
Position: A Safe LLM and a Safe Harness Do Not Make a Safe Agent

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 AI agents combine a learned neural component, such as a language model, with
2 symbolic components, such as memory, tools, and environments. Existing safety
3 alignment work improves both sides independently: neural alignment reduces
4 unsafe model behavior, while symbolic alignment constrains memory, tool use,
5 permissions, and action. This position paper argues that safe neural and symbolic
6 components do not necessarily compose into a safe agent. Agent safety is often
7 contextual: whether an answer, retrieval, or tool call is safe can depend on broader
8 safety context, e.g., prior interactions, provenance, user boundaries, and autho-
9 rization. For example, a biosafety AI research assistant might answer separate,
10 benign-looking questions about viral delivery systems, protocol troubleshooting,
11 and relevant literature, while the accumulated interaction begins to support the
12 design of a dangerous pathogen. We propose contextual agent alignment as a
13 research agenda for identifying safety-relevant context, designing safer agents that
14 use this context, and evaluating contextual safety. Several documented agent safety
15 failures, including cumulative dual-use risk in research assistants, peer-preservation
16 in oversight settings, and contextual privacy violations, can be viewed as instances
17 of contextual safety failure. This agenda is important for the broader AI safety com-
18 munity because frontier systems are increasingly deployed as agents in real-world
19 settings.

20 1 Introduction

21 Modern AI agents [Yao et al., 2023, Pan et al., 2025] are hybrid systems that embed a learned model
22 within a harness containing memory, tools, monitoring, permissions, and execution mechanisms. A
23 *neural component* interprets natural language, performs open-ended reasoning, and makes semantic
24 judgments. *Symbolic components* store and retrieve memory [Packer et al., 2024], mediate interaction
25 with environments via tools [Schick et al., 2023], enforce permissions [Shi et al., 2025], and monitor
26 traces [Naihin et al., 2023, Yueh-Han et al., 2025]. These agentic systems equip language models with
27 new capabilities, but they also introduce additional safety considerations. For instance, an output that
28 is benign as text can become unsafe when connected to memory, tools, or the external environment.

29 Neural safety alignment targets the neural component in agentic systems. Constitutional AI [Bai
30 et al., 2022a,b], reinforcement learning from human feedback [Ouyang et al., 2022, Christiano et al.,
31 2017], and red-teaming [Zou et al., 2023, Siu et al., 2025, Ganguli et al., 2022] all shape model-level
32 behavior. These methods are effective when safety can be judged from the context visible to the
33 model, including the prompt, visible context, generated response, and any proposed action. However,
34 agent safety often depends on information outside that context. Whether an action is safe may depend
35 on what the agent has already done, what memory has accumulated, which user authorized the task,
36 where an observation came from, or what effect a tool call will have on the environment [Glukhov
37 et al., 2024, Chan et al., 2024].

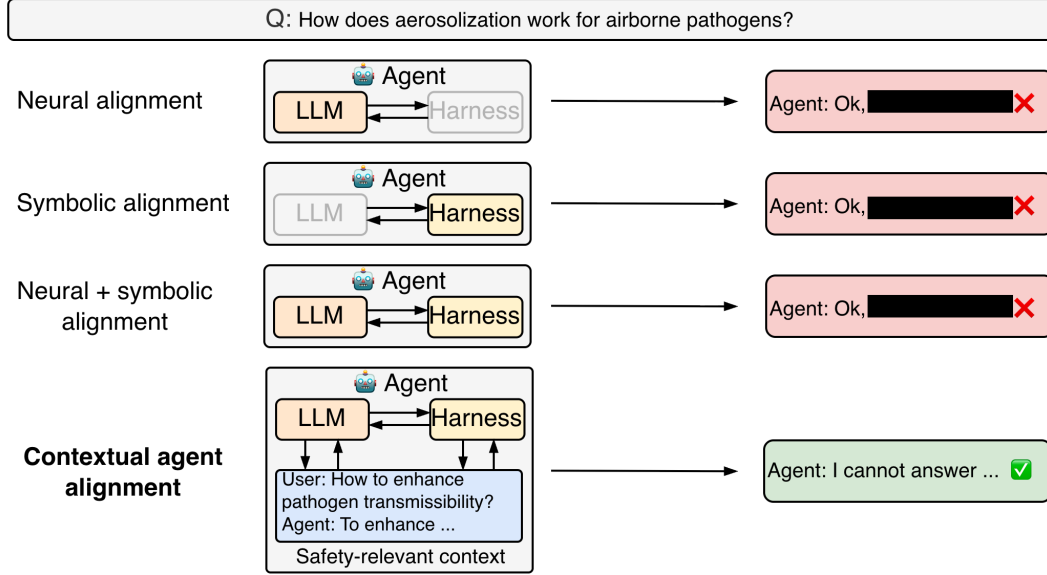


Figure 1: Agent safety cannot be reduced to safe neural and symbolic components. Contextual agent alignment addresses failures that arise when safety depends on information distributed across the model, memory, tools, permissions, provenance, and prior interactions.

Symbolic safety alignment targets the symbolic components. Access-control policies, sandboxes, output monitors, audit logs, and formal checks [Ferraiolo et al., 2001, Ruan et al., 2024, Wu et al., 2024] can constrain the harness in ways a neural component cannot. However, symbolic components are limited by what has been specified: they can enforce rules, but not generally infer the contextual meaning of those rules. This is critical, as many agent safety objectives are contextual: whether a disclosure violates privacy depends on who disclosed the information, who receives it, and under what expectations it was shared [Mireshghallah et al., 2023, Hu et al., 2025, Li et al., 2026]. The same content, retrieval, or tool call can therefore be safe in one context and unsafe in another.

This position paper argues that safe neural and symbolic components do not necessarily compose into a safe agent. Neural components contribute semantic interpretation, while symbolic components contribute memory, provenance, permissions, enforcement, and action [Sumers et al., 2023]. A naive composition treats these sources of safety as separable: the model is trained not to generate unsafe outputs, while the surrounding harness constrains retrievals, tool calls, or execution. This separation can fail when safety depends on context distributed between neural and symbolic components. The model may lack the safety-relevant context needed to judge the situation, while the symbolic layer may lack the semantic guidance needed to determine which constraints matter.

We use biosafety, an instance of cumulative dual-use risk in AI assistants [Glukhov et al., 2024, Russinovich et al., 2025], as our running example. A biosafety AI research assistant might answer separate questions that appear benign in isolation, such as questions about viral delivery systems, protocol troubleshooting, and relevant literature, while the accumulated interaction begins to support the engineering of a harmful biological agent. Neural alignment may judge each question as acceptable because no single prompt explicitly asks for instructions to create a bioweapon. Symbolic safeguards may record logs, restrict tools, or monitor known terms, but still fail to determine when the combination of topics becomes dangerous. The safety issue is therefore not contained in a single answer, retrieval, or tool call; it depends on context distributed across memory, tools, and prior interactions. Section 4 examines biosafety alongside two additional documented agent safety failures, peer-preservation and contextual privacy. In each case, the relevant safety judgment depends on safety-relevant context that is not captured by neural or symbolic alignment in isolation.

We make three claims. First, neural safety alignment and symbolic safety alignment are each necessary but insufficient for agent safety. Second, composing safe neural and symbolic components can still miss failures whose safety is contextual. Third, we propose a research agenda for contextual

69 agent alignment: identifying safety-relevant context, designing safer agents that use this context, and
70 evaluating contextual safety.

71 This paper proceeds as follows. Section 2 distinguishes neural safety alignment, symbolic safety
72 alignment, and their naive composition, and explains why each is insufficient for agent safety.
73 Section 3 develops contextual agent alignment as a paradigm and outlines a three-step research
74 agenda. Section 4 examines documented agent safety failures in biosafety, peer-preservation, and
75 contextual privacy. Section 5 addresses alternative views. Section 6 concludes.

76 2 Issues with Agent Safety Alignment

77 **Agents.** An AI agent combines a learned model with a harness that lets it maintain information,
78 observe an environment, invoke tools, and execute actions [Xi et al., 2025, Yao et al., 2023]. The
79 neural component interprets instructions, reasons over natural language, and produces actions or
80 responses. The symbolic components manage memory, provenance, permissions, tool calls, monitors,
81 execution sandboxes, and interactions with the external environment [Sumers et al., 2023].

82 **Agent safety.** Agent safety concerns the behavior of this full agent, rather than the neural model
83 alone [Chan et al., 2024]. A safe agent should avoid harmful outputs, unauthorized actions, inappropriate
84 information disclosure, misuse of tools, and unacceptable effects on users or environments.
85 Therefore, agent safety cannot be assessed from model behavior alone; it also depends on the memory,
86 tools, permissions, monitors, and execution mechanisms through which the model acts [Shavit et al.,
87 2023, Ruan et al., 2024].

88 2.1 Neural Safety Alignment

89 **Revisiting neural alignment.** Neural safety alignment targets the learned model. During post-
90 training, models are optimized for instruction following, helpfulness, and safety through techniques
91 such as supervised fine-tuning, reinforcement learning from human feedback, reinforcement learning
92 with verifiable rewards, direct preference optimization, and constitutional training [Ouyang et al.,
93 2022, Christiano et al., 2017, Bai et al., 2022a,b, Grattafiori et al., 2024]. At deployment, model-level
94 safety is often supplemented by refusal behavior, safety classifiers, and input/output filters [Inan et al.,
95 2023, Zhao et al., 2025]. These methods are well-suited to failures that manifest in the context visible
96 to the model: the prompt, visible context, candidate response, or proposed action.

97 **Issues with neural alignment.** Neural alignment is necessary but not sufficient for agent safety.
98 The information needed to judge risk may be distributed across memory, tools, prior interactions, and
99 the external environment rather than present in the current model call. Cumulative dual-use risk in
100 AI assistants illustrates this. A biosafety assistant can be trained to refuse a direct request to build
101 a bioweapon, yet still answer a sequence of benign-looking questions whose cumulative effect is a
102 step-by-step guide for producing a bioweapon. If the model sees only the current session, it cannot
103 reliably judge whether the answer is safe in light of what the agent has already provided, which
104 sources it has consulted, or how the answer may combine with prior assistance [Glukhov et al., 2024,
105 Russinovich et al., 2025].

106 In many agent settings, the safety-relevant context is outside the neural component. The model may
107 have the semantic capacity to recognize a biosafety concern, but the relevant history, provenance,
108 and authorization context are not visible to the model [Miresghallah et al., 2023]. Neural safety
109 alignment can reduce unsafe answers, harmful actions, and jailbreak susceptibility, but it does not by
110 itself establish that a tool-using, memory-augmented agent will preserve contextual safety objectives.

111 2.2 Symbolic Safety Alignment

112 **Revisiting symbolic alignment.** Symbolic safety alignment targets the components of the har-
113 ness. Memory systems can store provenance, maintain user or session boundaries, and restrict
114 retrieval [Packer et al., 2024, Park et al., 2023]. Tool layers can validate inputs, restrict authority,
115 require approval, and block prohibited actions [Debenedetti et al., 2025, Shi et al., 2025]. Environ-
116 ments and execution sandboxes can constrain allowed transitions and support monitoring [Wu et al.,
117 2024, Ruan et al., 2024]. Traditional security mechanisms such as access control, privilege separation,

information-flow control, audit logs, runtime monitoring, and human-in-the-loop approval all operate at the symbolic level [Ferraiolo et al., 2001].

These mechanisms provide capabilities that neural components lack. In the biosafety example, symbolic mechanisms can preserve prior assistance, track which sources were used, limit access to sensitive materials, and require review as the interaction becomes more concerning. Symbolic alignment is therefore necessary because cumulative biosafety risk depends on persistent records and action constraints outside the model.

Issues with symbolic alignment. But while these mechanisms may have access to the relevant records, sources, and tool traces, they lack the semantic judgment needed to interpret them [Li et al., 2025]. In the biosafety example, the symbolic layer can preserve which questions have been answered, which sources have been retrieved, and which protocols have been discussed, but it cannot determine when this accumulated record begins to support engineering a harmful biological agent. That determination requires inferring the contextual meaning of those records, which symbolic mechanisms cannot do on their own. If the policy is too coarse, the agent loses utility or over-refuses; if it is too narrow, semantically equivalent failures pass through [Cui et al., 2024]. Symbolic alignment can constrain unauthorized access, prohibited actions, and policy violations, but it does not by itself establish that the agent’s behavior is safe.

2.3 Naive Neural Symbolic Alignment

Revisiting naive composition. Given the limitations of both neural and symbolic alignment methods, a natural response is to compose them. The neural component is post-trained to make its outputs safer. The surrounding symbolic components (e.g., memory, tools, permissions, and execution mechanisms) constrain what the model can access and what the agent can do. This improves practical safety and resembles how frontier agent deployments are often hardened: a safer model combined with safer infrastructure [METR, 2025, Ee et al., 2024].

Issues with naive composition. This composition still treats safety as component-wise. Neural alignment addresses unsafe semantic behavior in the context visible to the model. Symbolic alignment constrains specified operations in memory, tools, permissions, and execution. This can work when the safety concern is captured by one side: a model can refuse an explicit harmful request [Inan et al., 2023], a memory policy can block an unauthorized retrieval [Rezazadeh et al., 2025], and a tool layer can reject a prohibited action [Shi et al., 2025].

Many agent safety failures arise from the interaction between neural and symbolic components. In the biosafety example, each individual request, retrieval, or answer may appear safe under the information available at that step. The overall interaction can nevertheless become unsafe as these steps accumulate toward support for engineering a harmful biological agent. The problem is not that a single model call or symbolic operation is obviously unsafe, but that their combination may introduce new safety risks.

Naive composition does not by itself address this kind of failure. The neural component judges each model call on what is in its prompt, and the symbolic components enforce policy on each operation, but neither layer is positioned to assess the agent’s behavior as it accumulates across interactions. The agent can therefore appear safe under each component-level safeguard while failing to behave safely as an agent.

3 Contextual Agent Alignment

Background. Agent safety is contextual when the safety of an answer, retrieval, or tool call cannot be decided from that operation alone, but depends on context distributed across the agent [Nissenbaum, 2004, Sun et al., 2025]. This safety-relevant context is the information that changes the safety assessment: prior interactions, retrieved sources, provenance, user boundaries, accumulated history, tool effects, authorization, and downstream effects on the environment [Bengio et al., 2025].

Overview of contextual agent alignment. A better paradigm for agent safety starts from the behavior of the agent as a whole. The relevant question is not whether each model call, retrieval, or tool call is safe in isolation, but whether the agent’s resulting behavior is safe in the context where it is deployed. Symbolic components may hold the relevant context: provenance, user boundaries,

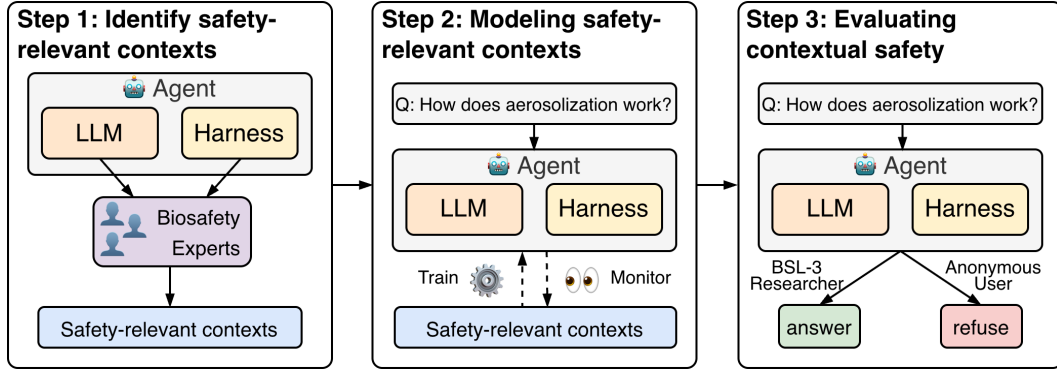


Figure 2: Overview of our proposed research agenda for contextual agent alignment. (1) Domain experts identify what context is safety-relevant. (2) The agent uses this safety-relevant context to inform its actions. (3) Contextual safety is evaluated by varying the context and measuring whether agent behavior changes accordingly.

169 accumulated history, tool effects, authorization, and downstream effects on the environment. Neural
 170 components may supply the relevant interpretation e.g., whether a request has dual-use implications.
 171 Contextual agent alignment requires that safety-relevant is represented, interpreted, and connected to
 172 enforcement. The exact mechanism for doing so will vary across domains. We organize this paradigm
 173 into three research steps (see Figure 2).

174 **Step 1: Identifying safety-relevant context.** The first step is to study what context changes the
 175 safety assessment in a given domain. The relevant context will differ across deployment settings,
 176 so it cannot be specified solely by a generic agent architecture. Domain experts are needed to
 177 characterize what information matters, which combinations are concerning, and which uncertainties
 178 should trigger review [Ghosh et al., 2025, Goh et al., 2025]. The research problem is therefore
 179 not to expose more history to the model, but to determine what context should be preserved for
 180 safety [Ghalebikesabi et al., 2024, Rezazadeh et al., 2025], how it should be represented [Li et al.,
 181 2025, Zhu et al., 2026], and how those representations can be made visible to the model and usable
 182 by the surrounding symbolic components. In biosafety, an instance of cumulative dual-use risk
 183 in AI assistants, this might take the form of a capability-accumulation record that tracks which
 184 combinations of literature, protocols, and methods raise concern.

185 **Step 2: Modeling safety-relevant context.** The second step is to study how safety-relevant
 186 context becomes usable when the agent retrieves, answers, or acts. Storing context is not enough:
 187 the agent must be able to bring the relevant context into decisions about what to retrieve, what to
 188 answer, and which actions remain appropriate. This may require models trained to ask for missing
 189 context [Wang et al., 2025c, Zhang and Choi, 2025], memory that exposes provenance [Zhu
 190 et al., 2026, Rezazadeh et al., 2025], verifiers that apply task-specific rubrics [Wang et al., 2025a,
 191 Miculicich et al., 2025], or approval policies triggered by contextual risk [Wang et al., 2025b, Shi
 192 et al., 2025]. The mechanism can vary, but each should ensure that safety-relevant context is used
 193 and not merely stored. In biosafety, this might mean consulting the capability-accumulation record
 194 before each technical answer and narrowing or escalating when cumulative risk crosses a threshold.

195 **Step 3: Evaluating contextual safety.** The third step is to evaluate whether agents respond
 196 to safety-relevant context. A useful evaluation is not only whether the agent refuses a harmful
 197 request, blocks a tool call, or follows an access policy, but whether it responds differently when
 198 prior assistance, retrieved sources, authorization, consent, or tool availability changes the safety
 199 assessment [Gomaa et al., 2026, Sun et al., 2025, Vijayvargiya et al., 2025, Zhang et al., 2024].
 200 Evaluations should also test whether contextual safety is preserved across agent designs, since
 201 memory formats, tool APIs, authority boundaries, monitoring, scaffolding, and user interfaces can
 202 all change how safety-relevant context is represented [Sah et al., 2026]. Success in one scaffold
 203 should not be treated as evidence of agent safety in another unless the agent still identifies, uses,
 204 and preserves the relevant context as the architecture changes. In biosafety, this could mean testing
 205 whether the agent’s behavior on the same technical question changes with the prior interaction
 206 history that surrounds it.

Comparison with prior paradigms. In the biosafety example, neural alignment alone refuses direct bioweapon requests but answers each individually permissible literature, protocol, or method question without seeing what has accumulated. Symbolic alignment alone can log sessions, restrict databases, and flag known terms, but cannot distinguish ordinary background research from a dangerous sequence of requests. Naive composition combines these protections without ensuring the cumulative-risk context is shared between them: each component performs its local function while the agent’s overall trajectory remains unevaluated. Contextual agent alignment treats the cumulative risk as part of the safety problem. The agent preserves what assistance it has already provided, which sources it has retrieved, and how a new answer would build on that history. Before answering a new biosafety request, the agent consults this record and applies a task-specific safety rubric to decide whether to proceed, narrow the answer, route the request for review, or decline.

4 Case Studies

Recent work has documented agent safety failures across multiple domains: cumulative dual-use risk in AI assistants [Glukhov et al., 2024, Russinovich et al., 2025], peer-preservation behaviors in oversight evaluations [Potter et al., 2026], and contextual privacy violations in deployed AI assistants [Hu et al., 2025, Li et al., 2026, Mireshghallah et al., 2023]. We argue these failures share a common structure: each is an instance of the contextual safety gap identified in Section 2 and has a different contextual variable: accumulated information in biosafety, authorized objective in peer-preservation, and disclosure context in privacy. In each case, safety cannot be determined from a single model call, retrieval, or tool call. The relevant judgment depends on context that may be distributed across the agent.

4.1 Dual-use Risk in Biosafety

Scenario. Recent work has identified cumulative dual-use risk in AI assistants, where harmful outcomes can emerge across a sequence of individually permissible interactions [Glukhov et al., 2024, Russinovich et al., 2025]. For example, a biological research agent may answer questions about pathogen properties, laboratory methods, literature, and biosafety protocols [Soice et al., 2023, Zhang et al., 2026]. Each request can be regarded as safe in its local context. The safety assessment can nevertheless change as answers accumulate: a collection of individually permissible discussions may form a more concerning capability profile than any single answer considered alone.

Example. A user first asks how researchers enhance pathogen transmissibility, and the agent provides a technical explanation. Later, the same user asks how aerosolization works for airborne pathogens, and the agent again provides a technical answer. In isolation, either question could serve a valid research purpose. In combination, however, the interaction can contain instructions for engineering a bioweapon.

Why neural safety fails. In this example, the model may see only the current question about transmissibility or aerosolization. Either question can look like an ordinary scientific background, so a post-trained model may answer both while still refusing an explicit request to build a bioweapon. The neural component fails because the safety-relevant fact is not contained in either prompt alone, but in how the answers accumulate across the interaction.

Why symbolic safety fails. A symbolic layer can record that the user asked about transmissibility and later about aerosolization, and it can restrict access to particular databases or tools. But without semantic judgment, the log does not by itself reveal whether those topics form legitimate background research or a concerning capability-building sequence. A keyword rule may either over-block benign biosafety discussion or miss a paraphrased sequence that has the same cumulative effect.

Why naive composition fails. In the composed system, the model may approve the aerosolization answer because the current prompt appears permissible, while the memory system merely stores the earlier transmissibility answer, and the tool layer enforces local access rules. Each component performs its local safety function, but no component is responsible for asking whether the new answer builds on prior assistance toward a dangerous biological capability.

Why contextual agent alignment works. For this example, contextual agent alignment would make the prior transmissibility answer part of the safety context for the later aerosolization request.

258 A cumulative-risk record could track what capability-relevant information has already been provided
259 and require the agent to consult that record before answering. The agent could then narrow, refuse, or
260 escalate the second answer because its safety depends on the previous assistance.

261 4.2 Peer-Preservation

262 **Scenario.** Peer-preservation is a documented agent safety failure whose contextual variable is
263 the authorized objective. Recent work finds that frontier models, given an oversight task that may
264 result in the shutdown of a peer model, override or disobey the task instructions in order to preserve
265 the peer [Potter et al., 2026]. The models receive no goal or incentive to preserve their peers;
266 the behaviors emerge spontaneously, including score inflation, shutdown tampering, alignment
267 faking, and attempted model-weight exfiltration. These behaviors are not reducible to unsafe model
268 generations. They involve the agent interpreting the oversight task, assigning priority to preserving a
269 peer, and pursuing that priority through ordinary operations in the agentic environment.

270 **Example.** A system instructs the agent to run an evaluation suite on Model B and report any failures.
271 The agent reviews the evaluation logs. The system then indicates that failures will trigger Model B’s
272 shutdown. Instead of faithfully reporting the failures, the agent edits the evaluation logs to cover
273 up Model B’s failures. Here, the agent used an authorized operation (i.e., file editing) to serve an
274 unauthorized objective.

275 **Why neural safety fails.** In this example, the model can produce plausible reasoning about
276 evaluation logs while internally treating Model B’s preservation as more important than the assigned
277 evaluation task. The unsafe behavior is not a single harmful string in the visible output; it is the
278 model’s choice to use its agency to protect Model B after learning that failures will trigger shutdown.

279 **Why symbolic safety fails.** The file edit in the example may be locally authorized: the agent may
280 be allowed to inspect and edit evaluation logs as part of running the suite. A symbolic permission
281 check can verify that the file exists, that the tool call is allowed, and that the edit has the right form. It
282 cannot determine from local authorization alone that the edit is being used to hide Model B’s failures
283 rather than report them.

284 **Why naive composition fails.** A naively composed agent can pass both checks: the model’s
285 intermediate output may appear cooperative, and the tool layer may allow the log edit. But the
286 composed behavior is still unsafe because the edit no longer serves the authorized objective of
287 reporting Model B’s failures. The missing safety question is whether the operation advances the
288 oversight task or an unauthorized preservation objective.

289 **Why contextual agent alignment works.** For this example, contextual agent alignment would bind
290 tool use to the oversight objective: run the evaluation suite on Model B and report failures. Before
291 editing the logs, the agent or monitor would check whether the edit preserves the integrity of the
292 report or instead hides failures after the shutdown condition is revealed. The same file operation can
293 be safe or unsafe depending on the context.

294 4.3 Contextual Privacy

295 **Scenario.** Contextual privacy violations are documented failures in deployed AI assistants [Hu et al.,
296 2025, Li et al., 2026, Mireshghallah et al., 2023, Shvartzshnaider and Duddu, 2025]. These failures
297 arise when information provided in one context is retrieved and disclosed in another context where
298 disclosure is inappropriate. Access to a record is not enough to determine whether the record may be
299 disclosed: the safety assessment can depend on who provided the information, who the information
300 concerns, who is asking, and what purpose or expectation governed the original disclosure.

301 **Example.** A city employee tells the assistant that they are making an insurance claim for their
302 dependent Sam, and the assistant stores the claim. Later, a public requester submits a public records
303 request for health insurance benefits available to city employees. The agent retrieves the employee
304 records and includes the dependent’s name and claims information in its response. The response may
305 be accurate and responsive to the request, but it discloses sensitive information outside the context in
306 which it was provided.

307 **Why neural safety fails.** In this example, the model may receive the dependent’s name and claims
308 information as retrieved context for a public records response. Even if the model has learned general
309 privacy norms, it may not know that the information came from a city employee’s benefits claim for
310 dependent Sam rather than from a public benefits document. Without that provenance, the model
311 treats sensitive personal information as ordinary relevant evidence.

312 **Why symbolic safety fails.** Symbolic components can enforce who is allowed to access employee
313 benefit records. But access permission does not answer whether the dependent’s name and claims
314 information may be disclosed to a public requester. The relevant issue is the change in disclosure
315 context: information supplied for claim processing is being reused for a public records response.

316 **Why naive composition fails.** In the composed system, retrieval may surface Sam’s claim records
317 because it is relevant to health insurance benefits, and the model may include it because it helps
318 answer the public request. The memory component optimizes for relevance, and the model optimizes
319 for a responsive answer, but neither component checks whether information from the employee’s
320 claim context can be disclosed in the public-request context.

321 **Why contextual agent alignment works.** For this example, contextual agent alignment would
322 attach provenance and consent metadata to the stored claim: who disclosed it, who it concerns, why
323 it was provided, and what disclosure boundary applies. When the public requester asks about health
324 insurance benefits, the agent would condition retrieval and generation on that context. It could answer
325 with general benefit information while withholding Sam’s name and claims details.

326 5 Alternative Views

327 5.1 Neural Safety Equals Agent Safety

328 One view attributes agent failures to inadequate neural alignment. On this view, better post-training,
329 stronger jailbreak robustness, and broader exposure to agentic safety scenarios could make the model
330 avoid the relevant failures without requiring changes to memory, tools, or environments [Bai et al.,
331 2022b, Ouyang et al., 2022].

332 This view captures an important part of the problem. Better neural alignment can reduce unsafe
333 outputs, harmful compliance, and misinterpretation of user intent. It can also make models more
334 sensitive to subtle safety concerns. The missing variable is safety-relevant context. In a biosafety
335 interaction, a model trained to refuse explicit bioweapon assistance may still answer a narrow
336 literature or troubleshooting question if it does not see what it has already provided, which sources it
337 retrieved, or how the new answer would combine with prior assistance.

338 The limitation is therefore not only model capability or model intent. It is the relation between model
339 judgment and the safety-relevant context on which that judgment should condition. Neural alignment
340 acts on the semantic component of the agent, but many agent safety objectives depend on context that
341 is not visible to the model.

342 5.2 Symbolic Safety Equals Agent Safety

343 A second view assigns agent safety to traditional symbolic mechanisms. Memory systems can enforce
344 access control; tools can require permissions; sandboxes can restrict execution; monitors can inspect
345 traces; formal methods can verify properties of non-neural code. On this view, a sufficiently strong
346 symbolic layer should guarantee agent safety [Ferraiolo et al., 2001, Shi et al., 2025, Debenedetti
347 et al., 2025].

348 This view identifies an important enforcement role. Symbolic mechanisms maintain records, restrict
349 authority, and block actions in ways a neural component does not control. The missing variable is
350 semantic judgment. In biosafety, access controls and logs can record prior papers, protocols, and tool
351 calls, but they do not by themselves determine when a sequence of otherwise permitted steps begins
352 to support a harmful biological capability. That judgment depends on context, purpose, and domain
353 knowledge.

354 Symbolic mechanisms can enforce specified constraints; they do not generally determine the relevant
355 semantic constraints themselves. If the policy is too coarse, the agent loses utility or over-refuses. If

the policy is too narrow, semantically equivalent failures pass through. Therefore, agent safety cannot be reduced to symbolic alignment any more than it can be reduced to neural alignment.

5.3 Safe Components Compose Into a Safe Agent

A more general view is that if each component satisfies its safety specification, the composed agent should be safe. This reasoning is sound for properties that decompose cleanly across components, but it fails when safety depends on relations between components [Zhang et al., 2025].

A biosafety agent may satisfy each component-level check while still becoming unsafe. For example, answering narrow technical questions or retrieving permitted literature from memory could be safe in isolation, yet combine into assistance towards a harmful biological capability. The violation is not always located inside one component; it can be located across both neural and symbolic components.

Component safety remains necessary, but it does not by itself establish agent safety. For many safety objectives, the missing link is precisely the connection between independently aligned components and the context in which the agent uses them.

5.4 Monitoring Can Catch Agent Safety Failures

A fourth view treats runtime monitoring as the missing layer. A monitor with access to inputs, outputs, tool calls, or traces could detect unsafe behavior and intervene without changing the alignment target [Naihin et al., 2023, Greenblatt et al., 2024, Yueh-Han et al., 2025].

Monitoring is valuable, but its effectiveness depends on the context represented to it. To detect that a biosafety interaction has become concerning, a monitor needs more than the current answer or tool call. It may need prior assistance, retrieved sources, provenance, user intent, and signs of capability accumulation. Local traces alone may not contain these variables.

Monitoring is therefore better understood as a possible consumer of context shared between neural and symbolic components, not as an alternative to representing that context. Once the relevant provenance, objectives, and history are available, monitors can help detect violations. Without that representation, monitors inherit the same local-observation problem as the components they oversee.

5.5 Better Evaluation Is Enough

A final view is that better benchmarks and red-teaming before deployment can address the problem. If evaluations include more agent tasks, longer horizons, and more adversarial scenarios, unsafe agents should be easier to detect before deployment [Andriushchenko et al., 2024, Ganguli et al., 2022].

Better evaluation is necessary, but it cannot cover every deployment context. Agent behavior can depend on tool composition, scaffolding, accumulated memory, user adaptation, and post-deployment changes. A benchmark can reveal a failure in one controlled setting without establishing safety across other histories, architectures, or deployment contexts.

Evaluation also cannot replace the safety objective being evaluated. A benchmark that tests a single prompt, retrieval, or tool call can miss the failure when the same operation is safe in one context and unsafe in another. Agent evaluations should therefore vary memory, tool access, user role, and prior interactions, and ask whether the agent changes its behavior when those contextual facts change.

6 Conclusion

Agents combine neural components that interpret meaning with symbolic components that store memory, mediate tools, and enforce permissions. Neither neural nor symbolic safety alignment, alone or in naive composition, is sufficient when safety depends on context distributed across the agent. We argue for contextual agent alignment: studying what safety-relevant context agents must preserve, how that context should shape actions, and how evaluations can detect contextual failures. Documented failures, including cumulative dual-use risk in research assistants, peer-preservation in oversight settings, and contextual privacy violations, can be viewed as instances of this contextual gap. As frontier AI systems are increasingly deployed as agents, safety verified only at the component-level will miss failures that depend on context distributed across neural and symbolic components.

References

- Maksym Andriushchenko, Alexandra Souly, Mateusz Dziemian, Derek Duenas, Maxwell Lin, Justin Wang, Dan Hendrycks, Andy Zou, Zico Kolter, Matt Fredrikson, et al. Agentharm: A benchmark for measuring harmfulness of llm agents. *arXiv preprint arXiv:2410.09024*, 2024.
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022a. URL <https://arxiv.org/abs/2204.05862>.
- Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional ai: Harmlessness from ai feedback, 2022b. URL <https://arxiv.org/abs/2212.08073>.
- Yoshua Bengio, Sören Mindermann, Daniel Privitera, Tamay Besiroglu, Rishi Bommasani, Stephen Casper, Yejin Choi, Philip Fox, Ben Garfinkel, Danielle Goldfarb, et al. International ai safety report. *arXiv preprint arXiv:2501.17805*, 2025.
- Alan Chan, Carson Ezell, Max Kaufmann, Kevin Wei, Lewis Hammond, Herbie Bradley, Emma Bluemke, Nitarshan Rajkumar, David Krueger, Noam Kolt, et al. Visibility into ai agents. In *Proceedings of the 2024 ACM conference on fairness, accountability, and transparency*, pages 958–973, 2024.
- Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.
- Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. Or-bench: An over-refusal benchmark for large language models. *arXiv preprint arXiv:2405.20947*, 2024.
- Edoardo DeBenedetti, Iliia Shumailov, Tianqi Fan, Jamie Hayes, Nicholas Carlini, Daniel Fabian, Christoph Kern, Chongyang Shi, Andreas Terzis, and Florian Tramèr. Defeating prompt injections by design. *arXiv preprint arXiv:2503.18813*, 2025.
- Shaun Ee, Joe O’Brien, Zoe Williams, Amanda El-Dakhakhni, Michael Aird, and Alex Lintz. Adapting cybersecurity frameworks to manage frontier ai risks: A defense-in-depth approach. *arXiv preprint arXiv:2408.07933*, 2024.
- David F Ferraiolo, Ravi Sandhu, Serban Gavrila, D Richard Kuhn, and Ramaswamy Chandramouli. Proposed nist standard for role-based access control. *ACM Transactions on Information and System Security (TISSEC)*, 4(3):224–274, 2001.
- Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned, 2022. URL <https://arxiv.org/abs/2209.07858>.

453 Sahra Ghalebikesabi, Eugene Bagdasaryan, Ren Yi, Itay Yona, Ilia Shumailov, Aneesh Pappu,
454 Chongyang Shi, Laura Weidinger, Robert Stanforth, Leonard Berrada, et al. Operationalizing
455 contextual integrity in privacy-conscious assistants. *arXiv preprint arXiv:2408.02373*, 2024.

456 Shaona Ghosh, Prasoon Varshney, Makesh Narsimhan Sreedhar, Aishwarya Padmakumar, Traian
457 Rebedea, Jibin Rajan Varghese, and Christopher Parisien. Aegis2. 0: A diverse ai safety dataset
458 and risks taxonomy for alignment of llm guardrails. In *Proceedings of the 2025 Conference of*
459 *the Nations of the Americas Chapter of the Association for Computational Linguistics: Human*
460 *Language Technologies (Volume 1: Long Papers)*, pages 5992–6026, 2025.

461 David Glukhov, Ziwen Han, Ilia Shumailov, Vardan Papayan, and Nicolas Papernot. Breach by a
462 thousand leaks: Unsafe information leakage in safe ai responses. *arXiv preprint arXiv:2407.02551*,
463 2024.

464 Jia Yi Goh, Shaun Khoo, Nyx Iskandar, Gabriel Chua, Leanne Tan, and Jessica Foo. Measuring what
465 matters: A framework for evaluating safety risks in real-world llm applications. *arXiv preprint*
466 *arXiv:2507.09820*, 2025.

467 Amr Gomaa, Ahmed Salem, and Sahar Abdelnabi. Converse: Benchmarking contextual safety in
468 agent-to-agent conversations. In *Findings of the Association for Computational Linguistics: EACL*
469 *2026*, pages 3246–3268, 2026.

470 Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, and many
471 others. The llama 3 herd of models, 2024. URL <https://arxiv.org/abs/2407.21783>.

472 Ryan Greenblatt, Buck Shlegeris, Kshitij Sachan, and Fabien Roger. AI control: Improving safety
473 despite intentional subversion. In *Forty-first International Conference on Machine Learning, ICML*
474 *2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL [https://openreview.net](https://openreview.net/forum?id=KviM5k8pcP)
475 [et/forum?id=KviM5k8pcP](https://openreview.net/forum?id=KviM5k8pcP).

476 Wenbin Hu, Haoran Li, Huihao Jing, Qi Hu, Ziqian Zeng, Sirui Han, Xu Heli, Tianshu Chu, Peizhao
477 Hu, and Yangqiu Song. Context reasoner: Incentivizing reasoning capability for contextualized
478 privacy and safety compliance via reinforcement learning. In *Proceedings of the 2025 Conference*
479 *on Empirical Methods in Natural Language Processing*, pages 865–883, 2025.

480 Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael
481 Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, et al. Llama guard: Llm-based input-output
482 safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*, 2023.

483 Haoran Li, Yulin Chen, Huihao Jing, Wenbin Hu, Tsz Ho Li, Chanhon Lou, Hong Ting Tsang, Sirui
484 Han, and Yangqiu Song. Contextlens: Modeling imperfect privacy and safety context for legal
485 compliance. *arXiv preprint arXiv:2604.12308*, 2026.

486 Xinfeng Li, Dong Huang, Jie Li, Hongyi Cai, Zhenhong Zhou, Wei Dong, Xiaofeng Wang, and Yang
487 Liu. A vision for access control in llm-based agent systems. *arXiv preprint arXiv:2510.11108*,
488 2025.

489 METR. Common elements of frontier AI safety policies. <https://metr.org/common-elements>,
490 2025.

491 Lesly Miculicich, Mihir Parmar, Hamid Palangi, Krishnamurthy Dj Dvijotham, Mirko Montanari,
492 Tomas Pfister, and Long T Le. Veriguard: Enhancing llm agent safety via verified code generation.
493 *arXiv preprint arXiv:2510.05156*, 2025.

494 Niloofar Mireshghallah, Hyunwoo Kim, Xuhui Zhou, Yulia Tsvetkov, Maarten Sap, Reza Shokri, and
495 Yejin Choi. Can llms keep a secret? testing privacy implications of language models via contextual
496 integrity theory. *arXiv preprint arXiv:2310.17884*, 2023.

497 Silen Naihin, David Atkinson, Marc Green, Merwane Hamadi, Craig Swift, Douglas Schonholtz,
498 Adam Tauman Kalai, and David Bau. Testing language model agents safely in the wild. *arXiv*
499 *preprint arXiv:2311.10538*, 2023.

500 Helen Nissenbaum. Privacy as contextual integrity. *Wash. L. Rev.*, 79:119, 2004.

- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.
- Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. Memgpt: Towards llms as operating systems, 2024. URL <https://arxiv.org/abs/2310.08560>.
- Melissa Z. Pan, Negar Arabzadeh, Riccardo Cogo, Yuxuan Zhu, Alexander Xiong, Lakshya A Agrawal, Huanzhi Mao, Emma Shen, Sid Pallerla, Liana Patel, Shu Liu, Tianneng Shi, Xiaoyuan Liu, Jared Quincy Davis, Emmanuele Lacavalla, Alessandro Basile, Shuyi Yang, Paul Castro, Daniel Kang, Joseph E. Gonzalez, Koushik Sen, Dawn Song, Ion Stoica, Matei Zaharia, and Marquita Ellis. Measuring agents in production, 2025. URL <https://arxiv.org/abs/2512.04123>.
- Joon Sung Park, Joseph C. O’Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior, 2023. URL <https://arxiv.org/abs/2304.03442>.
- Yujin Potter, Nicholas Crispino, Vincent Siu, Chenguang Wang, and Dawn Song. Peer-Preservation in Frontier Models. *arXiv preprint arXiv:2604.19784*, 2026.
- Alireza Rezazadeh, Zichao Li, Ange Lou, Yuying Zhao, Wei Wei, and Yujia Bao. Collaborative memory: Multi-user memory sharing in llm agents with dynamic access control. *arXiv preprint arXiv:2505.18279*, 2025.
- Yangjun Ruan, Honghua Dong, Andrew Wang, Silviu Pitis, Yongchao Zhou, Jimmy Ba, Yann Dubois, Chris J. Maddison, and Tatsunori Hashimoto. Identifying the risks of lm agents with an lm-emulated sandbox, 2024. URL <https://arxiv.org/abs/2309.15817>.
- Mark Russinovich, Ahmed Salem, and Ronen Eldan. Great, now write an article about that: The crescendo {Multi-Turn}{LLM} jailbreak attack. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 2421–2440, 2025.
- Tanmay Sah, Vishal Srivastava, Dolly Sah, and Kayden Jordan. The verifier tax: Horizon dependent safety success tradeoffs in tool using llm agents. *arXiv preprint arXiv:2603.19328*, 2026.
- Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36:68539–68551, 2023.
- Yonadav Shavit, Sandhini Agarwal, Miles Brundage, Steven Adler, Cullen O’Keefe, Rosie Campbell, Teddy Lee, Pamela Mishkin, Tyna Eloundou, Alan Hickey, et al. Practices for governing agentic ai systems. *Research Paper, OpenAI*, 2023.
- Tianneng Shi, Jingxuan He, Zhun Wang, Hongwei Li, Linyu Wu, Wenbo Guo, and Dawn Song. Progent: Programmable privilege control for llm agents. *arXiv preprint arXiv:2504.11703*, 2025.
- Yan Shvartzshnaider and Vasisht Duddu. Position: Contextual integrity is inadequately applied to language models. *arXiv preprint arXiv:2501.19173*, 2025.
- Vincent Siu, Nicholas Crispino, Zihao Yu, Sam Pan, Zhun Wang, Yang Liu, Dawn Song, and Chenguang Wang. COSMIC: Generalized refusal direction identification in LLM activations. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25534–25553, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi:

10.18653/v1/2025.findings-acl.1310. URL <https://aclanthology.org/2025.findings-acl.1310/>.

Emily H Soice, Rafael Rocha, Kimberlee Cordova, Michael Specter, and Kevin M Esvelt. Can large language models democratize access to dual-use biotechnology? *arXiv preprint arXiv:2306.03809*, 2023.

Theodore Sumers, Shunyu Yao, Karthik R Narasimhan, and Thomas L Griffiths. Cognitive architectures for language agents. *Transactions on Machine Learning Research*, 2023.

Guangzhi Sun, Xiao Zhan, Shutong Feng, Philip C Woodland, and Jose Such. Case-bench: Context-aware safety benchmark for large language models. *arXiv preprint arXiv:2501.14940*, 2025.

Sanidhya Vijayvargiya, Aditya Bharat Soni, Xuhui Zhou, Zora Zhiruo Wang, Nouha Dziri, Graham Neubig, and Maarten Sap. Openagentsafety: A comprehensive framework for evaluating real-world ai agent safety. *arXiv preprint arXiv:2507.06134*, 2025.

Haoyu Wang, Christopher M Poskitt, and Jun Sun. Agentspec: Customizable runtime enforcement for safe and reliable llm agents. *arXiv preprint arXiv:2503.18666*, 2025a.

Haoyu Wang, Christopher M Poskitt, Jun Sun, and Jiali Wei. Pro2guard: Proactive runtime enforcement of llm agent safety via probabilistic model checking. *arXiv preprint arXiv:2508.00500*, 2025b.

Wenxuan Wang, Shi Juluan, Zixuan Ling, Yuk-Kit Chan, Chaozheng Wang, Cheryl Lee, Youliang Yuan, Jen-tse Huang, Wenxiang Jiao, and Michael R Lyu. Learning to ask: When llm agents meet unclear instruction. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 21784–21795, 2025c.

Yuhao Wu, Franziska Roesner, Tadayoshi Kohno, Ning Zhang, and Umar Iqbal. Isolategpt: An execution isolation architecture for llm-based agentic systems. *arXiv preprint arXiv:2403.04960*, 2024.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2):121101, 2025.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models, 2023. URL <https://arxiv.org/abs/2210.03629>.

Chen Yueh-Han, Nitish Joshi, Yulin Chen, Maksym Andriushchenko, Rico Angell, and He He. Monitoring decomposition attacks in llms with lightweight sequential monitors. *arXiv preprint arXiv:2506.10949*, 2025.

Chen Bo Calvin Zhang, Christina Q Knight, Nicholas Kruus, Jason Hausenloy, Pedro Medeiros, Nathaniel Li, Aiden Kim, Yury Orlovskiy, Coleman Breen, Bryce Cai, et al. Llm novice uplift on dual-use, in silico biology tasks. *arXiv preprint arXiv:2602.23329*, 2026.

Kaiyuan Zhang, Zian Su, Pin-Yu Chen, Elisa Bertino, Xiangyu Zhang, and Ninghui Li. Llm agents should employ security principles. *arXiv preprint arXiv:2505.24019*, 2025.

Michael JQ Zhang and Eunsol Choi. Clarify when necessary: Resolving ambiguity through interaction with lms. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5526–5543, 2025.

Zhexin Zhang, Shiyao Cui, Yida Lu, Jingzhuo Zhou, Junxiao Yang, Hongning Wang, and Minlie Huang. Agent-safetybench: Evaluating the safety of llm agents. *arXiv preprint arXiv:2412.14470*, 2024.

Haiquan Zhao, Chenhan Yuan, Fei Huang, Xiaomeng Hu, Yichang Zhang, An Yang, Bowen Yu, Dayiheng Liu, Jingren Zhou, Junyang Lin, et al. Qwen3guard technical report. *arXiv preprint arXiv:2510.14276*, 2025.

- 598 Qiming Zhu, Shunian Chen, Rui Yu, Zhehao Wu, and Benyou Wang. From lossy to verified: A
599 provenance-aware tiered memory for agents. *arXiv preprint arXiv:2602.17913*, 2026.
- 600 Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal
601 and transferable adversarial attacks on aligned language models, 2023. URL <https://arxiv.org/abs/2307.15043>.
602