Frontier-AI til enterprise GPT-5.5 vs Opus 4.7

En hurtig sammenligning af de nye frontier-modeller — hvor de adskiller sig, og hvad I bør pilotere først.

På 14 dage har OpenAI og Anthropic landet to nye frontier-modeller. GPT-5.5 (kodenavn "Spud") kom 23. april. Claude Opus 4.7 kom 16. april. Begge claimer agentisk overlegenhed og time-besparelser. Begge er live for enterprise-kunder lige nu.

Men hvilken er bedst til hvad? Og — vigtigere — hvor klar er jeres organisation til at bruge dem rigtigt? Denne mini-guide er den hurtige version: hvad de kan, hvor de adskiller sig, hvilke use-cases passer til hvilken model, og hvad I bør have på plads inden I ruller ud.

Tempoet er absurd. Licensen kommer. Modellen kommer. Men governance, implementering og vaner mangler.

Du skal vide:

- **Opus 4.7**

  vinder de hårdeste kognitive benchmarks — HLE, SWE-Bench Pro, agentic reliability, long-context
- **GPT-5.5**

  vinder agentik (Terminal-Bench), pattern-reasoning (ARC-AGI), og er strukturelt billigere (72% færre tokens)
- Real-world reviewers er enige: ingen dominerer — det rigtige valg afhænger af workload-form
- Tidlige tests viser øget hallucinations-rate på GPT-5.5 — kræver verification i kritiske workflows
- Den største flaskehals er ikke modellen — det er governance og vaner

Hvad er der sket på 4 uger?

To frontier-lanceringer, syv dage fra hinanden. Et stykke kontekst for dem der ikke følger AI-cyklussen hver dag:

Dato	Begivenhed	Hvorfor det betyder noget
16. april	Anthropic lancerer **Claude Opus 4.7**	SWE-Bench Verified hopper fra 80,8% til 87,6%. Første Claude-model med high-res billed-input (3,75 MP).
23. april	OpenAI lancerer **GPT-5.5 "Spud"**	Første fuldt re-trænede base-model siden GPT-4.5. Stærkere agentisk evne. Bank of New York annonceret som tidlig tester.
5. maj	GPT-5.5 Instant til free-tier	Frontier-niveau er ikke længere kun for betalende kunder.
Primo maj	GPT-5.5 (Thinking + Instant) i Microsoft 365 Copilot	Enterprise-kunder med M365 Copilot får adgang uden ekstra køb. Thinking-varianten først, Instant fra 8. maj.

Resten af feltet (Google Gemini 3, Mistral Large 3, Meta Llama 4) har ikke leveret noget i samme klasse i denne periode. Slaget om enterprise-AI handler lige nu om GPT vs Claude.

1 GPT-5.5 ("Spud") — OpenAI

Lanceret 23. april 2026. OpenAI's første fuldt re-trænede base-model siden GPT-4.5. Tre tier: Instant (gratis), Pro (betalt), og 5.5 Pro Enterprise.

Styrker:

- **Terminal-Bench 2.0: 82,7%** — 13 procentpoint foran Opus 4.7. Bedst til at navigere kommandoer, værktøjer og workflows i terminalmiljøer.
- **ARC-AGI-2 Verified: 85,0%** — 9 procentpoint foran Opus 4.7 (75,8%). Bedst på abstrakt pattern-reasoning.
- **SWE-Bench Verified: 88,7%** — leder leaderboardet for "klassisk" kode-benchmark
- **1 million tokens context window** — markant større end Opus 4.7's 200K. Relevant for hele kodebaser, lange dokumenter, og komplekse multi-step workflows.
- **72% færre output-tokens** end Opus 4.7 på samme opgaver — strukturel cost-fordel ved skala (men Opus er 17% billigere per output-token, så netto-spread er ~55-60%)
- **Tighter Codex-integration** og hurtigere throughput
- Leder MMLU og MATH på ren reasoning

Svagheder:

- **HLE (Humanity's Last Exam): 41,4%** — 5,5 procentpoint bag Opus 4.7 på det hårdeste akademiske benchmark
- **SWE-Bench Pro: 58,6%** — taber til Opus 4.7 (64,3%) på de hårdere kode-tasks
- **Øget hallucinations-rate** rapporteret i tidlige tests — producerer overbevisende-men-forkerte svar oftere. Kræver verification i kritiske workflows.
- Image-håndtering bag Claude på opløsning og detalje

Adgang:

- ChatGPT Pro, Business, Enterprise
- API: GPT-5.5 og GPT-5.5 Pro (lanceret 24. april)
- Microsoft 365 Copilot — som "Think Deeper"-model siden 8. maj
- API til cybersikkerhedsbrug "very soon" — OpenAI venter på guardrails

Bedst til: agentiske workflows, terminal-automation, DevOps-orkestrering, tool-chaining over flere skridt. Bank of New York's CIO Leigh-Ann Russell rapporterer "meaningful improvements" i regulerede workflows — særligt på hallucinations-resistance og task completion.

2 Claude Opus 4.7 — Anthropic

Lanceret 16. april 2026. Inkremental opdatering af Opus 4.6, men benchmarks viser markant fremgang. Samme pris som forgængeren: $5 input / $25 output per million tokens.

Styrker:

- **HLE (Humanity's Last Exam): 46,9%** uden værktøjer, **54,7%** med — slår GPT-5.5 på det hårdeste akademiske benchmark. Bedst på dyb kognitiv reasoning.
- **SWE-Bench Pro: 64,3%** — slår GPT-5.5 (58,6%) på den hårdeste kode-benchmark. Faktiske GitHub-issues, ikke toy-problems.
- **SWE-Bench Verified: 87,6%** — fra 80,8% i Opus 4.6
- **Agentic coding reliability** — bedre instruction adherence på lange tasks, bevarer task coherence over lange action chains
- **Long-context reasoning** — verificerer ofte sit eget output før den rapporterer tilbage
- **Høj-opløsnings billed-input** — op til 2576px / 3,75 MP (3× mere end Opus 4.6)
- **CursorBench: 70%** — fra 58% i Opus 4.6. Stor fremgang i IDE-integrerede coding-agenter.

Svagheder:

- **Terminal-Bench 2.0: 69,4%** — taber 13 procentpoint til GPT-5.5 på agentik
- **ARC-AGI-2: 75,8%** — 9 procentpoint bag GPT-5.5 på abstrakt pattern-reasoning
- **Bruger 72% flere output-tokens** end GPT-5.5 på samme opgaver — markant højere driftsomkostning ved skala

Adgang:

- Claude.ai (Pro, Max, Team, Enterprise)
- Anthropic API direkte
- Amazon Bedrock
- Google Cloud Vertex AI
- Microsoft Foundry

Bedst til: kompleks kode-refactoring, kodebase-analyse, lange dokumenter, dokument-til-kode pipelines, billede-til-kode workflows, alt der kræver præcis instruktion over mange skridt.

3 Det øvrige felt — hvad I også bør kende

GPT vs Claude er hovedslaget, men der er tre andre spillere I bør have i baghovedet:

Claude Mythos (Anthropic, unreleased):

- Anthropic erkendte ved Opus 4.7-launch at deres egen interne Mythos-preview slår både Opus 4.7 og GPT-5.5 på flere benchmarks. Mythos SWE-Bench Pro: 77,8%.
- Ikke produktificeret endnu, men signalerer at næste generation er ca. 6 måneder ude.

Microsoft 365 Copilot:

- Ikke en model i sig selv — men en distribution. Drevet af GPT-5.5 i "Think Deeper" siden 8. maj.
- Enterprise-fordel: data forbliver i tenant, ingen træning på jeres prompts, M365-licens dækker.
- Den enkleste vej til frontier-AI for de fleste danske organisationer.

EU-suveræne alternativer:

- Mistral Large 3 — relevant for data-residens-cases hvor EU-jurisdiktion er kritisk
- Performance-gap til GPT/Claude er stadig betydelig, men reduceres kvartal for kvartal

Hold øje med: Mythos-release og GPT-6 i Q3 2026. Frontier-cyklussen er nu 6-8 uger, ikke 6 måneder. Det betyder noget for jeres roadmap-planlægning.

Benchmarks der betyder noget

Benchmarks er notorisk dårlige til at forudsige produktivitet i den virkelige verden. Brug dem som retning, ikke som sandhed. Her er hvad tallene siger lige nu:

Benchmark	GPT-5.5	Opus 4.7	Hvad det måler
HLE (no tools)	41,4%	**46,9%**	Humanity's Last Exam. Hårdest akademisk benchmark. Multi-modal, multi-domæne.
HLE (with tools)	52,2%	**54,7%**	HLE med værktøjs-adgang. Tester reasoning + tool-orkestrering.
ARC-AGI-2 Verified	**85,0%**	75,8%	Abstrakt pattern-reasoning. Fluid intelligence-test.
Terminal-Bench 2.0	**82,7%**	69,4%	Navigation og opgaveløsning i kommandolinje-miljøer. Agentik.
SWE-Bench Verified	**88,7%**	87,6%	Klassisk kode-benchmark. Open-source bug-fixing.
SWE-Bench Pro	58,6%	**64,3%**	Hårdere kode-benchmark. Faktiske GitHub-issues.
CursorBench	n/a	**70%**	IDE-integreret coding (Cursor). Opus 4.7 op fra 58%.
Token-effektivitet	**100% (baseline)**	+72% output	Output-tokens på samme task. Strukturel cost-fordel for GPT-5.5.
Max image res	1024×1024	**2576px / 3,75 MP**	Billed-input. Opus 4.7 ~3× mere end alle konkurrenter.
Context window	**1M tokens**	200K tokens	Relevant for hele kodebaser, lange dokumenter, multi-step workflows.
API-pris ($/1M tokens)	$5 in / $30 out	**$5 in / $25 out**	Samme input-pris. Opus 4.7 er 17% billigere på output.

Note: Claude Mythos Preview leder både SWE-Bench Pro (77,8%) og HLE (64,7%), men er ikke produktificeret endnu.

Hvad siger folk der faktisk bruger dem?

Læser man anmeldelser på tværs af DataCamp, Vellum, MindStudio, Tom's Guide og BuildFastWithAI fra de seneste uger, lander de alle sammen samme sted:

> *"Neither model dominates. Claude leads on cognitive depth and agentic reliability. GPT-5.5 leads on speed, breadth, and structural cost. The right call depends on workload shape — not benchmarks."*

Hvad reviewer-feltet er enige om:

- **Opus 4.7 vinder:** agentic coding reliability, instruction adherence på lange tasks, long-context discipline, hårdeste akademiske benchmarks (HLE)
- **GPT-5.5 vinder:** hastighed, tool-use breddde, ARC-AGI-2, Codex-økosystem, og cost ved skala
- **Bekymring rejst af flere:** GPT-5.5 hallucinerer mere end forventet trods bedre benchmark-tal. DeepLearning.AI flagger det eksplicit som en regression fra GPT-5.4.

Tom's Guide kørte 7 sammenlignende tests på tværs af logik, reasoning, domæneviden og praktisk anvendelighed — Claude vandt alle 7. Vellum og DataCamp landede på "tie depending on use-case". Den faktiske dom afhænger 100% af hvad du tester på.

Hvilken model passer til hvilken use-case?

Det vigtigste spørgsmål: hvilken er bedst til *jer*? Her er fem typiske enterprise-cases:

Agentiske workflows og tool-orkestrering → GPT-5.5.
Terminal-Bench-forspringet er for stort til at ignorere. Hvis I bygger AI-agenter der skal koordinere flere systemer, tjekke deres eget arbejde og handle over flere skridt — start her.

Kompleks kode-refactoring og legacy-system-vedligehold → Opus 4.7.
SWE-Bench Pro-forspringet kombineret med long-context reasoning gør Opus 4.7 til det bedre valg når koden er rodet, gammel, og kræver tålmodighed.

Knowledge work og dokumenthåndtering på tværs af Office → Microsoft 365 Copilot (GPT-5.5).
Hvis I allerede har M365, er adoptions-friktionen lavest her. Modellen er GPT-5.5 under emhætten, men leveret i en grænseflade brugerne allerede kender.

Billed-tung analyse — diagrammer, screenshots, scannede dokumenter → Opus 4.7.
3,75 MP billed-input giver Opus en reel praktisk fordel over GPT-5.5 på alt fra arkitektur-tegninger til håndskrevne notater.

EU-data-residens og compliance-første cases → Mistral Large 3 eller Opus 4.7 via Amazon Bedrock EU-region.
Performance-tabet er reelt men acceptabelt for cases hvor data-jurisdiktion er kontraktligt påkrævet.

Enterprise readiness — har I det der skal til?

Spørgsmålet I bør stille jer selv før I beslutter modeller:

"Har I overhovedet arbejdsprocesser, der kan udnytte en model, der planlægger, bruger værktøjer og tjekker sit eget arbejde?"

De fleste organisationer har det ikke. Ikke fordi medarbejderne er dårlige, men fordi processerne aldrig blev designet til at integrere et tredje, ikke-menneskeligt teammedlem. Her er en hurtig tjekliste:

- **Governance-struktur** — hvem ejer beslutninger om AI-anvendelse? Hvem godkender use-cases? Hvem stopper en pilot der går galt?
- **Data-katalog** — ved I hvad model'en må se? Hvor data-klassifikationen er? Hvad der er fortroligt vs. åbent?
- **Pilot-case identificeret** — én konkret workflow med målbart outcome. Ikke "vi vil bruge AI til at blive bedre".
- **Champion-team** — 5-10 superbrugere der bruger modellen dagligt og kan dele praksis videre
- **Vane-træning** — det er ikke et license-rollout, det er et adfærds-skift. Plan en 90-dages adoption-rytme.
- **ROI-måling** — definér KPI'er *før* I starter. Tid sparet, fejl reduceret, kunde-CSAT, revenue per ansat.

STRATEGI: Pilot én model på én use-case først. Ikke alle modeller på alle cases. Den hurtigste vej til at lære er at få første pilot i drift på 4 uger og samle data — ikke at lave en 6-måneders evaluering der bliver forældet før den er færdig.

Tips til enterprise rollout

STRATEGI: Test begge modeller. De er gode til forskellige ting. En enterprise-AI-stack er ikke "vi valgte X" — det er "vi har Opus 4.7 til kode, GPT-5.5 til agentik, og Copilot til alle de andre". Multi-model er det nye normal.

STRATEGI: Bind ikke kontrakter på multi-år. Frontier-cyklussen er nu 6-8 uger. En 3-årig forpligtelse til én leverandør gør jer fastlåste på en model der er to generationer bagud før kontrakten udløber. Forhandl 12-måneders fleksibilitet ind.

STRATEGI: Hold roadmap-buffer. Mythos kommer. GPT-6 kommer. Planlæg jeres 2026-strategi så modellen kan udskiftes uden at hele jeres workflow-arkitektur skal genopbygges. Abstraktioner over leverandører er nu en disciplin, ikke en luksus.

TIP: Start med Microsoft 365 Copilot hvis I er usikre. Lavest adoptions-friktion, accepterer-resistance, og de fleste medarbejdere har allerede licensen. Det er ikke det stærkeste valg på rå performance, men det er det stærkeste valg på rollout-hastighed.

TIP: Brug data-residens som filter, ikke som hindring. Mange organisationer afviser US-baserede modeller pga. compliance — men både Opus 4.7 og GPT-5.5 kan køres i EU-regioner via Bedrock/ Vertex/Foundry. Tjek leverandøren før I afviser modellen.

TIP: Måling før rollout, ikke efter. Hvis I ikke kan svare på "hvad er 1 times sparet tid værd for denne rolle?" inden I starter, kan I heller ikke svare på "var det det værd?" bagefter. Tag baseline-tal nu.

Quick reference: Links

Hvad	Link
GPT-5.5 announcement	openai.com/index/introducing-gpt-5-5
Claude Opus 4.7 announcement	anthropic.com/news/claude-opus-4-7
Terminal-Bench 2.0 leaderboard	tbench.ai/leaderboard/terminal-bench/2.0
SWE-Bench Pro leaderboard	labs.scale.com/leaderboard/swe_bench_pro_public
Microsoft 365 Copilot model picker	m365.cloud.microsoft/chat
Anthropic API pricing	anthropic.com/pricing