

DAYJOB: Can Agents Survive a 9 to 5?

PRIMARY

Publisher	Surge AI
Published	2026-10-04
Retrieved	2026-10-05
Cite key	surgeai2026dayjob
URL	https://surgehq.ai/blog/dayjob

Assessment. The primary actor's own write-up. Used for prompt lengths, input-file counts and task construction. Its per-task hour figures conflict with the paper's and were not used.

130 expert assignments across healthcare and finance.

- Finance tasks average 25.7 input files; healthcare 19.8.
- "The strongest models score less than 25% on both DAYJOB: Healthcare and DAYJOB: Finance." Highest: 24.7% healthcare, 23.9% finance.
- Task-specific rubrics, approximately 57.5 criteria in finance and 47.5 in healthcare.
- Prompts average 81 words in finance and 49 in healthcare, "far shorter than GDPval's typical 337 words". The brevity is deliberate: agents must scope and interpret the request themselves.
- Built by experts with direct industry experience, including a senior portfolio manager and a paediatric endocrinologist/epidemiologist. Three-layer expert review.
- Per-task hour estimates on this page: 21.6 h finance, 19.6 h healthcare (conflicts with the paper's abstract — not used).
- A claimed "\$25.7 million pricing error across 66 test runs" sits adjacent to the "25.7 input files" figure on the same page; treated as a possible conflation and not used.