

Source dossier

The skill file was the spec — agents satisfy every requirement of a procedure 12.6% to 40.9% of the time, and the live duty that asks about procedure is not in the AI Act

Edition 2026-10-16 · narrative: practitioner + EU regulatory lens · compiled 2026-10-16

Claims index

Claim	Statement	Status	Sources
CLAIM-2026-10-16-01	SkillForgeBench (arXiv 2610.14918 v1, submitted 10 October 2026) comprises 842 sandboxed tasks across five major professional domains and evaluates eight mainstream agents, using agent skills as unifying execution blueprints.	verified	arxiv2026skillforg, arxivexport2026metadata
CLAIM-2026-10-16-02	On SkillForgeBench, Opus 4.8 leads at 40.9% strict completion, and strict completion across the eight agents ranges from 12.6% to 40.9%. The top of that range equals the leader's score by construction, not by coincidence.	verified	arxiv2026skillforg
CLAIM-2026-10-16-03	The same agents' individual rubric pass rates run from 70.0% to 85.7%; the abstract states the 12.6%-40.9% strict range against that 70.0%-85.7% per-rubric range, so requiring all constraints at once is what produces the gap.	verified	arxiv2026skillforg
CLAIM-2026-10-16-04	Under external guidance from skills, top-tier agents obtain larger gains than low-tier agents: 19.2% for Opus 4.8 against 5.8% for Haiku 4.5, as the abstract states.	verified	arxiv2026skillforg
CLAIM-2026-10-16-05	Agents satisfy process rubrics far less frequently than completion rubrics, under 3% for Haiku 4.5, which the authors state proves that completion-only checks overestimate real-world agent reliability.	verified	arxiv2026skillforg
CLAIM-2026-10-16-06	AI Act Article 14 human oversight is a high-risk obligation; Annex III high-risk obligations were moved to 2 December 2027 by Regulation (EU) 2026/1744, in force 27 July 2026.	verified	euaiactexpl2026human, eurlex2026regulatio
CLAIM-2026-10-16-07	AI Act Article 50 transparency obligations have applied since 2 August 2026 and the marking-of-generated-content obligation lands 2 December 2026; both are duties to disclose and mark, not to inspect an AI system's procedure.	verified	euartificia2026transparent, eurlex2026regulatio

Claim	Statement	Status	Sources
CLAIM-2026-10-16-08	GDPR Article 32(1)(d) requires a process for regularly testing, assessing and evaluating the effectiveness of technical and organisational measures, and Article 5(2) requires the controller to be able to demonstrate compliance. Both are in force with no deferral, and both apply only where personal data is processed.	verified	eurlex2016regulation
CLAIM-2026-10-16-09	No author affiliations appear on the arXiv abstract record for 2610.14918 and arxiv.org/html/2610.14918v1 returned HTTP 429 with a do-not-retry instruction during this run, so the work is credited to the researchers and every figure published is quoted from the abstract's own wording rather than from a table in the full text.	verified	arxiv2026skillforg, arxivexport2026metadata

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

SkillForgeBench: Evaluating Agent Reliability Grounded in Agent Skill Specifications

PRIMARY

Publisher	arXiv
Published	2026-10-10
Retrieved	2026-10-16
Cite key	arxiv2026skillforg
URL	https://arxiv.org/abs/2610.14918

arXiv:2610.14918 v1, submitted 10 October 2026, cs.AI (cross-list cs.CL). 38 pages, 9 figures, 7 tables including appendices; the first two authors contributed equally. No author affiliations are shown on the arXiv abstract record, and arxiv.org/html/2610.14918v1 returned HTTP 429 with an instruction not to retry during this run, so the work is credited to the researchers rather than to an institution.

Abstract, verbatim in the parts quoted by this edition: "Comprising 842 sandboxed tasks across five major professional domains, SkillForgeBench reveals a steep capability hierarchy among eight mainstream agents, led by Opus 4.8 (40.9% strict completion). Requiring all constrains satisfaction uncovers severe capability bottlenecks across all model tiers that remain masked under individual rubric pass rates (70.0% to 85.7% vs. 12.6% to 40.9%). Crucially, under the external guidance from skills, top-tier agents obtain larger gains than low-tier agents (i.e., 19.2% for Opus 4.8 vs. 5.8% for Haiku 4.5). Finally, agents satisfy process rubrics far less frequently than completion rubrics (under 3% for Haiku 4.5), proving that completion-only checks overestimate real-world agent reliability."

The three co-designed stages named in the abstract are Resource Preparation, Skill-driven Task Synthesis and Holistic Evaluation Design; the last "translates skill-embedded requirements into fine-grained completion and process rubrics to enable strict requirement-level scoring".

Metadata record for 2610.14918

PRIMARY

Publisher	arXiv export API
Published	2026-10-10
Retrieved	2026-10-16
Cite key	arxivexport2026metadata
URL	https://export.arxiv.org/api/query?id_list=2610.14918

Used to verify the identifier against the story rather than the topic (research check 4) after `arxiv.org/html` returned HTTP 429. Returned title, published 2026-10-10T13:40:38Z, the nine author names, and the summary verbatim. Per D109 section 5 the export API is the default arXiv metadata route, not a fallback.

Human oversight — Article 14, Regulation (EU) 2024/1689

SECONDARY

Publisher	EU AI Act Explorer
Published	2026-01-01
Retrieved	2026-10-16
Cite key	euaiactexpl2026human
URL	https://artificialintelligenceact.eu/article/14/

Article 14 applies to high-risk AI systems. Annex III high-risk obligations, Article 14 among them, were moved to 2 December 2027 by Regulation (EU) 2026/1744.

Transparency obligations for providers and deployers of certain AI systems — Article 50

SECONDARY

Publisher	EU Artificial Intelligence Act
Published	2026-01-01
Retrieved	2026-10-16
Cite key	euartificia2026transpare
URL	https://artificialintelligenceact.eu/article/50/

Article 50 transparency obligations apply since 2 August 2026. The machine-readable marking of generated content obligation lands 2 December 2026. The duty is to disclose and mark, not to inspect how an AI system carried out a task.

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

PRIMARY

Publisher	EUR-Lex
Published	2026-07-27
Retrieved	2026-10-16
Cite key	eurlex2026regulatio
URL	https://eur-lex.europa.eu/eli/reg/2026/1744/oj

In force 27 July 2026. Amends Regulation (EU) 2024/1689. Moved Annex III high-risk obligations to 2 December 2027, Annex I to 2 August 2028 and member-state regulatory sandboxes to 2 August 2027.

Regulation (EU) 2016/679 — General Data Protection Regulation

PRIMARY

Publisher	EUR-Lex
Published	2016-04-27
Retrieved	2026-10-16
Cite key	eurlex2016regulatio
URL	https://eur-lex.europa.eu/eli/reg/2016/679/oj

Article 32(1)(d) requires, as appropriate, "a process for regularly testing, assessing and evaluating the effectiveness of technical and organisational measures for ensuring the security of the processing". Article 5(2) makes the controller responsible for, and able to demonstrate compliance with, the principles in Article 5(1). Both are in force and carry no deferral; both apply only where personal data is processed.