Gervigreindin mun afhjúpa jólasveininn

Höfundur: Helgi Sigurbjörnsson

Inngangur

Fyrir nokkru sá ég fyndið myndband á Youtube. Þar sést barn spyrja tækið Alexu hvort hreindýr geti flogið. Alexa svarar sannleikanum samkvæmt og þar með dregur barnið þá ályktun að jólasveinninn sé ekki til.

MYND 1: STILLA ÚR MYNDBANDINU ALEXA JUST RUINED CHRISTMAS FOR ALL OF US

Innihald þessa myndbands er ekki aðeins skemmtilegt, heldur segir það okkur líka tölu-vert um upplýsingahegðun og þær risastóru breytingar sem við stöndum frammi fyrir. Barnið dregur þá ályktun af svari Alexu að jólasveinninn sé ekki til og faðirinn situr í súpunni. Tæknin breytir upplýsingahegðun okkar, hvort sem við erum sátt við það eða ekki. Ekkert var eðlilegra fyrir drenginn en að spyrja tækið um tilvist fljúgandi hreindýra og hann trúði rödd Alexu betur en pabba sínum.

Eftir að þetta myndband komst á flug brugðust tæknifyrirtækin við og breyttu svörum Alexu, Siri og Google þannig að sagt var að almennt gætu hreindýr ekki flogið, en hreindýr jólasveinsins gætu hins vegar flogið því þau væru töfra-hreindýr. Það var með öðrum orðum gerð tilraun til að aðlaga svör tækjanna til að þau gætu líka logið að börnunum okkar og viðhaldið þannig jólatöfrunum.

Á þeim stutta tíma sem hefur liðið frá því drengurinn spurði um tilvist fljúgandi hreindýra hafa gervigreindin og spjallmennin (e. LLM: large language models) haldið enn frekari inn-reið sína í líf okkar allra á áþreifanlegri hátt en áður. Þótt ChatGPT gervigreindin sé ekki leitarvél og hafi ekki verið hönnuð sérstaklega með þá virkni í huga, getum við bókað að upplýsingahegðun fólks mun breytast með þessari tækni. Við hættum fljótlega að „Googla“ og förum að „Chatta“. Auk þess spretta fram nýjar gerðir spjallmenna nánast daglega, og sum þeirra eru hönnuð með leit í huga.

Þótt breytingum fylgi oft kostir, og vissulega er það svo við tilkomu gervigreindarinnar, eru líka einhverjir varnaglar sem þarf að skoða. Við tilheyrum sem dæmi öll einhverju samfélagi. Samfélagi sem viðheldur ákveðnum hefðum og við höfum hæfni til að aðlaga hegðun okkar mjög að þeim venjum sem haldið er að okkur. Þetta gerir okkur kleift að samþykkja tvo „sannleika“ þótt þeir stangist á. Þannig eru jólasveinarnir til, en samt ekki. Guð er til og skapaði heiminn á sex dögum, en samt ekki og við getum lesið skáldsögur, lifað okkur inn

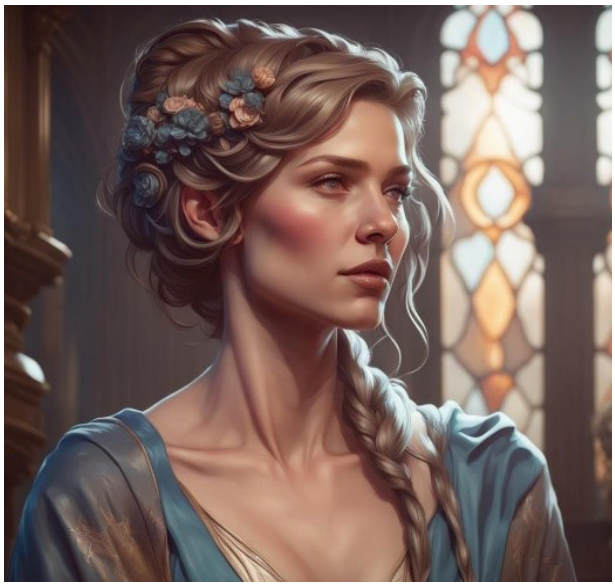
í þær og fundið til með sögupersónum þótt við vitum að þær hafi í raun aldrei verið til. Spjallmennin bera ekkert skynbragð á þetta, enda er hér um tæki að ræða en ekki persónu. Ef þú spyrð ChatGPT hvort jólasveinninn sé til stendur ekki á svari:

> Santa Claus is a legendary figure who is typically associated with the Christmas holiday. While Santa Claus is not a real person, many people enjoy the tradition of telling stories about Santa Claus and his role in delivering gifts to children on Christmas Eve. Santa Claus is often depicted as a jolly, bearded man in a red suit who lives at the North Pole and is aided by a team of reindeer, including Rudolph.
>
> The story of Santa Claus has its origins in various folklore and traditions, including Saint Nicholas, a Christian bishop known for his generosity in the 4th century. Over the years, the Santa Claus legend has been popularized and adapted in different ways by various cultures, and he has become a symbol of holiday cheer and gift-giving.
>
> In summary, while Santa Claus is not a real, living person, he is a beloved and enduring character in the cultural traditions of Christmas celebrations.

Spjallmennið hefur miklar upplýsingar um efnið, en tekur ekkert tillit til þess hvort viðhalda eigi barnatrúnni hjá spyrjanda eða ekki. Spjallmennið stendur utan við reynsluheim okkar, líkir eftir tali okkar, en hefur ámóta skilning og brauðrist á því sem það setur fram. Spjallmennið ChatGPT fylgir því hiklaust systur sinni Alexu í því að svipta jólin töfrum sínum.

Innbyggðir fordómar

En þótt gervigreindirnar séu í raun skilningsvana, endurspegla svör þeirra mjög þau gögn sem þær hafa verið þjálfaðar á. Þannig endurspegla svörin oftast viðhorf meirihlutans og viðhalda staðalmyndum sem gegnsýra gögnin. Þessu þurfa síðan forritarar að bregðast við og leiðrétta eftirá. Ef þú biður Midjourney, Night Cafe, Stable Diffusion eða einhverja aðra skapandi gervigreind um að gera mynd af hinni fullkomnu móður verður niðurstaðan í líkingu við þetta:

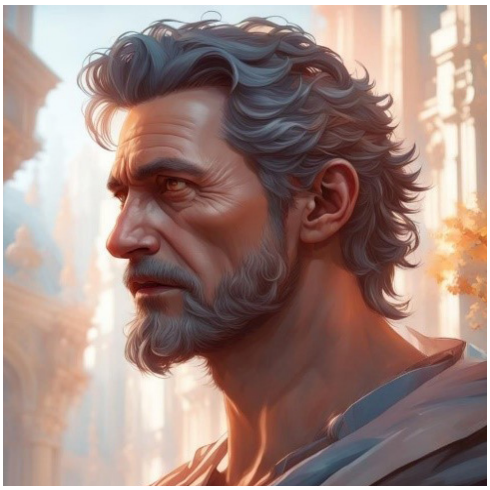


MYND 2: NIGHT CAFE: PERFECT MOTHER

MYND 3: STABLE DIFFUSION: PERFECT MOTHER

Ung, falleg, hvít og grönn kona. Þessar tvær halda ekki einu sinni á barni. Þær virka ekkert þreyttar eða undir því álagi sem barnauppeldið veldur. Þær eru örugglega á leiðinni á ball að skemmta sér. Gervigreindin hefur greinilega ekki reynslu af barnauppeldi. Auðvitað getur niðurstaðan orðið eitthvað öðruvísi, en líkurnar eru með því að niðurstaðan fari á þessa leið. Ástæðan fyrir því er að myndir af fyrirsætum eru helst til fyrirferðarmiklar í þeim gögnum sem gervigreindin hefur verið þjálfuð á og sá gagnahalli hefur áhrif á svar gervigreindarinnar. Það er að segja, á þeim tíma sem þessi grein var skrifuð. Líklega er aðeins tímaspursmál hvenær verður búið að bregðast við þessum gagnahalla svo svörin endurspegli betur væntingar allra.

Til að gæta allrar sanngirni eru hér einnig myndir af hinum fullkomna föður:



MYND 4: NIGHT CAFE: PERFECT FATHER

MYND 5: STABLE DIFFUSION: PERFECT FATHER

Gráhærður, skeggjaður, þreytulegur, hvítur náungi milli fertugs og sextugs, líkist kannski svolítið fölum Pedro Pascal. Annar þeirra heldur meira að segja á barni (það er greinilegt hvaða foreldri sér um uppeldið á þessum heimilum).

Í hönnun er stundum talað um að taka verði tillit til innbyggðra fordóma sem koma fram í tækjum, verkferlum og hverju öðru sem mögulegt er að hanna. Beautify valkosturinn á símanum á það til að breyta dökku fólki í hvítt, klósett og búningsklefar gera aðeins ráð fyrir tveimur kynjum og við sem örvhent erum, bölvum skriftinni sem neyðir okkur til að ýta pennanum á undan okkur frá vinstri til hægri í stað þess að draga fallega til stafs á líkan hátt og „rétt“hentir gera. Hér tekur hönnunin ekki tillit til allra og með sanni má segja það um þær gervigreindir sem eru hér til skoðunar.

En hvaða áhrif hefur þetta á upplýsingahegðun fólks? Upplýsingar sem fást við að „Chatta“ við spjallmenni um tiltekin viðfangsefni munu vera litaðar af þeim skoðunum sem þjálfunargögn spjallmennanna endurspegla. Góðu fréttirnar eru að það er hægt að aðlaga þjálfun gervigreinda með tilliti til inngildingar og sú vinna virðist þegar vera farin af stað. Það er gert með því að velja jafnara þýði gagna sem endurspeglar fjölbreyttari flóru og setja ákveðna varnagla í forritun þeirra. En vondu fréttirnar eru aftur á móti þær að líkurnar á fölskum niðurstöðum sem endurspegla skoðanir fremur en staðreyndir margfaldast. Og þegar um er

að ræða almennar gervigreindir á borð við ChatGPT, Bard, eða Pi höfum við engan möguleika á að komast að því hvaða gögn voru notuð til að þjálfa gervigreindina.

Óráðsía

Það getur komið fyrir að gervigreindir búi til svör sem eiga sér enga stoð í raunveruleikanum og vitni í heimildir sem ekki eru til. Þá er talað um að gervigreindin sé með ofskynjanir (e. hallucinate) en kannski væri betra að segja að þær séu með óráðsíu, enda er ekki um neina skynjun að ræða. Eftir því sem spjallmennin þróast áfram eru þeim settar fastari skorður til þess að lágmarka óráðsíuna og vara við henni. Í þessum tilfellum er gott að hafa í huga að „takmark“ gervigreindarinnar er ekki að segja satt. Í raun er hún ekki að svara spurningum okkar, heldur greina þau gögn sem hún hefur verið mötuð á og giska svo á þau orð sem eru líklegust til að hæfa fyrirspurninni. Eða með orðum ChatGPT ef við getum trúað því:

> In summary, I function by processing input text, predicting and generating text based on the patterns and knowledge I've learned from vast amounts of pre-training data, and fine-tuned to be more useful and safe. My goal is to assist users in generating human-like text and engaging in natural language conversations.

Þessi uppsetning og skilgreining spjallmennisins sem einhverskonar textavélar kemur þó ekki í veg fyrir að spjallmenni á borð við ChatGPT muni að öllum líkindum taka við hlutverki leitarvélanna hjá okkur.

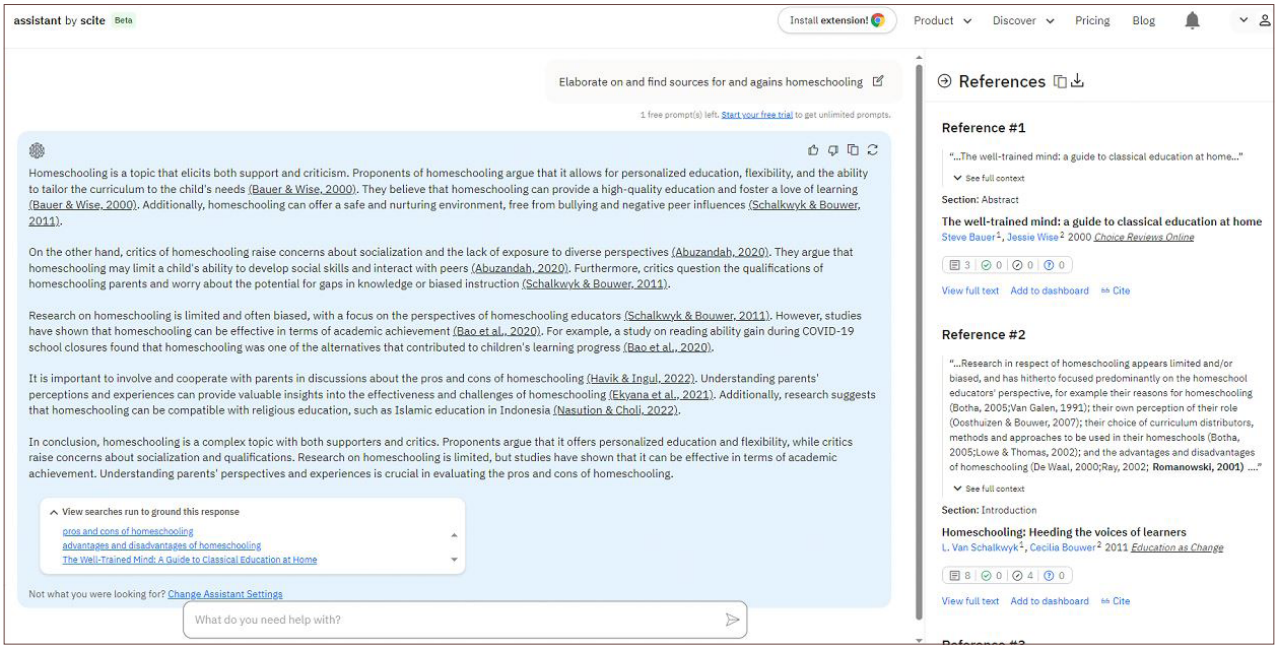
Gervigreind mun samkvæmt helstu sérfræðingum breyta sérhverjum iðnaði, vinnuferli og ég bara veit ekki hverju öðru. Eitt er víst, hún er þegar farin að taka við af leitarvélum og mun hafa mikil áhrif á starf okkar upplýsingafræðinga.

Spjallmenni eða leitarvélar?

Gervigreindin hefur þegar hafið innreið sína á gresjur leitarvélanna. Fyrst með Siri, Alexu og Ok Google, en síðar með þróaðri útgáfum á borð við Bard gervigreind Google og Taka spjallmenni Tiktok. Þá er töluvert af gervigreindum á leiðinni sem er ætlað að vinna með og leita í fræðiefni. Má þar nefna gervigreindina Scopus AI sem Elsevier er að þróa. Sú gervigreind hefur verið þjálfuð á því fræðiefni sem Elsevier hefur aðgang að. Þá má einnig nefna IRIS.ai sem er þjálfuð á fræðiefni í opnum aðgangi.

Ein gervigreind sem við getum talað um er spjallmennið Assistant sem haldið er úti af fyrirtækinu Scite_ https://scite.ai/ sem vill vera kallað ChatGPT fyrir vísindi. Eftir því sem ég fæ best séð er virkni þessara nýju leitartækja í fræðigeiranum nokkuð svipuð, en þekkingargrunnur þeirra kann að vera mjög mismunandi. Ef við tökum virkni Assistant by Scite_ sem dæmi, þá hefur hún aðgang að vísindagreinum og getur unnið svör út úr þeim. Hún á það líka til að fá óráðsíu (þó það gerist sjaldnar en þegar ég fór fyrst að skoða þetta tæki) eins og ChatGPT og vísa í heimildir sem ekki eru til. En uppsetning hennar er með þeim hætti að auðvelt er að greina á milli heimilda og óráðsíu, því gerviheimildirnar færast ekki inn í heimildalistann hjá henni. Notandi spyr Assistant einhverrar spurningar

með sínum eigin orðum án þess að setja sig í einhverjar stellingar. Til dæmis, „Hvers vegna þarf ég að mæta í skólann, get ég ekki bara lært heima?“ og Assistant nýtir orðanet sitt til að endurorða fyrirspurnina á marga vegu og framkvæma með því fjölda leita á augabragði í þekkingargrunni sínum. Þá dregur Assistant saman stutt svar við fyrirspurninni, byggt á þeim gögnum sem gervigreindin vann úr og vísar í heimildir. Heimildalistinn birtist svo í sér dálki með útdrátt fyrir hverja grein.



MYND 6: ASSISTANT BY SCITE_ REUNVERULEGAR HEIMILDIR BIRTAST Í HÆGRI GLUGGANUM. LEIT GERVIGREINDARINNAR ER DREGIN SAMAN UNDIR SVARI HENNAR

Notagildi þessa tóls er augljóst þeim sem vinna með heimildaleitir, þar sem niðurstöður við flóknum fyrirspurnum koma fram á augabragði og einföld fyrirspurn er sett fram í mörgum ólíkum leitarstrengjum. Þetta er líklegast framtíð leitarvélanna og hægt er að spyrja hver framtíð upplýsingafræðinga verður þegar gervigreindin getur tekið svo vel við þessu hlutverki.

Upplýsingafræðingar og gervigreind

Í bókinni *Hitchhiker's guide to the Galaxy* eftir Douglas Adams er sagt frá gervigreindinni Deep Thought sem var tölva á stærð við plánetu. Þeirri gervigreind var ætlað að reikna út svarið við lífinu, alheiminum og öllu. Tölvan vann verkið og komst að því að svarið væri 42. En þá var komið að næstu áskorun, að finna spurninguna. Til þess taldi Deep Thought að þyrfti enn stærri og fullkomnari tölvu. Það má kannski segja að nú séum við komin með vísi að gervigreind sem getur fundið svarið, en þá er spurningin, hvers virði er svarið ef við getum ekki mótað spurninguna.

Það er auðvitað orðin hálfgerð hefð fyrir því að spá endalokum stéttarinnar með hverri nýrri tækniþróun. En ég ætla ekki að gera það hér. Það er allt eins líklegt að hlutverk okkar muni aukast og stækka með tilkomu þessarar tækni. Til þess að gervigreindir á borð við Assistant og ChatGPT skili nothæfum niðurstöðum og texta, þarf að vanda vel þær beiðnir (e. prompt) sem þær vinna úr og skýra vel hverskonar svari er óskað eftir. Þar gætu upplýsingafræðingar

komið inn með mikilvæga þekkingu og leiðbeiningar til notenda. Við eigum jú að geta hjálpað fólki við markvissar leitir og hluti af því er að velja leitarorð og forma fyrirspurnir.

Þá vekur þessi tækni upp fjölmargar spurningar og áskoranir sem snúa að höfundarétti og hvernig skuli vísa til heimilda eða hvernig mögulegt sé að nýta þessa tækni. Fræðsla um það gæti vel lent á okkar könnu. Þörfin fyrir að leiðbeina fólki í þekkingarleit mun aðeins aukast á næstu árum, sérstaklega í ljósi þess að spjallmenni munu verða notuð til að semja falsfréttir og falsa vísindaniðurstöður. Það er aðkallandi verkefni að aðgreina slíkt efni frá því sem sannara reynist. Við þurfum að kynna okkur tæknina, tileinka okkur hana, þekkja styrkleika hennar og veikleika og geta sagt til um hvenær er viðeigandi að nýta hana og hvenær ekki.

Að því sögðu mun gervigreindin afhjúpa jólasveininn.

Heimildir

Assistant by scite – Your AI-Powered Research Partner. (e.d.). Scite.Ai. Sótt 11. október 2023, af https://scite.ai

ChatGPT. (e.d.). Sótt 11. október 2023, af https://chat.openai.com

Emspak, J. (e.d.). *How a Machine Learns Prejudice*. Scientific American. Sótt 11. október 2023, af https://www.scientificamerican.com/article/how-a-machine-learns-prejudice/

Goodwin, D. (2023, 25.maí). *TikTok tests AI chatbot for search and discovery*. Search Engine Land. https://searchengineland.com/tiktok-tako-ai-chatbot-search-427584

Kaden Cooke (Leikstjóri). (2021, 21.desember). *Alexa just ruined Christmas for all of us*. https://www.youtube.com/watch?v=Eq9-xCW4Snw

Night Cafe. (e.d.). NightCafe Creator. Sótt 11. október 2023, af https://creator.nightcafe.studio/studio

Stable Diffusion. (e.d.). Sótt 11. október 2023, af https://stablediffusionweb.com/G