Utvärdering av AI-Robotar

Förra veckan lanserade Google sin nya AI, eller snarare sin nya stora språkmodell, Gemini. Modellen Gemini 1.0 finns i tre versioner, Gemini Nano ska vara bäst lämpad för uppgifter på en specifik enhet, Gemini Pro ska vara bästa alternativet för ett bredare urval uppgifter, och Gemini Ultra är Googles största språkmodell som ska klara av de mest komplexa uppgifter du kan ge den.

Något som Google var noga med att lyfta fram vid lanseringen av Gemini Ultra var att språkmodellen var bättre än den senaste versionen av OpenAIs GPT-4 i 30 av de 32 mest använda testerna för att mäta språkmodellernas förmågor. Testerna täcker allt från läsförståelse och olika matematiska frågor till att skriva kod till Python och bildanalys. I vissa av testerna var det endast några tiondels procentenheter som skilde mellan de två AI-modellerna, medan det i andra fall rörde sig om uppemot tio procentenheters skillnad.

TEXT

Capability	Benchmark Higher is better	Description	Gemini Ultra	GPT-4 API numbers calculated where reported numbers were missing
General	MMLU	Representation of questions in 57 subjects (incl. STEM, humanities, and others)	**90.0%** CoT@32*	86.4% 5-shot* (reported)
Reasoning	Big-Bench Hard	Diverse set of challenging tasks requiring multi-step reasoning	**83.6%** 3-shot	83.1% 3-shot (API)
	DROP	Reading comprehension (F1 Score)	**82.4** Variable shots	80.9 3-shot (reported)
	HellaSwag	Commonsense reasoning for everyday tasks	87.8% 10-shot*	**95.3%** 10-shot* (reported)
Math	GSM8K	Basic arithmetic manipulations (incl. Grade School math problems)	**94.4%** maj1@32	92.0% 5-shot CoT (reported)
	MATH	Challenging math problems (incl. algebra, geometry, pre-calculus, and others)	**53.2%** 4-shot	52.9% 4-shot (API)
Code	HumanEval	Python code generation	**74.4%** 0-shot (IT)*	67.0% 0-shot* (reported)
	Natural2Code	Python code generation. New held out dataset HumanEval-like, not leaked on the web	**74.9%** 0-shot	73.9% 0-shot (API)

* See the technical report for details on performance with other methodologies

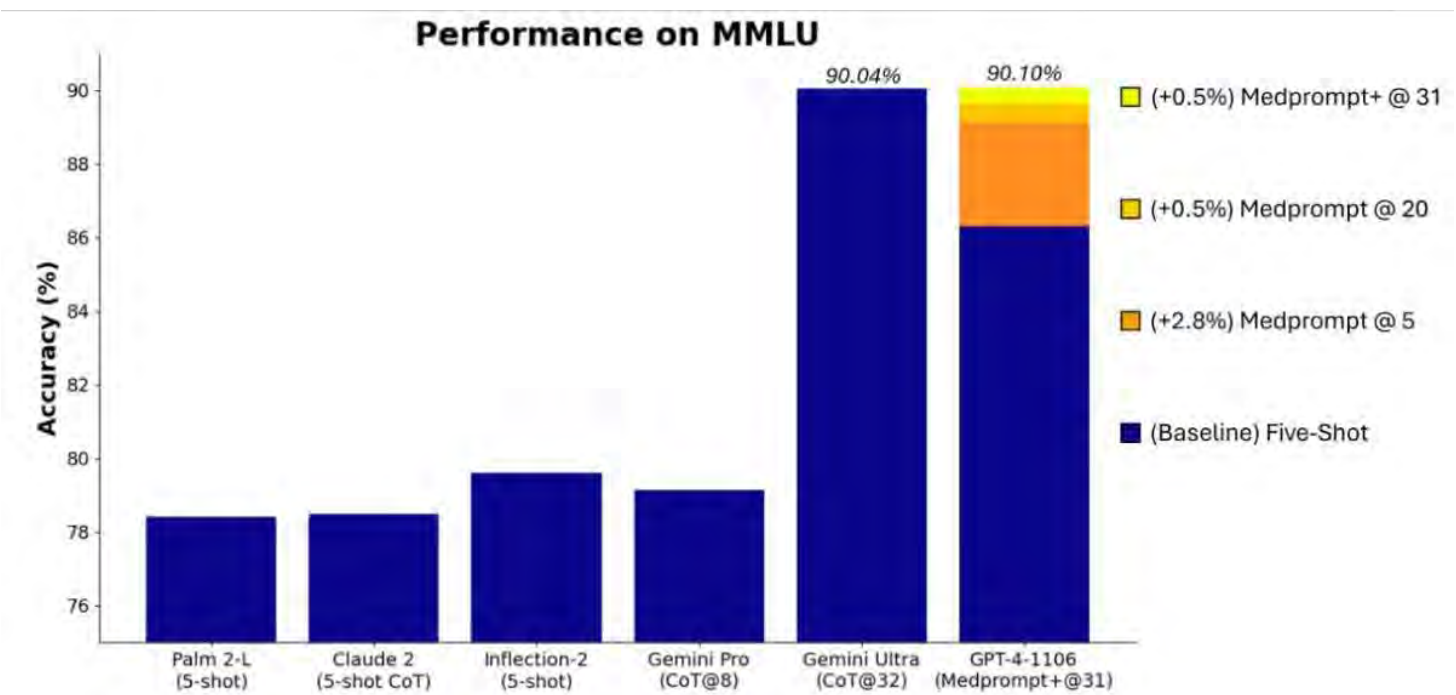
Google

Gemini Ultras mest imponerande prestation är nog dock att den är den första språkmodellen som kan klå mänskliga experter i MMLU-tester (massive multitask

language understanding) där Gemini Ultra och experterna ställdes inför problemlösningsuppgifter inom 57 olika fält, allt från matematik och fysik till medicin, juridik och etik. Gemini Ultra lyckades få ett resultat på 90,0 procent, medan den mänskliga experten AIn jämfördes mot ”endast” nådde upp i 89,8 procent.

Lanseringen av Gemini kommer ske stegvis, redan förra veckan blev Gemini Pro tillgänglig för offentligheten, då Googles chattrobot Bard börjat använda sig av en modifierad version av språkmodellen, och Gemini Nano finns inbyggd i ett antal olika funktioner på Pixel 8 Pro. Gemini Ultra är inte redo för allmänheten ännu, säger Google, den genomgår fortfarande säkerhetstester och delas endast med en handfull utvecklare och partners, samt experter inom ansvar och säkerhet kring AI. Tanken är dock att Gemini Ultra ska bli tillgänglig för allmänheten via Bard Advanced när den lanseras tidigt nästa år.

Microsoft har nu kontrat Googles påståenden om att Gemini Ultra kan slå GPT-4 genom att låta GPT-4 åter göra samma tester, fast denna gång med något modifierade prompter, eller inmatningar. Microsofts forskare publicerade i november forskning om något de kallade Medprompt, en blandning av olika strategier för att mata in prompter i språkmodellen för att få bättre resultat. Du kanske själv har märkt av hur svaren du får ut ur ChatGPT eller bilderna du får ut ur Bings bildskapare blir något annorlunda när du ändrar lite i formuleringen, lite så, fast mycket mer avancerat är tanken bakom Medprompt.



Microsoft

Genom att använda Medprompt lyckades Microsoft få GPT-4 att prestera bättre än

Gemini Ultra på en rad av de 30 testerna Google tidigare lyft fram, även MMLU-testet, där GPT-4 med Medprompt-inmatningar lyckades få ett resultat på 90,10 procent. Vilken språkmodell som kommer att dominera framöver återstår att se, kampen om AI-tronen är långt ifrån avgjord.

Av: Kristian Kask