

Source dossier

The sector is not the test — agents clear a quarter of real professional work, and Annex III never asked

Edition 2026-10-05 · narrative: A benchmark lands in the two sectors everybody calls high-risk, and the AI Act classifies by neither. · compiled 2026-10-05

Claims index

Claim	Statement	Status	Sources
CLAIM-2026-10-05-01	DAYJOB is a benchmark of 130 tasks built by working professionals, split 50 in healthcare and 80 in finance.	verified	arxiv2026dayjob, surgeai2026dayjob
CLAIM-2026-10-05-02	The strongest configuration cleared 24.7% on DAYJOB healthcare and 23.9% on DAYJOB finance.	verified	arxiv2026dayjob, surgeai2026healthcare, surgeai2026finance
CLAIM-2026-10-05-03	Across 30 model configurations from 13 developers, the median configuration cleared 0.6% in healthcare and 2.5% in finance.	verified	arxiv2026dayjob
CLAIM-2026-10-05-04	On the live DAYJOB healthcare board the leader at 24.7% is more than double the runner-up at 11.6%; on finance the top four sit within 4.1 points.	verified	surgeai2026healthcare, surgeai2026finance
CLAIM-2026-10-05-05	DAYJOB prompts average 81 words in finance and 49 in healthcare.	verified	surgeai2026dayjob
CLAIM-2026-10-05-06	DAYJOB grades each task against a rubric carrying a median of 57.5 binary criteria in finance and 47.5 in healthcare.	verified	arxiv2026dayjob, surgeai2026dayjob
CLAIM-2026-10-05-07	DAYJOB grading is performed by an agentic LLM judge running inside the task container; the harness is Apache-2.0.	verified	github2026surgeai-da
CLAIM-2026-10-05-08	AI Act Annex III point 5 is written as use cases about natural persons: 5(a) public authorities evaluating eligibility for essential public assistance benefits including healthcare, 5(b) creditworthiness of natural persons excluding fraud detection, 5(c) risk assessment and pricing for life and health insurance.	verified	euaiactexpl2026annex
CLAIM-2026-10-05-09	Article 6(3) takes an Annex III system out of the high-risk category where it performs a narrow procedural task, improves the result of a previously completed human activity, detects deviations without replacing human review, or performs a preparatory task — and never where it performs profiling of natural persons.	verified	euaiactexpl2026classification

Claim	Statement	Status	Sources
CLAIM-2026-10-05-10	Article 6(4) requires a provider relying on Article 6(3) to document that assessment before the system is placed on the market or put into service, to provide it to national competent authorities on request, and to register under Article 49(2).	verified	euaiactexpl2026classific, euaiactexpl2026registrat
CLAIM-2026-10-05-11	Article 49(2) requires providers who have concluded their Annex III system is not high-risk to register themselves and that system in the EU database referred to in Article 71.	verified	euaiactexpl2026registrat
CLAIM-2026-10-05-12	Regulation (EU) 2026/1744, the Digital Omnibus on AI, is dated 8 July 2026 and entered into force on 27 July 2026.	verified	eurlex2026regulation
CLAIM-2026-10-05-13	Under the Digital Omnibus, Article 50 transparency applies from 2 August 2026, the marking obligation from 2 December 2026, Annex III high-risk obligations from 2 December 2027 and Annex I from 2 August 2028.	verified	huntonandre2026eu, eurlex2026regulation
CLAIM-2026-10-05-14	The Bundesnetzagentur announced on 29 July 2026 that it takes the central role in implementing the AI Regulation in Germany, under the KI-MIG in force since the same date.	verified	bundesnetza2026bundesnet, bundesamt2026kimig
CLAIM-2026-10-05-15	The DAYJOB paper was submitted to arXiv as 2610.01306v1 on 1 October 2026.	verified	arxiv2026dayjob
CLAIM-2026-10-05-16	DAYJOB finance tasks span corporate finance, banking, credit, investing and real assets; healthcare tasks span clinical, pharmacy, operations, payer and compliance.	verified	surgeai2026finance, surgeai2026healthcare
CLAIM-2026-10-05-17	Per-task time estimates: the paper's abstract gives 13.6 hours for healthcare and 16.6 for finance; Surge AI's own page gives 19.6 and 21.6.	disputed	arxiv2026dayjob, surgeai2026dayjob
CLAIM-2026-10-05-18	The number of healthcare tasks: the paper states 50, the live healthcare leaderboard lists 65.	disputed	arxiv2026dayjob, surgeai2026healthcare
CLAIM-2026-10-05-19	Only one model caught a \$25.7 million pricing error, across 66 runs.	single-source	surgeai2026dayjob
CLAIM-2026-10-05-20	A comparable professional-work benchmark uses prompts averaging 337 words.	single-source	surgeai2026dayjob

Only claims marked **verified** may carry a post. **single-source** may appear as hedged context; **disputed** must not be published at all.

DAYJOB: A Benchmark for Long-Horizon Professional Work

PRIMARY

Publisher	arXiv
Published	2026-10-01
Retrieved	2026-10-05
Cite key	arxiv2026dayjob
URL	https://arxiv.org/abs/2610.01306

Assessment. The paper itself. arXiv HTML and PDF were both refused by the fetch proxy with HTTP 429 on 5 Oct 2026, so the abstract page is the archived text. Affiliations are not printed on the abs page; the work is published by Surge AI on its own site, which is the attribution used.

Abstract (abs page, v1, submitted 1 October 2026)

"Professional work often starts with a brief request that leaves the professional to work out what is needed, which documents matter, and whether the request's premise holds. We introduce DAYJOB, a benchmark of 130 tasks built by professionals in healthcare (50) and finance (80)."

Reported figures

- 130 tasks total: 50 healthcare, 80 finance.
- Median 47.5 binary criteria per healthcare task; 57.5 per finance task.
- Best configuration: 24.7% healthcare, 23.9% finance (Claude Opus 5.5).
- Median configuration: 0.6% healthcare, 2.5% finance.
- 30 model configurations from 13 developers evaluated.

Note on a figure NOT used

The abs extraction also reported per-task time estimates of 13.6 h (healthcare) and 16.6 h (finance). Surge AI's own page gives 19.6 h and 21.6 h. The two do not agree; the figure is treated as disputed and appears nowhere in the edition.

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.

Finance benchmark leaderboard

PRIMARY

Publisher	Surge AI
Published	2026-10-05
Retrieved	2026-10-05
Cite key	surgeai2026finance
URL	https://surgehq.ai/benchmarks/dayjob-finance

Assessment. Live board. Top-four spread of 4.1 points taken from here.

80 assignments built by finance professionals. Domains: corporate finance, banking, credit, investing, real assets. Scored on whether "analysis they hand back is usable by someone else".

#	Model	Score
1	Claude Opus 5.5 (Adaptive/Max)	23.9%
2	GPT 6 Astra (Max reasoning)	21.5%
3	Gemini 4 Argon (High reasoning)	20.3%
4	Claude Fable 5.1 (Adaptive/Max)	19.8%
5	Muse Spark 1.3 (Max reasoning)	14.8%
6	Grok 4.7 (xHigh reasoning)	14.5%
7	Grok 4.6 (xHigh reasoning)	12.8%
8	Claude Opus 5 (Adaptive/Max)	11.3%
9	GPT 6 Sol (Max reasoning)	9.3%
10	Claude Fable 5 (Adaptive/Max)	9%

Board continues to 27 rows. Top four span 23.9 to 19.8 = 4.1 points.

Healthcare benchmark leaderboard

PRIMARY

Publisher

Surge AI

Published

2026-10-05

Retrieved

2026-10-05

Cite key

surgeai2026healthcar

URL

<https://surgehq.ai/benchmarks/dayjob-healthcare>

Assessment. Live board, 32 rows. Lists 65 assignments against the paper's 50 — the set has grown since submission. The discrepancy is disclosed in the first comment; the paper's 50/80 split is what the edition quotes.

65 expert-built assignments across clinical, pharmacy, operations, payer and compliance domains. "One missed detail changes the right call."

#	Model	Org	Score
1	Claude Opus 5.5 (Adaptive/Max)	Anthropic	24.7%
2	GPT 6 Astra (Max reasoning)	OpenAI	11.6%
3	Claude Fable 5.1 (Adaptive/Max)	Anthropic	9.6%
4	Gemini 4 Argon (High reasoning)	Google	9.6%
5	Claude Opus 5 (Adaptive/Max)	Anthropic	8.4%
6	Grok 4.7 (xHigh reasoning)	xAI	8.4%
7	Muse Spark 1.3 (Max reasoning)	Meta	7.8%
8	Claude Fable 5 (Adaptive/Max)	Anthropic	5.6%
9	GPT 6 Sol (Max reasoning)	OpenAI	3.6%
10	Grok 4.6 (xHigh reasoning)	xAI	2.8%

Rows 11-32 run from 2% down to 0%. Thirteen distinct developers appear: Anthropic, OpenAI, Google, xAI, Meta, Alibaba, Moonshot AI, Tencent, Zhipu AI, DeepSeek, Thinking Machines, NVIDIA, Mistral AI. Leader is 2.13x the runner-up.

surge-ai/dayjob evaluation harness

PRIMARY

Publisher	GitHub
Published	2026-10-01
Retrieved	2026-10-05
Cite key	github2026surgeaida
URL	https://github.com/surge-ai/dayjob

Assessment. The harness. The one place that states what the grader actually is.

Evaluation harness for DAYJOB. Tasks distributed on Hugging Face under the Surge AI account.

Grading uses "Claude Opus 4.8 as an agentic LLM judge (OpenCode) running inside the task container". Tasks follow Harbor format and run through the OpenHands SDK. Licence: Apache-2.0.

The README carries no task counts, scores or developer counts — those come from the paper and the boards.

Classification Rules for High-Risk AI Systems

SECONDARY

Publisher EU AI Act Explorer

Published

Retrieved 2026-10-05

Cite key euaiactexpl2026classific

URL <https://artificialintelligenceact.eu/article/6/>

Assessment. Unofficial consolidated text of Regulation (EU) 2024/1689 Article 6. Reliable for wording; EUR-Lex is the authority.

Article 6(3) — derogation

An Annex III system is not high-risk where it "does not pose a significant risk of harm to the health, safety or fundamental rights of natural persons" and meets any of:

- "the AI system is intended to perform a narrow procedural task";
- "the AI system is intended to improve the result of a previously completed human activity";
- "the AI system is intended to detect decision-making patterns or deviations from prior decision-making patterns and is not meant to replace or influence the previously completed human assessment, without proper human review";
- "the AI system is intended to perform a preparatory task to an assessment relevant for the purposes of the use cases listed in Annex III".

The derogation does not apply where "the AI system performs profiling of natural persons".

Article 6(4) — what claiming it costs

A provider that considers a system is not high-risk "shall document its assessment before that system is placed on the market or put into service" and provide that documentation to national competent authorities on request. Such providers are "subject to the registration obligation set out in Article 49(2)".

Article 6(5)

The Commission was to provide guidelines with a comprehensive list of practical examples of high-risk and non-high-risk use cases by 2 February 2026.

Annex III: High-Risk AI Systems Referred to in Article 6(2)

SECONDARY

Publisher EU AI Act Explorer

Published

Retrieved 2026-10-05

Cite key euaiactexpl2026annex

URL <https://artificialintelligenceact.eu/annex/3/>

Assessment. The operative list. Point 5 is where the sector-versus-use-case argument is settled.

Point 5 — access to and enjoyment of essential private services and essential public services and benefits

(a) "AI systems intended to be used by public authorities or on behalf of public authorities to evaluate the eligibility of natural persons for essential public assistance benefits and services, including healthcare services, as well as to grant, reduce, revoke, or reclaim such benefits and services";

(b) "AI systems intended to be used to evaluate the creditworthiness of natural persons or establish their credit score, with the exception of AI systems used for the purpose of detecting financial fraud";

(c) "AI systems intended to be used for risk assessment and pricing in relation to natural persons in the case of life and health insurance";

(d) "AI systems intended to evaluate and classify emergency calls by natural persons or to be used to dispatch, or to establish priority in the dispatching of, emergency first response services, including by police, firefighters and medical aid, as well as of emergency healthcare patient triage systems".

Point 4 — employment

(a) recruitment or selection of natural persons, including targeted job advertisements, filtering applications, evaluating candidates; (b) decisions affecting terms of work-related relationships, promotion or termination, task allocation based on individual behaviour or traits, monitoring and evaluating performance and behaviour.

Every limb is written as an intended use, and point 5 is written about natural persons. Neither "healthcare" nor "finance" appears as a sector.

Registration of high-risk AI systems

SECONDARY

Publisher EU AI Act Explorer

Published

Retrieved 2026-10-05

Cite key euaiactexpl2026registrat

URL <https://artificialintelligenceact.eu/article/49/>

Assessment. Confirms the Article 6(4) cross-reference lands in the single EU database under Article 71.

Article 49(2): a provider (or its authorised representative) that considers an Annex III system is not high-risk under Article 6(3) "shall register themselves and that system in the EU database referred to in Article 71".

The registration is in the single EU database, not a national registry, and it applies precisely to systems the provider has self-classified OUT of the high-risk category.

Regulation (EU) 2026/1744 (Digital Omnibus on AI)

PRIMARY

Publisher	EUR-Lex
Published	2026-07-08
Retrieved	2026-10-05
Cite key	eurlex2026regulatio
URL	https://eur-lex.europa.eu/eli/reg/2026/1744/oj

Assessment. The authority for the instrument and its dates. The EUR-Lex page served metadata rather than the article text on this fetch, so the amended application dates are carried by the secondary summary below.

Title: "Regulation (EU) 2026/1744 of the European Parliament and of the Council of 8 July 2026 amending Regulations (EU) 2024/1689, (EU) 2018/1139 and (EU) 2023/1230..."

Subject: simplification of harmonised rules on artificial intelligence (Digital Omnibus on AI). Entry into force: 27 July 2026.

The fetch returned the ELI metadata header only; the substantive amending provisions were not in the served content. The new application dates are taken from the summary archived alongside this source.

EU Digital Omnibus on AI Enters Into Force

SECONDARY

Publisher	Hunton Andrews Kurth
Published	2026-07-28
Retrieved	2026-10-05
Cite key	huntonandre2026eu
URL	https://www.hunton.com/privacy-and-cybersecurity-law-blog/eu-digital-omnibus-on-ai-enters-into-force

Assessment. Law-firm summary, used only for the amended application dates, which match the dates held as settled by the pipeline.

New application dates under the Digital Omnibus:

- Article 50 transparency obligations: 2 August 2026 — "Transparency obligations enter into force."
- AI-generated content marking: 2 December 2026 — "Certain AI-generated content marking obligations will apply to systems already on the market before August 2, 2026."
- Annex III stand-alone high-risk systems: 2 December 2027 — "New application date for the main compliance obligations for stand-alone high-risk AI systems listed in Annex III, including systems used in employment and education."
- Annex I high-risk embedded in regulated products: 2 August 2028.

The post does not address whether the deferral also moves the Article 6(3) documentation duty or the Article 49 registration duty. The edition therefore makes no claim that it does: it states only that the Article 6(4) duty is keyed to placing on the market.

Bundesnetzagentur übernimmt zentrale Rolle bei der Umsetzung der KI-Verordnung

PRIMARY

Publisher Bundesnetzagentur

Published 2026-07-29

Retrieved 2026-10-05

Cite key bundesnetza2026bundesnet

URL https://www.bundesnetzagentur.de/SharedDocs/Pressemitteilungen/DE/2026/20260729_KI_VO.html

Assessment. The authority's own announcement of its role. Cited for the German addressee only.

Press release of 29 July 2026: the Bundesnetzagentur takes the central role in implementing the AI Regulation in Germany, including market surveillance.

KI-MIG

PRIMARY

Publisher	Bundesamt fuer Justiz
Published	2026-07-29
Retrieved	2026-10-05
Cite key	bundesamtfu2026kimig
URL	https://www.gesetze-im-internet.de/ki-mig/

Assessment. The German implementing act. Cited as the instrument behind the Bundesnetzagentur's designation.

Gesetz zur Marktueberwachung und Innovationsfoerderung von Kuenstlicher Intelligenz (KI-MIG), in force since 29 July 2026. Official text at [gesetze-im-internet.de](https://www.gesetze-im-internet.de).