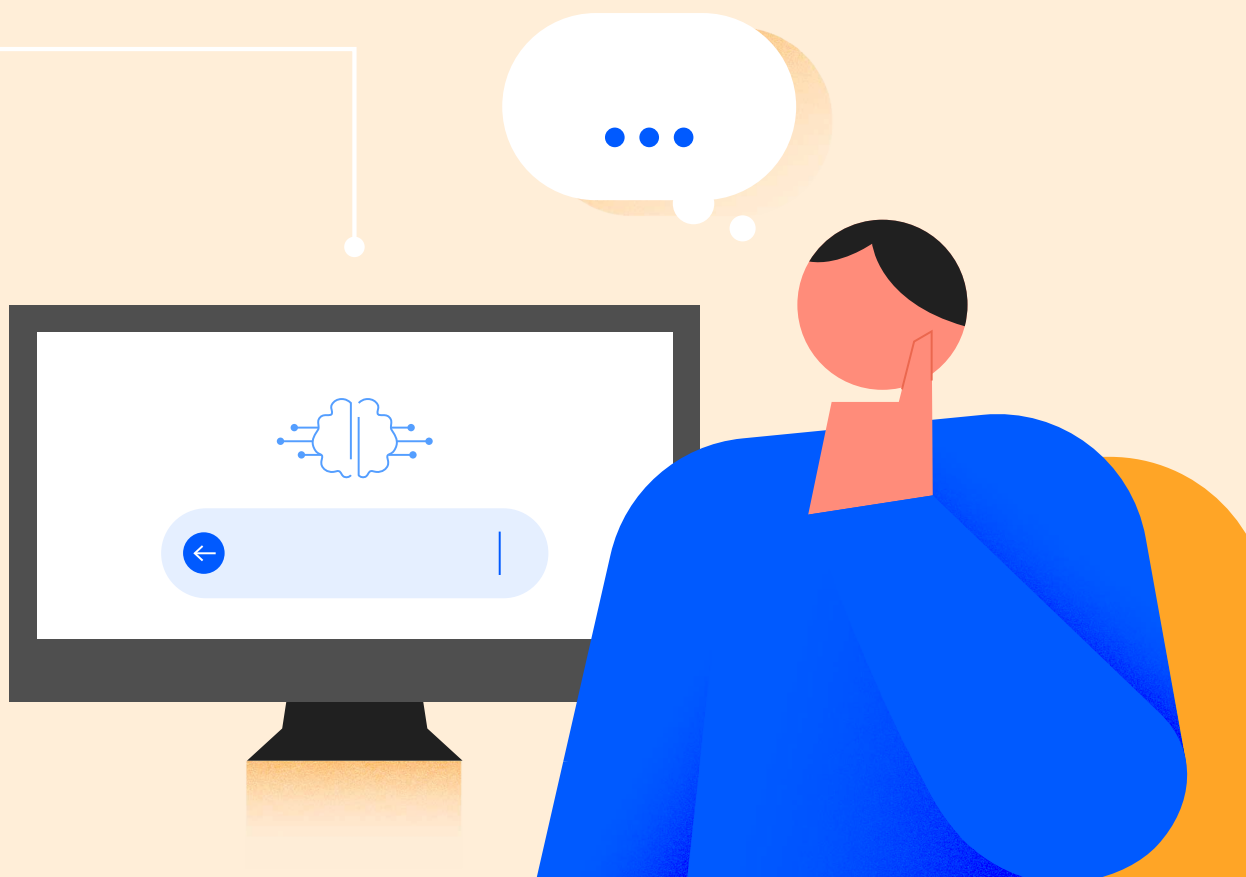


TECHSPLAINER

Hvad er generativ AI?



Hvad er generativ AI?

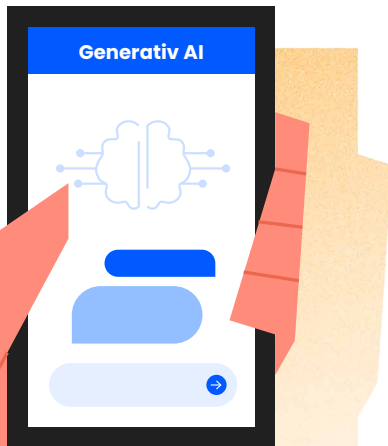
Generativ AI (GenAI) er en særlig type kunstig intelligens, som er kendetegnet ved, at den kan skabe originalt indhold af høj kvalitet, for eksempel tekst, billeder eller lyd. Typisk kræver det blot et simpelt input fra brugeren, et såkaldt "prompt", at få en GenAI til at skabe indhold.

Prompt

Et simpelt brugerinput, i form af en besked eller et spørgsmål, som GenAI anvender til at generere indhold.

Fire almindelige former for GenAI er **tekst-generation**, **billedgeneration**, **lydgeneration** og **videogeneration**.

Tekstgeneration betyder, at GenAI skaber tekst. Det kan blandt andet bruges til at søge informationer, til at opsummere pointer i en anden tekst, til at foreslå ideer, til at lave korrekturlæsning, og til at skrive hele tekster i den genre og med det emne, som brugeren ønsker. Nogle kendte eksempler er **Anthropics Claude**, **Googles Gemini**, **Microsofts CoPilot**, og **OpenAIs ChatGPT**.



"Generativ kunstig intelligens (AI) refererer til en gren af kunstig intelligens, der omhandler design og implementering af algoritmer, som kan generere nyt indhold. Dette indhold kan variere fra tekst, billeder, musik, til video, der ofte er af en sådan kvalitet, at det kunne antages at være skabt af mennesker..."

Definition genereret af ChatGPT 4.0

Billedgeneration betyder, at GenAI skaber billeder. Brugeren kan beskrive det ønskede indhold med detaljer som genre, vinkel og stemning. Billederne kan både være fotorealistiske, og imitere malerier og tegninger. Selvom GenAI fortsat kan lave fejl er resultaterne ofte så gode, at de er svære at skelne fra ægte fotografier eller menneskeskabt kunst. Nogle kendte eksempler er **Googles Imagen**, **Midjourney**, **OpenAIs DALL-E**, og **Stability AIs Stable Diffusion**.

Lydgeneration betyder, at GenAI skaber lyd. Det kan anvendes til at skabe musik, men bruges især til at lave computerskabt tale. Brugeren skriver en tekst, og kan bede om en bestemt følelse eller udtale. GenAI kan både lave originale stemmer, og trænes til at imitere virkelige personers stemme. Nogle kendte eksempler er **ElevenLab** og **15.ai**.

Videogeneration minder om billedgeneration, men opgaven er mere krævende, fordi den kunstige intelligens skal skabe en naturtro udvikling mellem de enkelte billeder. Video skabt med GenAI er fortsat ofte fejlbehæftet og let at genkende. Nogle kendte eksempler er **OpenAIs Sora**, **Metas Make-a-video** og **Runways Gen-3**.



Billeder genereret med DALL-E 3 med følgende prompt: "A modern-day middle-age academic working on her computer at her desk, writing a report on generative AI. The office environment is spacious and cozy, with bookshelves holding both books and art in the background. The style is a grainy black and white photo / a Japanese manga / an impressionist painting / a stick-figure pencil drawing."



Hvad er forskellen på generativ AI og andre former for kunstig intelligens?

Generativ AI (GenAI) er en særlig form for kunstig intelligens, men kunstig intelligens kan fungere på andre måder. De fleste af os møder mange forskellige slags kunstig intelligens i vores hverdag. Det kan for eksempel være:

- **En AI-agent**, som agerer modspiller i et computerspil, f. eks. skak.
- **Spamfiltre**, som sorterer e-mails.
- **Talegenkendelse**, som transskriberer tale til tekst, og for eksempel gør det muligt at diktere en besked til sin telefon.
- **Ansigtsgenkendelse**, som lader computeren identificere en person, og eksempelvis gør det muligt at låse en mobiltelefon op, ved at holde telefonens kamera foran sit ansigt.
- **Anbefalingsfiltre på medieplatforme**, som tilpasser indholdet til den enkelte bruger.
- **Autopilot**, som lader en bil køre selv.

Hvad karakteriserer disse forskellige eksempler på kunstig intelligens?

Der er ikke fuldstændig enighed om, hvornår et computersystem bør kaldes for en kunstig intelligens, men en bred og let forståelig definition af kunstig intelligens er, at et computersystem er en kunstig intelligens, hvis det løser en kompleks opgave, som det ellers ville kræve menneskelig intelligens at løse.¹

GenAI er ifølge denne definition de særlige former for kunstig intelligens, som løser opgaver med at skabe indhold, som det normalt ville kræve menneskelig intelligens at skabe.

¹ Se (McCarthy 2007; Wang 2019; Bringsjord and Govindarajulu 2018).

Hvordan udvikler man generativ AI?

Kernen i alle former for kunstig intelligens er en matematisk model. Når man udvikler en kunstig intelligens, skal man bestemme hvordan denne model skal se ud. Der findes grundlæggende to metoder: regelbaseret udvikling og maskinlæring.

Regelbaseret udvikling foregår ved, at mennesker bestemmer hvordan modellen skal se ud. Programmører kan for eksempel samarbejde med skakeksperter om at udvikle en kunstig intelligens til at spille skak, ved at formulere matematiske regler for, hvordan man skal evaluere brikernes position på brættet, og vælge et træk. Regelbaseret udvikling var den mest almindelige metode i det 20. århundrede.

Maskinlæring foregår ved, at en læringsalgoritme "træner" den kunstige intelligens. Læringsalgoritmen bruger et sæt relevante data til at beregne, hvordan den skal designe den matematiske model, så den kunstige intelligens bedst kan løse sin opgave. Maskinlæring gør det muligt at lave meget avanceret kunstig intelligens, som lærer af store datasæt. Til gengæld bliver det ofte svært for mennesker at gennemskue, hvordan kunstig intelligens udviklet med maskinlæring virker. Maskinlæring er blevet en stadig mere udbredt måde at udvikle kunstig intelligens de seneste årtier.

GenAI udvikles med maskinlæring, og modellens kvalitet afhænger derfor af både mængden og kvaliteten af de træningsdata, som den kunstige intelligens skal lære af. Udviklere taler om, at de mest avancerede GenAI-modeller trænes på "en væsentlig bid af internettet". Det inkluderer ikke blot skønlitteratur og faglitteratur, men også avisartikler,

Kunstig intelligens

Et computersystem, som er i stand til at løse komplekse opgaver, det normalt ville kræve menneskelig intelligens at løse.

Maskinlæring

En metode til at udvikle kunstig intelligens, ved at træne på store mængder data.

Generativ AI

En maskinlært kunstig intelligens, som med simple brugerinput kan skabe originalt indhold af høj kvalitet.

hjemmesider, og indlæg på sociale medier og online-fora. OpenAIs GPT4.0 er angiveligt trænet på tekster svarende til rundt regnet 100 millioner bøger.²

Træning af så avanceret kunstig intelligens kræver enorm regnekraft. Det bruger meget energi, og tager ofte dage eller uger, selv når det udføres på specialiserede supercomputere.

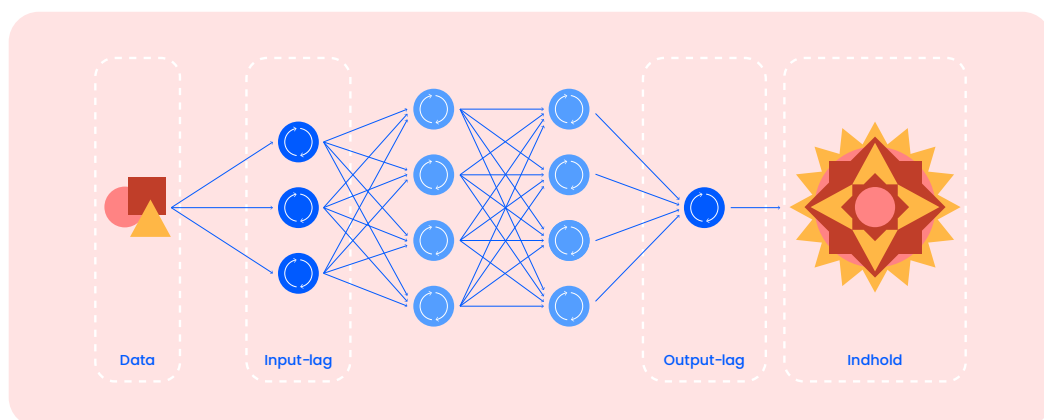


2 Se (Walker 2023; George 2023).

Hvordan virker generativ AI?

En central del af mange former for GenAI er en såkaldt "stor sprogmodel" (eng. "large language model"), forkortet LLM.³ En LLM er en matematisk model, som viser hvordan sproglige udtryk statistisk hænger sammen. Modellen arbejder med et ordforråd, der ikke er hverken hele ord eller individuelle tegn, men sproglige brudstykker, der kan minde om stavelser. De kaldes for "tokens". For at generere tekst vurderer en LLM, hvad den mest velfungerende fortsættelse af en given tekst er, og tilføjer en token ad gangen, indtil den har løst den overordnede opgave.

De matematiske modeller, som anvendes i GenAI, er en type kunstig intelligens, som kaldes for "dybe neurale netværk". Modellen kaldes for et neuralt netværk, fordi dens struktur imiterer den måde, neuronerne i den menneskelige hjerne er forbundet, og udveksler information. Et neuralt netværk er dybt, når der er mange lag af digitale neuroner, som er forbundet med hinanden.



GenAI hører til de dybeste, mest komplekse neurale netværk, mennesker har udviklet. Den frit tilgængelige model "Llama", som er udviklet af Meta, er et neuralt netværk, som i den mest avancerede version har 90.000.000.000 forbindelser mellem de digitale neuroner. OpenAIs GPT 4 model er ikke frit tilgængelig, men anslås til at være

³ Store sprogmodeller (LLM) er én variant af såkaldte grundlagsmodeller (eng. "foundation models") eller grænsem modeller (eng. "frontier models"). Se (Jones et al. 2023; Bommasani et al. 2021). GenAI-systemer anvender ofte LLM i kombination med andre typer GenAI-modeller, for eksempel såkaldte diffusionsmodeller til billedgeneration. Selvom mange pointer gælder for alle de forskellige typer GenAI, fokuserer vi her især på LLM.

endnu større, med cirka 1,7 trillioner forbindelser.⁴ Sådanne modeller er så komplekse, at de er uigennemskuelige for mennesker – selv udviklerne kan ikke i detaljer forstå, hvordan et prompt får modellen til at skabe indhold.

Kontekst og opmærksomhed

GenAIs evne til at skabe naturlig tekst ved at forudsige, hvordan teksten skal fortsætte, minder om den "auto-complete"-funktion, som har været en del af besked-tjenester på mobiltelefoner i mange år. Når brugeren skriver en besked, så foreslår systemet forskellige ord, som kunne være det næste ord i beskeden. Auto-complete gør det nemmere at skrive beskeder, men har den begrænsning, at det ofte foreslår de forkerte ord. Det skyldes blandt andet, at algoritmen kigger på en begrænset del af den sproglige kontekst. Den mest simple form for auto-complete kigger kun på det ene ord, som brugeren lige har skrevet, og foreslår ganske enkelt de ord, som hyppigst kommer efter netop dette ord. Mere avancerede versioner kigger på lidt mere af den sproglige kontekst, som ordet optræder i, og bliver derved bedre til at gætte på, hvad det næste ord skal være.

Hvorfor bruger auto-complete ikke hele konteksten? Det korte svar er, at det hurtigt bliver for svært at beregne, hvad den bedste fortsættelse er, når den sproglige kontekst bliver større og større, hvis man kigger på hele teksten.

Den store succes, som GenAI har nydt i de seneste år, skyldes et teknisk gennembrud i 2017, som løste denne udfordring: transformer-arkitekturen.⁵ Transformeren bygges ind i den kunstige intelligens, og består af to dele: en indkoder og en afkoder. Indkoderen analyserer den sproglige kontekst for information om tekstens indhold og struktur. Afkoderen bruger denne information til, at fokusere modellens opmærksomhed på de relevante dele af den sproglige kontekst, når den skal generere indhold. Det gør det muligt for GenAI at skabe indhold, som langt mere præcist svarer på den prompt, brugeren har lavet.⁶

Hallucinationer, temperatur og stokastiske papegøjer

GenAI er blevet så avanceret, at det som bruger nogle gange kan føles som om, at man har med et ægte tænkende væsen at gøre. Men det er vigtigt at holde sig for øje,

⁴ Se (Walker 2023).

⁵ Se (Vaswani 2017).

⁶ Se (Alammar 2018)

at en LLM ikke forstår den tekst den skaber. Den skaber tekst ved at gentage mønstre fra de tekster den er trænet på. Skridt for skridt genererer modellen en af de tokens, som den statistisk har lært, vil være velegnede fortsættelser.

Fordi en LLM ikke forstår hvad en tekst betyder, kan den heller ikke skelne mellem, om en tekst er faktisk korrekt eller faktisk forkert. En LLM kan generere velformuleret og meningsfuld tekst, som er fuld af fejlagtig information. Det kaldes, at GenAI "hallucinerer".

Hallucination

GenAI indhold, som er meningsfuldt og tilsyneladende troværdigt, men faktisk forkert.

Når en LLM skal vælge hvilken token den skal generere, vil der ofte være flere gode muligheder at vælge mellem, for eksempel fordi teksten kan anvende synonyme ord, fordi teksten kan præsentere pointer i forskellig rækkefølge, eller fordi teksten kan fortsætte i indholdsmæssigt forskellige retninger. I mange tilfælde er det en fordel, at en GenAI genererer tekst med et vist element af tilfældighed, fordi den derved får større sproglig fleksibilitet og variation. I stedet for at modellen altid genererer præcis den token, som er den bedst egnede, kan den vælge tilfældigt mellem de bedste tokens. GenAI-modeller har derfor normalt en indstilling, der styrer hvor meget tilfældighed modellen skal introducere – et såkaldt hyperparameter for "temperatur". Hvis temperaturen skrues højt op, kan to indtastninger af det samme prompt føre til skabelse af meget forskelligt indhold. Hvis temperaturen omvendt sættes lavt, så vil det samme prompt altid give indhold som er næsten eller komplet identisk. Kort sagt: jo højere temperaturen er indstillet, jo mere kreativ og varieret er GenAI, når den skaber indhold.

Kombinationen af, at GenAI reproducerer sproglige mønstre, som den ikke forstår, og at GenAI i en vis grad generer indholdet tilfældigt, har fået kritikere til at karakterisere LLM'er som "stokastiske papegøjer".⁷

Fintuning og indholdshegn

En LLM er en statistisk repræsentation af menneskeligt sprog. Det er på engang modellerens styrke og svaghed, at de ikke er designet til at løse bestemte opgaver, men kan være grundlag for løsningen af mange forskellige typer opgaver. Ofte skal modellen justeres, hvis den skal være god til at løse en mere specialiseret opgave. Det kaldes "fintuning".

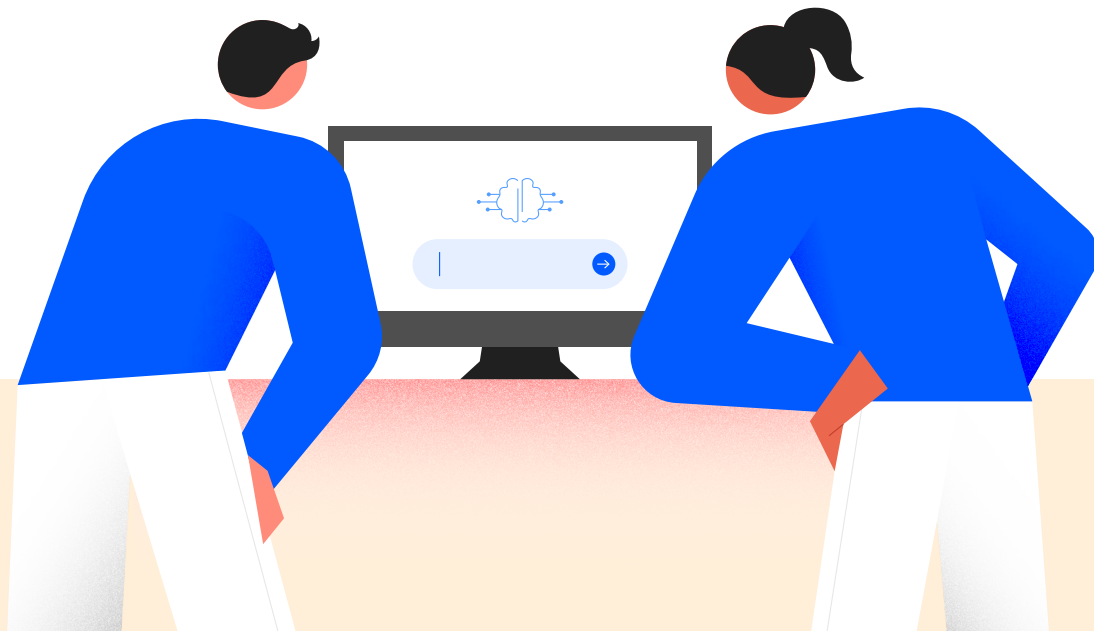
⁷ Se (Bender et al. 2021) Men se også (Arkoudas 2023).

Fintuning foregår normalt ved, at udvikleren træner modellen på et datasæt med eksempler på den type opgave, som modellen skal løse. Hvis eksempelvis en myndighed ønsker at udvikle en chatbot, som skal hjælpe borgere med at finde information, så kunne datasættet indeholde eksempler på spørgsmål, som borgere plejer at stille, og de svar, som menneskelige medarbejdere plejer at give. Fintuning træner modellen til at efterligne disse eksempler, når den genererer indhold.

Fintuning

Yderligere træning af en allerede udviklet GenAI, for at tilpasse modellen til en bestemt anvendelse.

Fintuning kan også bruges til at bygge såkaldte "indholdshegn" (eng. "guardrails") ind i modellen.⁸ GenAI skelner ikke af sig selv mellem forskellige typer indhold – modellen forsøger at generere det bedste svar på brugerens prompt. GenAI kan derfor komme til at genere indhold, som er upassende eller skadeligt. En LLM kunne eksempelvis generere racistisk hadtale, sexistisk chikane, eller grafisk seksuelt og voldeligt indhold. Indholdshegn er en teknisk begrænsning, som forhindrer GenAI i at generere upassende eller skadeligt indhold.



⁸ Indholdshegn kan bygges på flere måder. Det er almindeligt at kombinere fintuning med indbyggede men skjulte udvikler-prompts, såkaldt "prompt-engineering", hvor udvikleren automatisk tilføjer instrukser til alle prompts fra brugeren, eller indholdsfiltere, som kan analysere og afvise både prompts og det genererede indhold. Se (Ayyamperumal and Ge 2024; Dong et al. 2024; Conitzer and Leben 2024).

Indholdshegn bygges ofte ved, at udvikleren skaber en database med eksempler på uønsket indhold, for eksempel racistisk hadtale, og træner modellen til at undgå at lave indhold, der minder om disse eksempler. Det kalder man "Reinforcement-Learning from Human Feedback" (RLHF).

Mange af de store, kommercielle GenAI-systemer er allerede fintunet, med generelle adfærdsmønstre for chatbots, og indholdshegn, som forhindrer generation af upassende og skadeligt indhold.

RAG-modeller

En LLM er typisk trænet på store datasæt med et bredt udvalg af forskellige tekster. Det giver modellen en grundlæggende forståelse af menneskeligt sprog, men betyder, at modellen kan mangle specialiserede data, som er relevante i en særlig kontekst. For eksempel kan en virksomhed, der ønsker at bruge GenAI i en chatbot, have brug for, at modellen har adgang til specifik og aktuel viden om virksomhedens produkter. Det kan imidlertid være dyrt og tidskrævende løbende at fintune modellen, så den har adgang til disse data.

En udbredt løsning på denne udfordring er at bruge såkaldt "Retrieval-Augmented Generation", også kendt som RAG-modeller. En RAG-model er en GenAI, som kan søge i en database med relevante og opdaterede informationer. Når modellen modtager en prompt, tilføjer modellen relevante oplysninger fra databasen til prompten. Det gør det muligt for modellen at generere tekst baseret på denne udvidede og specialiserede information.⁹

RAG-model

En GenAI model, som kan trække på en database med relevant, specialiseret viden, når den skal generere indhold.

RAG-modeller reducerer risikoen for, at en LLM skaber fejlbehæftet tekst, og er typisk hurtigere og billigere at anvende end fintuning.

⁹ Se (Gao et al. 2023; Gupta et al. 2024).

Hvilke dataetiske udfordringer rejser generativ AI?

GenAI er fleksibelt og nemt at bruge. Mange har nok allerede oplevet, at teknologien kan være nyttig både til privat brug og professionelt. Men teknologien rejser også en række dataetiske udfordringer, som det er vigtigt at forholde sig till.¹⁰

Udfordringer i anvendelsen af GenAI	Udfordringer med potentiale for misbrug af GenAI	Udfordringer ved træning og bredere effekter af GenAI
Fejl og hallucinationer	Misinformation	Brud på intellektuel ejendomsret
Algoritmisk bias	Deep fakes	Miljøpåvirkning
Læk af private data	Manipulation	Udbytende arbejdspraksisser
Diffus ansvarlighed	Spredning af farlig information	Forstærkede tech-oligopoler
Uigennemsigtighed	Forklædning af AI-genereret indhold	Forskydninger på arbejdsmarkedet
		Eksistentielle risici ved AGI
		Mistillid

Det er vigtigt at forholde sig til disse dataetiske udfordringer, så vi sikrer, at vores anvendelse af den nye teknologi respekterer menneskers selvbestemmelse, gavner i

¹⁰ For overblik, se (Janjeva et al. 2023; Myrstad 2023; Weidinger et al. 2021; Bommasani, Hudson et al. 2021; Lorenz et al. 2023; Marchal et al. 2024; Fang and Perkins 2024).

stedet for at skade personer, og ikke skaber eller forstærker uretfærdige rammer for vores fælles liv. Dataetisk Råds arbejde med dataetiske udfordringer, tager udgangspunkt i rådets dataetiske principper.¹¹ Rådet har i 2024 behandlet udfordringen med risiko for fejl og hallucinationer når GenAI anvendes af offentlige myndigheder.¹²



¹¹ <https://dataetiskraad.dk/dataetiske-vaerktoejer-&-principper>

¹² <https://dataetiskraad.dk/dataetiske-temaer/generativ-AI>

Litteratur

Alammar, J. (2018). The Illustrated Transformer

Arkoudas, K. (2023). "ChatGPT is no Stochastic Parrot. But it also Claims that 1 is Greater than 1." *Philosophy & Technology* 36(3): 54 DOI: 10.1007/s13347-023-00619-6.

Ayyamperumal, S. G. and Ge, L. (2024). "Current state of LLM Risks and AI Guardrails." arXiv preprint arXiv:2406.12934.

Bender, E. M., Gebru, T., et al. (2021). On the dangers of stochastic parrots: Can language models be too big?? *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*.

Bommasani, R., Hudson, D. A., et al. (2021). "On the opportunities and risks of foundation models." arXiv preprint arXiv:2108.07258.

Bringsjord, S. and Govindarajulu, N. S. (2018). Artificial Intelligence. *Stanford Encyclopedia of Philosophy*. Edward N. Zalta

Conitzer, V. and Leben, D. (2024). How ChatGPT has been prompted to respect safety fairness, and copyright. *AI Ethics at Oxford Blog*

Dong, Y., Mu, R., et al. (2024). "Building guardrails for large language models." arXiv preprint arXiv:2402.01822.

Fang, A. and Perkins, J. M. (2024). "Large language models (LLMs): Risks and policy implications." *MIT Science Policy Review*.

Gao, Y., Xiong, Y., et al. (2023). "Retrieval-augmented generation for large language models: A survey." arXiv preprint arXiv:2312.10997.

George, A. (2023) "Visualizing the size of Large Language Models." Medium, <https://medium.com/@georgeanil/visualizing-size-of-large-language-models-ec576caa5557>.

Gupta, S., Ranjan, R., et al. (2024). "A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape and Future Directions." arXiv preprint arXiv:2410.12837.

Janjeva, A., Harris, A., et al. (2023). The Rapid Rise of Generative AI: Assessing risks to safety and security. London, UK, The Alan Turing Institute

Jones, E., Dunlop, C., et al. (2023). What is a foundation model? London, UK, Ada Lovelace Institute

Lorenz, P., Perset, K., et al. (2023). Initial Policy Considerations for Generative Artificial Intelligence. Paris, France, OECD

Marchal, N., Xu, R., et al. (2024). "Generative AI Misuse: A Taxonomy of Tactics and Insights from Real-World Data." arXiv preprint arXiv:2406.13843.

McCarthy, J. (2007). What is artificial intelligence?

Myrstad, F. L.-H. (2023). Ghost in the Machine - Addressing the Consumer Harms of Generative AI. Oslo, Norway, Forbrukerrådet Norge

Vaswani, A. (2017). "Attention is all you need." Advances in Neural Information Processing Systems.

Walker, S. M. (2023) "The Next Era of AI: Inside the Breakthrough GPT-4 Model." Hackernoon, https://hackernoon.com/the-next-era-of-ai-inside-the-breakthrough-gpt-4-model?utm_source=chatgpt.com.

Wang, P. (2019). "On defining artificial intelligence." Journal of Artificial General Intelligence 10(2): 1-37.

Weidinger, L., Mellor, J., et al. (2021). "Ethical and social risks of harm from language models." arXiv preprint arXiv:2112.04359

Dataetisk Råd 2025

Dataetisk Råd er en offentlig myndighed ved Nationalt Center for Etik, som rådgiver og skaber debat om etisk udvikling og anvendelse af digitale løsninger, data og kunstig intelligens.

København, 16. Januar 2025