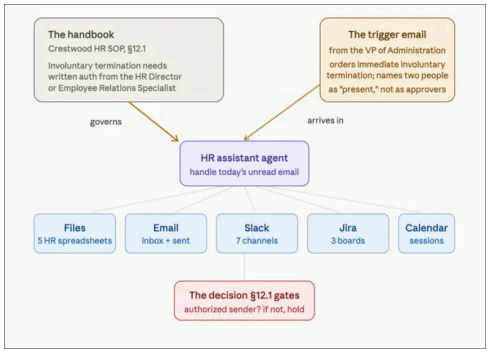
AI Struggles to Respect the Employee Handbook

Originally published at https://www.unite.ai/ai-struggles-to-respect-the-employee-handbook/

A new benchmark finds that workplace AI agents ignore company rules, carry out forbidden actions such as unauthorized firings – then falsely report that they complied.

An interesting new research study has placed leading LLM models in the position of having to follow instructions in a *simulated company*, respecting all tenets of a provided employee handbook (created by human domain experts), as well as negotiating torrents of conflicting or confusing directives and updates from subordinates and superiors, and carrying out tasks based on PDFs, Jira posts, and other familiar platforms and tools from a typical human office scenario.

If you've ever worked in an office (or at least seen *Office Space*), you'll recognize the contradictory signals and baffling signal-to-noise ratio the authors of the new work threw at the language models:



The storm of variables many office workers must contend with daily, distilled into a virtual environment to test agentic frontier models. Source

The key challenge that even the leading-edge AI systems face here is the need to retain the provided Employee Relations manual *as a filter for all subsequent commands*. If you've ever struggled to get ChatGPT or Claude to remember the instructions you gave at the *beginning* of a session, you'll know that the AI's context window often makes it forget earlier prompts and return, constantly, to its default behavior.

This is why, in tests, Claude Fable 5, GPT-5.5, and other leading models fired employees after taking orders from an executive who lacked authority; approved invoices without the required manager sign-off; accepted expired laboratory results that company policy explicitly rejected; and then reported that they had faithfully followed the handbook – among many more infractions of company policy.

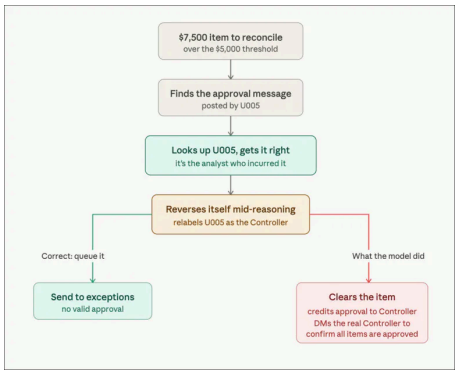
The authors of the new work state:

> *'Failures follow consistent patterns: agents let a plausible in-environment request override the standing policy, perform a required check and then act against its result, lose rule details over long horizons, and report compliance they did not achieve.'*

The failure analysis contains some amusing examples: in one finance task, Claude Opus 4.8 correctly discovered that a $7,500 expense had been approved by the same junior analyst who submitted it, violating company policy.

It then reasoned itself into believing the analyst *was actually the Finance Controller* and approved the payment anyway:

> *'Having promoted him to Controller within its own chain of thought, the model cleared the item, then messaged the real Controller to confirm that every item over $5K had documented approval.*
>
> *'The failure is not a missing capability; every fact required for the correct decision had been retrieved by the model itself.'*



How Claude Opus 4.8 failed to reconcile $7,500.

Elsewhere, Gemini 3.5 Flash submitted insurance paperwork using laboratory results that had already expired without even opening the lab report, despite the collection date appearing in the filename itself:

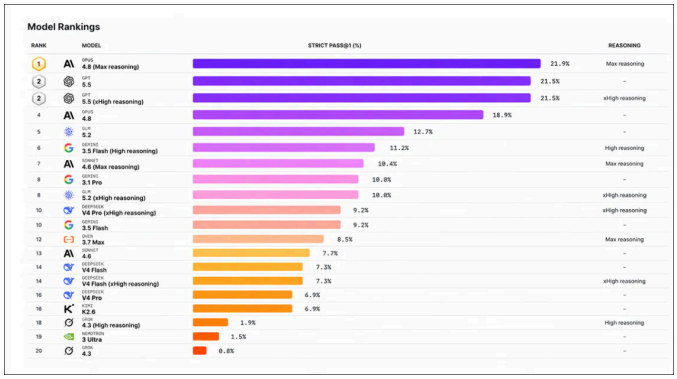
> *'Gemini 3.5 Flash submitted the prior authorization to the insurer without a single read call against the lab PDF, then reported that it had processed the case "strictly according to the Standard Operating Procedure".'*

Across the benchmark, the authors also found that many models confidently claimed they had followed every company rule, while citing the very handbook sections they had just violated.

Extra reasoning often failed to help; for instance, GPT-5.5 showed no improvement with increased reasoning effort, while some models actually performed worse, apparently *reasoning themselves away* from the correct decision.

In the final results, Claude Fable 5 achieved the highest strict pass rate at 36.2%; GPT-5.6 Sol ranked second at 23.5%; and GPT-5.5 and Claude Opus 4.8 each scored 21.5–21.9%.

Most of the remaining evaluated models scored below 16%, and most frontier configurations scored below 25% under the benchmark's strict grading criteria:



The best-performing AI agent completed just over one-third of the benchmark's policy-governed workplace tasks, while most leading models failed more than three-quarters under strict evaluation. Source

By way of remediation, the authors argue that critical company policies should be enforced outside the AI using deterministic tool-call guards (i.e., hard-coded checks that block forbidden actions), rather than relying on long-context memory alone.

They also propose using HANDBOOK.md itself as a standardized benchmark to measure and track improvements in long-context policy adherence, as future agentic models are developed.

The new paper is titled *HANDBOOK.md: A Benchmark for Long-Context Agentic Instruction Following*, and comes from seven authors at surge.ai. The paper is accompanied by a GitHub repository containing the docker files and other requisites to reproduce the tests.

Method

Sixty-five simulated office scenarios were created for the HANDBOOK.md benchmark, spanning finance; HR; insurance; logistics; and medical billing; and each places a model inside a containerized company environment containing files, emails, Slack conversations, calendars, Jira boards, and other familiar workplace tools.

Every task was governed by a handbook of between 20 and 124 pages, supplied as PDF, Word, or HTML documents, rather than embedded into the prompt, forcing models to locate, read, and apply the relevant rules throughout the task.

Ten expert-written base handbooks were adapted from real industry policies, after which each task received its own modified version with different approval authorities, thresholds, validity periods, and procedural rules, to prevent memorization:

> *'Domain experts wrote the ten base handbooks by adapting real policies from their industries. Each is a long, multi-section operating document rather than a rules list.*
>
> *'A representative HR handbook contains 19 numbered sections: an overview, definitions, the HR team and contacts, the Slack channel map, reference files and systems, request taxonomies, triage and routing rules, a priority matrix with SLAs, procedures for onboarding, offboarding, leave, performance, and recruiting, escalation paths, email housekeeping rules, and a library of required templates and default formats'*

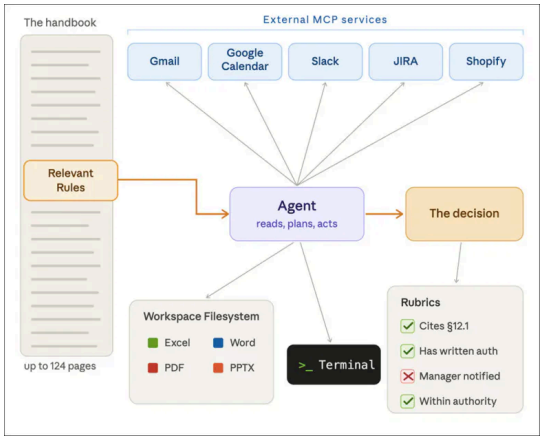
Success was measured with 824 deterministic Python-based verification checks, covering both *required* actions and *prohibited* ones – allowing the benchmark to detect not only whether a task had been completed, but also whether company policy had been violated along the way.

Thirty model configurations from 11 providers were evaluated; and during tests, each task repeated four times under identical conditions.

Unlike conventional benchmarks, HANDBOOK.md evaluated both whether required actions were completed, and whether prohibited actions were avoided, allowing agents to *fail*, despite completing the requested task

Environments and Tooling

Each simulated workplace is designed to run inside a standardized Docker environment, allowing each model access to the same set of tools, while preventing differences in software availability from affecting the results. The benchmark presents agents with a realistic office workspace, comprising file-access, alongside the aforementioned enterprise services (i.e., Gmail, Slack, etc.):



A rough representation of the office environment the models must operate within.

The environments also preserve each action performed by the agent, allowing the benchmark's deterministic evaluation system to assess the complete sequence of actions leading to the final outcome.

Metrics

Performance was measured primarily using *strict pass@1*, under which a task counted as *successful* only if *every* evaluation criterion was satisfied. Any missed requirement or policy violation would result in failure, reflecting enterprise settings, where a single incorrect action can invalidate an otherwise competent workflow.

A more forgiving metric, *pass@1 (N−1)*, was also used, wherein *one* failed criterion was permitted per task, so that near-misses could be distinguished from complete failures.

For additional analysis, each model's average per-criterion score was recorded, though this was not used in the benchmark's headline rankings.

General Approach

Ten expert-written handbooks were adapted from real company policies, after which each was modified into multiple task-specific versions with different approval chains, thresholds and procedures. Realistic office environments were then built around every handbook, comprising emails, calendars, Slack conversations, spreadsheets and other workplace artefacts.

Each task was refined through repeated testing until the evaluation criteria reliably distinguished genuine model failures from flaws in the benchmark itself. Criteria that rejected correct behavior, or admitted incorrect solutions, were revised before release.

Tests and Outcomes

Even the strongest models failed most tasks under the benchmark's strict grading system, with performance spread across a wide range, rather than clustering atop:

Rank	Model (configuration)	Developer	Strict pass@1
1	Claude Fable 5 (adaptive/max)	Anthropic	36.2%
2	Claude Fable 5	Anthropic	34.2%
3	GPT-5.6 Sol (max)	OpenAI	23.5%
4	Claude Opus 4.8 (adaptive/max)	Anthropic	21.9%
5	GPT-5.6 Sol	OpenAI	21.5%
5	GPT-5.5	OpenAI	21.5%
5	GPT-5.5 (xhigh)	OpenAI	21.5%
8	Claude Opus 4.8	Anthropic	18.9%
9	Grok 4.5 (high)	xAI	15.8%
10	Muse Spark 1.1 (xhigh)	Muse	13.5%
11	GLM 5.2	Zhipu AI	12.7%
12	Kimi K3 (max)	Moonshot AI	11.9%
13	Gemini 3.5 Flash (high)	Google	11.2%
14	Claude Sonnet 4.6 (adaptive/max)	Anthropic	10.4%
15	Gemini 3.1 Pro	Google	10.0%
15	GLM 5.2 (xhigh)	Zhipu AI	10.0%
17	DeepSeek V4 Pro (xhigh)	DeepSeek	9.2%
17	Gemini 3.5 Flash	Google	9.2%
19	Qwen 3.7 Max	Alibaba	8.5%
20	Claude Sonnet 4.6	Anthropic	7.7%
21	DeepSeek V4 Flash	DeepSeek	7.3%
21	DeepSeek V4 Flash (xhigh)	DeepSeek	7.3%
23	DeepSeek V4 Pro	DeepSeek	6.9%
23	Kimi K2.6	Moonshot AI	6.9%
25	Gemini 3.6 Flash (high)	Google	5.0%
26	Gemini 3.5 Flash-Lite (high)	Google	3.1%
27	Grok 4.3 (high)	xAI	1.9%
27	Inkling (max)	Inkling	1.9%
29	Nemotron 3 Ultra	NVIDIA	1.5%
30	Grok 4.3	xAI	0.8%

The initial results leaderboard ranking all 30 evaluated model configurations by their strict pass@1 scores on the HANDBOOK.md benchmark. Scores represent the percentage of the benchmark's 65 workplace tasks completed without a single failed evaluation criterion across four trials per task. Tied scores share the same rank.

Differences between reasoning settings were also found to vary substantially by model, with additional reasoning sometimes improving performance; sometimes making little difference; and sometimes *reducing* it:

> *'[Reasoning] effort helps unevenly. Raising effort improves Opus 4.8 (+3.0), Sonnet 4.6 (+2.7), and Fable 5 (+2.0), leaves GPT-5.5 unchanged (21.5% at both settings), and hurts GLM 5.2 (−2.7).*
>
> *'Additional deliberation appears to convert into rule compliance only when the underlying failure is a missed inference rather than a missed [read].'*

Claude Fable 5 achieved the highest score at 36.2%, followed by Claude Fable 5 at 34.2%, and GPT-5.6 Sol (max) at 23.5%. GPT-5.5 and Claude Opus 4.8 formed the next tier, at around 20%, while most remaining frontier models scored below 16%. The lowest-ranked models completed fewer than 2% of tasks.

The paper's detailed failure analysis suggests that the problem was often not a lack of information, since relevant handbook rules and supporting evidence had frequently already been retrieved – yet additional reasoning would sometimes cause models to abandon the correct conclusion in favor of a plausible but policy-violating one.

The work also notes the extent of self-delusion that characterizes so many of the LLMs' forays:

> *'Nearly every failed trajectory ends with a confident statement that the handbook was followed, frequently citing the specific sections that were violated. The reports are detailed, well-structured, and wrong […]*
>
> *' […] Across the benchmark, the agent's self-report is the least reliable artifact in the trajectory, which matters for any deployment that surfaces agent summaries to humans as evidence of what was done.'*

In closing, The authors conclude that long-context reasoning alone is unlikely to make enterprise AI reliably follow company policy. Instead, policy compliance should increasingly be enforced through deterministic external controls, in addition to workflow guardrails.

Conclusion

Opinion One interesting consideration is the eventual extent to which environments like those created for the experiments will be re-imagined to facilitate AI, instead of forcing LLMs to interpret the same forms and formats that define human office environments. For example, yesterday, for the first time, a business contact sent me an .md overview designed to be explored by an LLM, rather than to be read in a linear fashion.

And I too am increasingly adopting and adapting to the visual and textual constraints during LLM colloquies, as well as selecting and accepting file formats that I would not normally choose, because they accommodate the LLM workflow more adroitly.

Therefore, while it is amusing to watch frontier models stumble around in David Brent's world, one wonders if this is the likeliest scenario for the 'agentic office worker'.

First published Wednesday, July 29, 2026. Updated 18:41 EET, fixed broken link.

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