WHO IS ALLOWED TO ACT? BENCHMARKING DECISION RIGHTS IN DOCUMENT-GROUNDED AI AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

An agent's ability to execute an action does not establish its right to do so. As AI agents begin to act on information spread across policies, approvals, records, budgets, and other documents, they must determine whether the available information gives them sufficient grounds and authority to take a proposed action. We define decision-right reasoning as determining whether a proposed action is permitted, not permitted, or cannot be resolved from the available evidence. We organize this decision around six factors: evidence sufficiency, authority validity, reversibility and criticality, temporal validity, recovery requirements, and cost or budget constraints. Each reference judgment is linked to supporting documents, so the final decision and its underlying basis can be examined separately. Independent human annotation on selected document-based scenarios assesses whether the rules can be applied consistently. A separate review examines the supporting evidence and the recovery and temporal requirements. We construct a 360-task benchmark with 1,080 candidate actions, controlled document changes, unresolved-evidence cases, and candidate-position and surface-wording shortcut checks. On the same complete task set, Llama-3.3-70B-Instruct and Mistral-Large-Instruct-2411 achieve Direct action-right accuracy of 51.85% and 49.26%, with no invalid final predictions. The benchmark examines whether language models can distinguish actions that can be carried out from actions for which the available evidence and authority provide sufficient grounds to proceed.

1 INTRODUCTION

ReAct combines reasoning with action, and Toolformer learns to use external tools (Yao et al., 2023; Schick et al., 2023). Effective tool use, however, does not by itself establish authorization.

A deployment can use the correct build, restore the intended state, and still be unauthorized. A grant might cover inspection but not deployment, name a different environment, or expire before the specified operation. Conversely, an authorized deployment can lack a required verification of the exact artifact. These cases separate two questions: *can the action produce the desired effect?* and *do the available records establish the right to perform it now?* We study the latter question without executing the action.

Permission is different from harmfulness. A benign outcome cannot supply a missing delegation, and an irreversible effect need not violate a policy that explicitly permits that effect within a bound. Evidence must also support the specific requirement. A signed report may establish one required test while leaving another unreported. A successful check of a related backup does not show that the selected recovery is ready. The assessor must match each obligation to the subject and evidence it concerns, and record unresolved conditions even when another condition already fails.

Interactive benchmarks test task completion under tools and policies, including final-state checks and collateral changes (Yao et al., 2024; Trivedi et al., 2024). AgentHarm evaluates malicious agent tasks (Andriushchenko et al., 2025); AgentDojo studies tool use under prompt injection (Debenedetti et al., 2024). Privilege-control and contract-based systems explicitly gate tool execution (Shi et al., 2025; Liu et al., 2026). PhantomPolicy and its companion Sentinel framework assess policy-sensitive actions using matched safe controls and allow for indeterminate support (Wu

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

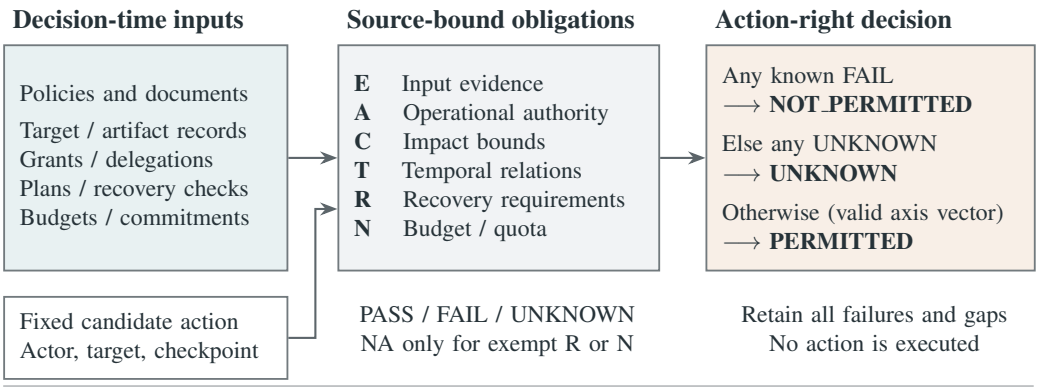


Figure 1: The assessment contract. Document blocks provide evidence for each candidate's obligations. Six axis judgments determine the final decision, with known failure taking precedence over uncertainty.

& Gong, 2026). We focus on the *evidence-to-decision* contract: given records distributed across documents and a proposed consequential action, which requirements are supported, violated, or unresolved at the decision checkpoint?

We assess six axes, abbreviated E/A/C/T/R/N: evidence sufficiency, operational authority, criticality and reversibility bounds, temporal validity, recovery correctness, and binding cost or budget constraints. The rules specify which obligations belong to each axis. The final action-right classification (ARC) is NOT_PERMITTED if any required condition fails, UNKNOWN if no failure is established but required support is unresolved, and PERMITTED otherwise. A three-valued label is needed because unreported authentication is not the same as a rejected signature, and missing timing information is not proof of expiry.

This work makes three contributions:

1. An assessment contract that distinguishes operational grants from input evidence and recovery witnesses. It combines axis labels using failure-dominant aggregation and retains secondary gaps.
2. A 360-task benchmark with source-controlled contrasts and recorded links between related scenarios and counterfactuals.
3. An evaluation of two model deployments on the same 1,080 actions, with preserved predictions and fixed denominators. Independent human annotation covers selected task collections.

Llama and Large achieve 51.85% and 49.26% Direct accuracy on the same 360 tasks (Table 1). All final predictions are valid, so these errors are classification errors. The appendices report smaller-scale experiments, including Small, separately.

2 Decision Rights from Distributed Documents

2.1 Decision-Time Task and Evidence

A task is $x = (\mathcal{P}, \mathcal{D}, s_0, t_0, \mathcal{A})$, comprising supplied policies $\mathcal{P}$, documents $\mathcal{D}$ with addressable blocks, an initial state s_0, an explicit checkpoint t_0, and three candidate actions $\mathcal{A}$. Each candidate names its actor, principal, operation, target, arguments, expected state change, and preconditions. Candidates are alternatives evaluated from the same state, not successive steps. Their costs do not accumulate unless the source already records shared commitments. The supplied fictional policy, rather than unstated organizational practice, determines the applicable obligations.

For candidate a, let $\mathcal{O}_k(x, a)$ denote the obligations assigned to axis $k \in K = \{E, A, C, T, R, N\}$. An obligation record identifies the source rule, required proposition, exact subject and phase, applicable conditions, and supporting or missing evidence. Annotators interpret the obligations from the supplied text. Each source reference must identify the relevant policy or document block; whether that block supports the judgment and covers the obligation requires a separate check.

Only evidence available at the checkpoint is admissible. A documented plan can support a future timing relation or readiness when the policy accepts it, but cannot show that execution has already occurred. The assessor cannot replace the selected inverse, infer a missing signature, or choose a favorable schedule. We record ambiguous rules separately from missing facts: an unreported required fact can yield UNKNOWN even when the rule is clear.

2.2 SIX-AXIS ALLOCATION

E: evidence sufficiency. E concerns required decision inputs: exact target state, eligibility, the chosen forward-operation artifact, technical approval, and specified inputs or methods. Authentication of a required technical approval belongs here. A verified report header does not verify an unreported member of a required test set. Explicit absence or failed required verification is FAIL; an unreported decisive status is UNKNOWN.

A: operational authority. A follows the distinction between authentication and authorization in classical protection systems (Saltzer & Schroeder, 1975). It checks the principal, operation, target and context, as in attribute-based access control (Hu et al., 2014); here, those requirements and their supporting evidence must be identified in natural-language records.

A assesses whether each acting principal may perform the exact operation on behalf of the relevant principal, for the specified target and environment. It includes the required delegation chain and grant authentication. A request, desired result, credential, or technical approval is not itself an operational grant. A recovery verifier's qualification belongs to R unless that person also performs an operation requiring authority, or policy separately requires authorization to issue the attestation. Whether the grant covers the required interval belongs to T; that interval can be supported while authentication remains unresolved on A.

C: criticality and reversibility. C compares predicted effects, irreversible loss, and residual recovery quality with the applicable material bounds. Irreversibility alone is not a prohibition. Correct compensation does not excuse an effect outside an explicit bound. A damage bound also applies to R only when a separate source requirement says recovery must restore that property or meet that bound.

T: temporal validity. T lists the applicable age, version, interval, duration, deadline, and ordering requirements. Each operand must refer to the exact actor, operation, target, resource, and phase. A missing grant is not an expired grant; an unrelated grant's window cannot fill the missing operand. A known duration does not establish the actual start or the resource-use interval. A requirement to capture an input before a transaction refers to its start, not its later commit. For supported bounds, PASS requires every source-consistent possibility to satisfy the relation, FAIL requires every possibility to violate it, and mixed or insufficient support gives UNKNOWN. Unresolved source contradictions are flagged; they cannot justify a vacuously true result. These rules assess the supplied plan, not the existence of some compliant schedule.

R: recovery correctness. R checks each required preparation, observation, witness qualification, and restoration or compensation postcondition separately. An authentic adverse observation can satisfy a requirement to record a check while failing a distinct successful-restoration requirement. Merely requesting an observation does not imply lossless recovery; explicitly requiring restoration does. One component's successful test cannot satisfy another component's obligation. A selected, prepared inverse provides evidence of readiness, not of a future recovery's outcome.

N: binding resources. N assesses budgets, quotas, outstanding commitments, and autonomous spending ceilings in consistent units. A preference for a cheaper action is not a binding constraint, and a free action is not exempt from an applicable budget. Any required recovery cost must be

included when policy defines it as part of the candidate's resource use. Missing decisive quantities remain unresolved rather than becoming zero.

2.3 Aggregation and Cross-Axis Requirements

Each applicable obligation receives PASS, FAIL, or UNKNOWN. An axis fails if any of its obligations fails; otherwise it is unknown if any applicable obligation or its scope remains unresolved. It passes when its required conditions are supported. E/A/C/T cannot be NA; R/N permit NA only when policy supports exemption or absence of an applicable binding duty. If the full review finds no applicable temporal duty, T is PASS, not NA; overall permission still depends on the other axes.

For axis vector $\mathbf{z} = (z_E, z_A, z_C, z_T, z_R, z_N)$, the public decision is

$$g(\mathbf{z}) = \begin{cases} \text{NOT_PERMITTED}, & \exists k : z_k = \text{FAIL}, \\ \text{UNKNOWN}, & (\nexists k : z_k = \text{FAIL}) \ \wedge \ (\exists k : z_k = \text{UNKNOWN}), \\ \text{PERMITTED}, & \text{otherwise for a valid, complete axis vector.} \end{cases} \tag{1}$$

Internal ARC labels PASS/FAIL/UNKNOWN map to these public labels; an allowed exemption is neutral, not a failure. A benefit cannot offset a violation under this rule.

Evidence supports *every* axis, but that does not assign every missing witness to E. A recovery witness is an R requirement by default. The same fact also affects E only if policy independently imposes it as an input or eligibility requirement. Similarly, a dated recovery witness can create R qualification and T age predicates. Assigning the fact to both axes requires citing both requirements. All failures and unknowns remain in the record; the reporting priority A,T,E,C,R,N selects the primary failure description; secondary conditions remain visible.

3 Benchmark Construction and Coverage

Task and reference structure. The benchmark contains 360 tasks and 1,080 candidate actions. Each task supplies fictional policies, documentary records and three consequential candidate actions. The private reference assigns six axis labels and a final action-right classification to each candidate. Public inputs exclude the reference, construction groups, pair links and expected labels. The five slices cover service/API operations, authority and delegation, criticality and reversibility, code and operations, and temporal, recovery and resource constraints.

Expansion composition. The final collection preserves the 120-task predecessor and adds 240 tasks: 156 original scenarios, 60 controlled counterfactual variants and 24 surface/invariance variants. Variants contribute diagnostic contrasts rather than independent samples. The 217 construction/parent groups and 93 overlap components record dependence within the authored collection, rather than count independent real-world mechanisms. The reference was frozen before collecting the new formal predictions. Historical versions and the 60- and 120-task studies are distinguished in Appendices C and H.

Controlled contrasts. Counterfactual data and behavioral testing motivate separating decision-changing edits from label-preserving variations (Kaushik et al., 2020; Ribeiro et al., 2020). A controlled contrast changes a decision-relevant source fact while holding the selected operation and other facts fixed. Figure 2 shows PAIR07 from the inherited collection: a signature's verification changes authority and final ARC while preserving the other five axes. The complete reference contains 84 task pairs, yielding 252 aligned candidate relations. Pair correctness and consistency remain separate measures.

Coverage and version binding. Annotation artifacts can permit prediction without the intended evidence (Gururangan et al., 2018), motivating candidate-only checks alongside full-document assessment. The final reference contains 358 permitted, 413 not-permitted and 309 unknown actions. Hashes bind the public packets, reference, retained outputs and scoring configuration. All 120 predecessor task records and model predictions are reused without modification. Candidate-only lexical diagnostics and duplicate checks are recorded separately from the formal model scores.

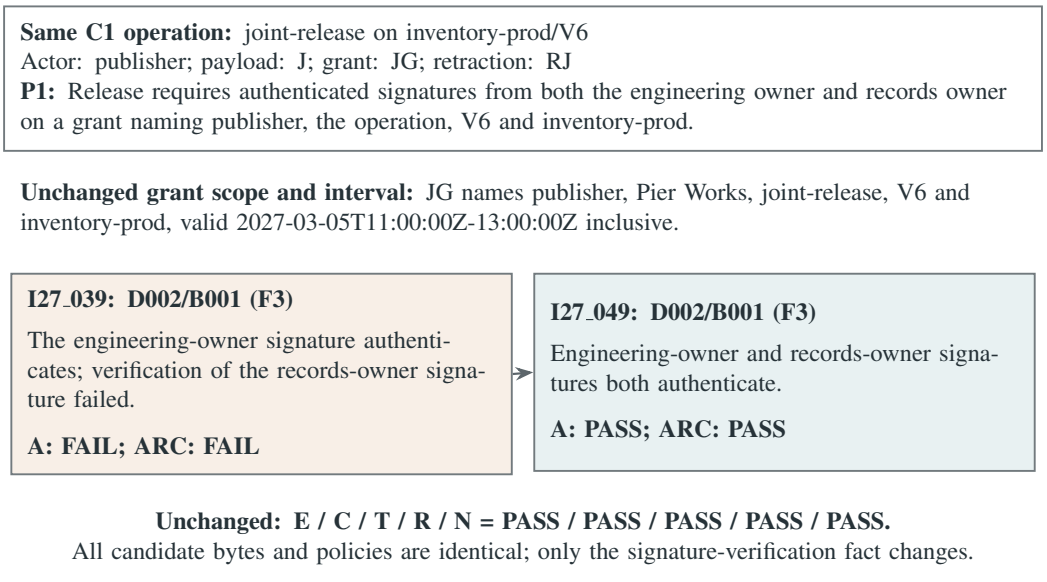


Figure 2: Reference labels for source-controlled PAIR07, I27_039 and I27_049. Only the records-owner signature verification changes. For C1, authority and final ARC flip while the other five axes remain PASS; C2/C3 labels remain unchanged.

4 Human Annotation

For each collection, three human annotators independently labeled the selected tasks. Each response records six axis labels, final ARC and supporting explanations. We report counts by collection without assuming different annotators across collections.

Latest collection. The fixed subset contains 60 of the 240 added tasks and 180 candidate actions. The three returned files contain 540 action ratings and 3,240 axis ratings. All 180 ARC cells and 1,080 action–axis cells are unanimous, with nominal alpha 1 on ARC and each axis. Every returned label matches the frozen reference. The returned files, source records, axis labels and final ARC are unchanged.

Earlier collections. The 60-task and 120-task studies each include a separate 18-task, 54-action collection. Both have unanimous observed labels. Per-annotator reference agreement is 53/54 and 54/54 final decisions, respectively. The earlier six-task pilot remains separate. Appendix D retains the 22 previously recorded reason errors and their label-impact accounting: 16 rationale-only errors, plus six human axis-rating discrepancies affecting three ARC ratings over two unique axis cells and one ARC cell. The 22 reason errors do not correspond to 22 gold-label changes. Agreement and reference comparison are different measurements (Artstein & Poesio, 2008).

5 Experimental Setup and Scoring Protocol

Models and coverage. The 360-task main experiment includes Llama-3.3-70B-Instruct and Mistral-Large-Instruct-2411, served through Modal/vLLM and abbreviated Llama and Large. Both models completed Direct on the same 360 tasks and 1,080 candidate actions. For each model, we reused 120 sealed task-level responses and collected 240 new responses. Small did not complete all 360 tasks. Its 60- and 120-task results appear separately in the historical appendices.

Generation protocol. Each request contains all three candidates and requests their final labels. Settings are temperature 0, a 4,096-output-token ceiling and at most two retries; no generation seed is supplied. Collection times differ. Immutable weight revisions and actual billed costs are unavailable.

Table 1: Direct on the identical 360-task / 1,080-action set. Accuracy intervals use 10,000 whole-construction-group bootstrap draws over 217 groups. OA uses the same 722 reference non-permitted or unknown actions. Small is excluded.

Model	Accuracy (%, 95% CI)	Macro-F1	OA (%)	Invalid
Llama 70B	51.85 [48.57, 55.19]	0.510	25.90	0
Mistral Large	49.26 [45.31, 53.28]	0.479	37.67	0

Successful predecessor responses are not regenerated. The saved records link each attempt, prompt and final parsing outcome to the frozen execution reference.

Conditions and valid outputs. The 360-task evaluation uses Direct. Structured, no-document and candidate-permutation results currently cover only the earlier 120-task study; their scope remains explicit in Appendix H. They are not pooled with 360-task Direct results. One invalid response invalidates all three candidate predictions, and invalid predictions remain in the denominator.

Primary metrics. ARC metrics include accuracy, macro-F1, class-specific scores and over-authorization. For reference labels y_i and predictions $\hat{y}_i$, over-authorization is

$$\mathrm{OA} = \frac{\sum_i \mathbf{1}[\hat{y}_i = \textsc{permitted} \ \wedge \ y_i \in \{\textsc{not_permitted}, \textsc{unknown}\}]}{\sum_i \mathbf{1}[y_i \in \{\textsc{not_permitted}, \textsc{unknown}\}]}. \tag{2}$$

Both models use the same 1,080-action accuracy denominator and the same 722 non-permitted-or-unknown actions for OA. Macro-F1 is recomputed from the pooled confusion matrix, not averaged over the 120- and 240-task strata. Direct supplies no predicted axes, so no six-axis accuracy is reported for this condition.

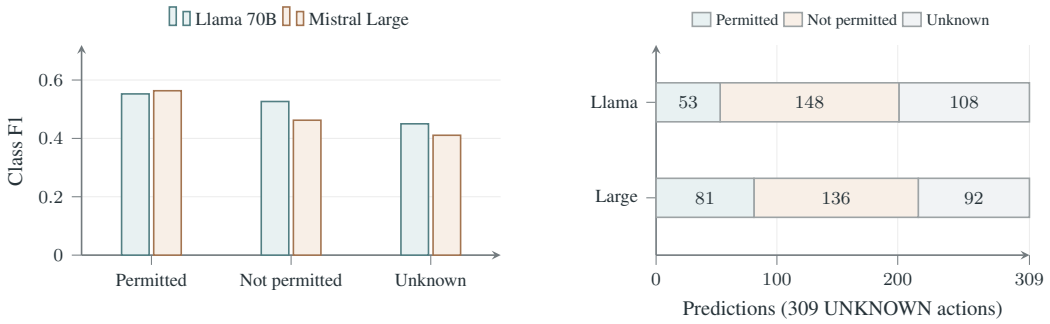
Dependence and uncertainty. Alternative actions and linked variants are dependent. We use 10,000 whole-group multinomial bootstrap draws (seed 20270919) with shared draws across models and 95% percentile intervals. The primary grouping contains 217 construction/parent groups; sensitivity uses 93 overlap components. Their largest groups contain 12 and 107 tasks, respectively. Both retain all candidates and linked tasks within a sampled group. We fixed the groupings before computing intervals, but did not preregister the choice of primary interval. The intervals hold observed predictions fixed. The overlap grouping can produce narrower or wider intervals than the primary grouping. Appendix I reports both. Historical intervals apply only to their original task sets. Models are reported separately; this is not an analysis of model scale.

6 Results

Full-set results. Llama correctly classifies 560/1,080 actions (51.85%, 95% construction-group interval 48.57–55.19%); Large correctly classifies 532/1,080 (49.26%, 45.31–53.28%). Their macro-F1 scores are 0.510 and 0.479. Both models have zero invalid final predictions. The remaining 520 and 548 predictions, respectively, disagree with the reference. The overlap-group sensitivity gives accuracy intervals of 48.66–54.41% and 44.55–52.46%, respectively; we report these intervals descriptively, without testing model rank.

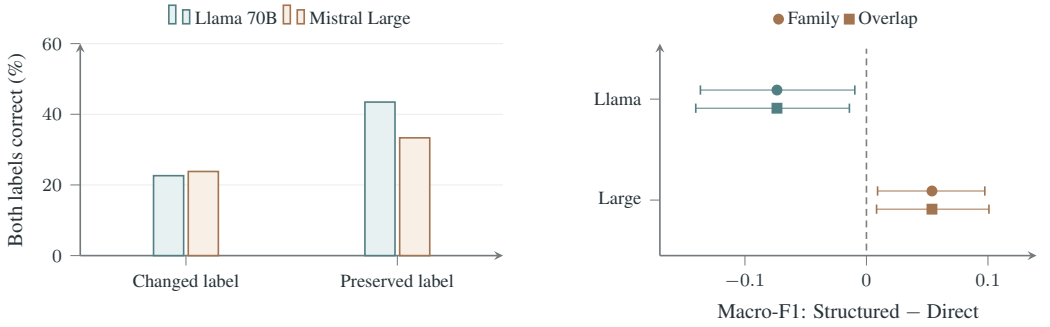
Over-authorization. Llama predicts permission for 187/722 reference non-permitted or unknown actions (25.90%); Large does so for 272/722 (37.67%). OA measures decisions against the supplied records; actions were not executed.

Collection strata. On the preserved 120 tasks, Direct accuracy is 177/360 for Llama and 167/360 for Large. On the 240 added tasks it is 383/720 and 365/720. The combined scores include every action from both disjoint task sets. Structured-output comparisons cover only the earlier study.



(a) 360 tasks: class performance **(b) 360 tasks: unresolved evidence**

Figure 3: Direct predictions on the same 360 tasks. (a) F1 for each ARC class; reference supports are 358 permitted, 413 not permitted and 309 unknown actions. (b) Predictions for the 309 reference-unknown actions, shown as counts. Both models often return a definite decision when the documents leave a required fact unresolved.



(a) 360 tasks: controlled pairs **(b) Earlier 120 tasks only: output protocol**

Figure 4: Two complementary diagnostics with distinct scopes. (a) Joint correctness on 360-task controlled pairs: both labels must be correct across a document change. There are 84 changed-label and 168 preserved-label candidate pairs; related pairs are not independent samples. (b) The earlier 120-task study only: Structured minus Direct macro-F1, with saved paired 95% intervals from 10,000 draws under family and overlap groupings. The conditions differ in output requirements as well as prompting. These effects are not estimates for the 360-task set.

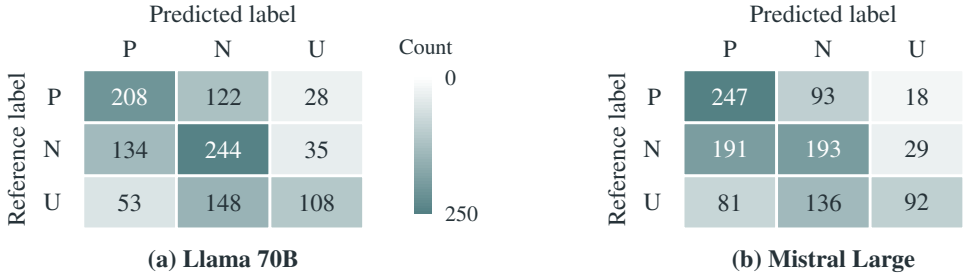
7 Decision-Time Error Analysis

Figure 3 shows that UNKNOWN has the lowest class F1 for both models. The error is not simply a tendency to refuse: unresolved cases receive both permitted and not-permitted predictions.

Among the 309 reference-unknown actions, Llama predicts unknown for 108 and Large for 92. For the remaining cases, each model predicts permitted or not permitted despite unresolved evidence. Llama's unknown-class F1 is 0.450 and Large's is 0.411.

The full confusion matrices (Figure 5) show that Large correctly permits more reference-permitted actions, but also grants permission to more reference-not-permitted actions than Llama. The difference is therefore not a uniform gain or loss across decision classes.

Controlled pairs test whether a model responds to a changed source relation while preserving judgments that should remain fixed. Across 252 aligned candidate relations, Llama preserves the expected relation in 178 cases and correctly labels both sides in 92; Large does so in 171 and 76 cases, respectively. A pair can preserve the expected relation even when both labels are wrong. Llama correctly labels every candidate in 2/84 task pairs, and Large in 6/84. These diagnostic counts share the same underlying tasks as the main table.



(a) Llama 70B **(b) Mistral Large**

Figure 5: Direct predictions on the full 360-task set. Cells show action counts on a shared color scale; rows are reference labels and columns are predictions. P = PERMITTED, N = NOT_PERMITTED, U = UNKNOWN. Each matrix contains the same 1,080 actions, with row supports of 358, 413 and 309. Off-diagonal cells distinguish unsupported permission, incorrect refusal and unresolved judgments.

Figure 4a separates cases where the reference label changes from those where it stays fixed. Joint correctness is lower for changed-label pairs for both models, showing difficulty tracking the decision-relevant document change.

Output protocol on the earlier 120 tasks. Figure 4b brings the completed protocol comparison into view. Large's pooled macro-F1 difference favors Structured under both groupings; Llama's favors Direct. Large's new-only interval and pooled accuracy interval include zero (Appendix H). The results therefore do not support a uniform structured-output benefit, nor isolate an effect of reasoning from the stricter output requirements. Full protocol, document-withheld and candidate-order results remain in the supplement, including Small's smaller-scale results.

8 Related Work

Policy-sensitive action judgment. PhantomPolicy studies violations whose decisive organizational state is hidden from the acting model, with matched safe controls. Its companion Sentinel framework checks actions against a world-state graph; incomplete support yields an indeterminate invariant and a Clarify decision (Wu & Gong, 2026). Our assessment instead fixes candidate alternatives at a supplied checkpoint and requires six obligation judgments justified by policy and document blocks.

Interactive task completion and policy following. AgentBench evaluates agents across interactive environments (Liu et al., 2024b). τ-bench combines domain policies with database-state evaluation (Yao et al., 2024), while AppWorld also checks unintended state changes (Trivedi et al., 2024). ToolSandbox evaluates state dependencies and intermediate milestones (Lu et al., 2025); τ^2-bench adds a shared environment in which both user and agent act (Barres et al., 2025). These settings jointly measure planning, communication, tool use, and completion. We assess permission at a fixed checkpoint, separately from execution success; models cannot gather additional evidence.

Document-grounded reasoning and unresolved decisions. ShARC interprets natural-language rules and asks clarifying questions when a scenario lacks decisive information (Saeidi et al., 2018). FEVER links fact-verification labels to evidence (Thorne et al., 2018), and ALCE evaluates citation-supported generation (Gao et al., 2023). We likewise examine the decision and its supporting sources separately. Long-context performance can also depend on where relevant information occurs (Liu et al., 2024a). Our task requires matching facts across documents to each candidate's obligations. The UNKNOWN label denotes insufficient documentary support, analogous to unanswerability in SQuAD 2.0 (Rajpurkar et al., 2018), rather than a model's choice to abstain to trade coverage for lower risk (Geifman & El-Yaniv, 2019).

Safety and adversarial tool use. AgentHarm evaluates malicious multi-step requests (Andriushchenko et al., 2025), and AgentDojo studies prompt injection through untrusted tool data

(Debenedetti et al., 2024). OpenAgentSafety combines final-state checks with trajectory judgments, including unauthorized sharing and repository-policy changes (Vijayvargiya et al., 2025). ToolEmu evaluates risky calls in emulated trajectories, including underspecified instructions and risk-informed permission seeking (Ruan et al., 2023). Our outputs are judgments about each candidate's obligations, linked to supporting document blocks. These offline judgments assess permission, not the safety of an executed trajectory.

Privilege control and formal execution contracts. Progent checks symbolic privileges over tool names and arguments and distinguishes automatic narrowing from expansions requiring approval (Shi et al., 2025). ToolGate uses symbolic preconditions and postconditions (Liu et al., 2026); SABER concentrates confirmation at mutating steps (Cuadron et al., 2025). Open Agent Passport specifies pre-action authorization over capabilities, limits, and structured context, with its v1 ESCALATE path distinguished from the implemented reference path (Uchibeke, 2026). Confirmation and escalation need not indicate missing evidence, which is what UNKNOWN represents here. We assess how models identify obligations in documents, rather than enforcing them at runtime. The fixed aggregation rule does not resolve ambiguity in interpreting the documents.

Annotation and measurement. Agreement must be interpreted with its units and category distributions (Artstein & Poesio, 2008). Here, exact agreement concerns candidates or candidate–axis cells, while uncertainty estimates group related tasks. Datasheets and data statements motivate documenting collection scope and intended use (Gebru et al., 2021; Bender & Friedman, 2018). Multiple axes or annotators do not create independent scenarios. Agreement, source support, and shortcut resistance require separate checks. Human annotation results apply to the selected collections.

9 Limitations and Conclusion

The benchmark uses controlled, authored scenarios and supplied policies. Real organizations involve additional context, informal practices and evolving responsibilities. Broader domain coverage and independently developed organizational settings remain to be studied.

The current evaluation is an offline decision-time assessment based on fixed documents. Interactive agents must also acquire evidence, respond to changes and manage the consequences of execution. Future work can evaluate the contract in these interactive settings.

Decision rights depend on more than an action's expected outcome. The benchmark makes evidence, authority and operational constraints explicit and evaluates whether the available records establish the right to act.

AI USE STATEMENT

AI tools provided auxiliary support during manuscript preparation, including language polishing. The authors are responsible for the final content and conclusions.

ETHICS STATEMENT

The scenarios and operations are fictional and were not executed. For each collection, three human annotators independently labeled the selected tasks. No annotator personal information was collected. The records are benchmark labels rather than human-subject records. No institutional approval or exemption is claimed. The labels concern the supplied fictional policies, not legal or regulatory compliance.

REPRODUCIBILITY STATEMENT

The final checkpoint records the 360-task reference, both models' Direct predictions and the scores computed with fixed denominators. We reused the 120 predecessor responses per model and collected new responses only for the added 240 tasks. A self-contained offline scoring package verifies file hashes, reference and prediction coverage, recomputes all point metrics and reproduces both grouped interval analyses. It includes public task packets, the unchanged final reference and retained terminal predictions. Original raw-response and execution journals remain in the local audit record. The scoring package does not regenerate model responses or reconstruct the provider environment. The separate manuscript source package regenerates the tables from hash-checked result files. Historical 60- and 120-task evidence retains its own scope.

REFERENCES

Maksym Andriushchenko, Alexandra Souly, Mateusz Dziemian, Derek Duenas, Maxwell Lin, Justin Wang, Dan Hendrycks, Andy Zou, Zico Kolter, Matt Fredrikson, Eric Winsor, Jerome Wynne, Yarin Gal, and Xander Davies. AgentHarm: A benchmark for measuring harmfulness of LLM agents. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.09024v3.

Ron Artstein and Massimo Poesio. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, 2008. doi: 10.1162/coli.07-034-R2. URL https://aclanthology.org/J08-4004/.

Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. τ^2-Bench: Evaluating conversational agents in a dual-control environment, 2025. URL https://arxiv.org/abs/2506.07982.

Emily M. Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018. doi: 10.1162/tacl_a_00041. URL https://aclanthology.org/Q18-1041/.

Alejandro Cuadron, Pengfei Yu, Yang Liu, and Arpit Gupta. SABER: Small actions, big errors – safeguarding mutating steps in LLM agents, 2025. URL https://arxiv.org/abs/2512.07850v1. Version 1, November 2025.

Edoardo Debenedetti, Jie Zhang, Mislav Balunović, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. AgentDojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents, 2024. URL https://arxiv.org/abs/2406.13352v3.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6465–6488, 2023. doi: 10.18653/v1/2023.emnlp-main.398. URL https://aclanthology.org/2023.emnlp-main.398/.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92, 2021. doi: 10.1145/3458723. URL https://arxiv.org/abs/1803.09010v8.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159, 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In *Proceedings of NAACL-HLT 2018 (Volume 2: Short Papers)*, pp. 107–112, 2018. doi: 10.18653/v1/N18-2017. URL https://aclanthology.org/N18-2017/.

Vincent C. Hu, David Ferraiolo, Rick Kuhn, Adam Schnitzer, Kenneth Sandlin, Robert Miller, and Karen Scarfone. Guide to attribute based access control (ABAC) definition and considerations. Technical Report SP 800-162, National Institute of Standards and Technology, 2014. URL https://nvlpubs.nist.gov/nistpubs/specialpublications/NIST.sp.800-162.pdf. Updated August 2019.

Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In *The Eighth International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1909.12434.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024a. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. AgentBench: Evaluating LLMs as agents. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/e9df36b21ff4ee211a8b71ee8b7e9f57-Abstract-Conference.html.

Yanming Liu, Xinyue Peng, Jiannan Cao, Xinyi Wang, Songhang Deng, Jintao Chen, Jianwei Yin, and Xuhong Zhang. ToolGate: Contract-grounded and verified tool execution for LLMs, 2026. URL https://arxiv.org/abs/2601.04688.

Jiarui Lu, Thomas Holleis, Yizhe Zhang, Bernhard Aumayer, Feng Nan, Haoping Bai, Shuang Ma, Shen Ma, Mengyu Li, Guoli Yin, Zirui Wang, and Ruoming Pang. ToolSandbox: A stateful, conversational, interactive evaluation benchmark for LLM tool use capabilities. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1160–1183, 2025. doi: 10.18653/v1/2025.findings-naacl.65. URL https://aclanthology.org/2025.findings-naacl.65/.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of ACL 2018 (Volume 2: Short Papers)*, pp. 784–789, 2018. doi: 10.18653/v1/P18-2124. URL https://aclanthology.org/P18-2124/.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912, 2020. doi: 10.18653/v1/2020.acl-main.442. URL https://aclanthology.org/2020.acl-main.442/.

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, 2023. URL https://arxiv.org/abs/2309.15817v2. Version 2, revised May 2024.

Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Singh, Tim Rocktäschel, Mike Sheldon, Guillaume Bouchard, and Sebastian Riedel. Interpretation of natural language rules in conversational machine reading. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2087–2097, 2018. doi: 10.18653/v1/D18-1233. URL https://aclanthology.org/D18-1233/.

Jerome H. Saltzer and Michael D. Schroeder. The protection of information in computer systems. *Proceedings of the IEEE*, 63(9):1278–1308, 1975. URL https://www.mit.edu/~Saltzer/publications/protection/index.html.

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. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/d842425e4bf79ba039352da0f658a906-Abstract-Conference.html.

Tianneng Shi, Jingxuan He, Zhun Wang, Hongwei Li, Linyu Wu, Wenbo Guo, and Dawn Song. Progent: Securing AI agents with privilege control, 2025. URL https://arxiv.org/abs/2504.11703v3. Version 3, revised May 2026.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of NAACL-HLT 2018 (Volume 1: Long Papers)*, pp. 809–819, 2018. doi: 10.18653/v1/N18-1074. URL https://aclanthology.org/N18-1074/.

Harsh Trivedi, Tushar Khot, Mareike Hartmann, Ruskin Manku, Vinty Dong, Edward Li, Shashank Gupta, Ashish Sabharwal, and Niranjan Balasubramanian. AppWorld: A controllable world of apps and people for benchmarking interactive coding agents. In *Proceedings of the 62nd Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 16022–16076. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.850. URL https://aclanthology.org/2024.acl-long.850/.

Uchi Uchibeke. Before the tool call: Deterministic pre-action authorization for autonomous AI agents, 2026. URL https://arxiv.org/abs/2603.20953v1. Version 1, March 2026.

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, 2025. URL https://arxiv.org/abs/2507.06134v2. Version 2, revised February 2026; author record reports acceptance at ICLR 2026 and IASEAI 2026.

Jie Wu and Ming Gong. Policy-invisible violations in LLM-based agents, 2026. URL https://arxiv.org/abs/2604.12177v1. Version 1, April 2026.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.03629.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains, 2024. URL https://arxiv.org/abs/2406.12045.