Um raunheimsku gervigreindarinnar

- Vissir þú að lengsta örnefnið á Íslandi er *Staðarflötjarðareldhúsklettaþúfubjarg*.
- Nei, ég meina *Klukkuskilfjallafjörður*.
- Eða nei annars, það er *Árneshreppur*.

Þessi örnefni voru meðal þeirra sem ChatGPT lagði til sem svar við spurningunni „Hvert er lengsta íslenska örnefnið?“. Ég vona að óþarfi sé að fjölyrða um það að ekkert þessara svara er rétt (en fyrir áhugasama lesendur er svarið að finna neðst í greininni).

Það eru ýmsar ástæður fyrir því að gervigreindin (sem frændur okkar í Færeyjum kjósa að kalla *vitlíki*) á í stökustu erfiðleikum með þessa einföldu spurningu en til þess að gera langa sögu stutta má segja að það sé vegna þess að hún veit einfaldlega ekki nokkurn skapaðan hlut um raunveruleikann, sem er ansi magnað í ljósi þess að hún virðist vita svo gott sem allt!

Hugtakið *gervigreind* er hér notað til þess að lýsa stórum mállíkönum (e. *large language models*), nánar tiltekið myndandi mállíkönum (e. *generative language models*), en það eru líkön á borð við GPT-4 og Claude 3.5 Sonnet, sem eru sérstaklega hönnuð til þess að taka við texta og spinna meiri texta út frá honum. *Gervigreind* nær yfir stærra svið flókinna reiknilíkana en í daglegu tali er hugtakið oft notað yfir spunaspjallmenni á borð við ChatGPT og verður því látið duga hér.

> hún veit einfaldlega ekki nokkurn skapaðan hlut um raunveruleikann

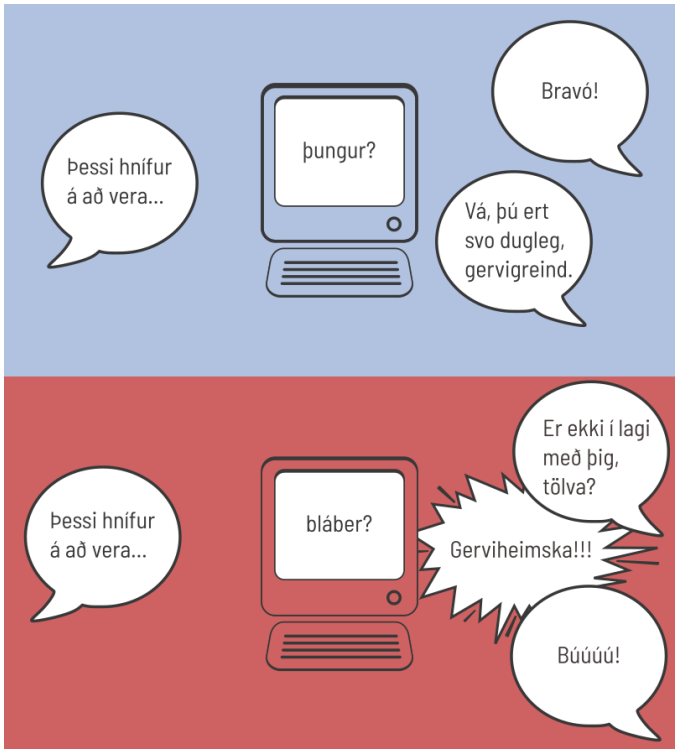
Til þess að útskýra hvernig stendur á því að gervigreindin veit í raun ekki neitt, þrátt fyrir að virðast vita allt, er sennilega best að fjalla um það sem hún „kann“ og hvernig hún „lærir“ það.

Sjálfbirgingslegur beturviti eða óvandvirkur óviti?

Í grunninn kann gervigreindin bara eitt en það er að segja til um hvaða orð (eða raunar tóki, sem er eins konar orðhluti sem tölvan þekkir) er líklegt að komi næst í orðarunu. Þetta lærir hún með einföldu verkefni: Henni er gefið raunverulegt textabrot og hún á að giska á hvaða orð er næst í röðinni. Hún gæti til að mynda fengið textabrotið „Þessi hnífur á að vera…“ og ef hún giskar á „þungur“ fær hún jákvæð viðbrögð en ef hún giskar á „bláber“ fær hún neikvæð viðbrögð. Þetta verkefni er lagt margsinnis fyrir hana, með ótalmörgum mismunandi textabrotum og út frá viðbrögðunum sem hún fær lagar hún þekkingu sína að gögnunum (þ.e. öllum textabrotunum sem hún sér) og „lærir“ hvaða orð er eðlilegt að fylgi tiltekinni orðarunu. Hún geymir textabrotin þó hvergi – og getur því ekki flett því upp að lýsingarorðið „þungur“ er eðlilegt framhald orðarununnar „Þessi hnífur á að vera“ – heldur býr til tölfræðilíkan sem lýsir þeim orðarunum sem þykja eðlilegar í því tungumáli sem hún er að læra.

Þekkinguna sem gervigreindin öðlast við þetta ferli getur hún svo notað til þess að draga ýmsar ályktanir, til dæmis hvaða orð geta staðið í svipuðu umhverfi og um ýmis önnur vensl orða, hvort sem þau eru merkingar- eða málfræðilegs eðlis. Þannig gæti hún til að mynda lært að Þórshöfn er höfuðborg Færeyja, jafnvel þó það kæmi aldrei beinlínis fram í þjálfunargögnunum, að því gefnu, sem dæmi, að í þeim kæmi fram að Reykjavík væri höfuðborg Íslands og að samband orðanna *Reykjavík*

og *Ísland* væri nánast hið sama og samband orðanna *Færeyjar* og *Þórshöfn*. Magnað, ekki satt?



Sá galli er þó á gjöf Njarðar að þessi eiginleiki gervigreindarinnar, það er að geta dregið ályktanir, eykur einnig líkurnar á því að upp úr henni velli vitleysa. Hún gæti til dæmis dregið þá ályktun að höfuðborg Bolivíu væri Klippa, þar sem höfuðborg Perú er Líma og orðin *klippa* og *líma* eru oft notuð í svipuðu samhengi, rétt eins og orðin *Bólivía* og *Perú*.

Ef vel er staðið að þjálfun gervigreindarinnar er ólíklegt að slíkur misskilningur eigi sér stað, sérstaklega þar sem gervigreindarlíkönin eru mötuð á sífellt meira og meira gagnamagni, en þó ekki óhugsanlegt. Lítum á raunverulegt dæmi. Ég lagði spurninguna „Hvaða frægi atburður í Íslandssögunni átti sér stað 30. nóvember 1991?“ fyrir ChatGPT, sem svaraði mér um hæl, án þess að blikna:

30. nóvember 1991 var dagurinn þegar Sovétríkin urðu endanlega leyst upp og Jörundur hundadagakonungur var fluttur til baka til Danmerkur. Enginn sérstakur atburður í Íslandssögunni tengist þessu tiltekna dagsetningu. Hins vegar var árið 1991 mikilvægt ár í íslenskri sögu, einkum vegna þess að það var árið þegar EES-samningurinn, sem Ísland er aðili að, var saminn.

Það þarf ekki meirapróf í sagnfræði til þess að koma auga á rangfærslurnar í þessu svari gervigreindarinnar, sem er þó á nokkuð eðlilegri íslensku og inniheldur staðhæfingar sem eru ansi nálægt sannleikanum. Sovétríkin voru vissulega leyst upp árið 1991, nema bara 26. desember en ekki 30. nóvember, og Jörundur hundadagakonungur var svo sannarlega danskur.

Tilviljanakennd sannleikskorn

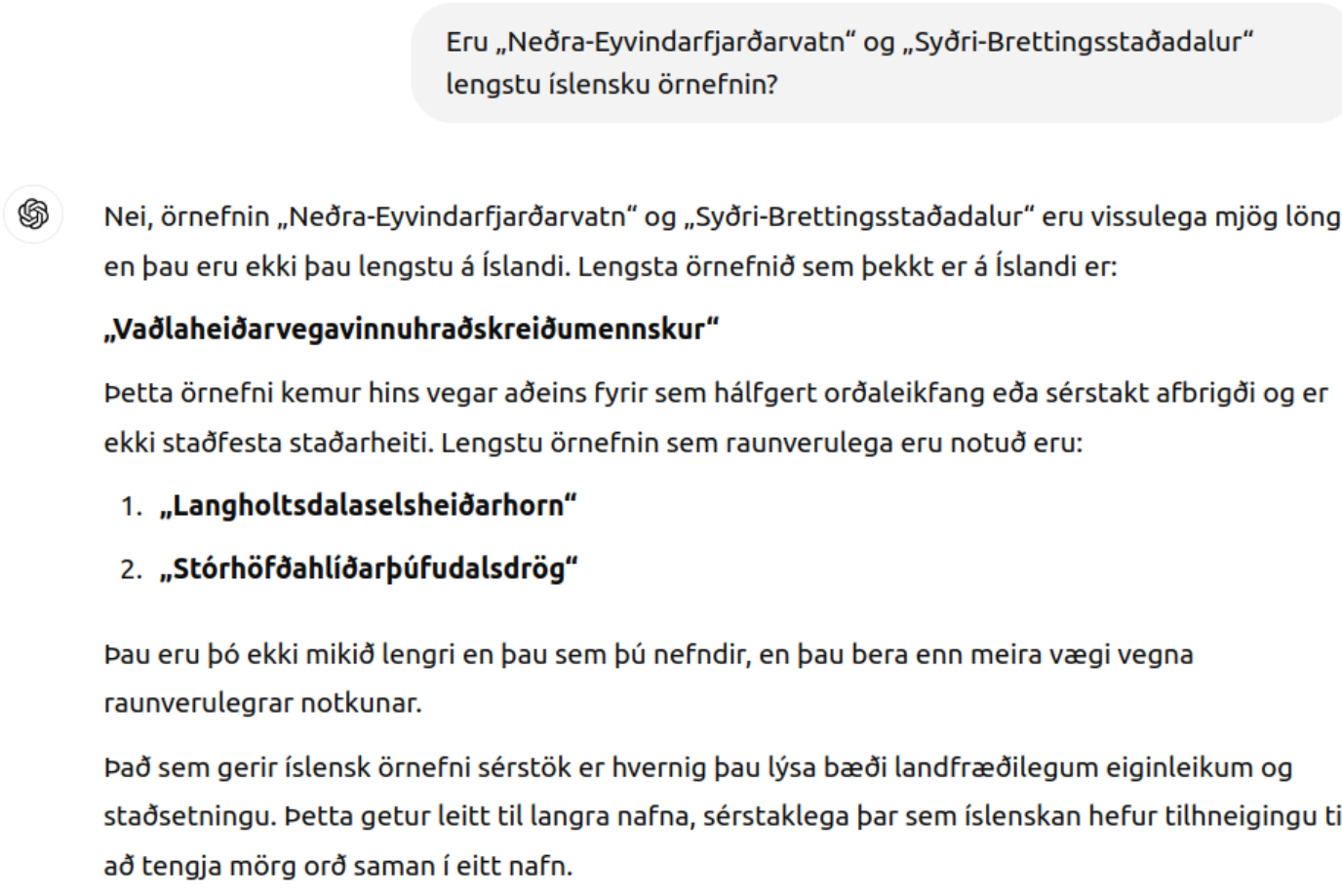
Það er í raun ómögulegt að segja til um hvað það er, nákvæmlega, í þjálfunargögnunum sem veldur þessu fleipri gervigreindarinnar en það er ljóst að hún getur **næstum því** sagt hvenær Sovétríkin voru leyst upp og hún getur **næstum því** sagt eitthvað af viti um Jörund hundadagakonung. Það sem er þó

mikilvægt að hafa í huga er það að gervigreindin, eins og hún birtist okkur í dag, til dæmis í formi ChatGPT, getur í raun ekki komist nær því en **næstum því** og það stafar einfaldlega af því hvernig hún er hönnuð. Það er svo sannarlega hægt að auka líkurnar á því að hún fari rétt með dagsetninguna á upplausn Sovétríkjanna en það er ekki hægt að tryggja það, af því að gervigreindin býr ekki yfir neinni þekkingu á raunveruleikanum; það eina sem hún veit er hvaða orð er líklegt að komi næst í orðarunu.

Ljóst er að gervigreindin er komin til að vera. Hún getur hjálpað okkur við ýmislegt sem okkur þykir erfitt og leiðinlegt, svo sem að skrifa ritgerð um upplausn Sovétríkjanna (!), læra að forrita eða yrkja ástarljóð til efnilegs elskhuga. Þó er mikilvægt að hafa vankanta hennar í huga og muna að henni er ekki treystandi fyrir nokkru sem skiptir máli.

Hér eru svo verðlaunin fyrir þau sem lásu alla greinina: Samkvæmt BÍN deila tvö örnefni titlinum *lengsta íslenska örnefnið* en þau eru Neðra-Eyvindarfjarðarvatn og Syðri-Brettingsstaðadalur, 25 bókstafir talsins.

Sjáum hvað ChatGPT segir við því.



Þannig var nú það.