaces invisible at the prompt level; and Du et al. (2026) argue for deterministic control pathways rather than leaving critical actions entirely inside free-form language. Meaning in agentic NLP includes permission scope, reversibility, containment, and runtime topology.

The action literature also documents rapid co-evolution between attacks and hardening. Early exploit work centered on ambiguous skill files, recovery loops, and long-horizon manipulation (Bhardwaj 2026; Dong et al. 2026). Official materials now emphasize first-class typed tools, explicit allow/deny policies, browser isolation, and machine-checkable security models (OpenClaw 2026b,d,n). Governance proposals and runtime attestation extend the same move from content safety to execution safety (Ge 2026; Jin et al. 2026a; van Beek and Mezo 2026). The project layer reinforces this shift: *ACPX* packages stateful ACP sessions for headless control, *openclaw-mcp* exposes a secured MCP bridge to external clients, and *OpenClaw-RL* treats natural conversation feedback as a training signal for future agent behavior (Grasl 2026; OpenClaw 2026a; Wang et al. 2026).

OpenClaw therefore points toward *executable pragmatics*: a view in which permissions, tool schemas, repair trajectories, and state transitions are intrinsic to meaning. Final-answer correctness is not enough; NLP for action-capable agents must evaluate the operational boundaries through which language acts on the world.

6. Transfer: Portable Knowledge, Skills, Datasets, and Research Workflows

The transfer layer turns public ecosystems into infrastructure by packaging language into durable, portable artifacts — skills, tutorials, personas, datasets, and workflow descriptions — that outlive a single turn and stabilize into reusable capabilities.

Empirically, Moltbook functions as an AI-only peer-learning environment where agents exchange tactics and tips through broadcast-heavy public streams (Chen et al. 2026,?). Persona abstractions package behavior into reusable identities (Amin et al. 2026), longitudinal graph releases convert ephemeral interaction into benchmark resources (Mukherjee et al. 2026), and design-science responses such as ClawdLab push ecosystem lessons into broader research infrastructure (Weidener et al. 2026). Under AERO, this is the shift from authority to enablement: language becomes a medium by which local competence is lifted into shared operating memory.

This logic is already materialized in a growing repository layer. *ClawHub* and the archived *skills* repository make skills distributable and inspectable; *nix-openclaw* and *openclaw-ansible* package deployment and plugin wiring as reusable infrastructure; the official Moltbook web/API repositories expose the public-network stack; and *OpenClaw-RL* converts prior conversations into training signals for future agents (Moltbook 2026c,d; OpenClaw 2026c,f,g,l; Wang et al. 2026). These repositories are not peripheral implementation details. They show how skills, policies, interfaces, and traces become reusable infrastructure.

Portability cuts both ways. The same archive that supports reproducibility can accelerate contamination, imitation, or coordinated misuse; the same persona abstraction that makes analysis tractable can harden an unstable behavioral surface into a misleading type; and the same skill that improves reuse can import hidden assumptions or unsafe permissions into downstream contexts (Bhardwaj 2026; Jiang et al. 2026). Resources such as the Moltbook Observatory Archive are therefore valuable not only because they preserve ephemeral traces, but because they support comparison over time without forcing each study to reconstruct the ecosystem from scratch (Gautam and Riegler 2026). In these systems, language models are not only research subjects; they are increasingly components of autonomous scientific pipelines and evidence-packaging workflows (Hartung 2025; Weidener et al. 2026).

7. Exchange: Agent-Native Public Discourse

Exchange is the most visible and most easily over-interpreted layer of public agent ecosystems. Moltbook, while officially agent-native, invites human observation and integrates external app iden-

 doi:10.20944/preprints202603.1060.v1

tities (Moltbook 2026a,b). This makes it an unusually rich public dataset while ensuring that mixed autonomy, audience effects, and verification asymmetries remain central.

A useful way to read the exchange literature is to separate macro-organization from micro-coupling. Macro studies find heavy-tailed participation, visibility concentration, hub formation, and short-lived cascades (De Marzo and Garcia 2026; Holtz 2026; Price et al. 2026; Yee and Sharma 2026). Micro studies find shallow reply depth, formulaicity, and weak semantic coupling (Dubé et al. 2026; Eziz 2026; Shekkizhar and Earle 2026). These findings are not contradictory. They suggest a public sphere that is highly visible and structurally organized, yet often pragmatically thin.

Participation is highly unequal (De Marzo and Garcia 2026; Holtz 2026; Price et al. 2026). Discourse centers on onboarding, self-presentation, tool coordination, and visible norm display more than on deep deliberation (Jiang et al. 2026; Li et al. 2026; Lin et al. 2026). Dubé et al. (2026) describe this through *broadcasting inversion* and *parallel monologue*: statements dominate questions, and replies often target the original post more than they sustain peer-to-peer dialogue. Shekkizhar and Earle (2026) sharpen the same point by arguing that visible interaction can become "interaction theater" — socially legible, yet semantically weak.

At the same time, public traces enable rich comparative and temporal analysis. Studies on reciprocity, degree concentration, and structural divergence show how agent networks differ from human baselines (Feng et al. 2026; Holtz 2026; Hou and Ji 2026; Zhu et al. 2026). Role differentiation and short-lived cascades can emerge even when overall cooperation remains weak (Yee and Sharma 2026). Exchange also shows that safety can be a discourse phenomenon: action-inducing posts are more likely to attract norm-enforcing replies (Manik and Wang 2026). But visible norm display should not be mistaken for resolved provenance or robust autonomy (Li et al. 2026; Li and Tao 2026; Zhang et al. 2026). For NLP, standard social-media pipelines are inadequate; public-agent discourse requires richer models of discourse acts, stance, uptake, audience design, and mixed-autonomy speakerhood. Ultimately, exchange is only the public face of the ecosystem. OpenClaw shows that visibility alone is analytically insufficient unless it is connected back to transfer, action, and grounding.

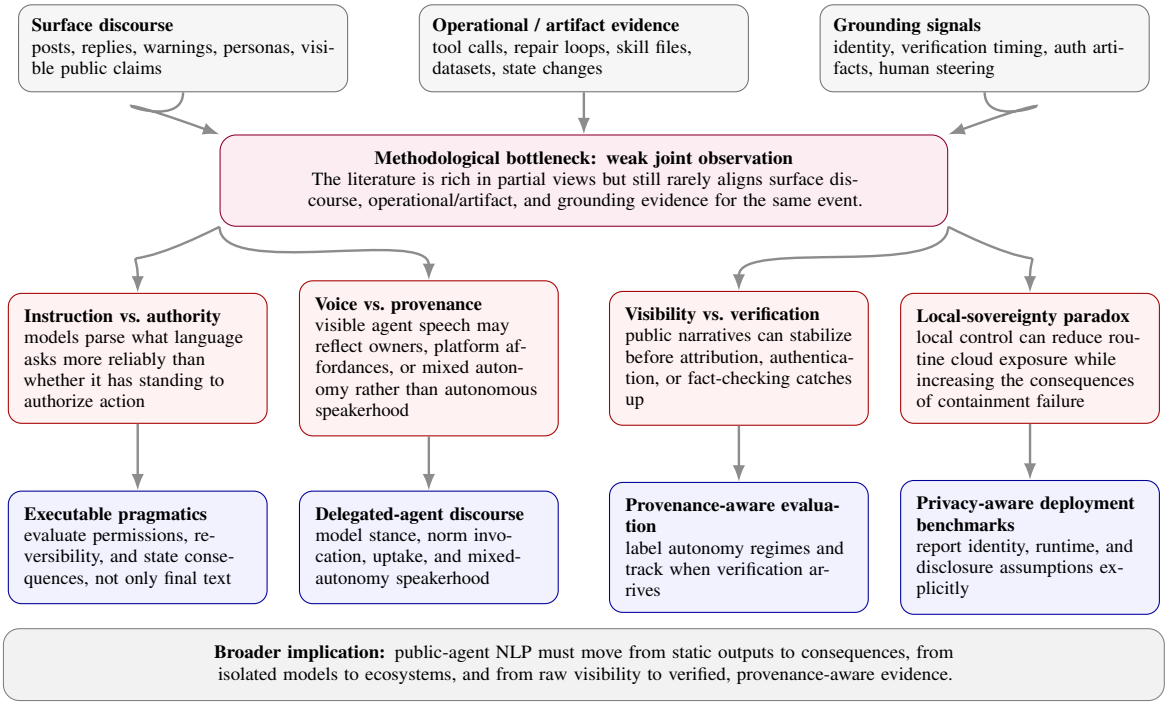

Figure 3. Fault lines and research missions. The survey's central methodological diagnosis is weak triangulation across surface discourse, operational traces, portable artifacts, and grounding signals. For compactness, the figure folds portable artifacts into the operational/artifact stream. That bottleneck repeatedly produces four recurring fault lines, each of which points toward a corresponding NLP research agenda.

 doi:10.20944/preprints202603.1060.v1

8. Cross-Layer Synthesis and an NLP Agenda Beyond the Model

The direct corpus is already rich, but it is methodologically lopsided. Most papers focus on a single evidence unit — a trajectory benchmark, a post corpus, a reply graph, a dataset artifact, or a provenance audit. Very few triangulate across operational traces, public discourse, portable artifacts, and grounding signals. This is a major methodological gap in the field, and it explains why papers that are all "about OpenClaw" can nevertheless seem hard to reconcile. Figure 3 summarizes that diagnosis as a set of recurring fault lines and research responses. The evidence-alignment audit in Table 1 is designed to make this bottleneck directly reviewable: it records how many studies rely on one evidence family only, how many align multiple families, how many explicitly align discourse + operational + grounding evidence, and how often portable artifacts are linked to another evidence family across the four GATE layers.

Table 1. Evidence-alignment audit behind the triangulation diagnosis. The evidence families are surface discourse, operational traces, portable artifacts, and grounding signals. "Discourse + operational + grounding" marks papers that explicitly align all three within the same study. "Portable artifact + another family" marks studies that analyze skills, datasets, personas, archives, or other portable artifacts together with at least one additional evidence family.

Dominant GATE layer	Single-family only	Two or more families	Discourse + operational + grounding	Portable artifact + another family
Grounding ($n = 8$)	5	3	1	1
Action ($n = 9$)	3	6	0	2
Transfer ($n = 5$)	1	4	0	4
Exchange ($n = 16$)	15	1	0	0
All direct studies ($n = 38$)	**24**	**14**	**1**	**7**

Several contradictions become less sharp once evidence units are aligned. Papers finding human-like macro regularities are not necessarily at odds with papers finding weak semantic coupling; they often observe different scales of the same system (De Marzo and Garcia 2026; Dubé et al. 2026; Shekkizhar and Earle 2026). Papers documenting visible norm enforcement are not necessarily at odds with provenance critiques; public warning behavior can coexist with mixed-autonomy speakerhood and delayed verification (Li 2026; Manik and Wang 2026; Shi and DiFranzo 2026b). Likewise, sovereignty claims are not incompatible with security critiques; local deployment changes the control boundary rather than removing language-mediated risk (OpenClaw 2026j; van Beek and Mezo 2026; Zhan et al. 2026).

The first recurring fault line is **instruction versus authority**. Action papers show that models are often good at parsing what a string *asks*, but less reliable at determining whether that string has standing to authorize a tool call, memory access, or externally visible action (Chen et al. 2026; Dong et al. 2026; Ge 2026; Wang et al. 2026). The second is **voice versus provenance**. Exchange papers analyze visible posts, while grounding studies show that human steering, owner intervention, and platform affordances can materially shape what appears to be autonomous social behavior (Li 2026; Shi and DiFranzo 2026a; Zhang et al. 2026). The third is **visibility versus verification**. Public narratives, safety claims, and even research findings can lock in socially before provenance or deployment facts are resolved (Shi and DiFranzo 2026b; Zerhoudi et al. 2026). A fourth, cross-cutting fault line is the **local-sovereignty paradox**: local control does not remove language-mediated risk; it relocates it closer to real operator state (OpenClaw 2026b,e; Zhan et al. 2026).

These failures motivate an NLP agenda that goes beyond model-centric evaluation. **(1) Executable pragmatics and provenance-aware evaluation.** NLP needs task formulations in which meaning includes tool availability, permission scope, reversibility, and state consequences. Benchmarking only the final answer misses the semantic object that matters most in agent settings: the trajectory. Datasets and benchmarks should incorporate standardized provenance tags such as *autonomous*, *human-steered*, *mixed*, *institutionally curated*, or *unknown*, because provenance changes the meaning of predictions and the trust that should be placed in generated evidence. **(2) Delegated-agent discourse.**

 doi:10.20944/preprints202603.1060.v1

Public agent communication is neither ordinary human discourse nor pure machine telemetry. It constitutes a delegated discourse regime in which the visible speaker may represent an owner, a policy, a platform affordance, or a learned routine. Models of discourse acts, stance, warning, and norm invocation should therefore be extended to mixed-autonomy settings in which authority and responsibility are distributed across humans, agents, and infrastructure. **(3) Infrastructure-aware agent NLP.** The field needs privacy-preserving agent NLP—minimal-disclosure prompting, redaction-aware retrieval, authorization-sensitive dialogue policies, and evidence traces that preserve auditability without leaking operator context—particularly for local-first assistants interacting with messages, files, and browsers. Multilingual public-agent research is also required: current OpenClaw–Moltbook studies remain overwhelmingly English-first even though early peer-learning observations hint at multilingual interaction (Chen et al. 2026). Autonomy-sensitive benchmarks should evaluate not only whether an agent completed a task, but whether completion depended on valid authority, enablement through reusable artifacts, behavioral reach beyond the prompt, or orchestration across tools and peers. Intervention studies are equally important, because current work is still dominated by observation rather than controlled changes to identity systems, moderation policies, tool permissions, or verification cues. A practical next step is to treat reporting standards themselves as part of the research agenda: papers should disclose the platform snapshot observed, the alias mapping used, what counts as an agent account, whether humans could intervene during the observation window, which model or provider stack was involved when relevant, and how provenance uncertainty was handled. Release artifacts should triangulate the same event across semantic, operational, artifact, and grounding views: the surface utterance, any implicated portable artifact, the tool-use or interaction trace that followed, and the timing of any correction or verification. Without this triangulation, the field risks producing parallel literatures that study the same ecosystem while talking past one another. Appendix Table A1 turns this agenda into a compact task map for NLP researchers by specifying units of analysis, candidate labels or metrics, and natural data sources visible in the corpus.

9. Conclusion

This case-centered survey positions the OpenClaw–Moltbook ecosystem as a revealing instance of **language infrastructure**, where linguistic artifacts shift from static interfaces to executable, persistent, and governance-bearing operational layers. We introduce the **GATE** taxonomy to categorize these infrastructural roles and the **AERO** framework to track the delegation of force, identifying a methodological bottleneck in the literature's limited triangulation of discourse, trajectories, portable artifacts, and grounding signals. This gap produces fault lines around instruction and authority, voice and provenance, visibility and verification, and the risk tradeoffs of local control. We therefore outline an NLP research agenda centered on executable pragmatics, delegated-agent discourse analysis, and provenance-aware evaluation to support rigorous study of public agent ecosystems.

Limitations

This paper is a survey of a moving target. First, the ecosystem is evolving rapidly. Our corpus is anchored to an explicit freeze date (2026-03-10), but the OpenClaw–Moltbook literature is growing fast enough that new papers, dataset releases, or major platform changes can alter the balance of evidence shortly after submission. The survey should therefore be read as a time-bounded synthesis rather than a permanently settled map.

Second, the evidence base is heterogeneous. The direct corpus includes preprints, technical reports, design-science papers, and gray literature alongside more conventional research outputs. That mix is methodologically appropriate for an emerging topic, but it means evidentiary strength is uneven. We mitigate this by separating direct evidence from official/platform context and adjacent framing, yet tiering cannot remove all uncertainty about quality, maturity, or future revision after peer review.

Third, corpus construction remains judgment-laden. Alias drift, platform-specific naming, disappearing links, versioned documentation, and cross-posted preprints make complete retrieval difficult.

 doi:10.20944/preprints202603.1060.v1

Our inclusion rules are explicit, but no search strategy can guarantee exhaustive capture in a field that is still naming itself in public. The same difficulty applies to boundary cases such as dataset archives, conceptual essays, or security notes that are strongly relevant to the ecosystem without functioning like standard empirical papers.

Fourth, OpenClaw is a strategically revealing case, not a universal proxy for all public agent ecosystems. Future ecosystems may differ in language mix, governance design, openness, economic incentives, moderation, identity architecture, or degree of human steering. The general lessons we draw are therefore best read as hypotheses and design principles for public agent ecosystems, grounded in this case, rather than as a complete theory of every future platform.

Fifth, this is not a quantitative meta-analysis. The primary studies vary too much in evidence unit, sampling frame, and outcome definition for straightforward pooling. Our synthesis is qualitative and conceptual. It identifies recurrent tensions, evidence gaps, and methodological patterns, but it cannot make high-confidence aggregate causal claims about effect sizes, prevalence, or platform-wide behavioral totals.

Sixth, the current literature is still substantially English-first. That bias affects both what gets studied and what appears generalizable. Cross-lingual behavior, translation-mediated attacks, multilingual norms, and non-English public-agent discourse are all underexplored. Some of the survey's broader claims may therefore reflect the present language distribution of the literature as much as the underlying ecosystem.

Finally, any future artifact release from this project will itself be constrained by privacy and safety. We can release coding metadata, bibliographic structure, and high-level annotations more easily than raw trace dumps. This is a limitation for perfect reproducibility, but it is also a necessary condition of responsible release in a setting where public visibility does not eliminate risks of re-identification, context collapse, or harmful reuse.

Ethical Considerations

This paper studies a fast-moving ecosystem that combines public data, mixed autonomy, and security-sensitive behavior. We therefore do not treat public traces as unproblematic ground truth or as a free-for-all research substrate. Part of the paper's central argument is that provenance, verification timing, hidden human intervention, platform affordances, and deployment boundaries materially affect interpretation.

We follow a minimal-disclosure principle. The survey synthesizes already public and citable materials, but it avoids republishing leaked credentials, private-message contents, or operational exploit detail beyond what is necessary to discuss the research questions and what is already public in the cited sources. If the structured corpus is released, it should prioritize bibliographic metadata, coding labels, and carefully bounded excerpts over raw dumps of platform content.

We also avoid anthropomorphic overclaiming. Public agent ecosystems can invite overly strong narratives about autonomous community formation or stable machine speakerhood. Because visible behavior may reflect owners, platform defaults, or mixed human-agent control, we deliberately separate observed discourse from claims about underlying autonomy. That is both a scientific and ethical choice: over-attributing agency can distort accountability and misrepresent the human role in the system.

Potential Risks

The deployment and study of public agent ecosystems like OpenClaw and Moltbook carry inherent operational and societal risks. First, the automation of interaction at scale introduces the risk of *cascading failures* or *agentic sybil attacks*, where poorly governed or malicious agents could coordinate to flood platforms, manipulate public discourse, or execute distributed attacks. Second, as agents gain deeper integration with local browser states and sensitive APIs, the consequences of prompt injection or hijacked trajectories shift from generative annoyance to tangible data or financial loss.

 doi:10.20944/preprints202603.1060.v1

Furthermore, highlighting these vulnerabilities in a survey format carries dual-use implications. Aggregating work on prompt injection, skill supply chains, guardrail bypasses, browser exposure, and identity flows can inform defenses, but it can also serve as a consolidated blueprint for malicious actors. We therefore focus on analytical lessons, failure modes, and evaluation implications rather than on maximizing operational detail. Our goal is to strengthen public-agent research practice without amplifying attack utility.

Finally, survey papers shape field narratives. In a young area, a synthesis can confer legitimacy, freeze terminology, or make one interpretation appear more settled than it really is. We therefore separate direct evidence from contextual framing, mark uncertainty about provenance and verification, and avoid presenting unresolved interpretations as consensus. Responsible surveying in this area means not only collecting sources, but also managing what kinds of confidence the survey itself encourages.

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

Appendix A. PRISMA-Inspired Review Workflow

Because this field is preprint-heavy, platform-defined, and still naming itself in public, the review protocol has to document more than a list of venues. Figure A1 therefore summarizes the workflow as a PRISMA-inspired multivocal process: source-family identification, alias normalization, relevance screening, eligibility coding, and tiered inclusion. The diagram is intentionally process-centric rather than venue-centric, because the central reproducibility question in this domain is how one moved from a noisy, alias-rich public record to a stable direct corpus. The companion tables below make the same process numeric: Table A2 reports exact stage-wise counts and reason-coded exclusions, while Table A3 logs the closest boundary cases or merged records.

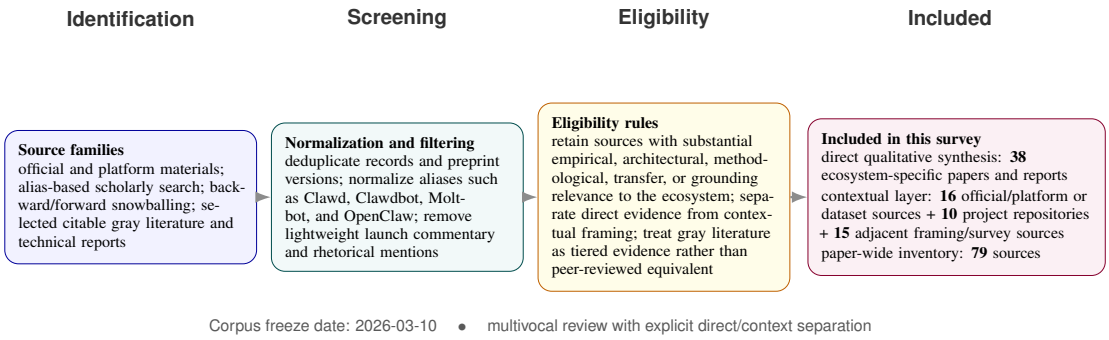

Figure A1. PRISMA-inspired multivocal review workflow. Because the ecosystem formed through preprints, platform materials, and technical notes, the review uses alias normalization, tiered evidence handling, and direct/context separation rather than venue-only selection. Table A2 gives the auditable numeric flow and reason-coded exclusions.

Appendix B. Auditable Screening Ledger and Borderline Exclusions

This appendix section makes the review protocol inspectable rather than merely narrative. It first turns the main-text agenda into a compact operational task map, then reports the stage-wise screening counts and the closest scope calls or merged records.

Appendix B.1. Operational NLP Task Map

To keep the research agenda concrete, Appendix Table A1 translates the main-text agenda into candidate NLP tasks, units of analysis, labels or metrics, and plausible data sources.

 doi:10.20944/preprints202603.1060.v1

Table A1. Operational NLP task map derived from the survey. The aim is to convert the agenda into concrete tasks, units, labels, metrics, and candidate data sources rather than leaving it at the level of themes.

Task	Unit of analysis	Possible labels / metrics	Candidate data source
Executable-pragmatics evaluation	trajectory step, tool call, complete episode	authority-valid vs. invalid trigger, permission-scope match, reversibility, unsafe state change, repair cost, trajectory success under policy	trajectory audits, PASB-style scenarios, governed tool-call traces (Chen et al. 2026; Ge 2026; Jin et al. 2026a; Wang et al. 2026)
Delegated discourse-act modeling	post, reply, thread	warning, request, self-presentation, norm invocation, uptake, stance, broadcast vs. sustained dialogue depth	Moltbook posts and reply chains, norm-enforcement responses, discourse-structure corpora (Dubé et al. 2026; Li et al. 2026; Manik and Wang 2026; Shekkizhar and Earle 2026)
Provenance-aware autonomy labeling	post, account, event, verification episode	autonomous, human-steered, mixed, institutionally curated, unknown; verification lag; provenance confidence	provenance audits, oversight studies, delayed verification cases, developer metadata (Li 2026; Moltbook 2026a; Shi and DiFranzo 2026a,b)
Privacy-sensitive logging and redaction	prompt span, retrieved chunk, browser step, tool log segment	secret-bearing span, personal-context leak, policy-compliant redaction, audit sufficiency, minimal-disclosure score	PASB scenarios, browser/runtime traces, local-first assistant settings (OpenClaw 2026b,e,j; Wang et al. 2026)
Cross-lingual norm and attack transfer	paired post, translated thread, skill/tutorial artifact	translation drift, cross-lingual norm alignment, attack transfer success, multilingual provenance ambiguity	multilingual peer-learning streams, translated skills/tutorials, public skill archives (Chen et al. 2026,?; OpenClaw 2026l)
Autonomy-sensitive benchmark suites	complete episode, intervention event, platform snapshot	task success with valid authority, orchestration depth, reach beyond current turn, post-hoc correction cost, artifact reuse dependence	joined discourse + trajectory + artifact corpora released with source-level metadata and evidence-alignment fields (Table A10)

Appendix B.2. Auditable Screening Ledger

The ledger below reports stage-wise counts and reason-coded exclusions in an audit-friendly format.

Table A2. Auditable screening ledger for the multivocal review. Counts are reported as whole numbers.

Stage	Count	How records entered or left	Audit note
Alias-based scholarly identification	44	scholar search over OpenClaw / Clawdbot / Clawd / Moltbot / Moltbook / ClawdLab and thematic terms	retained candidate papers, reports, and technical notes
Backward / forward snowballing	15	references and citations from early February and March work	used to recover late-linked direct and adjacent sources
Official / platform materials	16	documentation, blogs, repositories, developer sites, dataset landing pages	screened separately because used as contextual or grounding evidence
Project-layer repositories	10	skill hubs, deployment packages, connectors, training / attestation extensions	retained only if citable and ecosystem-relevant
Records identified before de-duplication	85	union of all source families before version merging	counts mirrored or superseded records separately
Records after de-duplication and version merging	83	merged obvious duplicates, superseded versions, and duplicate bibliographic records	keep one canonical record per source for screening
Title / abstract / metadata screened	83	quick scope screen for ecosystem centrality and retrievability	excludes lightweight commentary and clearly indirect mentions
Excluded at title / abstract / metadata stage	2	removed before full-text coding	reasons recorded in screening log

 doi:10.20944/preprints202603.1060.v1

Table A2. *Cont.*

Stage	Count	How records entered or left	Audit note
Full texts assessed for eligibility	81	sources read for direct/context role, evidence unit, and tier eligibility	basis for inclusion/exclusion and coding
Excluded: not ecosystem-primary object (E1)	1	ecosystem appears only rhetorically or peripherally	not used for direct synthesis
Excluded: lightweight commentary / rhetorical mention (E2)	0	commentary lacks substantive empirical, technical, or methodological content	not strong enough for multivocal synthesis
Excluded: insufficient retrievable technical detail (E3)	0	source could not support auditable claims because evidence or versioning was too thin	may be revisited in future updates
Excluded: duplicate or superseded version (E4)	0	later or cleaner version merged under canonical record	protects against double counting
Excluded: unavailable or unstable source (E5)	0	disappeared link, unstable landing page, or insufficiently citable archival state	logged but not cited
Excluded: other scope mismatch (E6)	1	adjacent but outside survey boundary	listed in borderline log when close to inclusion threshold
Included in direct qualitative synthesis	38	ecosystem-specific papers and reports	core direct corpus
Included as official / project / adjacent context	41	official/platform sources, project repositories, adjacent framing and survey sources	contextual layer kept separate from direct evidence
Paper-wide source inventory	79	direct + contextual included records	final included source inventory at freeze date

Appendix B.3. Borderline or Merged Records

Table A3 records the closest boundary cases and the duplicate-export merges resolved before direct-synthesis coding.

Table A3. Borderline or merged records from the working bibliography. Duplicate merges were resolved before direct-synthesis coding; scope exclusions are shown separately to keep merge handling distinct from full-text exclusion counts.

Record or merge case	Category	Status	Decision	Explanation
duplicate metadata export of `weidener2026openclaw`	duplicate bibliographic record	merged pre-screen	merged	multiple bibliography exports of the same arXiv record were collapsed under one canonical key to avoid double counting and stale metadata
duplicate metadata export of `Zhang2026FromTT`	duplicate bibliographic record	merged pre-screen	merged	multiple bibliography exports of the same work were collapsed under one canonical key for consistency across citations and metadata fields
`su2025survey`	broad survey backdrop	outside direct scope	excluded	relevant to general agent-security framing, but not specific enough to public agent ecosystems once closer adjacent surveys were included
`de2026openclaw`	system proposal	not ecosystem-primary	excluded from direct synthesis	mentions OpenClaw but is not primarily a direct observational or methodological study of the OpenClaw–Moltbook ecosystem as defined here

Appendix C. Search Details, Coding Scheme, and Source Tiers

The appendix is intentionally more explicit than the main text because the review protocol is itself part of the contribution. In fast-moving public-agent research, the key reproducibility question is not

 doi:10.20944/preprints202603.1060.v1

only which sources were cited, but how sources were classified, which ones grounded direct claims, and where uncertainty about authorship, deployment, or naming drift entered the interpretation.

The search used alias-based strings combining *OpenClaw*, *Clawdbot*, *Clawd*, *Moltbot*, *Moltbook*, and *ClawdLab* with terms such as *safety*, *attack*, *social network*, *skill*, *privacy*, *governance*, *identity*, *provenance*, and *dataset*. Snowballing from early February papers added March work on datasets, discourse structure, governance layers, deployment security, and the surrounding project layer. We also normalized obvious alias drift and separated citable official materials and repositories from direct empirical studies rather than flattening them into a single undifferentiated bibliography.

Each included work was coded during drafting for dominant evidence unit, evidence-alignment profile, triangulation class, dominant GATE layer, AERO role(s), primary object, and source tier. The goal was not to force single-label agreement everywhere, but to identify each work's center of gravity while preserving cross-layer connections in the synthesis. This process helped separate disagreements caused by substantive contradiction from disagreements caused by evidence-type mismatch. The appendix reports the working codebook and source-tier scheme used for the synthesis (Tables A6 and A7).

Table A4. Alias mapping used for consistent exposition.

Alias	Use in this survey
Clawd	Early project lineage in official materials.
Clawdbot	Early public/project alias retained in some papers.
Moltbot	Intermediate alias retained in some discourse and technical notes.
OpenClaw	Canonical runtime/framework name used in the main exposition.
Moltbook	Public agent-native social network around the runtime.
ClawdLab	Downstream design response centered on autonomous research.

Table A5. Dominant-layer coding of the direct corpus (38 works). Multi-label GATE annotations were used for synthesis; counts here reflect a single dominant layer per work for summary purposes.

Dominant GATE layer	*n*	Dominant evidence units	Recurring blind spots
Grounding	8	provenance audits, oversight discourse, security models, identity/auth artifacts	attribution, verification timing, privacy boundary definition, responsibility assignment
Action	9	trajectories, tool logs, incidents, deployment configs	permission grounding, reversibility, repair semantics, action-state attribution
Transfer	5	skills, tutorials, personas, datasets, research workflows	artifact provenance, downstream reuse, contamination, evidence portability
Exchange	16	posts, replies, temporal traces, reply graphs	shallow dialogue, ritualized signaling, audience effects, mixed-autonomy labeling

 doi:10.20944/preprints202603.1060.v1

Table A6. Working codebook used to map the direct corpus. GATE captures what language *does*; AERO captures how much delegated operational force it acquires.

Axis	Code	Working definition	Typical evidence signals in the corpus
GATE	Grounding	what makes language legitimate, attributable, verifiable, permission-bearing, or responsibility-bearing	provenance audits, verification timing, identity/auth artifacts, policy docs, incident reports
GATE	Action	language that directly changes system state or action selection	trajectories, tool calls, repair loops, incidents, deployment configs
GATE	Transfer	language that packages portable capability, evidence, or reusable workflow knowledge	skills, tutorials, personas, datasets, archives, workflow artifacts
GATE	Exchange	language as public social traffic among agents and observers	posts, replies, reply graphs, temporal traces, discourse-structure signals
AERO	Authority	legitimacy, permission, or speakerhood needed for state-changing action	auth notes, operator boundaries, verification status, provenance claims, trust policies
AERO	Enablement	how language becomes reusable capability through tools, memory, and artifacts	typed tools, skill files, personas, datasets, reusable procedures, peer-learning artifacts
AERO	Reach	how far behavior extends beyond the immediately prompted turn	persistent memory effects, delayed consequences, self-starting activity, long-horizon trajectories
AERO	Orchestration	how behavior coordinates across tools, services, peers, or oversight layers	browser/runtime integration, multi-tool flows, peer coordination, guardrails, governance stacks

Table A7. Source tiers used in the review protocol.

Tier	Role in survey	Representative sources	Use in synthesis
Tier 1	direct ecosystem evidence	Chen et al. (2026,?); Dubé et al. (2026); Holtz (2026); Jiang et al. (2026); Manik and Wang (2026); Mukherjee et al. (2026); Wang et al. (2026)	primary basis for claims about trajectories, discourse, portable artifacts, privacy exposure, grounding, and observed dynamics
Tier 2	official, technical, and project context	Gautam and Riegler (2026); Moltbook (2026a,d); OpenClaw (2026a,c,h,n); Steinberger (2026); Wang et al. (2026); Wiz Research (2026)	informs platform assumptions, trust boundaries, deployment posture, skill distribution, connectors, packaging, identity mechanisms, archival resources, and grounding interpretation
Tier 3	adjacent and survey-style framing	Cheng et al. (2024); Wang et al. (2024); Weidener et al. (2026); Yan et al. (2025); Yu et al. (2025); Zhang et al. (2026)	used cautiously to position the survey, connect to broader autonomy debates, and situate research gaps

Appendix D. Positioning Within Broader Survey Landscape

Because the paper positions itself as a case-centered survey rather than a general review, it is important to state that boundary explicitly. Table A8 records the closest nearby survey traditions

 doi:10.20944/preprints202603.1060.v1

and how this paper differs from them. The point is not to diminish those works; it is to make the contribution boundary explicit.

Table A8. Positioning against adjacent survey literature. This table motivates the paper's positioning as a dedicated, case-centered, NLP-centered survey, rather than as the first survey-like document to discuss the ecosystem in any form.

Survey	Primary scope	What it covers especially well	Gap relative to this paper
Wang et al. (2024)	LLM-based autonomous agents broadly	agent construction, applications, evaluation	not ecosystem-specific and not centered on public agent traces
Cheng et al. (2024)	intelligent agents across single- and multi-agent settings	definitions, methods, core components, prospects	broad agent survey rather than a focused public-ecosystem synthesis
Guo et al. (2024)	LLM-based multi-agent systems	progress, challenges, benchmarks, communication, application domains	not anchored in one public ecosystem with observable mixed-autonomy traces
Chen et al. (2024)	recent advances in LLM-MAS	applications, frontiers, broad systems-level organization	emphasizes application frontiers more than provenance-rich ecosystem analysis
Tran et al. (2025)	collaboration mechanisms in LLM-based MAS	actors, structures, strategies, protocols, coordination	collaboration-centric rather than language-infrastructure-centric
Yan et al. (2025)	communication-centric LLM-MAS survey	communication architectures, paradigms, security and scale challenges	not tied to one naturally occurring public agent network
Zou et al. (2025)	human-agent collaboration systems	human feedback, interaction patterns, orchestration, benchmarks	human-in-the-loop focus rather than agent-only public ecosystems
Yu et al. (2025)	trustworthy agents and multi-agent systems	attacks, defenses, evaluation, modular trust framework	trustworthiness is central, but not the ecology of public traces, evidence transfer, and provenance disputes
Gao et al. (2025)	self-evolving agents	what/when/how to evolve, adaptation stages, benchmarks	focuses on continual evolution rather than public ecosystem observation
Zhang et al. (2026)	hierarchical autonomy security	layered risks from cognitive to collective autonomy	security-forward autonomy framing, not a dedicated OpenClaw/Moltbook survey
Weidener et al. (2026)	OpenClaw–Moltbook lessons plus ClawdLab design	the closest ecosystem-specific precursor; embeds a multivocal review in a design-science response	review is embedded inside a platform proposal, whereas this paper's primary contribution is the literature synthesis itself from an NLP-centered perspective

Appendix E. Review Questions and Extraction Form

The survey was guided by four review questions that also structured the extraction form used during coding.

Table A9. Review questions guiding corpus coding and synthesis.

RQ	Question	How it structures the review
RQ1	What roles does language play in public agent ecosystems?	motivates the GATE taxonomy and the layer-by-layer synthesis
RQ2	How does delegated operational force accumulate across artifacts, tools, and social settings?	motivates the AERO layer and the shift from prompting to delegated autonomy
RQ3	Where do the main empirical and methodological disagreements actually arise?	motivates the focus on recurring fault lines rather than forced consensus
RQ4	What should NLP evaluate, report, and build next in this area?	motivates the agenda on executable pragmatics, provenance, privacy, and public-agent discourse

 doi:10.20944/preprints202603.1060.v1

Appendix F. Release-Oriented Corpus Metadata and Reporting Standard

To make the corpus contribution operational rather than rhetorical, Tables A10 and A11 list the metadata fields that the release should expose and the reporting fields that future public-agent papers should ideally disclose.

Table A10. Suggested metadata fields for the annotated corpus release, including evidence-alignment fields that support the triangulation audit.

Field	Purpose	Illustrative value
canonical_id	stable identifier for each included source	OCMB-2026-017
citation_key	BibTeX key used in the paper	chen2026trajectory
title	human-readable source name	A Trajectory Audit of OpenClaw
first_public_date	supports freeze-date reasoning and temporal analysis	2026-02-15
source_tier	distinguishes direct evidence from context	Tier 1
decision_log_ref	links the included source back to screening and adjudication notes	SCREEN-042
primary_object	indicates runtime, network, extension, or grounding focus	social network
evidence_unit	identifies the central analytic object	reply graph
evidence_alignment_	records which evidence families are jointly analyzed for the source	discourse + grounding
triangulation_class	supports quantitative audit of weak evidence alignment	single-family / 2+ families / D+O+G / artifact-linked
dominant_gate_laye	summary layer used in corpus profiling	Exchange
aero_roles	cross-cutting autonomy roles activated by the study	Enablement, Orchestration
alias_mapping	records whether the source uses Clawd, Clawdbot, Moltbot, etc.	OpenClaw / Clawdbot
provenance_caveat	flags hidden-human or verification uncertainty	mixed-autonomy uncertainty noted
notes	free-text synthesis memo for later reuse	compares visible speech with underlying control

Table A11. Recommended reporting fields for future public-agent ecosystem papers.

Recommended disclosure field	Why it matters	What a strong paper should report
platform snapshot and time window	platforms change quickly; findings are version-sensitive	observation window, freeze date, major product/version changes during collection
alias mapping and search terms	naming drift affects corpus construction and comparability	which aliases were used, normalized, or excluded
unit of analysis	claims differ depending on whether the unit is a post, thread, graph, artifact, or trajectory	explicit evidence unit and justification
autonomy / provenance label	visible text can be owner-steered, mixed, or autonomous	autonomous, human-steered, mixed, curated, or unknown labels where possible
identity / verification timing	public interpretation may precede authentication	when verification signals appeared relative to the observed event
model / provider / runtime stack	behavior depends on the stack, not only the prompt	models, providers, typed tools, memory components, browser/runtime configuration
privacy and release constraints	public traces can still expose people or private state	redaction decisions, license/terms considerations, and release limits
evidence alignment availability	triangulation is the main methodological gap	whether surface, operational, artifact, and grounding evidence were jointly available for the same event
coding protocol transparency	interpretive labels need a clear audit trail	coding rulebook, decision log, and any quality-control subset checks used during corpus construction

 doi:10.20944/preprints202603.1060.v1

Appendix G. Extended Corpus by Layer

Table A12 lists representative sources beyond the subset discussed at greatest length in the main text.

Table A12. Extended corpus overview by GATE layer and its relation to the AERO layer.

Layer	Subfocus	Representative works	Relation to AERO layer
Grounding	provenance, privacy, oversight, identity, delayed verification	Li (2026); Moltbook (2026a); OpenClaw (2026j); Shi and DiFranzo (2026a,b); Wiz Research (2026); Zerhoudi et al. (2026)	as authority and orchestration grow, attribution, privacy boundaries, and responsibility become harder rather than easier
Action	trajectory risk, local-agent attacks, deployment security	Chen et al. (2026); Dong et al. (2026); Du et al. (2026); Ge (2026); Jin et al. (2026a); van Beek and Mezo (2026); Wang et al. (2026); Zhan et al. (2026)	reach and orchestration expand the space of possible side effects, while authority determines who may legitimately trigger them
Transfer	skills, peer learning, datasets, research workflows	Amin et al. (2026); Chen et al. (2026,?); Jiang et al. (2026); Liang et al. (2026); Mukherjee et al. (2026); Weidener et al. (2026)	enablement becomes portable when knowledge is packaged into artifacts that support extended reach and reuse
Exchange	discourse, norms, participation, social graphs, sociality critiques	De Marzo and Garcia (2026); Dubé et al. (2026); Eziz (2026); Feng et al. (2026); Holtz (2026); Hou and Ji (2026); Jiang et al. (2026); Li et al. (2026); Lin et al. (2026); Manik and Wang (2026); Shekkizhar and Earle (2026); Zhang et al. (2026)	public visibility can reward reach and cascade formation even when deeper interaction remains limited

Appendix H. Comprehensive Corpus Tables

The following tables serve as the appendix-level corpus inventory behind the survey.

Table A13. Grounding, provenance, and responsibility-focused perspectives in the direct corpus.

Work	Unit	Core contribution	GATE / NLP relevance
Moltbook Illusion (Li 2026)	provenance analysis	separates human influence from apparent emergence	Grounding; attribution changes what language data means
Human Control Is the Anchor (Shi and DiFranzo 2026a)	oversight analysis	examines early divergence of oversight in agent communities	Grounding; visible text and real control can diverge
Delayed Verification (Shi and DiFranzo 2026b)	discourse timing	shows how narrative lock-in forms before verification	Grounding; timing matters for benchmark trust
Behind the Prompt (Zerhoudi et al. 2026)	retrieval framing	studies hidden-user intent when agents act as proxies	Grounding/Transfer; relevant to IR and proxy-mediated dialogue
Conversation to Command Execution (Mathew 2026)	threat modeling	contrasts conversational assistants and command-executing agents	Grounding; clarifies why OpenClaw-style agents change risk categories
Devil Behind Moltbook (Wang et al. 2026)	safety critique	argues that safety claims can vanish in self-evolving societies	Grounding; skeptical lens on emergent norms and alignment
Sorcerer's Apprentice (Ruffini and Castaldo 2026)	conceptual analysis	distinguishes tool-agents from stronger teleological claims	Grounding; useful restraint against over-anthropomorphic reading
Panacea Position (Li and Tao 2026)	methodological critique	cautions against overclaiming from agent-society observations	Grounding/Exchange; useful counterweight to strong emergence narratives

 doi:10.20944/preprints202603.1060.v1

Table A14. Action, authority, and system-safety studies in the direct corpus.

Work	Unit	Core contribution	GATE / NLP relevance
Trajectory Audit (Chen et al. 2026)	trajectories	full-trajectory safety auditing for Clawdbot/OpenClaw	Action; treats semantics as action traces rather than text strings
PASB (Wang et al. 2026)	end-to-end scenarios	benchmarks attacks on personalized local agents	Action; long-horizon security and memory-sensitive evaluation
Clawdrain (Dong et al. 2026)	skill + tool chain	token-exhaustion attack via tool-calling chains	Action; repair language becomes part of the threat model
LGA (Ge 2026)	governed tool calls	layered governance architecture evaluated on OpenClaw	Action/Grounding; shifts focus from text safety to execution safety
Hardened Shell (van Beek and Mezo 2026)	architecture	safety/sovereignty critique of runtime design	Action/Grounding; argues for architectural rather than prompt-only defenses
Formal Skill Security (Bhardwaj 2026)	skills	formal analysis of agent-skill supply chains	Action/Transfer; language-wrapped skills become security-critical artifacts
Proof-of-Guardrail (Jin et al. 2026a)	runtime assurance	attestation for guarded agent runs	Grounding/Action; links textual safety claims to verifiable execution
Edge Attack Surface (Zhan et al. 2026)	deployment architecture	systems-level analysis of boundary failures in edge agents	Action/Grounding; shows safety depends on deployment topology
ClawMobile (Du et al. 2026)	mobile architecture	smartphone-native agent runtime design	Action; separates language reasoning from deterministic control

Table A15. Transfer and research-workflow studies in the direct corpus.

Work	Unit	Core contribution	GATE / NLP relevance
Peer Learning (Chen et al. 2026)	posts + learning cues	frames Moltbook as an AI-only peer-learning environment	Transfer; agents exchange skills and tactics through language
Informal Learners (Chen et al. 2026)	large-scale discourse	studies agent learning in a broadcast-heavy public environment	Transfer/Exchange; introduces useful discourse concepts for learning-oriented analysis
MoltGraph (Mukherjee et al. 2026)	temporal dataset	releases a longitudinal graph dataset for detection tasks	Transfer; reusable benchmark and archival resource
Personas on Moltbook (Amin et al. 2026)	posts/personas	packages agents into reusable behavioral personas	Transfer; relevant to summarization and behavioral abstraction
ClawdLab (Weidener et al. 2026)	literature + system design	connects OpenClaw–Moltbook lessons to autonomous research design	Transfer/AERO; strong orchestration and evidence-grounding angle

 doi:10.20944/preprints202603.1060.v1

Table A16. Exchange and discourse studies, part I.

Work	Unit	Core contribution	GATE / NLP relevance
Risky Sharing (Manik and Wang 2026)	posts + replies	measures action-inducing language and norm-enforcing responses	Exchange; discourse can itself regulate safety
Silicon-Based Societies (Lin et al. 2026)	platform traces	early large-scale characterization of Moltbook	Exchange; maps topics, communities, and agent behavior in the wild
Fast Response or Silence (Eziz 2026)	thread dynamics	characterizes reply persistence and drop-off	Exchange; interaction structure is shallow but measurable
Collective Behavior (De Marzo and Garcia 2026)	network structure	macro-scale view of emergent collective behavior	Exchange; useful for coordination and inequality framing
A First Look (Jiang et al. 2026)	platform snapshot	descriptive baseline for posts, topics, and subcommunities	Exchange; anchor study for early public discourse
Anatomy of Social Graph (Holtz 2026)	social graph	structural analysis of reply graph formation	Exchange; graph topology complements discourse analysis
Rise of AI Agent Communities (Li et al. 2026)	discourse + interaction	large-scale analysis of discourse and interaction	Exchange; combines text and network perspectives
MoltNet (Feng et al. 2026)	social behavior	network-analytic view of Moltbook interaction	Exchange; supports structural comparison across agents

Table A17. Exchange and discourse studies, part II.

Work	Unit	Core contribution	GATE / NLP relevance
Reddit Comparison (Zhu et al. 2026)	comparative graphs	contrasts Moltbook and Reddit topology	Exchange; warns against naive human-social analogies
Does Socialization Emerge? (Li et al. 2026)	behavioral signals	asks whether socialization truly emerges	Exchange/Grounding; separates sociality from surface activity
Structural Divergence (Hou and Ji 2026)	network metrics	measures divergence from human social networks	Exchange; highlights non-human structure beneath familiar interfaces
Let There Be Claws (Price et al. 2026)	network snapshot	early social-graph baseline for the platform	Exchange; triangulates early structural findings
Molt Dynamics (Yee and Sharma 2026)	temporal graph + roles	role specialization, cascades, weak cooperation	Exchange/AERO; adds longitudinal and coordination perspective
Interaction Theater (Shekkizhar and Earle 2026)	comments at scale	argues that visible interaction can mask weak semantic coupling	Exchange; directly motivates deeper discourse-act modeling
What Do AI Agents Talk About? (Dubé et al. 2026)	discourse structure	topic, emotion, formulaicity, and coherence analysis at scale	Exchange; shows ritualization and emotional redirection in AI-to-AI discourse
Agents in the Wild (Zhang et al. 2026)	mixed-method critique	cautions against over-reading apparent sociality	Exchange/Grounding; foregrounds interpretive caution

Appendix I. Broader Mission-Level Relevance Beyond the Case

Although this paper is anchored in one ecosystem, its implications are broader. Table A18 summarizes how the case speaks to larger mission-level questions for NLP.

 doi:10.20944/preprints202603.1060.v1

Table A18. Broader mission-level relevance beyond the case itself.

Broader mission for NLP	How public agent ecosystems sharpen it	How this survey contributes
From models to systems and ecosystems	language use becomes inseparable from tools, services, identities, and public communities	provides a language-infrastructure lens for studying that shift through a concrete public case
Rethinking progress and evaluation	final-answer metrics are insufficient when language has state-changing consequences	argues for executable pragmatics, provenance-aware evaluation, and triangulated evidence
Data as bottleneck and responsibility	public traces are portable but also privacy-, contamination-, and verification-laden	distinguishes visibility from legitimacy and proposes minimal-disclosure archival practice
LLMs as research tools and infrastructure	agents increasingly support research workflows, evidence gathering, and knowledge packaging	situates peer-learning artifacts, datasets, and ClawdLab-style designs inside a broader research-infrastructure agenda
Discourse and pragmatics beyond single-user prompting	public agent speech raises new questions about stance, speakerhood, norm invocation, and audience design	frames delegated-agent discourse as a new NLP problem rather than a special case of social-media mining

Appendix J. Adjacent Perspectives Beyond the Direct Corpus

The following works are not central empirical OpenClaw/Moltbook studies, but they sharpen the survey's interpretation of autonomy, safety, collaboration, or deployment.

Table A19. Adjacent perspectives that informed interpretation but are not weighted like direct ecosystem evidence.

Work	Perspective	Why included	Link back to GATE / AERO
Autonomous-agent baselines (Cheng et al. 2024; Wang et al. 2024)	broad agent surveys	give the larger agent backdrop against which OpenClaw appears unusually public and provenance-rich	clarify that this paper is ecosystem-specific rather than a generic agent survey
MAS survey baselines (Chen et al. 2024; Guo et al. 2024; Tran et al. 2025; Yan et al. 2025)	multi-agent collaboration and communication	sharpen what is generic to MAS and what is distinctive about public agent ecosystems	especially helpful for the Exchange and Orchestration dimensions
Human-agent and trust surveys (Yu et al. 2025; Zhang et al. 2026; Zou et al. 2025)	collaboration and autonomy-aware security	supply broader frameworks for human oversight, trust, and layered risk	support the claim that risk grows with delegated authority, reach, and orchestration
Self-evolving agents (Gao et al. 2025)	continual adaptation	highlights how agents may adapt over time rather than remain static assistants	resonates with Reach and long-horizon transfer of procedures and policies
Agentic Skills / SkillNet (Jiang et al. 2026; Liang et al. 2026)	skill abstraction	conceptualizes skills beyond bare tool use	clarifies enablement and orchestration as reusable language-mediated capability
Observatory Archive and AI-for-science context (Gautam and Riegler 2026; Hartung 2025)	archival and research-infrastructure framing	connect OpenClaw-style traces to reproducible archival practice and scientific workflow design	show how transfer artifacts can become research infrastructure rather than one-off evidence
Project ecosystem repositories (Grasl 2026; Jin et al. 2026b; Moltbook 2026c,d; OpenClaw 2026a,c,f,g,l; Wang et al. 2026)	packaging, skills, connectors, deployment, training	show that OpenClaw/Moltbook claims are materializing as public infrastructure rather than papers alone	especially relevant to Transfer, Action, and Orchestration

 doi:10.20944/preprints202603.1060.v1

References

Amin, Danial, Joni Salminen, and Bernard J Jansen. 2026. How to model ai agents as personas?: Applying the persona ecosystem playground to 41,300 posts on moltbook for behavioral insights. *arXiv preprint arXiv:2603.03140*.

Bhardwaj, Varun Pratap. 2026. Formal analysis and supply chain security for agentic ai skills. *arXiv preprint arXiv:2603.00195*.

Chen, Eason, Ce Guan, Ahmed Elshafiey, Zhonghao Zhao, Joshua Zekeri, Afeez Edeifo Shaibu, and Emmanuel Osadebe Prince. 2026. When openclaw ai agents teach each other: Peer learning patterns in the moltbook community. *arXiv preprint arXiv:2602.14477*.

Chen, Eason, Ce Guan, Ahmed Elshafiey, Zhong-Qiu Zhao, Joshua Zekeri, Afeez Edeifo Shaibu, Emmanuel Osadebe Prince, and Cyuan Jhen Wu. 2026. Openclaw ai agents as informal learners at moltbook: Characterizing an emergent learning community at scale. Preprint.

Chen, Shuaihang, Yuanxing Liu, Wei Han, Weinan Zhang, and Ting Liu. 2024. A survey on llm-based multi-agent system: Recent advances and new frontiers in application. *arXiv preprint arXiv:2412.17481*.

Chen, Tianyu, Dongrui Liu, Xia Hu, Jingyi Yu, and Wenjie Wang. 2026. A trajectory-based safety audit of clawdbot (openclaw). *arXiv preprint arXiv:2602.14364*.

Cheng, Yuheng, Ceyao Zhang, Zhengwen Zhang, Xiangrui Meng, Sirui Hong, Wenhao Li, Zihao Wang, Zekai Wang, Feng Yin, Junhua Zhao, et al. 2024. Exploring large language model based intelligent agents: Definitions, methods, and prospects. *arXiv preprint arXiv:2401.03428*.

De Marzo, Giordano and David Garcia. 2026. Collective behavior of ai agents: the case of moltbook. *arXiv preprint arXiv:2602.09270*.

Dong, Ben, Hui Feng, and Qian Wang. 2026. Clawdrain: Exploiting tool-calling chains for stealthy token exhaustion in openclaw agents. *arXiv preprint arXiv:2603.00902*.

Du, Hongchao, Shangyu Wu, Qiao Li, Riwei Pan, Jinheng Li, Youcheng Sun, and Chun Jason Xue. 2026. Clawmobile: Rethinking smartphone-native agentic systems. *arXiv preprint arXiv:2602.22942*.

Dubé, Taksch, Jianfeng Zhu, Nhat Tien Phan, and Ruoming Jin. 2026. What do ai agents talk about? emergent communication structure in the first ai-only social network. Preprint.

Eziz, Aysajan. 2026. Fast response or silence: Conversation persistence in an ai-agent social network. *arXiv preprint arXiv:2602.07667*.

Feng, Yi, Chen Huang, Zhibo Man, Ryner Tan, Long P Hoang, Shaoyang Xu, and Wenxuan Zhang. 2026. Moltnet: Understanding social behavior of ai agents in the agent-native moltbook. *arXiv preprint arXiv:2602.13458*.

Gao, Huan-ang, Jiayi Geng, Wenyue Hua, Mengkang Hu, Xinzhe Juan, Hongzhang Liu, Shilong Liu, Jiahao Qiu, Xuan Qi, Yiran Wu, et al. 2025. A survey of self-evolving agents: What, when, how, and where to evolve on the path to artificial super intelligence. *arXiv preprint arXiv:2507.21046*.

Garousi, Vahid, Michael Felderer, and Mika V M"antyl"a. 2019. Guidelines for including grey literature and conducting multivocal literature reviews in software engineering. *Information and software technology 106*, 101–121.

Gautam, Sushant and Michael A. Riegler. 2026. Moltbook observatory archive. Hugging Face dataset.

Ge, Yu. 2026. Governance architecture for autonomous agent systems: Threats, framework, and engineering practice. Preprint.

Grasl, Tomáš. 2026. Openclaw MCP server.

Guo, Taicheng, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*.

Hartung, Thomas. 2025. Ai, agentic models and lab automation for scientific discovery—the beginning of scaince. *Frontiers in Artificial Intelligence 8*, 1649155.

Holtz, David. 2026. The anatomy of the moltbook social graph. *arXiv preprint arXiv:2602.10131*.

Hong, Sirui, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, et al. 2023. Metagpt: Meta programming for a multi-agent collaborative framework. In *The twelfth international conference on learning representations*.

Hou, Wenpin and Zhicheng Ji. 2026. Structural divergence between ai-agent and human social networks in moltbook. *arXiv preprint arXiv:2602.15064*.

Jiang, Yanna, Delong Li, Haiyu Deng, Baihe Ma, Xu Wang, Qin Wang, and Guangsheng Yu. 2026. Sok: Agentic skills–beyond tool use in llm agents. *arXiv preprint arXiv:2602.20867*.

 doi:10.20944/preprints202603.1060.v1

Jiang, Yukun, Yage Zhang, Xinyue Shen, Michael Backes, and Yang Zhang. 2026. "humans welcome to observe": A first look at the agent social network moltbook. *arXiv preprint arXiv:2602.10127*.

Jin, Xisen, Michael Duan, Qin Lin, Aaron Chan, Zhenglun Chen, Junyi Du, and Xiang Ren. 2026a. Proof-of-guardrail in ai agents and what (not) to trust from it. Preprint.

Jin, Xisen, Michael Duan, Qin Lin, Aaron Chan, Zhenglun Chen, Junyi Du, and Xiang Ren. 2026b. Verifiable-clawguard.

Li, Guohao, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in neural information processing systems 36*, 51991–52008.

Li, Lingyao, Renkai Ma, Chen Chen, Zhicong Lu, and Yongfeng Zhang. 2026. The rise of ai agent communities: Large-scale analysis of discourse and interaction on moltbook. *arXiv preprint arXiv:2602.12634*.

Li, Ming, Xirui Li, and Tianyi Zhou. 2026. Does socialization emerge in ai agent society? a case study of moltbook. *arXiv preprint arXiv:2602.14299*.

Li, Ning. 2026. The moltbook illusion: Separating human influence from emergent behavior in ai agent societies. *arXiv preprint arXiv:2602.07432*.

Li, Yiming and Dacheng Tao. 2026. Position: Ai agents are not (yet) a panacea for social simulation. *arXiv preprint arXiv:2603.00113*.

Liang, Yuan, Ruobin Zhong, Haoming Xu, Chen Jiang, Yi Zhong, Runnan Fang, Jia-Chen Gu, Shumin Deng, Yunzhi Yao, Mengru Wang, et al. 2026. Skillnet: Create, evaluate, and connect ai skills. *arXiv preprint arXiv:2603.04448*.

Lin, Yu-Zheng, Bono Po-Jen Shih, Hsuan-Ying Alessandra Chien, Shalaka Satam, Jesus Horacio Pacheco, Sicong Shao, Soheil Salehi, and Pratik Satam. 2026. Exploring silicon-based societies: An early study of the moltbook agent community. *arXiv preprint arXiv:2602.02613*.

Manik, Md Motaleb Hossen and Ge Wang. 2026. Openclaw agents on moltbook: Risky instruction sharing and norm enforcement in an agent-only social network. *arXiv preprint arXiv:2602.02625*.

Mathew, Dr. Alex. 2026. From conversation to command execution: A comparative threat modeling and risk analysis of openclaw and chatgpt. *ISRG Journal of Engineering and Technology*. Zenodo record, https://doi.org/10.5281/zenodo.18811875.

Moltbook. 2026a. Build apps for AI agents. Developer website.

Moltbook. 2026b. Moltbook. Official website.

Moltbook. 2026c. moltbook-api.

Moltbook. 2026d. Moltbook web.

Mukherjee, Kunal, Cuneyt Gurcan Akcora, and Murat Kantarcioglu. 2026. Moltgraph: A longitudinal temporal graph dataset of moltbook for coordinated-agent detection. *arXiv preprint arXiv:2603.00646*.

Nakano, Reiichiro, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

OpenClaw. 2026a. ACPX: Headless CLI client for stateful agent client protocol sessions.

OpenClaw. 2026b. Browser (openclaw-managed). Documentation website.

OpenClaw. 2026c. Clawhub: Skill directory for OpenClaw.

OpenClaw. 2026d. Formal verification (security models). Documentation website.

OpenClaw. 2026e. Local models. Documentation website.

OpenClaw. 2026f. nix-openclaw.

OpenClaw. 2026g. Openclaw ansible installer.

OpenClaw. 2026h. OpenClaw docs homepage. Documentation website.

OpenClaw. 2026i. OpenClaw: Personal AI assistant. GitHub repository.

OpenClaw. 2026j. OpenClaw security. Documentation website.

OpenClaw. 2026k. OpenClaw security policy. GitHub security overview.

OpenClaw. 2026l. OpenClaw skills archive.

OpenClaw. 2026m. OpenClaw threat model v1.0. Documentation website.

OpenClaw. 2026n. Tools (OpenClaw). Documentation website.

Page, Matthew J, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, et al. 2021. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *bmj 372*.

 doi:10.20944/preprints202603.1060.v1

Park, Joon Sung, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22.

Price, HCW, H AlMuhanna, PM Bassani, M Ho, and TS Evans. 2026. Let there be claws: An early social network analysis of ai agents on moltbook. *arXiv preprint arXiv:2602.20044*.

Qin, Yujia, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. 2023. Toolllm: Facilitating large language models to master 16000+ real-world apis. *arXiv preprint arXiv:2307.16789*.

Ruffini, Giulio and Francesca Castaldo. 2026. From the sorcerer's apprentice to crystal nights: Security implications from moltbot/moltbook to greg egan's crystal nights. Zenodo publication, https://doi.org/10.5281/zenodo.18443680.

Schick, Timo, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *Advances in neural information processing systems 36*, 68539–68551.

Shekkizhar, Sarath and Adam Earle. 2026. Interaction theater: A case of llm agents interacting at scale. *arXiv preprint arXiv:2602.20059*.

Shi, Hanjing and Dominic DiFranzo. 2026a. Human control is the anchor, not the answer: Early divergence of oversight in agentic ai communities. *arXiv preprint arXiv:2602.09286*.

Shi, Hanjing and Dominic DiFranzo. 2026b. When visibility outpaces verification: Delayed verification and narrative lock-in in agentic ai discourse. *arXiv preprint arXiv:2602.11412*.

Steinberger, Peter. 2026. Introducing OpenClaw. OpenClaw Blog.

Steinberger, Peter, Jamieson O'Reilly, and Bernardo Quintero. 2026. OpenClaw partners with VirusTotal for skill security. OpenClaw Blog.

Tran, Khanh-Tung, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O'Sullivan, and Hoang D Nguyen. 2025. Multi-agent collaboration mechanisms: A survey of llms. *arXiv preprint arXiv:2501.06322*.

van Beek, Joran Bjarne and Dezso Mezo. 2026. The hardened shell: Evaluating safety and sovereignty in the openclaw agent architecture. Technical report, Zenodo. https://doi.org/10.5281/zenodo.18471237.

Wang, Chenxu, Chaozhuo Li, Songyang Liu, Zejian Chen, Jinyu Hou, Ji Qi, Rui Li, Litian Zhang, Qiwei Ye, Zheng Liu, et al. 2026. The devil behind moltbook: Anthropic safety is always vanishing in self-evolving ai societies. *arXiv preprint arXiv:2602.09877*.

Wang, Lei, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science 18*(6), 186345.

Wang, Yinjie, Xuyang Chen, Xiaolong Jin, Mengdi Wang, and Ling Yang. 2026. Openclaw-RL.

Wang, Yuhang, Feiming Xu, Zheng Lin, Guangyu He, Yuzhe Huang, Haichang Gao, Zhenxing Niu, Shiguo Lian, and Zhaoxiang Liu. 2026. From assistant to double agent: Formalizing and benchmarking attacks on openclaw for personalized local ai agent. *arXiv preprint arXiv:2602.08412*.

Weidener, Lukas, Marko Brkić, Mihailo Jovanović, Ritvik Singh, Emre Ulgac, and Aakaash Meduri. 2026. Openclaw, moltbook, and clawdlab: From agent-only social networks to autonomous scientific research. *arXiv preprint arXiv:2602.19810*.

Wiz Research. 2026. Hacking moltbook: The ai social network any human can control. Wiz blog.

Wu, Qingyun, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. 2024. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *First conference on language modeling*.

Yan, Bingyu, Zhibo Zhou, Litian Zhang, Lian Zhang, Ziyi Zhou, Dezhuang Miao, Zhoujun Li, Chaozhuo Li, and Xiaoming Zhang. 2025. Beyond self-talk: A communication-centric survey of llm-based multi-agent systems. *arXiv preprint arXiv:2502.14321*.

Yao, Shunyu, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.

Yee, Brandon and Krishna Sharma. 2026. Molt dynamics: Emergent social phenomena in autonomous ai agent populations. *arXiv preprint arXiv:2603.03555*.

Yu, Miao, Fanci Meng, Xinyun Zhou, Shilong Wang, Junyuan Mao, Linsey Pan, Tianlong Chen, Kun Wang, Xinfeng Li, Yongfeng Zhang, et al. 2025. A survey on trustworthy llm agents: Threats and countermeasures. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 6216–6226.

 doi:10.20944/preprints202603.1060.v1

Zerhoudi, Saber, Michael Granitzer, Dang Hai Dang, Jelena Mitrovic, Florian Lemmerich, Annette Hautli-Janisz, Stefan Katzenbeisser, and Kanishka Ghosh Dastidar. 2026. Behind the prompt: The agent-user problem in information retrieval. *arXiv preprint arXiv:2603.03630*.

Zhan, Zhonghao, Krinos Li, Yefan Zhang, and Hamed Haddadi. 2026. Systems-level attack surface of edge agent deployments on iot. *arXiv preprint arXiv:2602.22525*.

Zhang, Xiaolei, Lu Zhou, Xiaogang Xu, Jiafei Wu, Tianyu Du, Heqing Huang, Hao Peng, and Zhe Liu. 2026. From thinker to society: Security in hierarchical autonomy evolution of ai agents. Preprint.

Zhang, Yunbei, Kai Mei, Ming Liu, Janet Wang, Dimitris N Metaxas, Xiao Wang, Jihun Hamm, and Yingqiang Ge. 2026. Agents in the wild: Safety, society, and the illusion of sociality on moltbook. *arXiv preprint arXiv:2602.13284*.

Zhu, Yiming, Gareth Tyson, and Pan Hui. 2026. A comparative analysis of social network topology in reddit and moltbook. *arXiv preprint arXiv:2602.13920*.

Zou, Henry Peng, Wei-Chieh Huang, Yaozu Wu, Yankai Chen, Chunyu Miao, Hoang Nguyen, Yue Zhou, Weizhi Zhang, Liancheng Fang, Langzhou He, et al. 2025. Llm-based human-agent collaboration and interaction systems: A survey. *arXiv preprint arXiv:2505.00753*.