

DeepSeek releases V4, a 1.6-trillion-parameter open-weights model with 1M context

Source copy kept by wheresthe.ai for the story "DeepSeek releases V4, a 1.6-trillion-parameter open-weights model with 1M context" (<https://wheresthe.ai/d/cn-deepseek-v4-open-weights-2026-04/>).

Source: <https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro>

Text saved 18 Oct 2026 (the page as published).

SHA-256 of the saved text: bd29b0993021deee6221abd0b0ace8b5b60282d9766c4f75ab2bf8906ffb33c1

```

{"__type":"AddedToken","content":"< | end of sentence | >","lstrip":false,"normalized":true,"rstrip":false,"single_word":false},"eos_token":{"__type":"AddedToken","content":"< | end of sentence | >","lstrip":false,"normalized":true,"rstrip":false,"single_word":false},"pad_token":{"__type":"AddedToken","content":"< | end of sentence | >","lstrip":false,"normalized":true,"rstrip":false,"single_word":false},"unk_token":null},"createdAt":"2026-04-22T06:04:created","downloads":373563,"downloadsAllTime":10895485,"id":"deepseek-ai/DeepSeek-V4-Pro","isLikedByUser":false,"availableInferenceProviders":[{"provider":"together","modelStatus":"error","providerStatus":"live","providerId":"deepseek-ai/DeepSeek-V4-Pro","task":"conversational","adapterWeightsPath":"model-00001-of-00064.safetensors","features":{"structuredOutput":true,"toolCalling":true},"isCheapestPricingOutput":false,"isFastestThroughput":false,"isModelAuthor":{"provider":"novita","modelStatus":"live","providerStatus":"live","providerId":"deepseek/deepseek-v4-pro","task":"conversational","adapterWeightsPath":"model-00001-of-00064.safetensors","features":{"structuredOutput":false,"toolCalling":true},"isCheapestPricingOutput":false,"isFastestThroughput":false,"isModelAuthor":{"provider":"featherless-ai","modelStatus":"live","providerStatus":"live","providerId":"deepseek-ai/DeepSeek-V4-Pro","task":"conversational","adapterWeightsPath":"model-00001-of-00064.safetensors","isCheapestPricingOutput":false,"isFastestThroughput":false,"isModelAuthor":{"provider":"deepinfra","modelStatus":"live","providerStatus":"live","providerId":"deepseek-ai/DeepSeek-V4-Pro","task":"conversational","adapterWeightsPath":"model-00001-of-00064.safetensors","features":{"structuredOutput":true,"toolCalling":true},"isCheapestPricingOutput":true,"isFastestThroughput":true,"isModelAuthor":{"provider":"baseten","modelStatus":"error","providerStatus":"live","providerId":"deepseek-ai/DeepSeek-V4-Pro","task":"conversational","adapterWeightsPath":"model-00001-of-00064.safetensors","features":{"structuredOutput":false,"toolCalling":true},"isCheapestPricingOutput":false,"isFastestThroughput":false,"isModelAuthor":{"generation","library_name":"transformers","librariesOther":[],"trackDownloads":true,"model-index":null,"evalResults":{"dataset":{"id":"TIGER-Lab/MMLU-Pro","isBenchmark":true,"task_id":"mmlu_pro"},"value":87.5,"source":{"url":"https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro","name":"Model Card","isExternal":false},"filename":"eval_results/mmlu-pro.yaml"},"verified":false,"pullRequest":110,"rank":5,"label":"Mmlu Pro"},"dataset":{"id":"openai/gsm8k","isBenchmark":true,"task_id":"gsm8k"},"value":92.6,"source":{"url":"https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro","name":"Model Card","isExternal":false},"filename":"eval_results/gsm8k.yaml"},"verified":false,"pullRequest":110,"rank":5,"label":"Gsm8k"},"dataset":{"id":"SWE-bench/SWE-bench_Verified","isBenchmark":true,"task_id":"swe_bench_%_resolved"},"value":80.6,"source":{"url":"https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro","name":"Model Card","isExternal":false},"filename":"eval_results/swe_bench_verified.yaml"},"verified":false,"pullRequest":110,"rank":6,"label":"Swe Bench Resolved"},"dataset":{"id":"agents-last-exam/agents-last-exam","isBenchmark":true,"task_id":"overall-pass-rate"},"value":12.390350877192983,"source":{"url":"https://agents-last-exam.org/leaderboard","name":"ALE official leaderboard","isExternal":true,"author":{"id":"69f848800d5d932e60c42eac"},"avatarUrl":"https://cdn-avatars.huggingface.co/v1/production/uploads/6582486355a1e6cdb3b5e19d/eCSzorSXXKmJ5srUhvq7H.png"},"fullName":"agents-last-exam","name":"agents-last-exam","type":"org","isHf":false,"isHfAdmin":false,"isMod":false,"followerCount":75,"isUserFollowing":false},"filename":"ale-openclaw-default.yaml"},"verified":false,"pullRequest":215,"rank":7,"label":"Overall Pass Rate","notes":"harness=openclaw; harness_variant=not reported; attempted=146/152; official leaderboard snapshot=2026-10-01T20:39:17.826877+00:00; values are percentages (0-100); equal-task aggregation; unattempted tasks contribute zero; pass threshold score>=0.999. This is an official leaderboard export, not an HF-verified evaluation."},"dataset":{"id":"agents-last-exam/agents-last-exam","isBenchmark":true,"task_id":"overall-score"},"value":27.62014846817647,"source":{"url":"https://agents-last-exam.org/leaderboard","name":"ALE official leaderboard","isExternal":true,"author":{"id":"69f848800d5d932e60c42eac"},"avatarUrl":"https://cdn-

```

avatars.huggingface.co/v1/production/uploads/6582486355a1e6cdb3b5e19d/eCSzSXXXKmj5srUhvq7H.png", "fullname": "agents-last-exam", "name": "agents-last-exam", "type": "org", "isHf": false, "isHfAdmin": false, "isMod": false, "followerCount": 75, "isUserFollowing": false}}, "filename": "ale-openclaw-default.yaml", "verified": false, "pullRequest": 215, "rank": 5, "label": "Overall Score", "notes": "harness=openclaw; harness_variant=not reported; attempted=146/152; official leaderboard snapshot=2026-10-01T20:39:17.826877+00:00; values are percentages (0-100); equal-task aggregation; unattempted tasks contribute zero; pass threshold score>=0.999. This is an official leaderboard export, not an HF-verified evaluation.", {"dataset": {"id": "agents-last-exam/agents-last-exam", "isBenchmark": true, "task_id": "ale-cli-pass-rate"}, "value": 14.126984126984127, "source": {"url": "https://agents-last-exam.org/leaderboard", "name": "ALE official leaderboard", "isExternal": true, "author": {"_id": "69f848800d5d932e60c42eac", "avatarUrl": "https://cdn-avatars.huggingface.co/v1/production/uploads/6582486355a1e6cdb3b5e19d/eCSzSXXXKmj5srUhvq7H.png", "fullname": "agents-last-exam", "name": "agents-last-exam", "type": "org", "isHf": false, "isHfAdmin": false, "isMod": false, "followerCount": 75, "isUserFollowing": false}}, "filename": "ale-openclaw-default.yaml", "verified": false, "pullRequest": 215, "rank": 8, "label": "Ale Cli Pass Rate", "notes": "harness=openclaw; harness_variant=not reported; attempted=104/105; official leaderboard snapshot=2026-10-01T20:39:17.826877+00:00; values are percentages (0-100); equal-task aggregation; unattempted tasks contribute zero; pass threshold score>=0.999. This is an official leaderboard export, not an HF-verified evaluation.", {"dataset": {"id": "agents-last-exam/agents-last-exam", "isBenchmark": true, "task_id": "ale-cli-score"}, "value": 30.84164065884846, "source": {"url": "https://agents-last-exam.org/leaderboard", "name": "ALE official leaderboard", "isExternal": true, "author": {"_id": "69f848800d5d932e60c42eac", "avatarUrl": "https://cdn-avatars.huggingface.co/v1/production/uploads/6582486355a1e6cdb3b5e19d/eCSzSXXXKmj5srUhvq7H.png", "fullname": "agents-last-exam", "name": "agents-last-exam", "type": "org", "isHf": false, "isHfAdmin": false, "isMod": false, "followerCount": 75, "isUserFollowing": false}}, "filename": "ale-openclaw-default.yaml", "verified": false, "pullRequest": 215, "rank": 4, "label": "Ale Cli Score", "notes": "harness=openclaw; harness_variant=not reported; attempted=104/105; official leaderboard snapshot=2026-10-01T20:39:17.826877+00:00; values are percentages (0-100); equal-task aggregation; unattempted tasks contribute zero; pass threshold score>=0.999. This is an official leaderboard export, not an HF-verified evaluation.", {"dataset": {"id": "agents-last-exam/agents-last-exam", "isBenchmark": true, "task_id": "near-term-pass-rate"}, "value": 19.900497512437813, "source": {"url": "https://agents-last-exam.org/leaderboard", "name": "ALE official leaderboard", "isExternal": true, "author": {"_id": "69f848800d5d932e60c42eac", "avatarUrl": "https://cdn-avatars.huggingface.co/v1/production/uploads/6582486355a1e6cdb3b5e19d/eCSzSXXXKmj5srUhvq7H.png", "fullname": "agents-last-exam", "name": "agents-last-exam", "type": "org", "isHf": false, "isHfAdmin": false, "isMod": false, "followerCount": 75, "isUserFollowing": false}}, "filename": "ale-openclaw-default.yaml", "verified": false, "pullRequest": 215, "rank": 4, "label": "Near Term Pass Rate", "notes": "harness=openclaw; harness_variant=not reported; attempted=67/67; official leaderboard snapshot=2026-10-01T20:39:17.826877+00:00; values are percentages (0-100); equal-task aggregation; unattempted tasks contribute zero; pass threshold score>=0.999. This is an official leaderboard export, not an HF-verified evaluation.", {"dataset": {"id": "agents-last-exam/agents-last-exam", "isBenchmark": true, "task_id": "near-term-score"}, "value": 43.81171518153468, "source": {"url": "https://agents-last-exam.org/leaderboard", "name": "ALE official leaderboard", "isExternal": true, "author": {"_id": "69f848800d5d932e60c42eac", "avatarUrl": "https://cdn-avatars.huggingface.co/v1/production/uploads/6582486355a1e6cdb3b5e19d/eCSzSXXXKmj5srUhvq7H.png", "fullname": "agents-last-exam", "name": "agents-last-exam", "type": "org", "isHf": false, "isHfAdmin": false, "isMod": false, "followerCount": 75, "isUserFollowing": false}}, "filename": "ale-openclaw-default.yaml", "verified": false, "pullRequest": 215, "rank": 4, "label": "Near Term Score", "notes": "harness=openclaw; harness_variant=not reported; attempted=67/67; official leaderboard snapshot=2026-10-01T20:39:17.826877+00:00; values are percentages (0-100); equal-task aggregation; unattempted tasks contribute zero; pass threshold score>=0.999. This is an official leaderboard export, not an HF-verified evaluation.", {"dataset": {"id": "agents-last-exam/agents-last-exam", "isBenchmark": true, "task_id": "full-spectrum-pass-rate"}, "value": 10.909090909090908, "source": {"url": "https://agents-last-exam.org/leaderboard", "name": "ALE official leaderboard", "isExternal": true, "author": {"_id": "69f848800d5d932e60c42eac", "avatarUrl": "https://cdn-avatars.huggingface.co/v1/production/uploads/6582486355a1e6cdb3b5e19d/eCSzSXXXKmj5srUhvq7H.png", "fullname": "agents-last-exam", "name": "agents-last-exam", "type": "org", "isHf": false, "isHfAdmin": false, "isMod": false, "followerCount": 75, "isUserFollowing": false}}, "filename": "ale-openclaw-default.yaml", "verified": false, "pullRequest": 215, "rank": 4, "label": "Full Spectrum Pass

Rate","notes":"harness=openclaw; harness_variant=not reported; attempted=54/55; official leaderboard snapshot=2026-10-01T20:39:17.826877+00:00; values are percentages (0-100); equal-task aggregation; unattempted tasks contribute zero; pass threshold score>=0.999. This is an official leaderboard export, not an HF-verified evaluation."},{ "dataset":{"id":"agents-last-exam/agents-last-exam","isBenchmark":true,"task_id":"full-spectrum-score"},"value":23.799048181818186,"source":{"url":"https://agents-last-exam.org/leaderboard"},"name":"ALE official leaderboard","isExternal":true,"author":{"_id":"69f848800d5d932e60c42eac","avatarUrl":"https://cdn-avatars.huggingface.co/v1/production/uploads/6582486355a1e6cdb3b5e19d/eCSzSXXKmJ5srUhvq7H.png","fullName":"agents-last-exam","name":"agents-last-exam","type":"org","isHf":false,"isHfAdmin":false,"isMod":false,"followerCount":75,"isUserFollowing":false}}, {"filename":"ale-openclaw-default.yaml","verified":false,"pullRequest":215,"rank":4,"label":"Full Spectrum Score","notes":"harness=openclaw; harness_variant=not reported; attempted=54/55; official leaderboard snapshot=2026-10-01T20:39:17.826877+00:00; values are percentages (0-100); equal-task aggregation; unattempted tasks contribute zero; pass threshold score>=0.999. This is an official leaderboard export, not an HF-verified evaluation."},{ "dataset":{"id":"agents-last-exam/agents-last-exam","isBenchmark":true,"task_id":"last-exam-pass-rate"},"value":0,"source":{"url":"https://agents-last-exam.org/leaderboard"},"name":"ALE official leaderboard","isExternal":true,"author":{"_id":"69f848800d5d932e60c42eac","avatarUrl":"https://cdn-avatars.huggingface.co/v1/production/uploads/6582486355a1e6cdb3b5e19d/eCSzSXXKmJ5srUhvq7H.png","fullName":"agents-last-exam","name":"agents-last-exam","type":"org","isHf":false,"isHfAdmin":false,"isMod":false,"followerCount":75,"isUserFollowing":false}}, {"filename":"ale-openclaw-default.yaml","verified":false,"pullRequest":215,"rank":4,"label":"Last Exam Pass Rate","notes":"harness=openclaw; harness_variant=not reported; attempted=33/38; official leaderboard snapshot=2026-10-01T20:39:17.826877+00:00; values are percentages (0-100); equal-task aggregation; unattempted tasks contribute zero; pass threshold score>=0.999. This is an official leaderboard export, not an HF-verified evaluation."},{ "dataset":{"id":"agents-last-exam/agents-last-exam","isBenchmark":true,"task_id":"last-exam-score"},"value":2.521315789473684,"source":{"url":"https://agents-last-exam.org/leaderboard"},"name":"ALE official leaderboard","isExternal":true,"author":{"_id":"69f848800d5d932e60c42eac","avatarUrl":"https://cdn-avatars.huggingface.co/v1/production/uploads/6582486355a1e6cdb3b5e19d/eCSzSXXKmJ5srUhvq7H.png","fullName":"agents-last-exam","name":"agents-last-exam","type":"org","isHf":false,"isHfAdmin":false,"isMod":false,"followerCount":75,"isUserFollowing":false}}, {"filename":"ale-openclaw-default.yaml","verified":false,"pullRequest":215,"rank":4,"label":"Last Exam Score","notes":"harness=openclaw; harness_variant=not reported; attempted=33/38; official leaderboard snapshot=2026-10-01T20:39:17.826877+00:00; values are percentages (0-100); equal-task aggregation; unattempted tasks contribute zero; pass threshold score>=0.999. This is an official leaderboard export, not an HF-verified evaluation."},{ "dataset":{"id":"ldavidrein/gpqa","isBenchmark":true,"task_id":"diamond"},"value":90.1,"source":{"url":"https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro"},"name":"Model Card","isExternal":false}, {"filename":"eval_results/gpqa.yaml","verified":false,"pullRequest":110,"rank":14,"label":"Diamond"}, {"dataset":{"id":"ScaleAI/SWE-bench_Pro","isBenchmark":true,"task_id":"SWE_Bench_Pro"},"value":55.4,"source":{"url":"https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro"},"name":"Model Card","isExternal":false}, {"filename":"eval_results/swe-bench_pro.yaml","verified":false,"pullRequest":110,"rank":27,"label":"SWE Bench Pro"}, {"dataset":{"id":"IntelligenceLab/Long-Horizon-Terminal-Bench","isBenchmark":true,"task_id":"lhtb_solved"},"value":3,"source":{"url":"https://zli12321.github.io/LHTB/leaderboard.html"},"name":"LHTB leaderboard","isExternal":true}, {"filename":"eval_results/lhtb.yaml","verified":false,"pullRequest":210,"rank":2,"label":"Lhtb Solved","notes":"3/46 tasks solved at reward >= 0.95 (pass@0.95); mean reward x100 = 30.7; official LHTB Harbor harness"}, {"dataset":{"id":"IntelligenceLab/Long-Horizon-Terminal-Bench","isBenchmark":true,"task_id":"lhtb2_solved"},"value":9,"source":{"url":"https://github.com/zli12321/LHTB/tree/lhtb-2.0#lhtb-20-results-october-2026"},"name":"LHTB 2.0 results","isExternal":true}, {"filename":"eval_results/lhtb2_solved.yaml","verified":false,"pullRequest":210,"rank":5,"label":"Lhtb2 Solved","notes":"9/77 tasks solved at reward >= 0.95 (pass@0.95); mean reward x100 = 52.9; LHTB 2.0 (77 tasks), terminus-2, 3-hour budget per task, official LHTB Harbor harness"}, {"dataset":{"id":"actava/chi-bench","isBenchmark":true,"task_id":"chi_bench"},"value":14.2,"source":{"url":"https://arxiv.org/abs/2605.16679"},"name":"CHI-Bench","isExternal":true}, {"filename":"eval_results/chi-bench.yaml","verified":false,"pullRequest":197,"rank":3,"label":"Chi Bench","notes":"Harness: OpenAI Agents SDK; Protocol: 75 tasks x 3 trials; Metric: pass@1 (%)"}, {"dataset":{"id":"actava/chi-

bench", "isBenchmark": true, "task_id": "prior_authorization", "value": 10.7, "source": {"url": "https://arxiv.org/abs/2605.16679", "name": "CHI-Bench", "isExternal": true}, "filename": ".eval_results/chi-bench.yaml", "verified": false, "pullRequest": 197, "rank": 3, "label": "Prior Authorization", "notes": "Harness: OpenAI Agents SDK; Protocol: 75 tasks x 3 trials; Metric: pass@1 (%)", {"dataset": {"id": "actava/chi-bench", "isBenchmark": true, "task_id": "utilization_management", "value": 28, "source": {"url": "https://arxiv.org/abs/2605.16679", "name": "CHI-Bench", "isExternal": true}, "filename": ".eval_results/chi-bench.yaml", "verified": false, "pullRequest": 197, "rank": 2, "label": "Utilization Management", "notes": "Harness: OpenAI Agents SDK; Protocol: 75 tasks x 3 trials; Metric: pass@1 (%)", {"dataset": {"id": "actava/chi-bench", "isBenchmark": true, "task_id": "care_management", "value": 4, "source": {"url": "https://arxiv.org/abs/2605.16679", "name": "CHI-Bench", "isExternal": true}, "filename": ".eval_results/chi-bench.yaml", "verified": false, "pullRequest": 197, "rank": 3, "label": "Care Management", "notes": "Harness: OpenAI Agents SDK; Protocol: 75 tasks x 3 trials; Metric: pass@1 (%)", {"dataset": {"id": "harborframework/terminal-bench-2.0", "isBenchmark": true, "task_id": "terminalbench_2", "value": 67.9, "source": {"url": "https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro", "name": "Model Card", "isExternal": false}, "filename": ".eval_results/terminal-bench-2.0.yaml", "verified": false, "pullRequest": 110, "rank": 6, "label": "Terminalbench 2", {"dataset": {"id": "joelniklaus/LEXam-hard", "isBenchmark": true, "task_id": "lexam_hard", "value": 38.93, "source": {"url": "https://huggingface.co/buckets/joelniklaus/SwissLegalEvals", "name": "SwissLegalEvals per-sample details (lighteval)", "isExternal": true, "author": {"_id": "5fae5f68b8423e1d80b8a988", "avatarUrl": "https://cdn-avatars.huggingface.co/v1/production/uploads/1656078368144-5fae5f68b8423e1d80b8a988.jpeg", "fullname": "Joel Niklaus", "name": "joelniklaus", "type": "user", "isPro": false, "isHf": true, "isHfAdmin": false, "isMod": false, "followerCount": 17: leexam-hard.yaml", "verified": false, "pullRequest": 214, "rank": 3, "label": "Lexam Hard", "notes": "lighteval, LEXam paper prompts, one response per question, no tools; DeepSeek-R1-0528 judge; mean of the German and English means over the 518 questions, 0-100", {"dataset": {"id": "SWE-bench/SWE-bench_Multilingual", "isBenchmark": true, "task_id": "swe_bench_multilingual_%_resolved", "value": 76.2, "source": {"url": "https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro", "name": "Model Card", "isExternal": false}, "filename": ".eval_results/swe-bench_multilingual.yaml", "verified": false, "pullRequest": 213, "rank": 9, "label": "Swe Bench Multilingual Resolved", {"dataset": {"id": "benchflow/skillsbench", "isBenchmark": true, "task_id": "skillsbench_v1_1", "value": 50.1, "source": {"url": "https://huggingface.co/datasets/benchflow/skillsbench-leaderboard/raw/main/leaderboard/skillsbench/v1.1/official.json", "name": "SkillsBench v1.1 official leaderboard", "isExternal": true, "author": {"_id": "69657f24ec1d157f1590a81d", "avatarUrl": "https://cdn-avatars.huggingface.co/v1/production/uploads/663fe2d26304d377fc253322/L6ik2VpL5iizADEXK4W-A.png", "fullname": "BenchFlow", "name": "benchflow", "type": "org", "isHf": false, "isHfAdmin": false, "isMod": false, "plan": "te: skillsbench.yaml", "verified": false, "pullRequest": 205, "rank": 8, "label": "Skillsbench V1 1", "notes": "with-skills; BenchFlow harness; OpenHands agent; 87 tasks x 3 trials; full 261/261 coverage", {"dataset": {"id": "internlm/WildClawBench", "isBenchmark": false, "task_id": "overall", "value": 43.7, "source": {"url": "https://internlm.github.io/WildClawBench", "name": "WildClawBench", "isExternal": true, "author": {"_id": "64a2d5fa81252883206f24c9", "avatarUrl": "https://cdn-avatars.huggingface.co/v1/production/uploads/6432683407bad11484a68457/Q3Y0dL79GcsnaBCGRMooZ.png", "fullname": "Intern Large Models", "name": "internlm", "type": "org", "isHf": false, "isHfAdmin": false, "isMod": false, "followerCount": 1347, "isUserFollow wildclawbench.yaml", "verified": false, "pullRequest": 187, "rank": 12, "label": "Overall", {"dataset": {"id": "internlm/WildClawBench", "isBenchmark": false, "task_id": "avg_time", "value": 605, "source": {"url": "https://internlm.github.io/WildClawBench", "name": "WildClawBench", "isExternal": true, "author": {"_id": "64a2d5fa81252883206f24c9", "avatarUrl": "https://cdn-avatars.huggingface.co/v1/production/uploads/6432683407bad11484a68457/Q3Y0dL79GcsnaBCGRMooZ.png", "fullname": "Intern Large Models", "name": "internlm", "type": "org", "isHf": false, "isHfAdmin": false, "isMod": false, "followerCount": 1347, "isUserFollow wildclawbench.yaml", "verified": false, "pullRequest": 187, "rank": 3, "label": "Avg Time", {"dataset": {"id": "internlm/WildClawBench", "isBenchmark": false, "task_id": "avg_cost", "value": 12, "source": {"url": "https://internlm.github.io/WildClawBench", "name": "WildClawBench", "isExternal": true, "author": {"_id": "64a2d5fa81252883206f24c9", "avatarUrl": "https://cdn-avatars.huggingface.co/v1/production/uploads/6432683407bad11484a68457/Q3Y0dL79GcsnaBCGRMooZ.png", "fullname": "Intern Large Models", "name": "internlm", "type": "org", "isHf": false, "isHfAdmin": false, "isMod": false, "followerCount": 1347, "isUserFollow wildclawbench.yaml", "verified": false, "pullRequest": 187, "rank": 7, "label": "Avg Cost", {"dataset": {"id": "InternScience/ResearchClawBench", "isBenchmark": false, "task_id": "overall", "value": 17.12, "source": {"url": "https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro", "name": "Model Card", "isExternal": false}, "filename": ".eval_results/

```

researchclawbench.yaml", "verified": false, "pullRequest": 186, "rank": 4, "label": "Overall", "notes": "ResearchHarness
evaluation with tools enabled, code execution, and a file-system workspace; completed 39/40
ResearchClawBench tasks.", {"dataset": {"id": "claw-eval/Claw-
Eval", "isBenchmark": true, "task_id": "general"}, "value": 58.4, "source": {"url": "https://claw-
eval.github.io", "name": "Claw-Eval Leaderboard", "isExternal": true, "author":
{"_id": "6038d6d0612f5eef3cc05ea9", "avatarUrl": "https://cdn-avatars.huggingface.co/v1/production/
uploads/6038d6d0612f5eef3cc05ea9/ryhvxAX5djQpD5OrlIZQ1f.jpeg", "fullname": "Lei
Li", "name": "tobiaslee", "type": "user", "isPro": false, "isHf": false, "isHfAdmin": false, "isMod": false, "followerCount": 30, "isUse
claw-eval.yaml", "verified": false, "pullRequest": 173, "rank": 16, "label": "General", "notes": "Pass3% | N=3 | 161 tasks"},
{"dataset": {"id": "claw-eval/Claw-Eval", "isBenchmark": true, "task_id": "multi_turn"}, "value": 65.8, "source":
{"url": "https://claw-eval.github.io", "name": "Claw-Eval Leaderboard", "isExternal": true, "author":
{"_id": "6038d6d0612f5eef3cc05ea9", "avatarUrl": "https://cdn-avatars.huggingface.co/v1/production/
uploads/6038d6d0612f5eef3cc05ea9/ryhvxAX5djQpD5OrlIZQ1f.jpeg", "fullname": "Lei
Li", "name": "tobiaslee", "type": "user", "isPro": false, "isHf": false, "isHfAdmin": false, "isMod": false, "followerCount": 30, "isUse
claw-eval.yaml", "verified": false, "pullRequest": 173, "rank": 2, "label": "Multi Turn", "notes": "Pass3% | N=3 | 38 tasks"},
{"dataset": {"id": "collinear-ai/yc-bench", "isBenchmark": true, "task_id": "medium"}, "value": 1066426, "source":
{"url": "https://github.com/collinear-ai/yc-bench", "name": "YC-Bench
eval", "isExternal": true}, {"filename": ".eval_results/yc-
bench.yaml", "verified": false, "pullRequest": 165, "rank": 4, "label": "Medium", "notes": "medium preset, avg final funds
(USD) across seeds 1,2,3"}, {"dataset": {"id": "cais/hle", "isBenchmark": false, "task_id": "hle"}, "value": 37.7, "source":
{"url": "https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro", "name": "Model
Card", "isExternal": false}, {"filename": ".eval_results/
hle.yaml", "verified": false, "pullRequest": 110, "rank": 28, "label": "Hle"}], "private": false, "repoType": "model", "gated": false, "t
["transformers", "safetensors", "deepseek_v4", "text-generation", "arxiv:2606.19348", "license:mit", "eval-
results", "endpoints_compatible", "8-bit", "fp8", "deploy:sagemaker", "region:us"], "tag_objs": [{"id": "text-
generation", "label": "Text Generation", "type": "pipeline_tag", "subType": "nlp"},
{"id": "transformers", "label": "Transformers", "type": "library"}, {"id": "safetensors", "label": "Safetensors", "type": "library"},
{"id": "deepseek_v4", "label": "deepseek_v4", "type": "other", "clickable": true}, {"id": "eval-results", "label": "Eval
Results", "type": "other", "clickable": true}, {"id": "endpoints_compatible", "label": "Inference
Endpoints", "type": "other", "clickable": true}, {"id": "8-bit", "label": "8-bit precision", "type": "other", "clickable": true},
{"id": "fp8", "label": "fp8", "type": "other", "clickable": true},
{"id": "arxiv:2606.19348", "label": "arxiv:2606.19348", "type": "arxiv", "extra": {"paperTitle": "DeepSeek-V4: Towards
Highly Efficient Million-Token Context Intelligence"}}, {"id": "license:mit", "label": "mit", "type": "license"},
{"id": "deploy:sagemaker", "label": "sagemaker", "type": "deploy"}, {"type": "region", "label": "🌐 Region:
US", "id": "region:us"}], "transformersInfo": {"auto_model": "AutoModelForCausalLM", "pipeline_tag": "text-
generation", "processor": "AutoTokenizer"}, "widgetData": [{"text": "Hi, what can you help me with?"}, {"text": "What is
84 * 3 / 2?"}, {"text": "Tell me an interesting fact about the universe!"}, {"text": "Explain quantum computing in simple
terms."}], "safetensors": {"parameters":
{"BF16": 2816899328, "I64": 2327040, "F32": 87776414, "F8_E4M3": 23169335296, "I8": 1572763336704, "total": 1598839
{"closed": 15, "open": 198, "total": 213, "query": {}, "inferenceContextData": {"billableEntities": [], "entityName2Providers":
{}}, "defaultProviders":
[{"isOriginalProvider": false, "name": "novita", "enabled": true, "position": 1, "isReleased": true, "accuratePricing": true},
{"isOriginalProvider": false, "name": "deepinfra", "enabled": true, "position": 16, "isReleased": true, "accuratePricing": true},
{"isOriginalProvider": false, "name": "featherless-
ai", "enabled": true, "position": 7, "isReleased": true, "accuratePricing": true},
{"isOriginalProvider": false, "name": "together", "enabled": true, "position": 5, "isReleased": true, "accuratePricing": false},
{"isOriginalProvider": false, "name": "baseten", "enabled": true, "position": 13, "isReleased": true, "accuratePricing": true}], "h
[]}>

```

DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence

Technical Report 

Introduction

We present a preview version of DeepSeek-V4 series, including two strong Mixture-of-Experts (MoE) language models — DeepSeek-V4-Pro with 1.6T parameters (49B activated) and DeepSeek-V4-Flash with 284B parameters (13B activated) — both supporting a context length of one million tokens .

DeepSeek-V4 series incorporate several key upgrades in architecture and optimization:

Hybrid Attention Architecture: We design a hybrid attention mechanism combining Compressed Sparse Attention (CSA) and Heavily Compressed Attention (HCA) to dramatically improve long-context efficiency. In the 1M-token context setting, DeepSeek-V4-Pro requires only 27% of single-token inference FLOPs and 10% of KV cache compared with DeepSeek-V3.2.

Manifold-Constrained Hyper-Connections (mHC): We incorporate mHC to strengthen conventional residual connections, enhancing stability of signal propagation across layers while preserving model expressivity.

Muon Optimizer: We employ the Muon optimizer for faster convergence and greater training stability.

We pre-train both models on more than 32T diverse and high-quality tokens, followed by a comprehensive post-training pipeline. The post-training features a two-stage paradigm: independent cultivation of domain-specific experts (through SFT and RL with GRPO), followed by unified model consolidation via on-policy distillation, integrating distinct proficiencies across diverse domains into a single model.

DeepSeek-V4-Pro-Max , the maximum reasoning effort mode of DeepSeek-V4-Pro, significantly advances the knowledge capabilities of open-source models, firmly establishing itself as the best open-source model available today. It achieves top-tier performance in coding benchmarks and significantly bridges the gap with leading closed-source models on reasoning and agentic tasks. Meanwhile, DeepSeek-V4-Flash-Max achieves comparable reasoning performance to the Pro version when given a larger thinking budget, though its smaller parameter scale naturally places it slightly behind on pure knowledge tasks and the most complex agentic workflows.

Model Downloads

Model #Total Params #Activated Params Context Length Precision Download

DeepSeek-V4-Flash-Base 284B 13B 1M FP8 Mixed [HuggingFace](#) | [ModelScope](#)

DeepSeek-V4-Flash 284B 13B 1M FP4 + FP8 Mixed* [HuggingFace](#) | [ModelScope](#)

DeepSeek-V4-Pro-Base 1.6T 49B 1M FP8 Mixed [HuggingFace](#) | [ModelScope](#)

DeepSeek-V4-Pro 1.6T 49B 1M FP4 + FP8 Mixed* [HuggingFace](#) | [ModelScope](#)

*FP4 + FP8 Mixed: MoE expert parameters use FP4 precision; most other parameters use FP8.

Evaluation Results

Base Model

Benchmark (Metric) # Shots DeepSeek-V3.2-Base DeepSeek-V4-Flash-Base DeepSeek-V4-Pro-Base

Architecture - MoE MoE MoE

Activated Params - 37B 13B 49B

Total Params - 671B 284B 1.6T

World Knowledge

AGIEval (EM) 0-shot 80.1 82.6 83.1

MMLU (EM) 5-shot 87.8 88.7 90.1

MMLU-Redux (EM) 5-shot 87.5 89.4 90.8

MMLU-Pro (EM) 5-shot 65.5 68.3 73.5

MMMLU (EM) 5-shot 87.9 88.8 90.3

C-Eval (EM) 5-shot 90.4 92.1 93.1

CMMLU (EM) 5-shot 88.9 90.4 90.8

MultiLoKo (EM) 5-shot 38.7 42.2 51.1

Simple-QA verified (EM) 25-shot 28.3 30.1 55.2

SuperGPQA (EM) 5-shot 45.0 46.5 53.9

FACTS Parametric (EM) 25-shot 27.1 33.9 62.6

TriviaQA (EM) 5-shot 83.3 82.8 85.6

Language & Reasoning

BBH (EM) 3-shot 87.6 86.9 87.5

DROP (F1) 1-shot 88.2 88.6 88.7

HellaSwag (EM) 0-shot 86.4 85.7 88.0

WinoGrande (EM) 0-shot 78.9 79.5 81.5

CLUEWSC (EM) 5-shot 83.5 82.2 85.2

Code & Math

BigCodeBench (Pass@1) 3-shot 63.9 56.8 59.2

HumanEval (Pass@1) 0-shot 62.8 69.5 76.8

GSM8K (EM) 8-shot 91.1 90.8 92.6

MATH (EM) 4-shot 60.5 57.4 64.5

MGSM (EM) 8-shot 81.3 85.7 84.4

CMath (EM) 3-shot 92.6 93.6 90.9

Long Context

LongBench-V2 (EM) 1-shot 40.2 44.7 51.5

Instruct Model

DeepSeek-V4-Pro and DeepSeek-V4-Flash both support three reasoning effort modes:

Reasoning Mode Characteristics Typical Use Cases Response Format

Non-think Fast, intuitive responses Routine daily tasks, low-risk decisions </think> summary

Think High Conscious logical analysis, slower but more accurate Complex problem-solving, planning <think> thinking </think> summary

Think Max Push reasoning to its fullest extent Exploring the boundary of model reasoning capability Special system prompt + <think> thinking </think> summary

DeepSeek-V4-Pro-Max vs Frontier Models

Benchmark (Metric) Opus-4.6 Max GPT-5.4 xHigh Gemini-3.1-Pro High K2.6 Thinking GLM-5.1 Thinking DS-V4-Pro Max

Knowledge & Reasoning

MMLU-Pro (EM) 89.1 87.5 91.0 87.1 86.0 87.5

SimpleQA-Verified (Pass@1) 46.2 45.3 75.6 36.9 38.1 57.9

Chinese-SimpleQA (Pass@1) 76.4 76.8 85.9 75.9 75.0 84.4

GPQA Diamond (Pass@1) 91.3 93.0 94.3 90.5 86.2 90.1

HLE (Pass@1) 40.0 39.8 44.4 36.4 34.7 37.7

LiveCodeBench (Pass@1) 88.8 - 91.7 89.6 - 93.5

Codeforces (Rating) - 3168 3052 - - 3206

HMMT 2026 Feb (Pass@1) 96.2 97.7 94.7 92.7 89.4 95.2
 IMOAnswerBench (Pass@1) 75.3 91.4 81.0 86.0 83.8 89.8
 Apex (Pass@1) 34.5 54.1 60.9 24.0 11.5 38.3
 Apex Shortlist (Pass@1) 85.9 78.1 89.1 75.5 72.4 90.2
 Long Context
 MRCR 1M (MMR) 92.9 - 76.3 - - 83.5
 CorpusQA 1M (ACC) 71.7 - 53.8 - - 62.0
 Agentic
 Terminal Bench 2.0 (Acc) 65.4 75.1 68.5 66.7 63.5 67.9
 SWE Verified (Resolved) 80.8 - 80.6 80.2 - 80.6
 SWE Pro (Resolved) 57.3 57.7 54.2 58.6 58.4 55.4
 SWE Multilingual (Resolved) 77.5 - - 76.7 73.3 76.2
 BrowseComp (Pass@1) 83.7 82.7 85.9 83.2 79.3 83.4
 HLE w/ tools (Pass@1) 53.1 52.0 51.6 54.0 50.4 48.2
 GDPval-AA (Elo) 1619 1674 1314 1482 1535 1554
 MCPAtlas Public (Pass@1) 73.8 67.2 69.2 66.6 71.8 73.6
 Toolathlon (Pass@1) 47.2 54.6 48.8 50.0 40.7 51.8
 Comparison across Modes
 Benchmark (Metric) V4-Flash Non-Think V4-Flash High V4-Flash Max V4-Pro Non-Think V4-Pro High V4-Pro Max
 Knowledge & Reasoning
 MMLU-Pro (EM) 83.0 86.4 86.2 82.9 87.1 87.5
 SimpleQA-Verified (Pass@1) 23.1 28.9 34.1 45.0 46.2 57.9
 Chinese-SimpleQA (Pass@1) 71.5 73.2 78.9 75.8 77.7 84.4
 GPQA Diamond (Pass@1) 71.2 87.4 88.1 72.9 89.1 90.1
 HLE (Pass@1) 8.1 29.4 34.8 7.7 34.5 37.7
 LiveCodeBench (Pass@1) 55.2 88.4 91.6 56.8 89.8 93.5
 Codeforces (Rating) - 2816 3052 - 2919 3206
 HMMT 2026 Feb (Pass@1) 40.8 91.9 94.8 31.7 94.0 95.2
 IMOAnswerBench (Pass@1) 41.9 85.1 88.4 35.3 88.0 89.8
 Apex (Pass@1) 1.0 19.1 33.0 0.4 27.4 38.3
 Apex Shortlist (Pass@1) 9.3 72.1 85.7 9.2 85.5 90.2
 Long Context
 MRCR 1M (MMR) 37.5 76.9 78.7 44.7 83.3 83.5
 CorpusQA 1M (ACC) 15.5 59.3 60.5 35.6 56.5 62.0
 Agentic
 Terminal Bench 2.0 (Acc) 49.1 56.6 56.9 59.1 63.3 67.9
 SWE Verified (Resolved) 73.7 78.6 79.0 73.6 79.4 80.6

SWE Pro (Resolved) 49.1 52.3 52.6 52.1 54.4 55.4

SWE Multilingual (Resolved) 69.7 70.2 73.3 69.8 74.1 76.2

BrowseComp (Pass@1) - 53.5 73.2 - 80.4 83.4

HLE w/ tools (Pass@1) - 40.3 45.1 - 44.7 48.2

MCPAtlas (Pass@1) 64.0 67.4 69.0 69.4 74.2 73.6

GDPval-AA (Elo) - - 1395 - - 1554

Toolathlon (Pass@1) 40.7 43.5 47.8 46.3 49.0 51.8

Chat Template

This release does not include a Jinja-format chat template. Instead, we provide a dedicated encoding folder with Python scripts and test cases demonstrating how to encode messages in OpenAI-compatible format into input strings for the model, and how to parse the model's text output. Please refer to the encoding folder for full documentation.

A brief example:

```
from encoding_dsv4 import encode_messages, parse_message_from_completion_text

messages = [ { "role" : "user" , "content" : "hello" }, { "role" : "assistant" , "content" : "Hello! I am DeepSeek." ,
"reasoning_content" : "thinking..." }, { "role" : "user" , "content" : "1+1=?" } ]

# messages -> string prompt = encode_messages(messages, thinking_mode= "thinking" )

# string -> tokens import transformers tokenizer = transformers.AutoTokenizer.from_pretrained( "deepseek-ai/
DeepSeek-V4-Pro" ) tokens = tokenizer.encode(prompt)
```

How to Run Locally

Please refer to the inference folder for detailed instructions on running DeepSeek-V4 locally, including model weight conversion and interactive chat demos.

For local deployment, we recommend setting the sampling parameters to temperature = 1.0, top_p = 1.0 . For the Think Max reasoning mode, we recommend setting the context window to at least 384K tokens.

License

This repository and the model weights are licensed under the MIT License .

Citation

@misc{deepseekai2026deepseekv4, title={DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence}, author={DeepSeek-AI}, year={2026}, }

Contact

If you have any questions, please raise an issue or contact us at service@deepseek.com .

Downloads last month 373,563

```
", "lstrip": false, "normalized": true, "rstrip": false, "single_word": false, "eos_token": { "__type": "AddedToken", "content": "<
| end of sentence | >", "lstrip": false, "normalized": true, "rstrip": false, "single_word": false, "pad_token":
{ "__type": "AddedToken", "content": "< | end of sentence |
>", "lstrip": false, "normalized": true, "rstrip": false, "single_word": false, "unk_token": null }, "createdAt": "2026-04-22T06:04:
created", "downloads": 373563, "downloadsAllTime": 10895485, "id": "deepseek-ai/DeepSeek-V4-
Pro", "isLikedByUser": false, "availableInferenceProviders":
[ { "provider": "together", "modelStatus": "error", "providerStatus": "live", "providerId": "deepseek-ai/DeepSeek-V4-
Pro", "task": "conversational", "adapterWeightsPath": "model-00001-of-00064.safetensors", "features":
{ "structuredOutput": true, "toolCalling": true, "isCheapestPricingOutput": false, "isFastestThroughput": false, "isModelAuth
{ "provider": "novita", "modelStatus": "live", "providerStatus": "live", "providerId": "deepseek/deepseek-v4-
pro", "task": "conversational", "adapterWeightsPath": "model-00001-of-00064.safetensors", "features":
```

```

{"structuredOutput":false,"toolCalling":true},"isCheapestPricingOutput":false,"isFastestThroughput":false,"isModelAutl
{"provider":"featherless-ai","modelStatus":"live","providerStatus":"live","providerId":"deepseek-ai/DeepSeek-V4-
Pro","task":"conversational","adapterWeightsPath":"model-00001-
of-00064.safetensors","isCheapestPricingOutput":false,"isFastestThroughput":false,"isModelAuthor":false},
{"provider":"deepinfra","modelStatus":"live","providerStatus":"live","providerId":"deepseek-ai/DeepSeek-V4-
Pro","task":"conversational","adapterWeightsPath":"model-00001-of-00064.safetensors","features":
{"structuredOutput":true,"toolCalling":true},"isCheapestPricingOutput":true,"isFastestThroughput":true,"isModelAuthor
{"provider":"baseten","modelStatus":"error","providerStatus":"live","providerId":"deepseek-ai/DeepSeek-V4-
Pro","task":"conversational","adapterWeightsPath":"model-00001-of-00064.safetensors","features":
{"structuredOutput":false,"toolCalling":true},"isCheapestPricingOutput":false,"isFastestThroughput":false,"isModelAutl
generation","library_name":"transformers","librariesOther":[],"trackDownloads":true,"model-
index":null,"evalResults":[{"dataset":{"id":"TIGER-Lab/MMLU-
Pro"},"isBenchmark":true,"task_id":"mmlu_pro"},"value":87.5,"source":{"url":"https://huggingface.co/deepseek-ai/
DeepSeek-V4-Pro"},"name":"Model Card","isExternal":false},"filename":"eval_results/mmlu-
pro.yaml","verified":false,"pullRequest":110,"rank":5,"label":"Mmlu Pro"},{"dataset":{"id":"openai/
gsm8k"},"isBenchmark":true,"task_id":"gsm8k"},"value":92.6,"source":{"url":"https://huggingface.co/deepseek-ai/
DeepSeek-V4-Pro"},"name":"Model Card","isExternal":false},"filename":"eval_results/
gsm8k.yaml","verified":false,"pullRequest":110,"rank":5,"label":"Gsm8k"},{"dataset":{"id":"SWE-bench/SWE-
bench_Verified"},"isBenchmark":true,"task_id":"swe_bench_%_resolved"},"value":80.6,"source":{"url":"https://
huggingface.co/deepseek-ai/DeepSeek-V4-Pro"},"name":"Model Card","isExternal":false},"filename":"eval_results/
swe-bench_verified.yaml","verified":false,"pullRequest":110,"rank":6,"label":"Swe Bench Resolved"},{"dataset":
{"id":"agents-last-exam/agents-last-exam"},"isBenchmark":true,"task_id":"overall-pass-
rate"},"value":12.390350877192983,"source":{"url":"https://agents-last-exam.org/leaderboard"},"name":"ALE official
leaderboard","isExternal":true,"author":{"_id":"69f84880d5d932e60c42eac"},"avatarUrl":"https://cdn-
avatars.huggingface.co/v1/production/uploads/6582486355a1e6cdb3b5e19d/
eCSzSXXXKmj5srUhvq7H.png"},"fullName":"agents-last-exam","name":"agents-last-
exam","type":"org","isHf":false,"isHfAdmin":false,"isMod":false,"followerCount":75,"isUserFollowing":false}},{"filename"
ale-openclaw-default.yaml","verified":false,"pullRequest":215,"rank":7,"label":"Overall Pass
Rate","notes":"harness=openclaw; harness_variant=not reported; attempted=146/152; official leaderboard
snapshot=2026-10-01T20:39:17.826877+00:00; values are percentages (0-100); equal-task aggregation;
unattempted tasks contribute zero; pass threshold score>=0.999. This is an official leaderboard export, not an HF-
verified evaluation."},"dataset":{"id":"agents-last-exam/agents-last-exam"},"isBenchmark":true,"task_id":"overall-
score"},"value":27.62014846817647,"source":{"url":"https://agents-last-exam.org/leaderboard"},"name":"ALE official
leaderboard","isExternal":true,"author":{"_id":"69f84880d5d932e60c42eac"},"avatarUrl":"https://cdn-
avatars.huggingface.co/v1/production/uploads/6582486355a1e6cdb3b5e19d/
eCSzSXXXKmj5srUhvq7H.png"},"fullName":"agents-last-exam","name":"agents-last-
exam","type":"org","isHf":false,"isHfAdmin":false,"isMod":false,"followerCount":75,"isUserFollowing":false}},{"filename"
ale-openclaw-default.yaml","verified":false,"pullRequest":215,"rank":5,"label":"Overall
Score","notes":"harness=openclaw; harness_variant=not reported; attempted=146/152; official leaderboard
snapshot=2026-10-01T20:39:17.826877+00:00; values are percentages (0-100); equal-task aggregation;
unattempted tasks contribute zero; pass threshold score>=0.999. This is an official leaderboard export, not an HF-
verified evaluation."},"dataset":{"id":"agents-last-exam/agents-last-exam"},"isBenchmark":true,"task_id":"ale-cli-
pass-rate"},"value":14.126984126984127,"source":{"url":"https://agents-last-exam.org/leaderboard"},"name":"ALE
official leaderboard","isExternal":true,"author":{"_id":"69f84880d5d932e60c42eac"},"avatarUrl":"https://cdn-
avatars.huggingface.co/v1/production/uploads/6582486355a1e6cdb3b5e19d/
eCSzSXXXKmj5srUhvq7H.png"},"fullName":"agents-last-exam","name":"agents-last-
exam","type":"org","isHf":false,"isHfAdmin":false,"isMod":false,"followerCount":75,"isUserFollowing":false}},{"filename"
ale-openclaw-default.yaml","verified":false,"pullRequest":215,"rank":8,"label":"Ale Cli Pass
Rate","notes":"harness=openclaw; harness_variant=not reported; attempted=104/105; official leaderboard
snapshot=2026-10-01T20:39:17.826877+00:00; values are percentages (0-100); equal-task aggregation;
unattempted tasks contribute zero; pass threshold score>=0.999. This is an official leaderboard export, not an HF-
verified evaluation."},"dataset":{"id":"agents-last-exam/agents-last-exam"},"isBenchmark":true,"task_id":"ale-cli-
score"},"value":30.84164065884846,"source":{"url":"https://agents-last-exam.org/leaderboard"},"name":"ALE official
leaderboard","isExternal":true,"author":{"_id":"69f84880d5d932e60c42eac"},"avatarUrl":"https://cdn-
avatars.huggingface.co/v1/production/uploads/6582486355a1e6cdb3b5e19d/
eCSzSXXXKmj5srUhvq7H.png"},"fullName":"agents-last-exam","name":"agents-last-

```

exam", "type": "org", "isHf": false, "isHfAdmin": false, "isMod": false, "followerCount": 75, "isUserFollowing": false}}, "filename": "ale-openclaw-default.yaml", "verified": false, "pullRequest": 215, "rank": 4, "label": "Ale Cli Score", "notes": "harness=openclaw; harness_variant=not reported; attempted=104/105; official leaderboard snapshot=2026-10-01T20:39:17.826877+00:00; values are percentages (0-100); equal-task aggregation; unattempted tasks contribute zero; pass threshold score>=0.999. This is an official leaderboard export, not an HF-verified evaluation.", {"dataset": {"id": "agents-last-exam/agents-last-exam", "isBenchmark": true, "task_id": "near-term-pass-rate"}, "value": 19.900497512437813, "source": {"url": "https://agents-last-exam.org/leaderboard", "name": "ALE official leaderboard", "isExternal": true, "author": {"_id": "69f848800d5d932e60c42eac", "avatarUrl": "https://cdn-avatars.huggingface.co/v1/production/uploads/6582486355a1e6cdb3b5e19d/eCSzorSXXKmJ5srUhvq7H.png", "fullName": "agents-last-exam", "name": "agents-last-exam", "type": "org", "isHf": false, "isHfAdmin": false, "isMod": false, "followerCount": 75, "isUserFollowing": false}}, "filename": "ale-openclaw-default.yaml", "verified": false, "pullRequest": 215, "rank": 4, "label": "Near Term Pass Rate", "notes": "harness=openclaw; harness_variant=not reported; attempted=67/67; official leaderboard snapshot=2026-10-01T20:39:17.826877+00:00; values are percentages (0-100); equal-task aggregation; unattempted tasks contribute zero; pass threshold score>=0.999. This is an official leaderboard export, not an HF-verified evaluation.", {"dataset": {"id": "agents-last-exam/agents-last-exam", "isBenchmark": true, "task_id": "near-term-score"}, "value": 43.81171518153468, "source": {"url": "https://agents-last-exam.org/leaderboard", "name": "ALE official leaderboard", "isExternal": true, "author": {"_id": "69f848800d5d932e60c42eac", "avatarUrl": "https://cdn-avatars.huggingface.co/v1/production/uploads/6582486355a1e6cdb3b5e19d/eCSzorSXXKmJ5srUhvq7H.png", "fullName": "agents-last-exam", "name": "agents-last-exam", "type": "org", "isHf": false, "isHfAdmin": false, "isMod": false, "followerCount": 75, "isUserFollowing": false}}, "filename": "ale-openclaw-default.yaml", "verified": false, "pullRequest": 215, "rank": 4, "label": "Near Term Score", "notes": "harness=openclaw; harness_variant=not reported; attempted=67/67; official leaderboard snapshot=2026-10-01T20:39:17.826877+00:00; values are percentages (0-100); equal-task aggregation; unattempted tasks contribute zero; pass threshold score>=0.999. This is an official leaderboard export, not an HF-verified evaluation.", {"dataset": {"id": "agents-last-exam/agents-last-exam", "isBenchmark": true, "task_id": "full-spectrum-pass-rate"}, "value": 10.909090909090908, "source": {"url": "https://agents-last-exam.org/leaderboard", "name": "ALE official leaderboard", "isExternal": true, "author": {"_id": "69f848800d5d932e60c42eac", "avatarUrl": "https://cdn-avatars.huggingface.co/v1/production/uploads/6582486355a1e6cdb3b5e19d/eCSzorSXXKmJ5srUhvq7H.png", "fullName": "agents-last-exam", "name": "agents-last-exam", "type": "org", "isHf": false, "isHfAdmin": false, "isMod": false, "followerCount": 75, "isUserFollowing": false}}, "filename": "ale-openclaw-default.yaml", "verified": false, "pullRequest": 215, "rank": 4, "label": "Full Spectrum Pass Rate", "notes": "harness=openclaw; harness_variant=not reported; attempted=54/55; official leaderboard snapshot=2026-10-01T20:39:17.826877+00:00; values are percentages (0-100); equal-task aggregation; unattempted tasks contribute zero; pass threshold score>=0.999. This is an official leaderboard export, not an HF-verified evaluation.", {"dataset": {"id": "agents-last-exam/agents-last-exam", "isBenchmark": true, "task_id": "full-spectrum-score"}, "value": 23.799048181818186, "source": {"url": "https://agents-last-exam.org/leaderboard", "name": "ALE official leaderboard", "isExternal": true, "author": {"_id": "69f848800d5d932e60c42eac", "avatarUrl": "https://cdn-avatars.huggingface.co/v1/production/uploads/6582486355a1e6cdb3b5e19d/eCSzorSXXKmJ5srUhvq7H.png", "fullName": "agents-last-exam", "name": "agents-last-exam", "type": "org", "isHf": false, "isHfAdmin": false, "isMod": false, "followerCount": 75, "isUserFollowing": false}}, "filename": "ale-openclaw-default.yaml", "verified": false, "pullRequest": 215, "rank": 4, "label": "Full Spectrum Score", "notes": "harness=openclaw; harness_variant=not reported; attempted=54/55; official leaderboard snapshot=2026-10-01T20:39:17.826877+00:00; values are percentages (0-100); equal-task aggregation; unattempted tasks contribute zero; pass threshold score>=0.999. This is an official leaderboard export, not an HF-verified evaluation.", {"dataset": {"id": "agents-last-exam/agents-last-exam", "isBenchmark": true, "task_id": "last-exam-pass-rate"}, "value": 0, "source": {"url": "https://agents-last-exam.org/leaderboard", "name": "ALE official leaderboard", "isExternal": true, "author": {"_id": "69f848800d5d932e60c42eac", "avatarUrl": "https://cdn-avatars.huggingface.co/v1/production/uploads/6582486355a1e6cdb3b5e19d/eCSzorSXXKmJ5srUhvq7H.png", "fullName": "agents-last-exam", "name": "agents-last-exam", "type": "org", "isHf": false, "isHfAdmin": false, "isMod": false, "followerCount": 75, "isUserFollowing": false}}, "filename": "ale-openclaw-default.yaml", "verified": false, "pullRequest": 215, "rank": 4, "label": "Last Exam Pass Rate", "notes": "harness=openclaw; harness_variant=not reported; attempted=33/38; official leaderboard

snapshot=2026-10-01T20:39:17.826877+00:00; values are percentages (0-100); equal-task aggregation; unattempted tasks contribute zero; pass threshold score>=0.999. This is an official leaderboard export, not an HF-verified evaluation.},{\"dataset\":{\"id\":\"agents-last-exam/agents-last-exam\",\"isBenchmark\":true,\"task_id\":\"last-exam-score\"},\"value\":2.521315789473684,\"source\":{\"url\":\"https://agents-last-exam.org/leaderboard\",\"name\":\"ALE official leaderboard\",\"isExternal\":true,\"author\":{\"_id\":\"69f848800d5d932e60c42eac\",\"avatarUrl\":\"https://cdn-avatars.huggingface.co/v1/production/uploads/6582486355a1e6cdb3b5e19d/eCSzorSXXKmJ5srUhvq7H.png\",\"fullName\":\"agents-last-exam\",\"name\":\"agents-last-exam\",\"type\":\"org\",\"isHf\":false,\"isHfAdmin\":false,\"isMod\":false,\"followerCount\":75,\"isUserFollowing\":false}},\"filename\":\"ale-openclaw-default.yaml\",\"verified\":false,\"pullRequest\":215,\"rank\":4,\"label\":\"Last Exam Score\",\"notes\":\"harness=openclaw; harness_variant=not reported; attempted=33/38; official leaderboard snapshot=2026-10-01T20:39:17.826877+00:00; values are percentages (0-100); equal-task aggregation; unattempted tasks contribute zero; pass threshold score>=0.999. This is an official leaderboard export, not an HF-verified evaluation.\"},{\"dataset\":{\"id\":\"ldavidrein/gpqa\",\"isBenchmark\":true,\"task_id\":\"diamond\"},\"value\":90.1,\"source\":{\"url\":\"https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro\",\"name\":\"Model Card\",\"isExternal\":false},\"filename\":\".eval_results/gpqa.yaml\",\"verified\":false,\"pullRequest\":110,\"rank\":14,\"label\":\"Diamond\"},{\"dataset\":{\"id\":\"ScaleAI/SWE-bench_Pro\",\"isBenchmark\":true,\"task_id\":\"SWE_Bench_Pro\"},\"value\":55.4,\"source\":{\"url\":\"https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro\",\"name\":\"Model Card\",\"isExternal\":false},\"filename\":\".eval_results/swe-bench_pro.yaml\",\"verified\":false,\"pullRequest\":110,\"rank\":27,\"label\":\"SWE Bench Pro\"},{\"dataset\":{\"id\":\"IntelligenceLab/Long-Horizon-Terminal-Bench\",\"isBenchmark\":true,\"task_id\":\"lhtb_solved\"},\"value\":3,\"source\":{\"url\":\"https://zli12321.github.io/LHTB/leaderboard.html\",\"name\":\"LHTB leaderboard\",\"isExternal\":true},\"filename\":\".eval_results/lhtb.yaml\",\"verified\":false,\"pullRequest\":210,\"rank\":2,\"label\":\"Lhtb Solved\",\"notes\":\"3/46 tasks solved at reward >= 0.95 (pass@0.95); mean reward x100 = 30.7; official LHTB Harbor harness\"},{\"dataset\":{\"id\":\"IntelligenceLab/Long-Horizon-Terminal-Bench\",\"isBenchmark\":true,\"task_id\":\"lhtb2_solved\"},\"value\":9,\"source\":{\"url\":\"https://github.com/zli12321/LHTB/tree/lhtb-2.0#lhtb-20-results-october-2026\",\"name\":\"LHTB 2.0 results\",\"isExternal\":true},\"filename\":\".eval_results/lhtb2_solved.yaml\",\"verified\":false,\"pullRequest\":210,\"rank\":5,\"label\":\"Lhtb2 Solved\",\"notes\":\"9/77 tasks solved at reward >= 0.95 (pass@0.95); mean reward x100 = 52.9; LHTB 2.0 (77 tasks), terminus-2, 3-hour budget per task, official LHTB Harbor harness\"},{\"dataset\":{\"id\":\"actava/chi-bench\",\"isBenchmark\":true,\"task_id\":\"chi_bench\"},\"value\":14.2,\"source\":{\"url\":\"https://arxiv.org/abs/2605.16679\",\"name\":\"CHI-Bench\",\"isExternal\":true},\"filename\":\".eval_results/chi-bench.yaml\",\"verified\":false,\"pullRequest\":197,\"rank\":3,\"label\":\"Chi Bench\",\"notes\":\"Harness: OpenAI Agents SDK; Protocol: 75 tasks x 3 trials; Metric: pass@1 (%)\"},{\"dataset\":{\"id\":\"actava/chi-bench\",\"isBenchmark\":true,\"task_id\":\"prior_authorization\"},\"value\":10.7,\"source\":{\"url\":\"https://arxiv.org/abs/2605.16679\",\"name\":\"CHI-Bench\",\"isExternal\":true},\"filename\":\".eval_results/chi-bench.yaml\",\"verified\":false,\"pullRequest\":197,\"rank\":3,\"label\":\"Prior Authorization\",\"notes\":\"Harness: OpenAI Agents SDK; Protocol: 75 tasks x 3 trials; Metric: pass@1 (%)\"},{\"dataset\":{\"id\":\"actava/chi-bench\",\"isBenchmark\":true,\"task_id\":\"utilization_management\"},\"value\":28,\"source\":{\"url\":\"https://arxiv.org/abs/2605.16679\",\"name\":\"CHI-Bench\",\"isExternal\":true},\"filename\":\".eval_results/chi-bench.yaml\",\"verified\":false,\"pullRequest\":197,\"rank\":2,\"label\":\"Utilization Management\",\"notes\":\"Harness: OpenAI Agents SDK; Protocol: 75 tasks x 3 trials; Metric: pass@1 (%)\"},{\"dataset\":{\"id\":\"actava/chi-bench\",\"isBenchmark\":true,\"task_id\":\"care_management\"},\"value\":4,\"source\":{\"url\":\"https://arxiv.org/abs/2605.16679\",\"name\":\"CHI-Bench\",\"isExternal\":true},\"filename\":\".eval_results/chi-bench.yaml\",\"verified\":false,\"pullRequest\":197,\"rank\":3,\"label\":\"Care Management\",\"notes\":\"Harness: OpenAI Agents SDK; Protocol: 75 tasks x 3 trials; Metric: pass@1 (%)\"},{\"dataset\":{\"id\":\"harborframework/terminal-bench-2.0\",\"isBenchmark\":true,\"task_id\":\"terminalbench_2\"},\"value\":67.9,\"source\":{\"url\":\"https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro\",\"name\":\"Model Card\",\"isExternal\":false},\"filename\":\".eval_results/terminal-bench-2.0.yaml\",\"verified\":false,\"pullRequest\":110,\"rank\":6,\"label\":\"Terminalbench 2\"},{\"dataset\":{\"id\":\"joelniklaus/LEXam-hard\",\"isBenchmark\":true,\"task_id\":\"lexam_hard\"},\"value\":38.93,\"source\":{\"url\":\"https://huggingface.co/buckets/joelniklaus/SwissLegalEvals\",\"name\":\"SwissLegalEvals per-sample details (lighteval)\",\"isExternal\":true,\"author\":{\"_id\":\"5fae5f68b8423e1d80b8a988\",\"avatarUrl\":\"https://cdn-avatars.huggingface.co/v1/production/uploads/1656078368144-5fae5f68b8423e1d80b8a988.jpeg\",\"fullName\":\"Joel Niklaus\",\"name\":\"joelniklaus\",\"type\":\"user\",\"isPro\":false,\"isHf\":true,\"isHfAdmin\":false,\"isMod\":false,\"followerCount\":17},\"filename\":\"lexam-hard.yaml\",\"verified\":false,\"pullRequest\":214,\"rank\":3,\"label\":\"Lexam Hard\",\"notes\":\"lighteval, LEXam paper

prompts, one response per question, no tools; DeepSeek-R1-0528 judge; mean of the German and English means over the 518 questions, 0-100"}, {"dataset": {"id": "SWE-bench/SWE-bench_Multilingual", "isBenchmark": true, "task_id": "swe_bench_multilingual_%_resolved"}, "value": 76.2, "source": {"url": "https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro", "name": "Model Card", "isExternal": false, "filename": ".eval_results/swe-bench_multilingual.yaml", "verified": false, "pullRequest": 213, "rank": 9, "label": "Swe Bench Multilingual Resolved"}, {"dataset": {"id": "benchflow/skillsbench", "isBenchmark": true, "task_id": "skillsbench_v1_1"}, "value": 50.1, "source": {"url": "https://huggingface.co/datasets/benchflow/skillsbench-leaderboard/raw/main/leaderboard/skillsbench/v1.1/official.json", "name": "SkillsBench v1.1 official leaderboard", "isExternal": true, "author": {"_id": "69657f24ec1d157f1590a81d", "avatarUrl": "https://cdn-avatars.huggingface.co/v1/production/uploads/663fe2d26304d377fc253322/L6ik2VpL5iizADEXK4W-A.png", "fullName": "BenchFlow", "name": "benchflow", "type": "org", "isHf": false, "isHfAdmin": false, "isMod": false, "plan": "te: skillsbench.yaml", "verified": false, "pullRequest": 205, "rank": 8, "label": "Skillsbench V1 1", "notes": "with-skills; BenchFlow harness; OpenHands agent; 87 tasks x 3 trials; full 261/261 coverage"}, {"dataset": {"id": "internlm/WildClawBench", "isBenchmark": false, "task_id": "overall"}, "value": 43.7, "source": {"url": "https://internlm.github.io/WildClawBench", "name": "WildClawBench", "isExternal": true, "author": {"_id": "64a2d5fa81252883206f24c9", "avatarUrl": "https://cdn-avatars.huggingface.co/v1/production/uploads/6432683407bad11484a68457/Q3Y0dL79GcsnaBCGRMooZ.png", "fullName": "Intern Large Models", "name": "internlm", "type": "org", "isHf": false, "isHfAdmin": false, "isMod": false, "followerCount": 1347, "isUserFollow wildclawbench.yaml", "verified": false, "pullRequest": 187, "rank": 12, "label": "Overall"}, {"dataset": {"id": "internlm/WildClawBench", "isBenchmark": false, "task_id": "avg_time"}, "value": 605, "source": {"url": "https://internlm.github.io/WildClawBench", "name": "WildClawBench", "isExternal": true, "author": {"_id": "64a2d5fa81252883206f24c9", "avatarUrl": "https://cdn-avatars.huggingface.co/v1/production/uploads/6432683407bad11484a68457/Q3Y0dL79GcsnaBCGRMooZ.png", "fullName": "Intern Large Models", "name": "internlm", "type": "org", "isHf": false, "isHfAdmin": false, "isMod": false, "followerCount": 1347, "isUserFollow wildclawbench.yaml", "verified": false, "pullRequest": 187, "rank": 3, "label": "Avg Time"}, {"dataset": {"id": "internlm/WildClawBench", "isBenchmark": false, "task_id": "avg_cost"}, "value": 12, "source": {"url": "https://internlm.github.io/WildClawBench", "name": "WildClawBench", "isExternal": true, "author": {"_id": "64a2d5fa81252883206f24c9", "avatarUrl": "https://cdn-avatars.huggingface.co/v1/production/uploads/6432683407bad11484a68457/Q3Y0dL79GcsnaBCGRMooZ.png", "fullName": "Intern Large Models", "name": "internlm", "type": "org", "isHf": false, "isHfAdmin": false, "isMod": false, "followerCount": 1347, "isUserFollow wildclawbench.yaml", "verified": false, "pullRequest": 187, "rank": 7, "label": "Avg Cost"}, {"dataset": {"id": "InternScience/ResearchClawBench", "isBenchmark": false, "task_id": "overall"}, "value": 17.12, "source": {"url": "https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro", "name": "Model Card", "isExternal": false, "filename": ".eval_results/researchclawbench.yaml", "verified": false, "pullRequest": 186, "rank": 4, "label": "Overall", "notes": "ResearchHarness evaluation with tools enabled, code execution, and a file-system workspace; completed 39/40 ResearchClawBench tasks."}, {"dataset": {"id": "claw-eval/Claw-Eval", "isBenchmark": true, "task_id": "general"}, "value": 58.4, "source": {"url": "https://claw-eval.github.io", "name": "Claw-Eval Leaderboard", "isExternal": true, "author": {"_id": "6038d6d0612f5eef3cc05ea9", "avatarUrl": "https://cdn-avatars.huggingface.co/v1/production/uploads/6038d6d0612f5eef3cc05ea9/ryhvAX5djQpD5OrlIZQ1f.jpeg", "fullName": "Lei Li", "name": "tobiaslee", "type": "user", "isPro": false, "isHf": false, "isHfAdmin": false, "isMod": false, "followerCount": 30, "isUse claw-eval.yaml", "verified": false, "pullRequest": 173, "rank": 16, "label": "General", "notes": "Pass3% | N=3 | 161 tasks"}, {"dataset": {"id": "claw-eval/Claw-Eval", "isBenchmark": true, "task_id": "multi_turn"}, "value": 65.8, "source": {"url": "https://claw-eval.github.io", "name": "Claw-Eval Leaderboard", "isExternal": true, "author": {"_id": "6038d6d0612f5eef3cc05ea9", "avatarUrl": "https://cdn-avatars.huggingface.co/v1/production/uploads/6038d6d0612f5eef3cc05ea9/ryhvAX5djQpD5OrlIZQ1f.jpeg", "fullName": "Lei Li", "name": "tobiaslee", "type": "user", "isPro": false, "isHf": false, "isHfAdmin": false, "isMod": false, "followerCount": 30, "isUse claw-eval.yaml", "verified": false, "pullRequest": 173, "rank": 2, "label": "Multi Turn", "notes": "Pass3% | N=3 | 38 tasks"}, {"dataset": {"id": "collinear-ai/yc-bench", "isBenchmark": true, "task_id": "medium"}, "value": 1066426, "source": {"url": "https://github.com/collinear-ai/yc-bench", "name": "YC-Bench eval", "isExternal": true, "filename": ".eval_results/yc-bench.yaml", "verified": false, "pullRequest": 165, "rank": 4, "label": "Medium", "notes": "medium preset, avg final funds (USD) across seeds 1,2,3"}, {"dataset": {"id": "cais/hle", "isBenchmark": false, "task_id": "hle"}, "value": 37.7, "source": {"url": "https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro", "name": "Model Card", "isExternal": false, "filename": ".eval_results/

```
hle.yaml", "verified": false, "pullRequest": 110, "rank": 28, "label": "Hle"}, {"private": false, "repoType": "model", "gated": false, "t
["transformers", "safetensors", "deepseek_v4", "text-generation", "arxiv:2606.19348", "license:mit", "eval-
results", "endpoints_compatible", "8-bit", "fp8", "deploy:sagemaker", "region:us"], "tag_objs": [{"id": "text-
generation", "label": "Text Generation", "type": "pipeline_tag", "subType": "nlp"},
{"id": "transformers", "label": "Transformers", "type": "library"}, {"id": "safetensors", "label": "Safetensors", "type": "library"},
{"id": "deepseek_v4", "label": "deepseek_v4", "type": "other", "clickable": true}, {"id": "eval-results", "label": "Eval
Results", "type": "other", "clickable": true}, {"id": "endpoints_compatible", "label": "Inference
Endpoints", "type": "other", "clickable": true}, {"id": "8-bit", "label": "8-bit precision", "type": "other", "clickable": true},
{"id": "fp8", "label": "fp8", "type": "other", "clickable": true},
{"id": "arxiv:2606.19348", "label": "arxiv:2606.19348", "type": "arxiv", "extra": {"paperTitle": "DeepSeek-V4: Towards
Highly Efficient Million-Token Context Intelligence"}}, {"id": "license:mit", "label": "mit", "type": "license"},
{"id": "deploy:sagemaker", "label": "sagemaker", "type": "deploy"}, {"type": "region", "label": "🌐 Region:
US", "id": "region:us"}], "transformersInfo": {"auto_model": "AutoModelForCausalLM", "pipeline_tag": "text-
generation", "processor": "AutoTokenizer", "widgetData": [{"text": "Hi, what can you help me with?"}, {"text": "What is
84 * 3 / 2?"}, {"text": "Tell me an interesting fact about the universe!"}, {"text": "Explain quantum computing in simple
terms."}], "safetensors": {"parameters":
{"BF16": 2816899328, "I64": 2327040, "F32": 87776414, "F8_E4M3": 23169335296, "I8": 1572763336704}, "total": 1598839
{"billableEntities": [], "entityName2Providers": {}, "defaultProviders":
[{"isOriginalProvider": false, "name": "novita", "enabled": true, "position": 1, "isReleased": true, "accuratePricing": true},
{"isOriginalProvider": false, "name": "deepinfra", "enabled": true, "position": 16, "isReleased": true, "accuratePricing": true},
{"isOriginalProvider": false, "name": "featherless-
ai", "enabled": true, "position": 7, "isReleased": true, "accuratePricing": true},
{"isOriginalProvider": false, "name": "together", "enabled": true, "position": 5, "isReleased": true, "accuratePricing": false},
{"isOriginalProvider": false, "name": "baseten", "enabled": true, "position": 13, "isReleased": true, "accuratePricing": true}]]}, "c
Safetensors Model size 1.6T params
```

Tensor type BF16 · I64 · F32 · F8_E4M3 · I8 ·

Files info

```
", "lstrip": false, "normalized": true, "rstrip": false, "single_word": false}, {"eos_token": {"__type": "AddedToken", "content": "<
| end of sentence | >", "lstrip": false, "normalized": true, "rstrip": false, "single_word": false}, {"pad_token":
{"__type": "AddedToken", "content": "< | end of sentence |
>", "lstrip": false, "normalized": true, "rstrip": false, "single_word": false}, {"unk_token": null}], "createdAt": "2026-04-22T06:04:
created", "downloads": 373563, "downloadsAllTime": 10895485, "id": "deepseek-ai/DeepSeek-V4-
Pro", "isLikedByUser": false, "availableInferenceProviders":
[{"provider": "together", "modelStatus": "error", "providerStatus": "live", "providerId": "deepseek-ai/DeepSeek-V4-
Pro", "task": "conversational", "adapterWeightsPath": "model-00001-of-00064.safetensors", "features":
{"structuredOutput": true, "toolCalling": true}, {"isCheapestPricingOutput": false, "isFastestThroughput": false, "isModelAuth
{"provider": "novita", "modelStatus": "live", "providerStatus": "live", "providerId": "deepseek/deepseek-v4-
pro", "task": "conversational", "adapterWeightsPath": "model-00001-of-00064.safetensors", "features":
{"structuredOutput": false, "toolCalling": true}, {"isCheapestPricingOutput": false, "isFastestThroughput": false, "isModelAutl
{"provider": "featherless-ai", "modelStatus": "live", "providerStatus": "live", "providerId": "deepseek-ai/DeepSeek-V4-
Pro", "task": "conversational", "adapterWeightsPath": "model-00001-
of-00064.safetensors", "isCheapestPricingOutput": false, "isFastestThroughput": false, "isModelAuthor": false},
{"provider": "deepinfra", "modelStatus": "live", "providerStatus": "live", "providerId": "deepseek-ai/DeepSeek-V4-
Pro", "task": "conversational", "adapterWeightsPath": "model-00001-of-00064.safetensors", "features":
{"structuredOutput": true, "toolCalling": true}, {"isCheapestPricingOutput": true, "isFastestThroughput": true, "isModelAuthor
{"provider": "baseten", "modelStatus": "error", "providerStatus": "live", "providerId": "deepseek-ai/DeepSeek-V4-
Pro", "task": "conversational", "adapterWeightsPath": "model-00001-of-00064.safetensors", "features":
{"structuredOutput": false, "toolCalling": true}, {"isCheapestPricingOutput": false, "isFastestThroughput": false, "isModelAutl
generation", "library_name": "transformers", "librariesOther": [], "trackDownloads": true, "model-
index": null, "evalResults": [{"dataset": {"id": "TIGER-Lab/MMLU-
Pro", "isBenchmark": true, "task_id": "mmlu_pro"}, {"value": 87.5, "source": {"url": "https://huggingface.co/deepseek-ai/
DeepSeek-V4-Pro", "name": "Model Card", "isExternal": false}, {"filename": ".eval_results/mmlu-
pro.yaml", "verified": false, "pullRequest": 110, "rank": 5, "label": "Mmlu Pro"}, {"dataset": {"id": "openai/
gsm8k", "isBenchmark": true, "task_id": "gsm8k"}, {"value": 92.6, "source": {"url": "https://huggingface.co/deepseek-ai/
DeepSeek-V4-Pro", "name": "Model Card", "isExternal": false}, {"filename": ".eval_results/
```


term-score"},"value":43.81171518153468,"source":{"url":"https://agents-last-exam.org/leaderboard","name":"ALE official leaderboard","isExternal":true,"author":{"_id":"69f848800d5d932e60c42eac","avatarUrl":"https://cdn-avatars.huggingface.co/v1/production/uploads/6582486355a1e6cdb3b5e19d/eCSzorSXXKmJ5srUHVq7H.png","fullName":"agents-last-exam","name":"agents-last-exam","type":"org","isHf":false,"isHfAdmin":false,"isMod":false,"followerCount":75,"isUserFollowing":false},"filename":"ale-openclaw-default.yaml","verified":false,"pullRequest":215,"rank":4,"label":"Near Term Score","notes":"harness=openclaw; harness_variant=not reported; attempted=67/67; official leaderboard snapshot=2026-10-01T20:39:17.826877+00:00; values are percentages (0-100); equal-task aggregation; unattempted tasks contribute zero; pass threshold score>=0.999. This is an official leaderboard export, not an HF-verified evaluation."},"dataset":{"id":"agents-last-exam/agents-last-exam","isBenchmark":true,"task_id":"full-spectrum-pass-rate"},"value":10.909090909090908,"source":{"url":"https://agents-last-exam.org/leaderboard","name":"ALE official leaderboard","isExternal":true,"author":{"_id":"69f848800d5d932e60c42eac","avatarUrl":"https://cdn-avatars.huggingface.co/v1/production/uploads/6582486355a1e6cdb3b5e19d/eCSzorSXXKmJ5srUHVq7H.png","fullName":"agents-last-exam","name":"agents-last-exam","type":"org","isHf":false,"isHfAdmin":false,"isMod":false,"followerCount":75,"isUserFollowing":false},"filename":"ale-openclaw-default.yaml","verified":false,"pullRequest":215,"rank":4,"label":"Full Spectrum Pass Rate","notes":"harness=openclaw; harness_variant=not reported; attempted=54/55; official leaderboard snapshot=2026-10-01T20:39:17.826877+00:00; values are percentages (0-100); equal-task aggregation; unattempted tasks contribute zero; pass threshold score>=0.999. This is an official leaderboard export, not an HF-verified evaluation."},"dataset":{"id":"agents-last-exam/agents-last-exam","isBenchmark":true,"task_id":"full-spectrum-score"},"value":23.799048181818186,"source":{"url":"https://agents-last-exam.org/leaderboard","name":"ALE official leaderboard","isExternal":true,"author":{"_id":"69f848800d5d932e60c42eac","avatarUrl":"https://cdn-avatars.huggingface.co/v1/production/uploads/6582486355a1e6cdb3b5e19d/eCSzorSXXKmJ5srUHVq7H.png","fullName":"agents-last-exam","name":"agents-last-exam","type":"org","isHf":false,"isHfAdmin":false,"isMod":false,"followerCount":75,"isUserFollowing":false},"filename":"ale-openclaw-default.yaml","verified":false,"pullRequest":215,"rank":4,"label":"Full Spectrum Score","notes":"harness=openclaw; harness_variant=not reported; attempted=54/55; official leaderboard snapshot=2026-10-01T20:39:17.826877+00:00; values are percentages (0-100); equal-task aggregation; unattempted tasks contribute zero; pass threshold score>=0.999. This is an official leaderboard export, not an HF-verified evaluation."},"dataset":{"id":"agents-last-exam/agents-last-exam","isBenchmark":true,"task_id":"last-exam-pass-rate"},"value":0,"source":{"url":"https://agents-last-exam.org/leaderboard","name":"ALE official leaderboard","isExternal":true,"author":{"_id":"69f848800d5d932e60c42eac","avatarUrl":"https://cdn-avatars.huggingface.co/v1/production/uploads/6582486355a1e6cdb3b5e19d/eCSzorSXXKmJ5srUHVq7H.png","fullName":"agents-last-exam","name":"agents-last-exam","type":"org","isHf":false,"isHfAdmin":false,"isMod":false,"followerCount":75,"isUserFollowing":false},"filename":"ale-openclaw-default.yaml","verified":false,"pullRequest":215,"rank":4,"label":"Last Exam Pass Rate","notes":"harness=openclaw; harness_variant=not reported; attempted=33/38; official leaderboard snapshot=2026-10-01T20:39:17.826877+00:00; values are percentages (0-100); equal-task aggregation; unattempted tasks contribute zero; pass threshold score>=0.999. This is an official leaderboard export, not an HF-verified evaluation."},"dataset":{"id":"agents-last-exam/agents-last-exam","isBenchmark":true,"task_id":"last-exam-score"},"value":2.521315789473684,"source":{"url":"https://agents-last-exam.org/leaderboard","name":"ALE official leaderboard","isExternal":true,"author":{"_id":"69f848800d5d932e60c42eac","avatarUrl":"https://cdn-avatars.huggingface.co/v1/production/uploads/6582486355a1e6cdb3b5e19d/eCSzorSXXKmJ5srUHVq7H.png","fullName":"agents-last-exam","name":"agents-last-exam","type":"org","isHf":false,"isHfAdmin":false,"isMod":false,"followerCount":75,"isUserFollowing":false},"filename":"ale-openclaw-default.yaml","verified":false,"pullRequest":215,"rank":4,"label":"Last Exam Score","notes":"harness=openclaw; harness_variant=not reported; attempted=33/38; official leaderboard snapshot=2026-10-01T20:39:17.826877+00:00; values are percentages (0-100); equal-task aggregation; unattempted tasks contribute zero; pass threshold score>=0.999. This is an official leaderboard export, not an HF-verified evaluation."},"dataset":{"id":"ldavidrein/gpqa","isBenchmark":true,"task_id":"diamond"},"value":90.1,"source":{"url":"https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro","name":"Model Card","isExternal":false},"filename":"eval_results/gpqa.yaml","verified":false,"pullRequest":110,"rank":14,"label":"Diamond"},"dataset":{"id":"ScaleAI/SWE-bench_Pro","isBenchmark":true,"task_id":"SWE_Bench_Pro"},"value":55.4,"source":{"url":"https://huggingface.co/

deepseek-ai/DeepSeek-V4-Pro", "name": "Model Card", "isExternal": false, "filename": ".eval_results/swe-bench_pro.yaml", "verified": false, "pullRequest": 110, "rank": 27, "label": "SWE Bench Pro", {"dataset": {"id": "IntelligenceLab/Long-Horizon-Terminal-Bench", "isBenchmark": true, "task_id": "lhtb_solved", "value": 3, "source": {"url": "https://zli12321.github.io/LHTB/leaderboard.html", "name": "LHTB leaderboard", "isExternal": true, "filename": ".eval_results/lhtb.yaml", "verified": false, "pullRequest": 210, "rank": 2, "label": "Lhtb Solved", "notes": "3/46 tasks solved at reward >= 0.95 (pass@0.95); mean reward x100 = 30.7; official LHTB Harbor harness"}, {"dataset": {"id": "IntelligenceLab/Long-Horizon-Terminal-Bench", "isBenchmark": true, "task_id": "lhtb2_solved", "value": 9, "source": {"url": "https://github.com/zli12321/LHTB/tree/lhtb-2.0#lhtb-20-results-october-2026", "name": "LHTB 2.0 results", "isExternal": true, "filename": ".eval_results/lhtb2_solved.yaml", "verified": false, "pullRequest": 210, "rank": 5, "label": "Lhtb2 Solved", "notes": "9/77 tasks solved at reward >= 0.95 (pass@0.95); mean reward x100 = 52.9; LHTB 2.0 (77 tasks), terminus-2, 3-hour budget per task, official LHTB Harbor harness"}, {"dataset": {"id": "actava/chi-bench", "isBenchmark": true, "task_id": "chi_bench", "value": 14.2, "source": {"url": "https://arxiv.org/abs/2605.16679", "name": "CHI-Bench", "isExternal": true, "filename": ".eval_results/chi-bench.yaml", "verified": false, "pullRequest": 197, "rank": 3, "label": "Chi Bench", "notes": "Harness: OpenAI Agents SDK; Protocol: 75 tasks x 3 trials; Metric: pass@1 (%)"}, {"dataset": {"id": "actava/chi-bench", "isBenchmark": true, "task_id": "prior_authorization", "value": 10.7, "source": {"url": "https://arxiv.org/abs/2605.16679", "name": "CHI-Bench", "isExternal": true, "filename": ".eval_results/chi-bench.yaml", "verified": false, "pullRequest": 197, "rank": 3, "label": "Prior Authorization", "notes": "Harness: OpenAI Agents SDK; Protocol: 75 tasks x 3 trials; Metric: pass@1 (%)"}, {"dataset": {"id": "actava/chi-bench", "isBenchmark": true, "task_id": "utilization_management", "value": 28, "source": {"url": "https://arxiv.org/abs/2605.16679", "name": "CHI-Bench", "isExternal": true, "filename": ".eval_results/chi-bench.yaml", "verified": false, "pullRequest": 197, "rank": 2, "label": "Utilization Management", "notes": "Harness: OpenAI Agents SDK; Protocol: 75 tasks x 3 trials; Metric: pass@1 (%)"}, {"dataset": {"id": "actava/chi-bench", "isBenchmark": true, "task_id": "care_management", "value": 4, "source": {"url": "https://arxiv.org/abs/2605.16679", "name": "CHI-Bench", "isExternal": true, "filename": ".eval_results/chi-bench.yaml", "verified": false, "pullRequest": 197, "rank": 3, "label": "Care Management", "notes": "Harness: OpenAI Agents SDK; Protocol: 75 tasks x 3 trials; Metric: pass@1 (%)"}, {"dataset": {"id": "harborframework/terminal-bench-2.0", "isBenchmark": true, "task_id": "terminalbench_2", "value": 67.9, "source": {"url": "https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro", "name": "Model Card", "isExternal": false, "filename": ".eval_results/terminal-bench-2.0.yaml", "verified": false, "pullRequest": 110, "rank": 6, "label": "Terminalbench 2"}, {"dataset": {"id": "joelniklaus/LEXam-hard", "isBenchmark": true, "task_id": "lexam_hard", "value": 38.93, "source": {"url": "https://huggingface.co/buckets/joelniklaus/SwissLegalEvals", "name": "SwissLegalEvals per-sample details (lighteval)", "isExternal": true, "author": {"id": "5fae5f68b8423e1d80b8a988", "avatarUrl": "https://cdn-avatars.huggingface.co/v1/production/uploads/1656078368144-5fae5f68b8423e1d80b8a988.jpeg", "fullName": "Joel Niklaus", "name": "joelniklaus", "type": "user", "isPro": false, "isHf": true, "isHfAdmin": false, "isMod": false, "followerCount": 17}, "lexam-hard.yaml", "verified": false, "pullRequest": 214, "rank": 3, "label": "Lexam Hard", "notes": "lighteval, LEXam paper prompts, one response per question, no tools; DeepSeek-R1-0528 judge; mean of the German and English means over the 518 questions, 0-100"}, {"dataset": {"id": "SWE-bench/SWE-bench_Multilingual", "isBenchmark": true, "task_id": "swe_bench_multilingual_%_resolved", "value": 76.2, "source": {"url": "https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro", "name": "Model Card", "isExternal": false, "filename": ".eval_results/swe-bench_multilingual.yaml", "verified": false, "pullRequest": 213, "rank": 9, "label": "Swe Bench Multilingual Resolved"}, {"dataset": {"id": "benchflow/skillsbench", "isBenchmark": true, "task_id": "skillsbench_v1_1", "value": 50.1, "source": {"url": "https://huggingface.co/datasets/benchflow/skillsbench-leaderboard/raw/main/leaderboard/skillsbench/v1.1/official.json", "name": "SkillsBench v1.1 official leaderboard", "isExternal": true, "author": {"id": "69657f24ec1d157f1590a81d", "avatarUrl": "https://cdn-avatars.huggingface.co/v1/production/uploads/663fe2d26304d377fc253322/L6ik2VpL5iizADEXK4W-A.png", "fullName": "BenchFlow", "name": "benchflow", "type": "org", "isHf": false, "isHfAdmin": false, "isMod": false, "plan": "te"}, "skillsbench.yaml", "verified": false, "pullRequest": 205, "rank": 8, "label": "Skillsbench V1 1", "notes": "with-skills; BenchFlow harness; OpenHands agent; 87 tasks x 3 trials; full 261/261 coverage"}, {"dataset": {"id": "internlm/WildClawBench", "isBenchmark": false, "task_id": "overall", "value": 43.7, "source": {"url": "https://internlm.github.io/WildClawBench", "name": "WildClawBench", "isExternal": true, "author": {"id": "64a2d5fa81252883206f24c9", "avatarUrl": "https://cdn-avatars.huggingface.co/v1/production/

uploads/6432683407bad11484a68457/Q3Y0dL79GcsnaBCGRMooZ.png", "fullname": "Intern Large Models", "name": "internlm", "type": "org", "isHf": false, "isHfAdmin": false, "isMod": false, "followerCount": 1347, "isUserFollow", "wildclawbench.yaml", "verified": false, "pullRequest": 187, "rank": 12, "label": "Overall"}, {"dataset": {"id": "internlm/WildClawBench", "isBenchmark": false, "task_id": "avg_time", "value": 605, "source": {"url": "https://internlm.github.io/WildClawBench", "name": "WildClawBench", "isExternal": true, "author": {"_id": "64a2d5fa81252883206f24c9", "avatarUrl": "https://cdn-avatars.huggingface.co/v1/production/uploads/6432683407bad11484a68457/Q3Y0dL79GcsnaBCGRMooZ.png", "fullname": "Intern Large Models", "name": "internlm", "type": "org", "isHf": false, "isHfAdmin": false, "isMod": false, "followerCount": 1347, "isUserFollow", "wildclawbench.yaml", "verified": false, "pullRequest": 187, "rank": 3, "label": "Avg Time"}, {"dataset": {"id": "internlm/WildClawBench", "isBenchmark": false, "task_id": "avg_cost", "value": 12, "source": {"url": "https://internlm.github.io/WildClawBench", "name": "WildClawBench", "isExternal": true, "author": {"_id": "64a2d5fa81252883206f24c9", "avatarUrl": "https://cdn-avatars.huggingface.co/v1/production/uploads/6432683407bad11484a68457/Q3Y0dL79GcsnaBCGRMooZ.png", "fullname": "Intern Large Models", "name": "internlm", "type": "org", "isHf": false, "isHfAdmin": false, "isMod": false, "followerCount": 1347, "isUserFollow", "wildclawbench.yaml", "verified": false, "pullRequest": 187, "rank": 7, "label": "Avg Cost"}, {"dataset": {"id": "InternScience/ResearchClawBench", "isBenchmark": false, "task_id": "overall", "value": 17.12, "source": {"url": "https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro", "name": "Model Card", "isExternal": false, "filename": ".eval_results/researchclawbench.yaml", "verified": false, "pullRequest": 186, "rank": 4, "label": "Overall", "notes": "ResearchHarness evaluation with tools enabled, code execution, and a file-system workspace; completed 39/40 ResearchClawBench tasks."}, {"dataset": {"id": "claw-eval/Claw-Eval", "isBenchmark": true, "task_id": "general", "value": 58.4, "source": {"url": "https://claw-eval.github.io", "name": "Claw-Eval Leaderboard", "isExternal": true, "author": {"_id": "6038d6d0612f5eef3cc05ea9", "avatarUrl": "https://cdn-avatars.huggingface.co/v1/production/uploads/6038d6d0612f5eef3cc05ea9/ryhvAX5djQpD5OrIIQZ1f.jpeg", "fullname": "Lei Li", "name": "tobiaslee", "type": "user", "isPro": false, "isHf": false, "isHfAdmin": false, "isMod": false, "followerCount": 30, "isUserFollow", "claw-eval.yaml", "verified": false, "pullRequest": 173, "rank": 16, "label": "General", "notes": "Pass³% | N=3 | 161 tasks"}, {"dataset": {"id": "claw-eval/Claw-Eval", "isBenchmark": true, "task_id": "multi_turn", "value": 65.8, "source": {"url": "https://claw-eval.github.io", "name": "Claw-Eval Leaderboard", "isExternal": true, "author": {"_id": "6038d6d0612f5eef3cc05ea9", "avatarUrl": "https://cdn-avatars.huggingface.co/v1/production/uploads/6038d6d0612f5eef3cc05ea9/ryhvAX5djQpD5OrIIQZ1f.jpeg", "fullname": "Lei Li", "name": "tobiaslee", "type": "user", "isPro": false, "isHf": false, "isHfAdmin": false, "isMod": false, "followerCount": 30, "isUserFollow", "claw-eval.yaml", "verified": false, "pullRequest": 173, "rank": 2, "label": "Multi Turn", "notes": "Pass³% | N=3 | 38 tasks"}, {"dataset": {"id": "collinear-ai/yc-bench", "isBenchmark": true, "task_id": "medium", "value": 1066426, "source": {"url": "https://github.com/collinear-ai/yc-bench", "name": "YC-Bench eval", "isExternal": true, "filename": ".eval_results/yc-bench.yaml", "verified": false, "pullRequest": 165, "rank": 4, "label": "Medium", "notes": "medium preset, avg final funds (USD) across seeds 1,2,3"}, {"dataset": {"id": "cais/hle", "isBenchmark": false, "task_id": "hle", "value": 37.7, "source": {"url": "https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro", "name": "Model Card", "isExternal": false, "filename": ".eval_results/hle.yaml", "verified": false, "pullRequest": 110, "rank": 28, "label": "Hle"}, {"private": false, "repoType": "model", "gated": false, "tags": ["transformers", "safetensors", "deepseek_v4", "text-generation", "arxiv:2606.19348", "license:mit", "eval-results", "endpoints_compatible", "8-bit", "fp8", "deploy:sagemaker", "region:us"], "tag_objs": [{"id": "text-generation", "label": "Text Generation", "type": "pipeline_tag", "subType": "nlp"}, {"id": "transformers", "label": "Transformers", "type": "library"}, {"id": "safetensors", "label": "Safetensors", "type": "library"}, {"id": "deepseek_v4", "label": "deepseek_v4", "type": "other", "clickable": true}, {"id": "eval-results", "label": "Eval Results", "type": "other", "clickable": true}, {"id": "endpoints_compatible", "label": "Inference Endpoints", "type": "other", "clickable": true}, {"id": "8-bit", "label": "8-bit precision", "type": "other", "clickable": true}, {"id": "fp8", "label": "fp8", "type": "other", "clickable": true}, {"id": "arxiv:2606.19348", "label": "arxiv:2606.19348", "type": "arxiv", "extra": {"paperTitle": "DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence"}}, {"id": "license:mit", "label": "mit", "type": "license"}, {"id": "deploy:sagemaker", "label": "sagemaker", "type": "deploy"}, {"type": "region", "label": "🌐 Region: US", "id": "region:us"}, {"transformersInfo": {"auto_model": "AutoModelForCausalLM", "pipeline_tag": "text-generation", "processor": "AutoTokenizer"}, "widgetData": [{"text": "Hi, what can you help me with?"}, {"text": "What is 84 * 3 / 2?"}, {"text": "Tell me an interesting fact about the universe!"}, {"text": "Explain quantum computing in simple terms."}], "safetensors": {"parameters": {"BF16": 2816899328, "I64": 2327040, "F32": 87776414, "F8_E4M3": 23169335296, "I8": 1572763336704, "total": 1598839

```
{}, "inferenceContextData": {"billableEntities": [], "entityName2Providers": {}, "defaultProviders": [{"isOriginalProvider": false, "name": "novita", "enabled": true, "position": 1, "isReleased": true, "accuratePricing": true}, {"isOriginalProvider": false, "name": "deepinfra", "enabled": true, "position": 16, "isReleased": true, "accuratePricing": true}, {"isOriginalProvider": false, "name": "featherless-ai", "enabled": true, "position": 7, "isReleased": true, "accuratePricing": true}, {"isOriginalProvider": false, "name": "together", "enabled": true, "position": 5, "isReleased": true, "accuratePricing": false}, {"isOriginalProvider": false, "name": "baseten", "enabled": true, "position": 13, "isReleased": true, "accuratePricing": true}]], "li {"url": "/spaces/huggingface/InferenceSupport/discussions/9467", "numReactions": 81}, {"isLoggedIn": false, "canWrite": false, "examples": [{"text": "Hi, what can you help me with?"}, {"text": "What is 84 * 3 / 2?"}, {"text": "Tell me an interesting fact about the universe!"}, {"text": "Explain quantum computing in simple terms."}], "initialSelectedProvider": "novita", "widgetType": "conversational", "isMaximized": false, "isMac": true}}">
Inference Providers NEW Novita
```

+2

Text Generation

Examples

Input a message to start chatting with deepseek-ai/DeepSeek-V4-Pro .

Send

View Code Snippets

Maximize

Model tree for deepseek-ai/DeepSeek-V4-Pro

Finetunes

15 models

Quantizations

27 models

Spaces using deepseek-ai/DeepSeek-V4-Pro 100

Collection including deepseek-ai/DeepSeek-V4-Pro 11 items • Updated 1 day ago • 932

Paper for deepseek-ai/DeepSeek-V4-Pro Paper • 2606.19348 • Published Apr 26 • 47

```
"", "lstrip": false, "normalized": true, "rstrip": false, "single_word": false}, {"eos_token": {"__type": "AddedToken", "content": "< | end of sentence | >", "lstrip": false, "normalized": true, "rstrip": false, "single_word": false}, {"pad_token": {"__type": "AddedToken", "content": "< | end of sentence | >", "lstrip": false, "normalized": true, "rstrip": false, "single_word": false}, {"unk_token": null}], {"createdAt": "2026-04-22T06:04:created", "downloads": 373563, "downloadsAllTime": 10895485, "id": "deepseek-ai/DeepSeek-V4-Pro", "isLikedByUser": false, "availableInferenceProviders": [{"provider": "together", "modelStatus": "error", "providerStatus": "live", "providerId": "deepseek-ai/DeepSeek-V4-Pro", "task": "conversational", "adapterWeightsPath": "model-00001-of-00064.safetensors", "features": {"structuredOutput": true, "toolCalling": true}, {"isCheapestPricingOutput": false, "isFastestThroughput": false, "isModelAuth {"provider": "novita", "modelStatus": "live", "providerStatus": "live", "providerId": "deepseek/deepseek-v4-pro", "task": "conversational", "adapterWeightsPath": "model-00001-of-00064.safetensors", "features": {"structuredOutput": false, "toolCalling": true}, {"isCheapestPricingOutput": false, "isFastestThroughput": false, "isModelAutl {"provider": "featherless-ai", "modelStatus": "live", "providerStatus": "live", "providerId": "deepseek-ai/DeepSeek-V4-Pro", "task": "conversational", "adapterWeightsPath": "model-00001-of-00064.safetensors", "isCheapestPricingOutput": false, "isFastestThroughput": false, "isModelAuthor": false}, {"provider": "deepinfra", "modelStatus": "live", "providerStatus": "live", "providerId": "deepseek-ai/DeepSeek-V4-Pro", "task": "conversational", "adapterWeightsPath": "model-00001-of-00064.safetensors", "features": {"structuredOutput": true, "toolCalling": true}, {"isCheapestPricingOutput": true, "isFastestThroughput": true, "isModelAuthor {"provider": "baseten", "modelStatus": "error", "providerStatus": "live", "providerId": "deepseek-ai/DeepSeek-V4-Pro", "task": "conversational", "adapterWeightsPath": "model-00001-of-00064.safetensors", "features":
```

```

{"structuredOutput":false,"toolCalling":true},"isCheapestPricingOutput":false,"isFastestThroughput":false,"isModelAutl
generation","library_name":"transformers","librariesOther":[],"trackDownloads":true,"model-
index":null,"evalResults":[{"dataset":{"id":"TIGER-Lab/MMLU-
Pro","isBenchmark":true,"task_id":"mmlu_pro"},"value":87.5,"source":{"url":"https://huggingface.co/deepseek-ai/
DeepSeek-V4-Pro","name":"Model Card","isExternal":false},"filename":"eval_results/mmlu-
pro.yaml","verified":false,"pullRequest":110,"rank":5,"label":"Mmlu Pro"},"dataset":{"id":"openai/
gsm8k","isBenchmark":true,"task_id":"gsm8k"},"value":92.6,"source":{"url":"https://huggingface.co/deepseek-ai/
DeepSeek-V4-Pro","name":"Model Card","isExternal":false},"filename":"eval_results/
gsm8k.yaml","verified":false,"pullRequest":110,"rank":5,"label":"Gsm8k"},"dataset":{"id":"SWE-bench/SWE-
bench_Verified","isBenchmark":true,"task_id":"swe_bench_%_resolved"},"value":80.6,"source":{"url":"https://
huggingface.co/deepseek-ai/DeepSeek-V4-Pro","name":"Model Card","isExternal":false},"filename":"eval_results/
swe-bench_verified.yaml","verified":false,"pullRequest":110,"rank":6,"label":"Swe Bench Resolved"},"dataset":
{"id":"agents-last-exam/agents-last-exam","isBenchmark":true,"task_id":"overall-pass-
rate"},"value":12.390350877192983,"source":{"url":"https://agents-last-exam.org/leaderboard","name":"ALE official
leaderboard","isExternal":true,"author":{"_id":"69f848800d5d932e60c42eac","avatarUrl":"https://cdn-
avatars.huggingface.co/v1/production/uploads/6582486355a1e6cdb3b5e19d/
eCSzorSXXKmJ5srUhvq7H.png","fullName":"agents-last-exam","name":"agents-last-
exam","type":"org","isHf":false,"isHfAdmin":false,"isMod":false,"followerCount":75,"isUserFollowing":false}},"filename"
ale-openclaw-default.yaml","verified":false,"pullRequest":215,"rank":7,"label":"Overall Pass
Rate","notes":"harness=openclaw; harness_variant=not reported; attempted=146/152; official leaderboard
snapshot=2026-10-01T20:39:17.826877+00:00; values are percentages (0-100); equal-task aggregation;
unattempted tasks contribute zero; pass threshold score>=0.999. This is an official leaderboard export, not an HF-
verified evaluation."},"dataset":{"id":"agents-last-exam/agents-last-exam","isBenchmark":true,"task_id":"overall-
score"},"value":27.62014846817647,"source":{"url":"https://agents-last-exam.org/leaderboard","name":"ALE official
leaderboard","isExternal":true,"author":{"_id":"69f848800d5d932e60c42eac","avatarUrl":"https://cdn-
avatars.huggingface.co/v1/production/uploads/6582486355a1e6cdb3b5e19d/
eCSzorSXXKmJ5srUhvq7H.png","fullName":"agents-last-exam","name":"agents-last-
exam","type":"org","isHf":false,"isHfAdmin":false,"isMod":false,"followerCount":75,"isUserFollowing":false}},"filename"
ale-openclaw-default.yaml","verified":false,"pullRequest":215,"rank":5,"label":"Overall
Score","notes":"harness=openclaw; harness_variant=not reported; attempted=146/152; official leaderboard
snapshot=2026-10-01T20:39:17.826877+00:00; values are percentages (0-100); equal-task aggregation;
unattempted tasks contribute zero; pass threshold score>=0.999. This is an official leaderboard export, not an HF-
verified evaluation."},"dataset":{"id":"agents-last-exam/agents-last-exam","isBenchmark":true,"task_id":"ale-cli-
pass-rate"},"value":14.126984126984127,"source":{"url":"https://agents-last-exam.org/leaderboard","name":"ALE
official leaderboard","isExternal":true,"author":{"_id":"69f848800d5d932e60c42eac","avatarUrl":"https://cdn-
avatars.huggingface.co/v1/production/uploads/6582486355a1e6cdb3b5e19d/
eCSzorSXXKmJ5srUhvq7H.png","fullName":"agents-last-exam","name":"agents-last-
exam","type":"org","isHf":false,"isHfAdmin":false,"isMod":false,"followerCount":75,"isUserFollowing":false}},"filename"
ale-openclaw-default.yaml","verified":false,"pullRequest":215,"rank":8,"label":"Ale Cli Pass
Rate","notes":"harness=openclaw; harness_variant=not reported; attempted=104/105; official leaderboard
snapshot=2026-10-01T20:39:17.826877+00:00; values are percentages (0-100); equal-task aggregation;
unattempted tasks contribute zero; pass threshold score>=0.999. This is an official leaderboard export, not an HF-
verified evaluation."},"dataset":{"id":"agents-last-exam/agents-last-exam","isBenchmark":true,"task_id":"ale-cli-
score"},"value":30.84164065884846,"source":{"url":"https://agents-last-exam.org/leaderboard","name":"ALE official
leaderboard","isExternal":true,"author":{"_id":"69f848800d5d932e60c42eac","avatarUrl":"https://cdn-
avatars.huggingface.co/v1/production/uploads/6582486355a1e6cdb3b5e19d/
eCSzorSXXKmJ5srUhvq7H.png","fullName":"agents-last-exam","name":"agents-last-
exam","type":"org","isHf":false,"isHfAdmin":false,"isMod":false,"followerCount":75,"isUserFollowing":false}},"filename"
ale-openclaw-default.yaml","verified":false,"pullRequest":215,"rank":4,"label":"Ale Cli
Score","notes":"harness=openclaw; harness_variant=not reported; attempted=104/105; official leaderboard
snapshot=2026-10-01T20:39:17.826877+00:00; values are percentages (0-100); equal-task aggregation;
unattempted tasks contribute zero; pass threshold score>=0.999. This is an official leaderboard export, not an HF-
verified evaluation."},"dataset":{"id":"agents-last-exam/agents-last-exam","isBenchmark":true,"task_id":"near-
term-pass-rate"},"value":19.900497512437813,"source":{"url":"https://agents-last-exam.org/
leaderboard","name":"ALE official leaderboard","isExternal":true,"author":
{"_id":"69f848800d5d932e60c42eac","avatarUrl":"https://cdn-avatars.huggingface.co/v1/production/

```

uploads/6582486355a1e6cdb3b5e19d/eCSzorSXXKmj5srUhvq7H.png", "fullname": "agents-last-exam", "name": "agents-last-exam", "type": "org", "isHf": false, "isHfAdmin": false, "isMod": false, "followerCount": 75, "isUserFollowing": false}, {"filename": "ale-openclaw-default.yaml", "verified": false, "pullRequest": 215, "rank": 4, "label": "Near Term Pass Rate", "notes": "harness=openclaw; harness_variant=not reported; attempted=67/67; official leaderboard snapshot=2026-10-01T20:39:17.826877+00:00; values are percentages (0-100); equal-task aggregation; unattempted tasks contribute zero; pass threshold score>=0.999. This is an official leaderboard export, not an HF-verified evaluation."}, {"dataset": {"id": "agents-last-exam/agents-last-exam", "isBenchmark": true, "task_id": "near-term-score"}, "value": 43.81171518153468, "source": {"url": "https://agents-last-exam.org/leaderboard", "name": "ALE official leaderboard", "isExternal": true, "author": {"_id": "69f848800d5d932e60c42eac", "avatarUrl": "https://cdn-avatars.huggingface.co/v1/production/uploads/6582486355a1e6cdb3b5e19d/eCSzorSXXKmj5srUhvq7H.png", "fullname": "agents-last-exam", "name": "agents-last-exam", "type": "org", "isHf": false, "isHfAdmin": false, "isMod": false, "followerCount": 75, "isUserFollowing": false}}, {"filename": "ale-openclaw-default.yaml", "verified": false, "pullRequest": 215, "rank": 4, "label": "Near Term Score", "notes": "harness=openclaw; harness_variant=not reported; attempted=67/67; official leaderboard snapshot=2026-10-01T20:39:17.826877+00:00; values are percentages (0-100); equal-task aggregation; unattempted tasks contribute zero; pass threshold score>=0.999. This is an official leaderboard export, not an HF-verified evaluation."}, {"dataset": {"id": "agents-last-exam/agents-last-exam", "isBenchmark": true, "task_id": "full-spectrum-pass-rate"}, "value": 10.909090909090908, "source": {"url": "https://agents-last-exam.org/leaderboard", "name": "ALE official leaderboard", "isExternal": true, "author": {"_id": "69f848800d5d932e60c42eac", "avatarUrl": "https://cdn-avatars.huggingface.co/v1/production/uploads/6582486355a1e6cdb3b5e19d/eCSzorSXXKmj5srUhvq7H.png", "fullname": "agents-last-exam", "name": "agents-last-exam", "type": "org", "isHf": false, "isHfAdmin": false, "isMod": false, "followerCount": 75, "isUserFollowing": false}}, {"filename": "ale-openclaw-default.yaml", "verified": false, "pullRequest": 215, "rank": 4, "label": "Full Spectrum Pass Rate", "notes": "harness=openclaw; harness_variant=not reported; attempted=54/55; official leaderboard snapshot=2026-10-01T20:39:17.826877+00:00; values are percentages (0-100); equal-task aggregation; unattempted tasks contribute zero; pass threshold score>=0.999. This is an official leaderboard export, not an HF-verified evaluation."}, {"dataset": {"id": "agents-last-exam/agents-last-exam", "isBenchmark": true, "task_id": "full-spectrum-score"}, "value": 23.799048181818186, "source": {"url": "https://agents-last-exam.org/leaderboard", "name": "ALE official leaderboard", "isExternal": true, "author": {"_id": "69f848800d5d932e60c42eac", "avatarUrl": "https://cdn-avatars.huggingface.co/v1/production/uploads/6582486355a1e6cdb3b5e19d/eCSzorSXXKmj5srUhvq7H.png", "fullname": "agents-last-exam", "name": "agents-last-exam", "type": "org", "isHf": false, "isHfAdmin": false, "isMod": false, "followerCount": 75, "isUserFollowing": false}}, {"filename": "ale-openclaw-default.yaml", "verified": false, "pullRequest": 215, "rank": 4, "label": "Full Spectrum Score", "notes": "harness=openclaw; harness_variant=not reported; attempted=54/55; official leaderboard snapshot=2026-10-01T20:39:17.826877+00:00; values are percentages (0-100); equal-task aggregation; unattempted tasks contribute zero; pass threshold score>=0.999. This is an official leaderboard export, not an HF-verified evaluation."}, {"dataset": {"id": "agents-last-exam/agents-last-exam", "isBenchmark": true, "task_id": "last-exam-pass-rate"}, "value": 0, "source": {"url": "https://agents-last-exam.org/leaderboard", "name": "ALE official leaderboard", "isExternal": true, "author": {"_id": "69f848800d5d932e60c42eac", "avatarUrl": "https://cdn-avatars.huggingface.co/v1/production/uploads/6582486355a1e6cdb3b5e19d/eCSzorSXXKmj5srUhvq7H.png", "fullname": "agents-last-exam", "name": "agents-last-exam", "type": "org", "isHf": false, "isHfAdmin": false, "isMod": false, "followerCount": 75, "isUserFollowing": false}}, {"filename": "ale-openclaw-default.yaml", "verified": false, "pullRequest": 215, "rank": 4, "label": "Last Exam Pass Rate", "notes": "harness=openclaw; harness_variant=not reported; attempted=33/38; official leaderboard snapshot=2026-10-01T20:39:17.826877+00:00; values are percentages (0-100); equal-task aggregation; unattempted tasks contribute zero; pass threshold score>=0.999. This is an official leaderboard export, not an HF-verified evaluation."}, {"dataset": {"id": "agents-last-exam/agents-last-exam", "isBenchmark": true, "task_id": "last-exam-score"}, "value": 2.521315789473684, "source": {"url": "https://agents-last-exam.org/leaderboard", "name": "ALE official leaderboard", "isExternal": true, "author": {"_id": "69f848800d5d932e60c42eac", "avatarUrl": "https://cdn-avatars.huggingface.co/v1/production/uploads/6582486355a1e6cdb3b5e19d/eCSzorSXXKmj5srUhvq7H.png", "fullname": "agents-last-exam", "name": "agents-last-exam", "type": "org", "isHf": false, "isHfAdmin": false, "isMod": false, "followerCount": 75, "isUserFollowing": false}}, {"filename": "ale-openclaw-default.yaml", "verified": false, "pullRequest": 215, "rank": 4, "label": "Last Exam

Score", "notes": "harness=openclaw; harness_variant=not reported; attempted=33/38; official leaderboard snapshot=2026-10-01T20:39:17.826877+00:00; values are percentages (0-100); equal-task aggregation; unattempted tasks contribute zero; pass threshold score>=0.999. This is an official leaderboard export, not an HF-verified evaluation.", {"dataset": {"id": "ldavidrein/gpqa", "isBenchmark": true, "task_id": "diamond", "value": 90.1, "source": {"url": "https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro", "name": "Model Card", "isExternal": false}, "filename": ".eval_results/gpqa.yaml", "verified": false, "pullRequest": 110, "rank": 14, "label": "Diamond"}, {"dataset": {"id": "ScaleAI/SWE-bench_Pro", "isBenchmark": true, "task_id": "SWE_Bench_Pro", "value": 55.4, "source": {"url": "https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro", "name": "Model Card", "isExternal": false}, "filename": ".eval_results/swe-bench_pro.yaml", "verified": false, "pullRequest": 110, "rank": 27, "label": "SWE Bench Pro"}, {"dataset": {"id": "IntelligenceLab/Long-Horizon-Terminal-Bench", "isBenchmark": true, "task_id": "lhtb_solved", "value": 3, "source": {"url": "https://zli12321.github.io/LHTB/leaderboard.html", "name": "LHTB leaderboard", "isExternal": true}, "filename": ".eval_results/lhtb.yaml", "verified": false, "pullRequest": 210, "rank": 2, "label": "Lhtb Solved", "notes": "3/46 tasks solved at reward >= 0.95 (pass@0.95); mean reward x100 = 30.7; official LHTB Harbor harness"}, {"dataset": {"id": "IntelligenceLab/Long-Horizon-Terminal-Bench", "isBenchmark": true, "task_id": "lhtb2_solved", "value": 9, "source": {"url": "https://github.com/zli12321/LHTB/tree/lhtb-2.0#lhtb-20-results-october-2026", "name": "LHTB 2.0 results", "isExternal": true}, "filename": ".eval_results/lhtb2_solved.yaml", "verified": false, "pullRequest": 210, "rank": 5, "label": "Lhtb2 Solved", "notes": "9/77 tasks solved at reward >= 0.95 (pass@0.95); mean reward x100 = 52.9; LHTB 2.0 (77 tasks), terminus-2, 3-hour budget per task, official LHTB Harbor harness"}, {"dataset": {"id": "actava/chi-bench", "isBenchmark": true, "task_id": "chi_bench", "value": 14.2, "source": {"url": "https://arxiv.org/abs/2605.16679", "name": "CHI-Bench", "isExternal": true}, "filename": ".eval_results/chi-bench.yaml", "verified": false, "pullRequest": 197, "rank": 3, "label": "Chi Bench", "notes": "Harness: OpenAI Agents SDK; Protocol: 75 tasks x 3 trials; Metric: pass@1 (%)"}, {"dataset": {"id": "actava/chi-bench", "isBenchmark": true, "task_id": "prior_authorization", "value": 10.7, "source": {"url": "https://arxiv.org/abs/2605.16679", "name": "CHI-Bench", "isExternal": true}, "filename": ".eval_results/chi-bench.yaml", "verified": false, "pullRequest": 197, "rank": 3, "label": "Prior Authorization", "notes": "Harness: OpenAI Agents SDK; Protocol: 75 tasks x 3 trials; Metric: pass@1 (%)"}, {"dataset": {"id": "actava/chi-bench", "isBenchmark": true, "task_id": "utilization_management", "value": 28, "source": {"url": "https://arxiv.org/abs/2605.16679", "name": "CHI-Bench", "isExternal": true}, "filename": ".eval_results/chi-bench.yaml", "verified": false, "pullRequest": 197, "rank": 2, "label": "Utilization Management", "notes": "Harness: OpenAI Agents SDK; Protocol: 75 tasks x 3 trials; Metric: pass@1 (%)"}, {"dataset": {"id": "actava/chi-bench", "isBenchmark": true, "task_id": "care_management", "value": 4, "source": {"url": "https://arxiv.org/abs/2605.16679", "name": "CHI-Bench", "isExternal": true}, "filename": ".eval_results/chi-bench.yaml", "verified": false, "pullRequest": 197, "rank": 3, "label": "Care Management", "notes": "Harness: OpenAI Agents SDK; Protocol: 75 tasks x 3 trials; Metric: pass@1 (%)"}, {"dataset": {"id": "harborframework/terminal-bench-2.0", "isBenchmark": true, "task_id": "terminalbench_2", "value": 67.9, "source": {"url": "https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro", "name": "Model Card", "isExternal": false}, "filename": ".eval_results/terminal-bench-2.0.yaml", "verified": false, "pullRequest": 110, "rank": 6, "label": "Terminalbench 2"}, {"dataset": {"id": "joelniklaus/LEXam-hard", "isBenchmark": true, "task_id": "lexam_hard", "value": 38.93, "source": {"url": "https://huggingface.co/buckets/joelniklaus/SwissLegalEvals", "name": "SwissLegalEvals per-sample details (lighteval)", "isExternal": true, "author": {"id": "5fae5f68b8423e1d80b8a988", "avatarUrl": "https://cdn-avatars.huggingface.co/v1/production/uploads/1656078368144-5fae5f68b8423e1d80b8a988.jpeg", "fullName": "Joel Niklaus", "name": "joelniklaus", "type": "user", "isPro": false, "isHf": true, "isHfAdmin": false, "isMod": false, "followerCount": 17}, "filename": "lexam-hard.yaml", "verified": false, "pullRequest": 214, "rank": 3, "label": "Lexam Hard", "notes": "lighteval, LEXam paper prompts, one response per question, no tools; DeepSeek-R1-0528 judge; mean of the German and English means over the 518 questions, 0-100"}, {"dataset": {"id": "SWE-bench/SWE-bench_Multilingual", "isBenchmark": true, "task_id": "swe_bench_multilingual_%_resolved", "value": 76.2, "source": {"url": "https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro", "name": "Model Card", "isExternal": false}, "filename": ".eval_results/swe-bench_multilingual.yaml", "verified": false, "pullRequest": 213, "rank": 9, "label": "Swe Bench Multilingual Resolved"}, {"dataset": {"id": "benchflow/skillsbench", "isBenchmark": true, "task_id": "skillsbench_v1_1", "value": 50.1, "source": {"url": "https://huggingface.co/datasets/benchflow/skillsbench-leaderboard/raw/main/leaderboard/skillsbench/v1.1/official.json", "name": "SkillsBench v1.1 official leaderboard", "isExternal": true, "author":

```

{"_id":"69657f24ec1d157f1590a81d","avatarUrl":"https://cdn-avatars.huggingface.co/v1/production/
uploads/663fe2d26304d377fc253322/L6ik2VpL5iizADEXK4W-
A.png","fullname":"BenchFlow","name":"benchflow","type":"org","isHf":false,"isHfAdmin":false,"isMod":false,"plan":"te:
skillsbench.yaml","verified":false,"pullRequest":205,"rank":8,"label":"Skillsbench V1 1","notes":"with-skills;
BenchFlow harness; OpenHands agent; 87 tasks x 3 trials; full 261/261 coverage"},{"dataset":{"id":"internlm/
WildClawBench","isBenchmark":false,"task_id":"overall"},"value":43.7,"source":{"url":"https://internlm.github.io/
WildClawBench","name":"WildClawBench","isExternal":true,"author":
{"_id":"64a2d5fa81252883206f24c9","avatarUrl":"https://cdn-avatars.huggingface.co/v1/production/
uploads/6432683407bad11484a68457/Q3Y0dL79GcsnaBCGRMooZ.png","fullname":"Intern Large
Models","name":"internlm","type":"org","isHf":false,"isHfAdmin":false,"isMod":false,"followerCount":1347,"isUserFollow
wildclawbench.yaml","verified":false,"pullRequest":187,"rank":12,"label":"Overall"},{"dataset":{"id":"internlm/
WildClawBench","isBenchmark":false,"task_id":"avg_time"},"value":605,"source":{"url":"https://internlm.github.io/
WildClawBench","name":"WildClawBench","isExternal":true,"author":
{"_id":"64a2d5fa81252883206f24c9","avatarUrl":"https://cdn-avatars.huggingface.co/v1/production/
uploads/6432683407bad11484a68457/Q3Y0dL79GcsnaBCGRMooZ.png","fullname":"Intern Large
Models","name":"internlm","type":"org","isHf":false,"isHfAdmin":false,"isMod":false,"followerCount":1347,"isUserFollow
wildclawbench.yaml","verified":false,"pullRequest":187,"rank":3,"label":"Avg Time"},{"dataset":{"id":"internlm/
WildClawBench","isBenchmark":false,"task_id":"avg_cost"},"value":12,"source":{"url":"https://internlm.github.io/
WildClawBench","name":"WildClawBench","isExternal":true,"author":
{"_id":"64a2d5fa81252883206f24c9","avatarUrl":"https://cdn-avatars.huggingface.co/v1/production/
uploads/6432683407bad11484a68457/Q3Y0dL79GcsnaBCGRMooZ.png","fullname":"Intern Large
Models","name":"internlm","type":"org","isHf":false,"isHfAdmin":false,"isMod":false,"followerCount":1347,"isUserFollow
wildclawbench.yaml","verified":false,"pullRequest":187,"rank":7,"label":"Avg Cost"},{"dataset":{"id":"InternScience/
ResearchClawBench","isBenchmark":false,"task_id":"overall"},"value":17.12,"source":{"url":"https://
huggingface.co/deepseek-ai/DeepSeek-V4-Pro","name":"Model Card","isExternal":false},"filename":"eval_results/
researchclawbench.yaml","verified":false,"pullRequest":186,"rank":4,"label":"Overall","notes":"ResearchHarness
evaluation with tools enabled, code execution, and a file-system workspace; completed 39/40
ResearchClawBench tasks."},"dataset":{"id":"claw-eval/Claw-
Eval","isBenchmark":true,"task_id":"general"},"value":58.4,"source":{"url":"https://claw-
eval.github.io","name":"Claw-Eval Leaderboard","isExternal":true,"author":
{"_id":"6038d6d0612f5eef3cc05ea9","avatarUrl":"https://cdn-avatars.huggingface.co/v1/production/
uploads/6038d6d0612f5eef3cc05ea9/ryhvAX5djQpD5OrlIZQ1f.jpeg","fullname":"Lei
Li","name":"tobiaslee","type":"user","isPro":false,"isHf":false,"isHfAdmin":false,"isMod":false,"followerCount":30,"isUse
claw-eval.yaml","verified":false,"pullRequest":173,"rank":16,"label":"General","notes":"Pass3% | N=3 | 161 tasks"},
{"dataset":{"id":"claw-eval/Claw-Eval","isBenchmark":true,"task_id":"multi_turn"},"value":65.8,"source":
{"url":"https://claw-eval.github.io","name":"Claw-Eval Leaderboard","isExternal":true,"author":
{"_id":"6038d6d0612f5eef3cc05ea9","avatarUrl":"https://cdn-avatars.huggingface.co/v1/production/
uploads/6038d6d0612f5eef3cc05ea9/ryhvAX5djQpD5OrlIZQ1f.jpeg","fullname":"Lei
Li","name":"tobiaslee","type":"user","isPro":false,"isHf":false,"isHfAdmin":false,"isMod":false,"followerCount":30,"isUse
claw-eval.yaml","verified":false,"pullRequest":173,"rank":2,"label":"Multi Turn","notes":"Pass3% | N=3 | 38 tasks"},
{"dataset":{"id":"collinear-ai/yc-bench","isBenchmark":true,"task_id":"medium"},"value":1066426,"source":
{"url":"https://github.com/collinear-ai/yc-bench","name":"YC-Bench
eval","isExternal":true},"filename":"eval_results/yc-
bench.yaml","verified":false,"pullRequest":165,"rank":4,"label":"Medium","notes":"medium preset, avg final funds
(USD) across seeds 1,2,3"},"dataset":{"id":"cais/hle","isBenchmark":false,"task_id":"hle"},"value":37.7,"source":
{"url":"https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro","name":"Model
Card","isExternal":false},"filename":"eval_results/
hle.yaml","verified":false,"pullRequest":110,"rank":28,"label":"Hle"},"private":false,"repoType":"model","gated":false,"t
["transformers","safetensors","deepseek_v4","text-generation","arxiv:2606.19348","license:mit","eval-
results","endpoints_compatible","8-bit","fp8","deploy:sagemaker","region:us"],"tag_objs":[{"id":"text-
generation","label":"Text Generation","type":"pipeline_tag","subType":"nlp"},
{"id":"transformers","label":"Transformers","type":"library"},{"id":"safetensors","label":"Safetensors","type":"library"},
{"id":"deepseek_v4","label":"deepseek_v4","type":"other","clickable":true},{"id":"eval-results","label":"Eval
Results","type":"other","clickable":true},{"id":"endpoints_compatible","label":"Inference
Endpoints","type":"other","clickable":true},{"id":"8-bit","label":"8-bit precision","type":"other","clickable":true},
{"id":"fp8","label":"fp8","type":"other","clickable":true},

```

`{"id":"arxiv:2606.19348","label":"arxiv:2606.19348","type":"arxiv","extra":{"paperTitle":"DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence"}}, {"id":"license:mit","label":"mit","type":"license"}, {"id":"deploy:sagemaker","label":"sagemaker","type":"deploy"}, {"type":"region","label":"🌐 Region: US","id":"region:us"}], "transformersInfo":{"auto_model":"AutoModelForCausalLM","pipeline_tag":"text-generation","processor":"AutoTokenizer"},"widgetData":[{"text":"Hi, what can you help me with?"}, {"text":"What is 84 * 3 / 2?"}, {"text":"Tell me an interesting fact about the universe!"}, {"text":"Explain quantum computing in simple terms."}], "safetensors":{"parameters":{"BF16":2816899328,"I64":2327040,"F32":87776414,"F8_E4M3":23169335296,"I8":1572763336704},"total":1598839}`
Evaluation results TIGER-Lab/MMLU-Pro · Mmlu Pro View evaluation results leaderboard 87.5

openai/gsm8k · Gsm8k View evaluation results leaderboard 92.6

SWE-bench/SWE-bench_Verified · Swe Bench Resolved View evaluation results leaderboard 80.6

agents-last-exam/agents-last-exam leaderboard

Overall Pass Rate View evaluation results source =0.999. This is an official leaderboard export, not an HF-verified evaluation." class="@max-md:hidden mr-1 max-w-36 truncate rounded-sm bg-gray-50 px-1 py-0.5 text-xs text-gray-600 dark:bg-gray-900 dark:text-gray-400">harness=openclaw; harness_variant=not reported; attempted=146/152; official leaderboard snapshot=2026-10-01T20:39:17.826877+00:00; values are percentages (0-100); equal-task aggregation; unattempted tasks contribute zero; pass threshold score>=0.999. This is an official leaderboard export, not an HF-verified evaluation. 12.39 *

Overall Score View evaluation results source =0.999. This is an official leaderboard export, not an HF-verified evaluation." class="@max-md:hidden mr-1 max-w-36 truncate rounded-sm bg-gray-50 px-1 py-0.5 text-xs text-gray-600 dark:bg-gray-900 dark:text-gray-400">harness=openclaw; harness_variant=not reported; attempted=146/152; official leaderboard snapshot=2026-10-01T20:39:17.826877+00:00; values are percentages (0-100); equal-task aggregation; unattempted tasks contribute zero; pass threshold score>=0.999. This is an official leaderboard export, not an HF-verified evaluation. 27.62 *

Ale Cli Pass Rate View evaluation results source =0.999. This is an official leaderboard export, not an HF-verified evaluation." class="@max-md:hidden mr-1 max-w-36 truncate rounded-sm bg-gray-50 px-1 py-0.5 text-xs text-gray-600 dark:bg-gray-900 dark:text-gray-400">harness=openclaw; harness_variant=not reported; attempted=104/105; official leaderboard snapshot=2026-10-01T20:39:17.826877+00:00; values are percentages (0-100); equal-task aggregation; unattempted tasks contribute zero; pass threshold score>=0.999. This is an official leaderboard export, not an HF-verified evaluation. 14.13 *

+7 more

Expand 13 benchmarks