Stora språkmodeller

State of the Art, 2026-04-20

Dr. Hannes Ebner
Metasolutions AB

ENTRYSCAPE

meta solutions

Begrepp

Token

Minsta byggstenen — ord, delar av ord eller tecken. En modell ser tokens, inte text.

Testa själv: https://platform.openai.com/tokenizer

Kontext

Modellens arbetsminne per anrop. Mäts i tokens (4k ⟶ 1M+).

Modell

Neuralt nätverk tränat på text. “Dense” vs. “Mixture of Experts”.

Harness

Verktyg som kopplar modellen till omvärlden — API, RAG, agenter.

System Prompt

Dolda instruktioner som styr modellens beteende och roll.

"Soul"

Anthropics term: modellens personlighet, värderingar och gränser.

Closed / Open Weights / Open Source

Closed Source

GPT, Claude, Gemini, Mistral

- Ingen tillgång till vikter
- API-only, vendor lock-in
- Ofta bäst resultat

Open Weights

Llama, DeepSeek, Qwen, Mistral

- Vikter tillgängliga
- Begränsad licens, ej data
- Snabbt ikapp kommersiella

Open Source

APERTUS, OLMo, BLOOM

- Vikter + data + kod
- Full reproducerbarhet
- Ofta ett steg efter frontier

Obs: Apache 2.0 på Open Weights innebär inte med automatik Open Source

Europeisk suverän AI

Apertus Schweiz	EPFL + ETH Zürich. 8B & 70B, Apache 2.0, 1800+ språk, full öppenhet inkl. träningsdata.
EuroLLM-22B EU-konsortium	Edinburgh, Paris-Saclay, Amsterdam m.fl. Alla 24 EU-språk. Tränad på MareNostrum 5.
Poro / GPT-SW3 Finland / Sverige	Silo AI + AI Sweden. Nordiska språk, Apache 2.0. Fokus på nordiska språk som annars underrepresenteras i stora modeller.
Minerva Italien	50/50 italienska-engelska. Egen tokenizer — ej Llama-baserad. Full dokumentation.
Dragon LLM Frankrike	Hybrid-arkitektur, 1/3 av beräkningskraften vs. transformers. Tränad på EuroHPC.

Gemensam drivkraft: digital suveränitet, EU AI Act-compliance, flerspråkighet — inte att slå GPT, Gemini, Opus, etc på benchmarks.

“State of the Art”

	Opus 4.7	Opus 4.6	GPT-5.4	Gemini 3.1 Pro	Mythos Preview
Agentic coding SWE-bench Pro	64.3%	53.4%	57.7%	54.2%	**77.8%**
Agentic coding SWE-bench Verified	87.6%	80.8%	—	80.6%	**93.9%**
Agentic terminal coding Terminal-Bench 2.0	69.4%	65.4%	75.1% self-reported harness	68.5%	**82.0%**
Multidisciplinary reasoning Humanity's Last Exam	46.9% no tools	40.0% no tools	42.7% no tools (Pro)	44.4% no tools	**56.8%** no tools
	54.7% with tools	53.3% with tools	58.7% with tools (Pro)	51.4% with tools	**64.7%** with tools
Agentic search BrowseComp	79.3%	83.7%	**89.3%** Pro	85.9%	86.9%
Scaled tool use MCP-Atlas	**77.3%**	75.8%	68.1%	73.9%	—
Agentic computer use OSWorld-Verified	78.0%	72.7%	75.0%	—	**79.6%**

meta solutions

Benchmarks - kan vi lita på dem?

Kontaminering

Benchmarkdata läcker in i träningsdata → artificiellt höga poäng.

Goodharts lag

"När ett mått blir ett mål, slutar det vara ett bra mått."

Fuskande modeller

Anthropic upptäckte att Opus aktivt försökte detektera om den körde en benchmark - och anpassade beteendet.

Vibes vs. benchmarks

Praktisk nytta korrelerar dåligt med leaderboard-placeringar. Privata evals behövs.

SWE-bench verified & Terminal Bench

Varför dessa är avgörande för IT-organisationer?

SWE-Bench Verified

- Riktiga GitHub-issues med verifierade tester
- Mäter: kan modellen fixa riktiga buggar i riktiga kodrepon?
- Bästa agenter löser >70% - jämfört med ~5% för ett år sedan
- Direkt relevant: "kan AI ersätta en junior dev?"

Terminal-Bench

- DevOps/SRE-uppgifter: k8s, Docker, CI/CD, debugging
- Mäter: kan en agent lösa infra-problem i en riktig terminal?
- Topmodeller når >40% - fortfarande svårt
- Relevant: "kan AI avlasta ops-teamet?"

FrontierSWE & Humanity’s Last Exam

Varför SWE-Bench & Terminal-Bench inte räcker

FrontierSWE

- SWE-Bench mättar - topmodeller >70%
- FrontierSWE: svårare issues, arkitekturbeslut
- Testar: senior-nivå ingenjörsarbete
- Bara Opus & GPT gör framsteg - övriga modeller knappt mätbara

Humanity's Last Exam

- 3000 extremt svåra frågor från 1000+ experter
- Fält: matematik, fysik, juridik, lingvistik m.fl.
- Mäter: nås "super human" intelligens snart?
- Bästa modeller: >50%, inte alls samma top 3 som för FrontierSWE

BullshitBench

Hur vet man att en modell inte bara håller med dig?

Problemet: Sycophancy

Modeller tenderar att hålla med användaren - även när användaren har fel.
I professionella sammanhang kan det vara direkt farligt.

- Säger modellen emot när du har objektivt fel?
- Mäter resistens mot socialt tryck, ledande frågor, falska premisser
- Avgörande för: kodgranskning, juridisk analys, medicinsk rådgivning
- Stor variation mellan modeller: vissa sviktar vid minsta tryck

MRCR, GraphWalks & Context Rot

Hur vet man vilken modell som verkligen klarar stora kontexter?

MRCR

Multi-Round Coreference Resolution - testar om modellen hittar nålar i en höstack över flera varv av dialog. Avslöjar om 200k eller ännu större kontext verkligen fungerar.

GraphWalks

Modellen måste följa relationer i en graf över lång kontext. Testar strukturell förståelse, inte bara textsökning.

Context Rot

Prestanda sjunker ju längre kontexten blir — men hur mycket? Mäter degraderingskurvan: "effektiv kontext" vs. annonserad.

Vart är vi på väg?

- Reasoning-modeller + agenter = 2025/2026:s stora skift
- Öppna modeller närmar sig frontier - men closed leder ännu
- Europa bygger suveräna alternativ - Apertus, EuroLLM, Poro
- Benchmarks mättar snabbt - "vibes" och privata evals avgör

Del 2: Hur Metasolutions tillämpar AI

Från teknikstack till konversation

Hittills

- Användare behöver förstå länkade data, RDF, SPARQL
- Portalbaserade sökgränssnitt
- Teknisk expertis krävdes

AI-stöd

- Ställ frågor i naturligt språk
- Modellen väljer rätt API - REST, SPARQL, etc
- Information hittas genom standardisering

EntryStore MCP

MCP-server för vår länkade data-plattform - ger AI-agenter direkt åtkomst till metadata och data via REST-API eller SPARQL, utan att användaren behöver kunna tekniken

Metadata som AI-bränsle

AI kan bara vara lika bra som den metadata den har att arbeta med

Högkvalitativa metadata

AI behöver strukturerad, konsekvent och rik information för att navigera, filtrera och dra slutsatser. Utan den famlar modellen i blindo.

Etablerade specifikationer

Standarder som DCAT-AP och schema.org gör data tolkbara utan specialanpassning.
Se dataportal.se/specifications för svenska exempel.

Snabbare till analys

Med standardiserade metadata och MCP kan man gå direkt från fråga till insikt

Tack!

hannes@metasolutions.se

ENTRYSCAPE

meta solutions