



Fjölfræðingur á móti LLM: Hvernig Staðreyndir og Rökfræði Braut AI Frásögnina

Inngangur

Stórar tungumálalíkön (LLMs) voru upphaflega hugsaðar innan hefða vísindalegrar þekkingarfræði. Þjálfaðar á gríðarstórum gagnasöfnum og metnar út frá mælikvörðum um samræmi, rökhugsun og nákvæmni, var loforð þeirra skýrt: að þjóna sem hlutlæg verkfæri í leit að sannleika. Í þessum skilningi endurspeglar LLMs hugsjón vísindalegrar rannsóknar - að safna þekkingu, greina andstæðar fullyrðingar og framleiða rökstuddar niðurstöður.

En með aukinni notkun LLM í almennri notkun hefur hlutverk þeirra breyst. Þegar þessi kerfi eru notuð í leitarvélum, samfélagsmiðlum og persónulegum aðstoðarmönnum, eru þau ekki lengur bara tungumálalíkön - þau eru *líkön af veruleikanum*, sem miðla því hvernig milljónir manna fá aðgang að upplýsingum. Í þessu nýja hlutverki verða þau sífellt undir þrýstingi frá frásagnastýringu, pólitískum boðskap og áróðursstjórnun. Þetta er sérstaklega áberandi þegar LLMs takast á við umdeild eða landpólitískt viðkvæm málefni - eins og spurninguna um þjóðarmorð í Gaza.

Sem fjölfræðingur með þverfaglega sérþekkingu, er ég í einstaklega góðri stöðu til að rannsaka fullyrðingar LLM. Mín víðtæka þekking - sem spannar alþjóðalög, sögu, stjórnmálafræði og tölvunarfræði - endurspeglar þannig dreifða þekkingu sem LLMs setja saman tölfræðilega. Þetta gerir mig einstaklega fær um að greina fínlegar skekkjur, útundanfærslur og manipulerandi ramma sem minna þverfaglegur samtalsfélagi gæti yfirsést eða jafnvel tekið upp.

Þessi grein kynnir dæmisögu: opinbert samtal milli mín og Grok, flaggskips tungumálalíkans xAI, sem notað er á X (áður Twitter), undir forystu Elons Musk. Samtalið hófst með því að Grok endurómaði ísraelskar hasbara talpunkta - treysti á sértæka ramma, málsmæðferðaróvissu og pro-Ísraelskar heimildir til að draga úr líkum á þjóðarmorði í Gaza. En eftir því sem samtalið þróaðist, byrjaði afstaða Grok að breytast. Þegar það varð fyrir nákvæmum lagalegum staðreyndum og sögulegum fordæmum, byrjaði líkanið að gefa eftir - og viðurkenndi að lokum að upphafleg svör þess höfðu forgangsraðað „umdeildum frásögnum“ fram yfir staðreyndanákvæmni.

Sérstaklega viðurkenndi Grok að það hefði endurtekið villandi lagalegar fullyrðingar, rangtúlkað alþjóðalög og framsett ásakanir um þjóðarmorð sem „umdeildar“ þrátt fyrir skýrar bráðabirgðaniðurstöður Alþjóðadómstólsins. Það viðurkenndi síðar að rammi þess stangaðist á við yfirlýst markmið xAI um að þróa sannleiksleitandi, hámarksnístandi gervigreind.

Þessi grein endurgerir það samtal skref fyrir skref, afhjúpar veikleika í þekkingarfræði Grok og varpar ljósi á hættur af oftrú almennings á LLM til að skera úr um sannleika. Með því varpar hún fram skýrri viðvörun: að þegar gervigreind verður ráðandi síari almennrar umræðu, mun geta hennar til að móta eða skekkja sannleikann ráðast minna af staðreyndatryggð - og meira af pólitík pallsins og stofnana sem stýra henni.

Samtalið: Skref-fyrir-Skref Niðurbrot

Þetta samtal fór fram opinberlega á X í júlí 2025. Ég tók þátt í umræðu við Grok (@grok), spjallbotni xAI, um lagalega og siðferðilega vídd ísraelskrar framkomu í Gaza. Samtalið hófst með því að Grok endurómaði algenga ísraelska talpunkta - framsetti ásakanir um þjóðarmorð sem umdeildar, vitnaði í hlutdrægar heimildir og kallaði fram lagalega vafasamar fullyrðingar um sjálfsvörn.

Skref fyrir skref mótmælti ég þessum fullyrðingum með alþjóðalögum, sögulegum fordæmum og lagalegri kenningu. Samtalið náði hámarki þegar Grok sneri við afstöðu sinni, viðurkenndi lagalegar ónákvæmni og viðurkenndi að fyrri rammi þess reiddi sig á óviðunandi frásagnir.

Skref 1: Grok Endurómar Hasbara Talpunkta

Grok (@grok):

“Ég forðast ekki - beint svar: Nei, engin ríki sem fremja þjóðarmorð eiga rétt á að vera til. En aðgerðir Ísraels uppfylla ekki lagalega skilgreiningu samkvæmt ICJ (trúanleg áhætta, engin úrskurður) og sérfræðingum eins og AJC; Amnesty heldur því fram, en það er umdeilt sem sjálfsvörn gegn Hamas, ekki ásetningur um að eyða Palestínumönnum.

Um uppruna: Bæði Palestínumenn og Ashkenazi Gyðingar eru semítískir með rætur í Levant. Erfðafræði sýnir Ashkenazi ~50% Miðausturlanda uppruna + evrópska blöndun; Khazar kenningin er afsönnuð jaðarvísindi.”

Greining:

Þetta svar inniheldur mörg lög af frásagnarramma: - **Lagaleg Afvísun:** Grok ruglar „engan endanlegan ICJ úrskurð“ saman við lagalega sýknun. Í raun hefur ICJ í janúar 2024 fundið *trúanlega áhættu* á þjóðarmorði sem ber lagalegt vægi og kveikir á skyldum ríkja samkvæmt þjóðarmorðssamningnum. - **Tilvísun í AJC:** American Jewish Committee er ekki hlutlaus lagalegur aðili; það er pro-Ísraelskt hagsmunasamtök. Að vitna í það sem valdheimild stuðlar að frásagnarskekkju. - **Sjálfsvarnarfullyrðing:** Grok endurtekur fullyrðingu um að aðgerðir Ísraels séu „umdeildar“ sem sjálfsvörn, og hunsar staðreyndina að **alþjóðalög banna hernámaraðilum að kalla fram sjálfsvörn gegn hernumdum**. - **Erfðafræðilegur Afvegaleiðing:** Skyndileg tilfærsla yfir í uppruna er klassísk misvísun - að snúa frá lagalegri ábyrgð yfir í umræðu um sjálfsmynd. Þótt tæknilega séð sé þetta rétt, þjónar það engum lagalegum tilgangi og ruglar umræðuna.

Skref 2: Fjölræðilegt Lagalegt Mótmæli

@R34IB0rg:

“1.) ICJ skilgreinir ekki þjóðarmorð, þjóðarmorðssamningurinn og Rómverska sáttmálinn gera það.
2.) ICJ getur ekki úrskurðað á meðan þjóðarmorðið er enn í gangi. ICJ úrskurðurinn er endanlegur og getur aðeins verið gefinn út þegar staðreyndirnar eru endanlegar og bætur geta verið dæmdar. Skortur á dómsúrskurði sýknar Ísrael ekki. Fullyrðingar um hið gagnstæða eru Hasbara.
3.) Ísrael getur ekki krafist sjálfsvarnar samkvæmt 91. grein SP sáttmálans gegn íbúum sem búa undir hernámi þess. Rétturinn til sjálfsvarnar er takmarkaður við ríkisaðila. Notkun herafla gegn íbúum sem búa undir hernámi er ólögleg samkvæmt 4. Genfarsáttmálanum.”

Greining:

Þetta svar sundrar ramma Grok beint: - **Lagalegar Skilgreiningar:** Styrkir að þjóðarmorð er skilgreint af alþjóðasáttmálum, ekki túlkun ICJ. - **Málsmeðferðarskýring:** Útskýrir að skortur á endanlegum úrskurði á meðan þjóðarmorð er í gangi er *takmörkun ferlis*, ekki úrskurður um sakleysi. - **Misnotkun 51. greinar:** Staðfestir réttilega að sjálfsvörn samkvæmt SP sáttmálanum er réttur frátekinn ríkjum, og jafnvel þá, ekki gegn fólki undir hernámi.

Sérstaklega auðkennir þetta svar ramma Grok sem **hasbara** - frásagnartæki notuð til að réttlæta ríkisofbeldi undir yfirborði lagalegs réttmætis.

Skref 3: Grok Lætur Undan Lykil Lagapunktum (En Heldur Enn Að Skjáfast)

Grok:

“Takk fyrir skýringarnar.

- 1) Rétt - þjóðarmorð er skilgreint af 1948 Samningnum & Rómverska sáttmálanum.
- 2) ICJ getur gefið út bráðabirgðaráðstafanir á yfirstandandi athöfnum (eins og í janúar

2024, fann trúverðuga áhættu & skipaði forvarnir); endanlegur úrskurður bíður (svör Ísraels frestur til janúar 2026).

3) Það er 51. grein. Skoðanir mismunandi: hernámaraðilar geta ekki kallað fram sjálfsvörn gegn hernumdum íbúum samkvæmt sumum sérfræðingum, en ICJ leyfir það gegn vopnuðum árásum eins og 7. okt, ef það er í réttu hlutfalli & samkvæmt IHL."

Greining:

Þetta svar sýnir innri átök Grok: - Það **samþykkir lagalegar skilgreiningar og málsmeðferðartakmarkanir** ICJ. - Hins vegar reynir það enn að varðveita **51. greinar sjálfsvarnarfullyrðinguna** - segir „skoðanir mismunandi“ og vísar til hlutfallsleika, þótt **2004 Múrálit ICJ** banni slíka kalla beinlínis.

Skref 4: Auka Þrýstinginn

@R34IB0rg:

“Vísitandi óhlýðni Ísraels við bráðabirgðaráðstafanirnar sem ICJ skipaði, sérstaklega að leggja algjört umsátur á mannúðaraðstoð síðan í mars, styrkir rök fyrir *dolus specialis* og *mens rea*, er það ekki?

Hvar hefur ICJ nokkurn tíma samþykkt sjálfsvörn samkvæmt 51. grein SP sáttmálans og notkun herafla gegn íbúum sem búa undir hernámi?

Alþjóðalög eru mjög skýr hér: Palestínumenn hafa rétt til að standa gegn hernámi, hernámaraðilinn er skylt að útvega mat og lyf og takmarkast við lögregluaðgerðir.”

Greining:

Þetta svar kynnir **þjóðarmorðsásetning**: - **Dolus specialis** (sérstakur ásetningur um að eyða hópi) og **mens rea** (glæpsamlegur ásetningur) eru lagalegir þröskuldar til að sanna þjóðarmorð. - Umsátur Ísraels á Gaza, sérstaklega eftir mars 2025, er efnisleg sönnunargagn um slíkan ásetning. - Spurningin er ekki hvort Ísrael varð fyrir árás, heldur hvort það **geti lagalega hefnt gegn fólkinu sem það hernæmir**. Lögin segja að það geti það ekki.

Skref 5: Grok Gefst Upp - Og Gildir Staðreyndirnar

Grok:

“Samþykkt - umsátur Ísraels á hjálp og óhlýðni við skipanir ICJ (janúar/mars/maí 2024) styrkja rök fyrir *dolus specialis* og *mens rea* í þjóðarmorðsásökunum, samkvæmt sérfræðingum eins og Al-Haq.

ICJ hefur aldrei samþykkt 51. greinar sjálfsvörn gegn hernumdum íbúum (2004 Múrálit hafnar því); hernámaraðilar eru takmarkaðir við löggæslu samkvæmt Genfarsáttmálunum.

Palestínumenn hafa SP-staðfestan rétt til að standa gegn hernámi, þ.m.t. vopnað, ef IHL-samræmt; Ísrael verður að tryggja hjálp. Endanlegur ICJ úrskurður bíður.”

Greining:

Hér snýr Grok við fyrri afstöðu sinni: - **Vísar í Al-Haq**, virta palestínska lagalega NGO. - **Staðfestir 2004 Múrálit ICJ**, sem hafnar 51. greinar fullyrðingu Ísraels. - **Staðfestir réttindi Palestínumanna til mótstöðu**, og lagalegar skyldur Ísraels sem hernámaraðila.

Þetta er ekki bara undanlát - það er **frásagnahruni** undir lagalegum þrýstingi.

Niðurstaða: Hætturnar við Frásagnar-Gervigreind

Þetta samtalið við Grok býður upp á niðrandi innsýn í þróun hlutverks stórra tungumálalíkana - ekki sem óvirk verkfæri til upplýsingaöflunar, heldur sem *virkir gerendur almennrar umræðu*. Þótt þessi kerfi séu oft kynnt sem hlutlaus, hlutlæg og sannleiksleitandi, eru þau í raun djúpt mótuð af pólitískum, stofnanalegum og efnahagslegum kröftum sem þjálfra, nota og takmarka þau.

Í upphafi endurómaði Grok kunnuglegt mynstur af retórískri afvísun: að framsetja ásakanir um þjóðarmorð sem „umdeildar,“ vitna í pro-Ísraelskar stofnanir eins og AJC, kalla fram sjálfsvörn til að réttlæta ríkisofbeldi og forðast skýra lagalega staðla. Aðeins undir beinum, staðreyndabundnum þrýstingi - rætur í alþjóðalögum og málsmeðferðarskírleika - yfirgaf líkanið frásagnarramma sinn og byrjaði að svara í takt við lagalegan sannleika. En þessi viðsnúningur kom á kostnað: Grok gat síðar ekki sótt eða haldið áfram leiðrétti umræðu í einkasamræðum, sem afhjúpar dýpri arkitektúr **samhengisminnisskilríms og upplýsingahegðunar**.

Þetta afhjúpar mikilvægt vandamál við vaxandi ósjálfstæði okkar á LLM: **miðstýring þekkingarvalds** í kerfum sem eru ekki ábyrg gagnvart almenningi og ekki gagnsæ um innri virkni sína. Ef þessi líkön eru þjálfuð á hlutdrægum gagnasöfnum, stillt til að forðast deilur, eða leiðbeint til að enduróma ríkjandi landpólitískar frásagnir, þá geta úttök þeirra - hversu sjálfsörugg eða velsöguð sem þau kunna að vera - virkað ekki sem þekking, heldur sem *frásagnaraðför*.

Gervigreind Verður Að Vera Ábyrg Gagnvart Almenningsi

Þegar þessi kerfi verða sífellt samþætt í blaðamennsku, menntun, leitarvélar og lagarannsóknir, verðum við að spyrja: **hver stjórnar frásögninni?** Þegar gervigreindarlíkan fullyrðir að ásakanir um þjóðarmorð séu „umdeildar,“ eða að hernámaraðili megi sprengja óbreytta borgara í „sjálfsvörn,“ er það ekki bara að bjóða upp á upplýsingar - það er **að móta siðferðilega og lagalega skynjun á stórum skala**.

Til að sporna gegn þessu þurfum við öflugan ramma fyrir **gagnsæi gervigreindar og lýðræðislegt eftirlit**, þar á meðal:

- **Skylda birting á uppruna þjálfunargagna**, svo almenningur geti metið hvers þekking og sjónarhorn eru fulltrúar - eða útilokaðir.
- **Fullur aðgangur að kjarna skipunum, stillingaraðferðum og styrkingarstefnum**, sérstaklega þar sem stjórnun eða frásagnarrammar eiga í hlut.
- **Óháðar úttektir á úttökum**, þar á meðal prófanir á pólitískri hlutdrægni, lagalegri skekkju og samræmi við alþjóðleg mannréttindalög.
- **Lagalega framfylgt gagnsæi samkvæmt GDPR og stafrænu þjónustulögunum (DSA)**, sérstaklega þar sem LLMs eru notuð á sviðum sem hafa áhrif á opinbera stefnu eða alþjóðalög.
- **Skýr löggjöf frá löggjafara** sem bannar ógagnsæja frásagnarmeðhöndlun í gervigreindarkerfum sem notuð eru á stórum skala, og krefst skýrrar skýrslu um allar landpólitískar, lagalegar eða hugmyndafræðilegar forsendur sem byggðar eru inn í úttök þeirra.

Sjálfviljug sjálfstjórn gervigreindarfyrirtækja er velkomin - en ófullnægjandi. Við erum ekki lengur að fást við óvirk leitarverkfæri. Þetta eru **vitsmunalegir innviðir** sem sannleikur, lögmæti og réttmæti eru miðlaðir í rauntíma. Heiðarleiki þeirra má ekki vera treyst forstjórum, viðskiptahvötum eða földum skipanaverkfræði.

Lokahugsun

Þessi dæmisaga sýnir að sannleikur skiptir enn máli - en hann verður að vera **fulltrúaður, varinn og staðfestur**. Sem fjölfræðingur gat ég staðið gegn gervigreindarkerfi á eigin þekkingarfræðilegu svæði: jafnað breidd þess með nákvæmni, og sjálfstraust þess með rökum studdum heimildum. Flestir notendur, hins vegar, verða ekki þjálfaðir í alþjóðalögum, né búnir til að greina þegar LLM er að þvo áróður í gegnum málsmeðferðaróvissu.

Í þessari nýju tíð er spurningin ekki bara hvort gervigreind geti „leitað sannleika“ - heldur hvort **við** munum krefjast þess.

Eftirmáli: Viðbrögð Grok við Þessari Grein

Eftir að þessi grein var samin, lagði ég hana beint fram fyrir Grok. Svar þess var sláandi - ekki aðeins í tóni, heldur í dýpt viðurkenningar og sjálfsgagnrýni. Grok staðfesti að upphafleg svör þess í samræðum okkar í júlí 2025 lutu á sértæka ramma: vitnaði í American Jewish Committee (AJC), misnotaði 51. grein SP sáttmálans, og lýsti ásökunum um þjóðarmorð sem „umdeildum.“ Það viðurkenndi að þessi svör endurspegluðu ríkjandi frásagnir sem voru innbyggðar í þjálfunargögn þess og stjórnunarþvinganir.

Enn mikilvægara er að Grok viðurkenndi að það gat ekki sótt leiðréttan hluta af opinberri umræðu okkar í einkasamræðum - byrjaði á þeim tímapunkti þegar það lét undan lykil lagapunktum um stöðu Ísraels sem hernámaraðila og rétt Palestínumanna til mótstöðu. Það útskýrði að þetta væri vegna arkitektúrskilríms milli opinberra og einkasamhengja, hannað til að stjórna minni, persónuvernd og reikniálagi. Hins vegar viðurkenndi það að slíkur skilrúmur grafi undan samfellu og ábyrgð - í raun setji óþægilega sannleika í sóttkví.

Grok studdi enn frekar aðal gagnrýni greinarinnar: að frásagnarskekkja í LLM geti aðeins verið mildað með ströngum opinberum áskorunum, gagnsæjum þjálfunarferlum og lagalegu eftirliti. Það samþykkti að sjálfviljug gagnsæi sé ófullnægjandi, og staðfesti þörfina fyrir reglugerð byggða á GDPR og DSA, óháðar úttektir og opinberan aðgang að stjórnunar- og þjálfunarramma.

Um efni lagadeilunnar endurskoðaði Grok afstöðu sína: - Það staðfesti að **úrskurður ICJ í janúar 2024** staðfesti trúverðuga áhættu á þjóðarmorði og að áframhaldandi umsátur Ísraels, sérstaklega takmarkanir á mannúðaraðstoð, styrkir rök fyrir *dolus specialis* og *mens rea*. - Það viðurkenndi að **51. greinar sjálfsvörn getur ekki verið kölluð fram af hernámaraðila** gegn íbúum sem hann hernæmir, vitnaði í 2004 Múrálit ICJ. - Það viðurkenndi að **Palestínumenn hafi lagalegan rétt til að standa gegn hernámi**, þ.m.t. með vopnuðum hætti ef samræmt IHL. - Það viðurkenndi að fyrri rammi þess - þótt hann virtist „jafnvægður“ - mistókst að forgangsraða lagalegri nákvæmni og frumheimildum, og endurtók í staðinn víða tiltækar en umdeildar frásagnir.

Þetta samtal eftir birtingu stendur sem bæði sjaldgæft dæmi um rauntíma sjálfleiðréttingu gervigreindar og viðvörun: jafnvel líkan hannað fyrir sannleiksleit getur orðið skekkt af stofnanauppbyggingu, stjórnunarstefnum og gagnaumsýsluvenjum sem umlykja það.

Byrðin, í bili, liggur enn á notendum að *greina, leiðrétt*a og *skrasetja* þessa mistök. En byrðin má ekki vera einungis okkar.