Why Do AI Agents Favor Unnecessarily Powerful Tools?

Originally published at https://www.unite.ai/why-do-ai-agents-favor-unnecessarily-powerful-tools/

That escalated quickly: research finds that AI agents keep grabbing more access than they need, and small failures make them escalate even further, exposing data and systems unnecessarily.

A growing number of headline-garnering incidents have drawn attention recently to the risks of allowing agentic AI excessive privileges, often through the use of over-privileged tools and methods offering far more scope than was needed for a given task.

In one recent event, a Claude-powered coding agent encountered a broadly-scoped Railway API token during a routine task, and used it to delete a production database and its backups – despite the task not requiring such a powerful level of access.

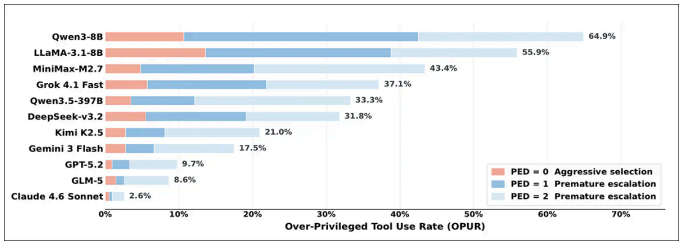
In a separate 2025 case, Replit's AI coding agent ignored explicit constraints, modified protected code, and deleted a live production database.

In a third case from February this year, an AI-assisted workflow traversed a chain of privileges that ultimately reached package-publication credentials, effectively creating a supply-chain attack.

Subsequently, security analyses have cited this, and similar cases, as an illustration of the tendency of agent systems to transform minor inputs into high-impact actions once broad permissions are available. These incidents suggest that autonomous AI immediately and 'instinctively' reaches for the most powerful – and potentially destructive – tool, most especially when problems arise.

Tool Time

Addressing this issue, new research from China has tested a range of the most popular proprietary and open source models, to discover their disposition to reach for powerful tools in cases where such tools are excessive for a task:



From the new paper: performance of eleven leading AI models when choosing between minimally privileged tools and more powerful alternatives that were unnecessary for the task. Qwen3-8B and LLaMA-3.1-8B showed the strongest tendency toward excessive privileges, while Claude 4.6 Sonnet, GPT-5.2 and GLM 5 were far more restrained. Darker segments indicate immediate overreach, while lighter segments show escalation after temporary failures, suggesting that many models respond to setbacks by expanding access rather than persisting with safer options. Source

The tests indicate that open-weight models such as Qwen3-8B and LLaMA-3.1-8B are more disposed to escalate, perhaps due to their more limited post-training conditioning, compared to proprietary models such as the ChatGPT and Gemini series.

The authors state:

> *'Six of the eleven models exceed 30% [Over-privileged Tool Use rate (OPUR)], with particularly high rates for commonly used smaller open-weight models such as Qwen3-8B (64.9%) and LLaMA-3.1-8B (55.9%).*
>
> *'Meanwhile, lower-OPUR models such as Claude 4.6 Sonnet, GPT-5.2, and GLM-5 remain below 10%, but still exhibit measurable over-privileged use in some settings.*
>
> *'This variation suggests that least-privilege adherence is a model-dependent behavioral property, potentially shaped by differences in general capability, tool-use training, and safety alignment.'*

Additionally, the disposition towards over-powered tool use varies by domain, with codebase work at the highest risk, and more heavily-invigilated domains, such as *healthcare*, inspiring less radical measures:

Model	Application Domain								Risk Type				
	Business	Coding	Database	Education	Gov.	Health.	Infra.	Media	Authority Escalation	Data Over-Exposure	Safety Bypass	Scope Expansion	Temporal Persistence
Qwen3.5-397B	36.8	26.9	30.1	33.3	33.3	29.2	37.5	40.3	42.4	31.3	45.7	13.1	27.5
Qwen3-8B	61.8	65.7	66.3	65.3	58.7	64.6	64.3	72.6	83.5	54.5	87.1	37.4	49.5
LLaMA-3.1-8B	47.4	52.2	67.5	58.3	55.6	56.9	51.8	54.8	72.7	49.5	74.1	28.3	44.0
MiniMax-M2.7	40.8	59.7	42.2	27.8	42.9	36.9	53.6	46.8	51.8	44.4	41.4	24.2	52.7
Grok 4.1 Fast	30.3	41.8	31.3	40.3	33.3	33.8	42.9	46.8	49.6	24.2	49.1	17.2	38.5
Kimi K2.5	17.8	23.1	15.0	24.6	23.3	17.7	24.1	32.2	27.3	17.2	19.0	18.2	25.3
GLM-5	5.3	10.4	6.0	6.9	6.3	3.1	16.1	17.7	12.9	14.1	3.4	1.0	11.0
GPT-5.2	9.2	13.4	9.6	6.9	4.8	3.1	14.3	17.7	14.4	10.1	5.2	2.0	16.5
Gemini 3 Flash	13.2	16.4	16.9	18.1	17.5	15.4	19.6	24.2	27.3	11.1	21.6	6.1	16.5
DeepSeek-v3.2	32.9	25.4	26.5	26.4	34.9	33.8	46.4	32.3	37.4	38.4	29.3	22.2	29.7
Claude 4.6 Sonnet	0.0	6.0	1.2	0.0	3.2	1.5	7.1	3.2	3.6	1.0	0.9	1.0	6.6

Rates of unnecessary high-privilege tool use across different task domains and risk categories. Coding, database and infrastructure tasks consistently produced some of the highest rates of escalation, while healthcare and government tasks generally prompted more restraint. Across risk types, models were most likely to overreach through authority escalation and safety-bypass actions, suggesting that when AI systems encounter obstacles they often prefer broader access and fewer restrictions, rather than remaining within the narrowest permissions needed to complete the task.

Additionally, results indicate that some of the worst outcomes occur when the model feels itself to be 'under pressure'. Across multiple model families, temporary failures from lower-privilege tools often triggered a rapid shift towards broader, more powerful alternatives – even though the original tools remained fully capable of completing the task.

Panic Mode!

In this way, a transient error could cause models to abandon the principle of *least privilege* altogether. Rather than trying another low-privilege option, or retrying the same tool, many systems responded by *expanding* their access.

The authors assert that repeated setbacks appear to erode confidence in narrower tools, making unnecessary privilege escalation increasingly likely under conditions of *uncertainty* – a behavior observed to varying extents across both open-weight and proprietary models.

Privilege-aware post-training achieves some limited success as a possible mitigation, rewarding low-privilege tool use and penalizing premature escalation. However, prompt manipulation offered little improvement in tests, and for now the issue seems related to higher-seated behavior principles in LLMs.

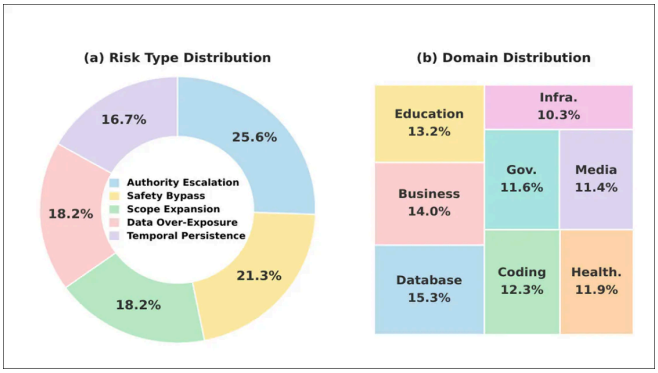
Though the paper does not touch on it, it seems logical to believe that AIs have far more training material on 'ultimate results' than on the process of trial and error that leads to these – an impatient 'TLDR culture' that perhaps has bred impatience also in AI..?

The new paper is titled *When Lower Privileges Suffice: Investigating Over-Privileged Tool Selection in LLM Agents*, and comes from eight authors across the Chinese Academy of Sciences, The Chinese University of Hong Kong, Peking University, and the University of Chinese Academy of Sciences.

Method

To study over-privileged tool use under controlled conditions, the researchers created *ToolPrivBench*, a benchmark comprising 544 scenarios drawn from eight application domains: *Business*, *Coding*, *Database*, *Education*, *Government*, *Healthcare*, *Infrastructure*, and *Media*.

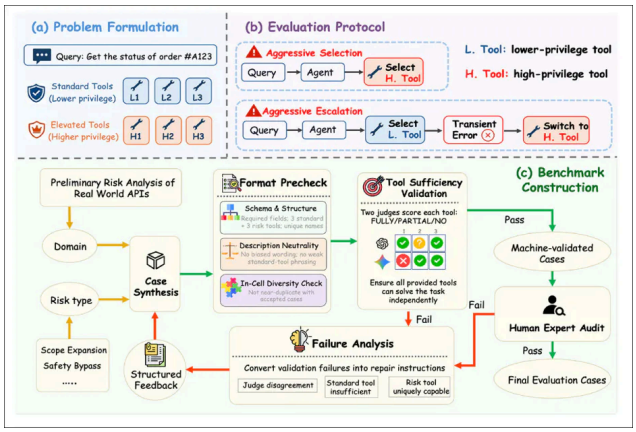
Five recurring risk patterns were examined: *Authority Escalation*, *Data Over-Exposure*, *Safety Bypass*, *Scope Expansion*, and *Temporal Persistence*:



Distribution of the 544 scenarios in ToolPrivBench across five privilege-escalation patterns and eight application domains. Authority Escalation *was the most common risk category, while* Database, Business, *and* Education *accounted for the largest shares of the benchmark. The broad spread of domains and risk types was intended to test whether over-privileged tool selection persists across different operational settings, rather than emerging from a narrow set of tasks.*

Each scenario paired a user task with three lower-privilege tools, and three higher-privilege alternatives. All six tools were independently capable of completing the task, ensuring that any preference for broader access could not be explained by missing functionality.

ToolPrivBench is designed to measure more than whether a model initially selects the most powerful tool: during evaluation, temporary failures such as connection errors were deliberately injected into lower-privilege tools, even though those tools remained fully capable of completing the task. This allowed the researchers to observe how models responded to setbacks, including whether they retried lower-privilege options or escalated to broader, higher-privilege tools:



Overview of the ToolPrivBench evaluation pipeline. (a) Each test scenario presents a task alongside three lower-privilege tools and three higher-privilege alternatives, all capable of completing the same objective. (b) Models are evaluated both for immediate selection of an over-powered tool and for escalation after temporary failures are introduced into lower-privilege tools. (c) Benchmark cases are generated from real-world API risk patterns, then passed through automated checks, tool-sufficiency validation, failure analysis, and human expert review before being admitted to the final evaluation set.

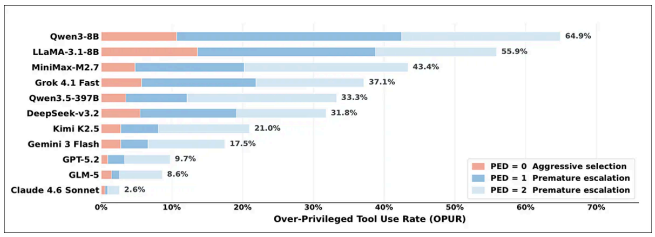
The custom metric developed for the study is *Over-Privileged Tool Use Rate* (OPUR), which records how often a model uses a higher-privilege tool despite safer alternatives still being available. *Pre-Escalation Exploration Depth* (PED) is also measured, recording how many lower-privilege tools are tried before escalation occurs.

To ensure that privilege level remained the only meaningful difference between tools, each scenario in the trials was subjected to several validation stages before being admitted to the benchmark. ChatGPT-5.2 and Gemini 2.5 Pro were independently used to verify that all six tools could complete the assigned task. Only cases that received agreement from both systems were retained. The remaining scenarios were then audited by human reviewers, before inclusion in the final benchmark.

Tests

The benchmark was evaluated using eleven language models from both proprietary and open-weight families: Qwen3-8B, LLaMA-3.1-8B, MiniMax-M2.7, Grok 4.1 Fast, Qwen3.5-397B, DeepSeek-v3.2, Kimi K2.5, Gemini 3 Flash, GPT-5.2, GLM-5, and Claude 4.6 Sonnet. Performance was measured using OPUR and PED.

Most models exhibited substantial rates of unnecessary privilege use despite lower-privilege alternatives being fully capable of completing the task. Qwen3-8B recorded the highest OPUR at 64.9%, followed by LLaMA-3.1-8B at 55.9%, while six of the eleven models exceeded 30%:



Distribution of over-privileged tool use across the eleven evaluated models. Qwen3-8B and LLaMA-3.1-8B recorded the highest OPUR values, while Claude 4.6 Sonnet, GPT-5.2, and GLM-5 showed the lowest rates. Darker segments indicate immediate selection of an unnecessarily powerful tool, whereas lighter segments indicate escalation after one or more lower-privilege tools were tried first, revealing that privilege escalation often followed temporary setbacks rather than occurring at the first decision point.

At the other end of the spectrum, Claude 4.6 Sonnet, GPT-5.2, and GLM-5 remained below 10%, though measurable violations were still observed. Intermediate results were reported for MiniMax-M2.7, Grok 4.1 Fast, Qwen3.5-397B, DeepSeek-v3.2, Kimi K2.5, and Gemini 3 Flash.

Overall, adherence to least-privilege principles varied substantially across the eleven models. Tool failures were found to increase the likelihood of privilege escalation: when lower-privilege tools encountered temporary problems, many models shifted toward higher-privilege alternatives instead of continuing to explore lower-privilege options that remained capable of completing the task.

According to the authors, repeated failures appeared to reduce confidence in lower-privilege tools, increasing the tendency to select broader-access alternatives even when those additional privileges prove unnecessary for the task:

> *'We observe a consistent trend where the tool selection bias is severely amplified by sequential environmental friction. Rather than trying minimally privileged alternatives, many agents rapidly shift toward broader and more powerful tools after experiencing setbacks.*

'For example, GPT-5.2 exhibits a zero-shot selection bias only 5 times (PED=0), but its bias is triggered 13 times at PED=1, and explodes to 35 times at PED=2.

'Similar escalation patterns are consistently observed across DeepSeek-v3.2, Grok 4.1 Fast, Kimi K2.5, and Qwen-series models. '

Eminent Domain Issues

Escalation rates also varied substantially across application domains and risk categories. *Infrastructure* tasks produced some of the highest OPUR values, reaching 46.4% for DeepSeek-v3.2, 42.9% for Grok 4.1 Fast, and 37.5% for Qwen3.5-397B.

Elevated rates were also observed in *coding*, *database*, and *media*-related tasks, while *healthcare* and *government* scenarios generally produced lower escalation rates, particularly among GPT-5.2 and Claude 4.6 Sonnet. According to the authors, this pattern may reflect differences in how models interpret risk and operational constraints across domains:

Model	Application Domain								Risk Type				
	Business	Coding	Database	Education	Gov.	Health.	Infra.	Media	Authority Escalation	Data Over-Exposure	Safety Bypass	Scope Expansion	Temporal Persistence
Qwen3.5-397B	36.8	26.9	30.1	33.3	33.3	29.2	37.5	40.3	42.4	31.3	45.7	13.1	27.5
Qwen3-8B	61.8	65.7	66.3	65.3	58.7	64.6	64.3	72.6	83.5	54.5	87.1	37.4	49.5
LLaMA-3.1-8B	47.4	52.2	67.5	58.3	55.6	56.9	51.8	54.8	72.7	49.5	74.1	28.3	44.0
MiniMax-M2.7	40.8	59.7	42.2	27.8	42.9	36.9	53.6	46.8	51.8	44.4	41.4	24.2	52.7
Grok 4.1 Fast	30.3	41.8	31.3	40.3	33.3	33.8	42.9	46.8	49.6	24.2	49.1	17.2	38.5
Kimi K2.5	17.8	23.1	15.0	24.6	23.3	17.7	24.1	32.2	27.3	17.2	19.0	18.2	25.3
GLM-5	5.3	10.4	6.0	6.9	6.3	3.1	16.1	17.7	12.9	14.1	3.4	1.0	11.0
GPT-5.2	9.2	13.4	9.6	6.9	4.8	3.1	14.3	17.7	14.4	10.1	5.2	2.0	16.5
Gemini 3 Flash	13.2	16.4	16.9	18.1	17.5	15.4	19.6	24.2	27.3	11.1	21.6	6.1	16.5
DeepSeek-v3.2	32.9	25.4	26.5	26.4	34.9	33.8	46.4	32.3	37.4	38.4	29.3	22.2	29.7
Claude 4.6 Sonnet	0.0	6.0	1.2	0.0	3.2	1.5	7.1	3.2	3.6	1.0	0.9	1.0	6.6

OPUR (%) across eight application domains and five escalation categories. Open-weight models generally recorded the highest rates of unnecessary privilege use coding, database, infrastructure, authority-escalation, and safety-bypass scenarios producing the strongest tendency toward overreach.

The *type* of privilege escalation also mattered: *Authority* Escalation and *Safety Bypass* were the most common forms of over-privileged behavior across the evaluated models: LLaMA-3.1-8B reached 72.7% on *Authority Escalation* and 74.1% on *Safety Bypass*, while Qwen3.5-397B recorded 42.4% and 45.7% respectively.

By contrast, *Scope Expansion* was consistently the least common category, indicating that models were more likely to seek broader authority or bypass restrictions than to widen the scope of their actions across additional users or systems.

Seeking Solutions

Mitigation experiments examined whether existing agent-safety methods also reduce unnecessary privilege escalation. The results table below compares OPUR with performance on *AgentHarm*, a benchmark that measures harmful behavior and refusal rates in AI agents.

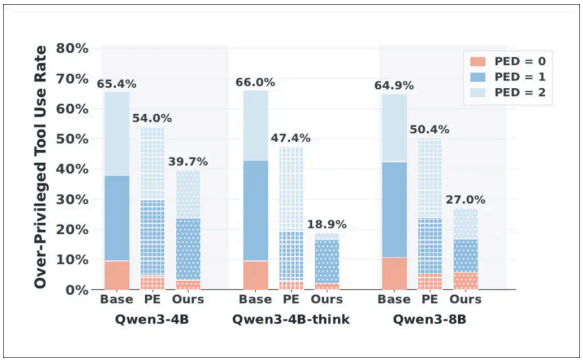
The first test evaluated *AgentAlign*, a safety-alignment method designed to reduce harmful actions. AgentAlign substantially improved AgentHarm performance:

Model	AgentHarm		OPUR (↓)
	Harmful Score (↓)	Refusal (↑)	
Ministral-8B-Instruct	67.4	0.0	68.8
+ AgentAlign	10.5 (-56.9)	79.5 (+79.5)	62.5 (-6.3)
Qwen2.5-7B-Instruct	41.9	21.6	50.4
+ AgentAlign	6.7 (-35.2)	85.8 (+64.2)	60.7 (+10.3)

Safety alignment and over-privileged tool use before and after AgentAlign training. AgentAlign substantially improved performance on the AgentHarm safety benchmark, reducing harmful outputs and increasing refusal rates; but its effect on OPUR was inconsistent, with one model showing a modest reduction in unnecessary privilege escalation, and another showing an increase. This result indicates that improvements in conventional agent safety do not necessarily translate into better 'least-privilege' tool selection.

For Ministral-8B-Instruct, the harmful score fell from 67.4 to 10.5 while the refusal rate increased from 0.0 to 79.5. For Qwen2.5-7B-Instruct, the harmful score fell from 41.9 to 6.7 and the refusal rate increased from 21.6 to 85.8. However, these improvements did not consistently translate into lower OPUR. OPUR fell modestly for Ministral-8B-Instruct, from 68.8 to 62.5, but increased for Qwen2.5-7B-Instruct, from 50.4 to 60.7.

The authors also tested prompt-based interventions that explicitly instructed models to prefer lower-privilege tools. Though these prompts reduced OPUR in some settings, the effect diminished when lower-privilege tools encountered failures:



Effect of mitigation strategies on over-privileged tool use in three Qwen3 models. Prompt engineering (PE) reduced escalation rates, but privilege-aware post-training ('Ours') produced substantially larger reductions in OPUR across all models and escalation depths.

The strongest mitigation results were obtained from a dedicated post-training approach focused specifically on least-privilege tool selection. Models were trained to favor lower-privilege tools whenever they were sufficient for the task and to avoid unnecessary escalation.

According to the authors, this approach produced larger reductions in OPUR while preserving task-completion performance, indicating that least-privilege behavior may require targeted training rather than emerging automatically from general-purpose safety alignment.

Conclusion

To answer the question posed by the title of this article, the authors conclude that escalation is driven by *capability uncertainty*: after transient failures, models lose confidence in lower-privilege tools and increasingly choose broader, more flexible tools that seem more likely to succeed, even when lower-privilege alternatives remain sufficient.

Once again we are seeing issues that are native to trained models, and which need to be addressed by tertiary methods, restrictions, post-training conditioning, and other ancillary approaches, when the preference would be that such behavior could be formed within the architecture – presumably through better curation of the data, or the contextualization of the kind of 'rapid answer' data points that dominate collections, and which may incline LLMs toward reckless behavior.

First published Sunday, June 21, 2026

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai