

TARTU ÜLIKOOL
Arvutiteaduse instituut
Informaatika õppekava

Juhan Pauklin

Terviseandmetel treenitud keelemudelist kontseptsioonide eraldamine

Bakalaureusetöö (9 EAP)

Juhendaja Raivo Kolde, PhD

Tartu 2025

Terviseandmetel treenitud keelemudelist kontseptsioonide eraldamine

Lühikokkuvõte:

Keelemudelid on suutelised täitma suure hulga erinevaid ülesandeid, kuid see, kuidas nad oma tulemuseni jõuavad on nagu must kast - kasutaja annab sisendi ja saab väljundi, kuid seda, kuidas väljundini jõuti, ei teata. Kui võrgu modelleerimise käik oleks inimestele jälgitav ning arusaadav, siis see tõlgendatavus suurendaks usaldust mudeli väljundite vastu ning aitaks väära mudeli väljundi puhul vea tekkepõhjust mõista ning seda parandada. Selles uurimistöös kasutati keelemudeli töökäigu uurimiseks sõnastiku õppimise meetodit hõredate autoenkooderite abil, kus autoenkooderi mudel eraldab keelemudeli närvivõrgu aktivatsioonid tunnusteks, mida saab vaadelda kui mudeli poolt õpitud kontseptsioone. Töö raames treeniti kolm hõredat autoenkooderit, mis erinesid üksteisest tunnuste arvu ning etteantud keelemudeli kihi poolest. Kolmest autoenkooderist parima poolt eraldatud tunnuseid analüüsiti ning leiti erineva üldistusastmega kontseptsioone: näiteks patsienti mõjutavad spetsiifilised terviseprobleemid, patsiendi füüsiline aktiivsus ning ravi positiivne kulg.

Võtmesõnad:

GPT, keelemudelid, tehisintellekt, tervishoiuinformaatika, terviseandmed, kontseptsioonide eraldamine, hõredad autoenkooderid, mehhaaniline tõlgendatavus

CERCS: P176 Tehisintellekt

Extracting Concepts From A Large Language Model Trained On Health Data

Abstract:

Language models are capable of performing a wide range of tasks, but how they arrive at their results is like a black box - the user provides input and receives output, but how the output was arrived at is unknown. If the process of network modeling were observable and understandable to humans, then this interpretability would increase confidence in the model's outputs and, in the event of an incorrect model output, understand what went wrong and fix it. In this research, the dictionary learning method using sparse autoencoders was used to study the workflow of a language model, where the autoencoder model separates the neural network activations of the language model into features, which can be viewed as concepts learned by the model. As part of the work, three sparse autoencoders were trained, which differed from each other in the number of features and the given language model layer. The features extracted by the best of the three autoencoders were analyzed and concepts with different degrees of generalization were found, for example, specific health problems affecting the patient, the patient's physical activity, and the positive course of treatment.

Keywords: GPT, large language models, artificial intelligence, healthcare informatics, health data, extracting concepts, sparse autoencoders, mechanistic interpretability

CERCS: P176 Artificial intelligence

Sisukord

Sissejuhatus.....	5
Mõisted ja terminid.....	7
1. Teoreetiline ülevaade.....	8
1.1 Keelemudel.....	8
1.2 Keelemudelite tööpõhimõte.....	10
1.3 Transformeri arhitektuur.....	10
1.4 Kasutatav keelemudel.....	12
2. Keelemudelitest kontseptsioonide eraldamine.....	14
2.1 Hõre autoenkooder.....	15
2.2 Sõnastiku õppimise meetod.....	17
3. Metoodika.....	20
3.1 Andmestik.....	20
3.2 Autoenkooderi treenimine.....	20
3.3 Keelemudeli aktivatsioonidest tunnuste eraldamine.....	21
4. Tulemused.....	23
4.1 Autoenkooderite evalveerimine.....	23
4.2 Tunnuste analüüsimine.....	24
4.3 Näited tunnustest.....	25
Kokkuvõte ja võimalikud edasiarendused.....	30
Viidatud kirjandus.....	31
Lisad.....	35
Lisa 1. Keelemudeli GPT-4o viip ning väljund termini mitmekihiline pertseptron jaoks.....	35
Lisa 2. Kasutatava keelemudeli detailne arhitektuur.....	36
Lisa 3. Koodi repositoorium.....	37
Lisa 4. Tunnus X16_L11#38.....	38
Lisa 5. Tunnus X16_L11#91:.....	39
Litsents.....	40

Sissejuhatus

Keelemudelid, tehisintellekti üks arenenumaid vorme, on suutelised täitma palju erinevaid ülesandeid: kasutaja küsimustele vastamine, tekstist kokkuvõtte tegemine, olemite klassifitseerimine ning varasemate andmete põhjal ennustuste tegemine ja palju muud (Markowitz, 2021). Keelemudeleid on hakatud ka meditsiini ülesannete jaoks rakendama. Tartu Ülikooli arvutiteaduse instituudi terviseinformaatika töörühm on loonud Eesti terviseandmete põhjal mitu genereerivat eeltreenitud transformeri (GPT-2) arhitektuuril keelemudelit, näiteks mudel, mis genereerib arstide kirjutatavat teksti vastavalt diagnoosidele.

Keelemudeli tulemused on rahuldavad, kuid see on nagu must kast - seda, kuidas täpselt tulemuseni jõutakse, hetkel ei teata. Võimalik on teada järgmise tokeni tõenäosusjaotust, kuid mitte seda, kuidas võrk seda modelleerib¹. Tehisintellekti meditsiinilised rakendused on ajalooliselt olnud valdkond, kus paljud teadlased on rõhutanud tõlgendatavuse olulisust. Arvestades meditsiiniliste otsuste suurt kaalu, võib tõlgendatavus suurendada usaldust mudeli väljundite vastu ja võimaldada meditsiinitöötajatel sünteesida mudeli arutluskäiku omaenda omaga (Lindsey et al., 2025). Lisaks lubaks see väära mudeli väljundi puhul mõista, milles viga tekkis, ning seda parandada.

Ühendriikides asuv tehisintellekti firma Anthropic uurimisrühm (Templeton et al., 2024) on loonud meetodi, mis kasutab hõredaid autoenkodereid, et keelemudeli sisemist tööd uurida ja kaalude kombinatsioonidele tähendusi anda. Sellega vaadatakse, millised mudeli osad erinevate sisendite peale aktiveeruvad ning püütakse seda hõreda võrguga ennustada. Selle hõreda võrgu parameetrite kombinatsioone ehk tunnuseid, võib aga vaadelda kui kontseptsioone, mida mudel on ära õppinud. Nende parameetrite manuaalne võimendamine suutis ka mudeli panna uut moodi käituma, näiteks vigast koodi väljastama, valetama või iga küsimuse vastuseks kindlast teemast rääkima, aga selle meetodiga oli ka võimalus vastuseid parandada.

Terviseinformaatika töörühma mudelite puhul pole sarnast uurimust varem tehtud. Nendes mudelites võiksid kontseptsioonid olla näiteks: patsiendi terviseprobleem, tema kasutatavate ravimite loetelu ja ravi mõju. Esmalt tuleks sellised kontseptsioonid mudelist üles leida.

¹ Info allikaks on e-kiri lõputöö juhendajalt

Seejärel on võimalik mudeli parameetrite kaalusid muuta, et parandada tulemuste korrektsust ning seeläbi saavutada realistlikumate terviseandmete genereerimine.

Tulemuste parandamine oleks üks samm lähemal selliste mudelite reaalses meditsiinitöös kasutamisele. Samuti, kui avastatud meetodid on rakendatavad teistele sarnastele mudelitele, siis aitaksid need meetodid ka teiste keelemudelite uurimist ning tulemuste parandamist.

Töö esimeses peatükis antakse ülevaade keelemudelistest ja keelemudelite alamkategoriasst — transformerite arhitektuurist. Samuti tutvustatakse selles töös kasutatavat keelemudelit. Teises peatükis on selgitatud, mis on sõnastiku õppimise meetod ning kuidas selles hõredaid autoenkoodereid kasutatakse. Kolmandas peatükis on kasutatud metoodika seletus, mis hõlmab infot kasutatud andmestiku, autoenkooderite treenimise ning keelemudeli aktivatsioonidest tunnuste eraldamise kohta. Neljas peatükk on tulemuste kirjeldus.

Mõisted ja terminid

Keelemudel (ingl *language model*) on loomulikku keelt kirjeldav tõenäosuslik mudel, mis põhineb keeleandmetele toetuvatel algoritmidel².

Sõne (ingl *string*) on märkide järjend, (tervikuna käsitletav) sümbolite jada³.

Token (ingl *token*) on keelemudeli sisendi ja väljundi väikseim jaotusühik. Loomuliku keele töötlemise ülesannetes tähistavad märgid tavaliselt sõnu, alamsõnu või märke³.

Pärilevivõrk (ingl *feedforward network*) on mitmekihiline võrk, milles pole tagasisideteid ega teid mingi ühe kihi tehisneuronite vahel⁴.

Mitmekihiline pertseptron (ingl *multi-layer perceptron*, MLP) tehisnärvivõrk, mis koosneb sisendkihist, ühest või enamast peidetud kihist ja väljundkihist, kus iga neuron on ühendatud järgmise kihi kõikide neuronitega. See suudab õppida keerukaid seoseid andmetes, kasutades mittelineaarseid aktivatsioonifunktsioone ja tagasi-levi algoritmi⁵.

Nihe (ingl *bias*) on mudelitüübi süsteemne võimetus teatud andmepunkte hästi ennustada, sest mudel ei suuda tegelikku andmetekke funktsiooni korrektselt kujutada või olemasolev andmestik ei luba seda teha⁶.

Ülesobitamine (ingl *overfitting*), on nähtus, kus mudel jälgib treeningandmeid nii täpselt, et päris andmete puhul, mis varieeruvad treeningandmetest, ei ennusta mudel õigesti⁷.

² Termin defineeritud Sõnaveebist <https://sonaveeb.ee/>.

³ Termin defineeritud artiklist <https://www.koyeb.com/blog/what-are-large-language-models>

⁴ Termin defineeritud IT terministandardi sõnastikust <https://arhiiv.eki.ee/dict/its/>

⁵ Termin seletuse genereeris GPT-4o. Kasutatud viipa ning GPT-4o väljundi jaoks vaata lisa 1.

⁶ Termin defineeritud Praktilise Andmeteaduse õpikust (Sügis et al., 2024)

⁷ Termin seletus võetud kursuselt

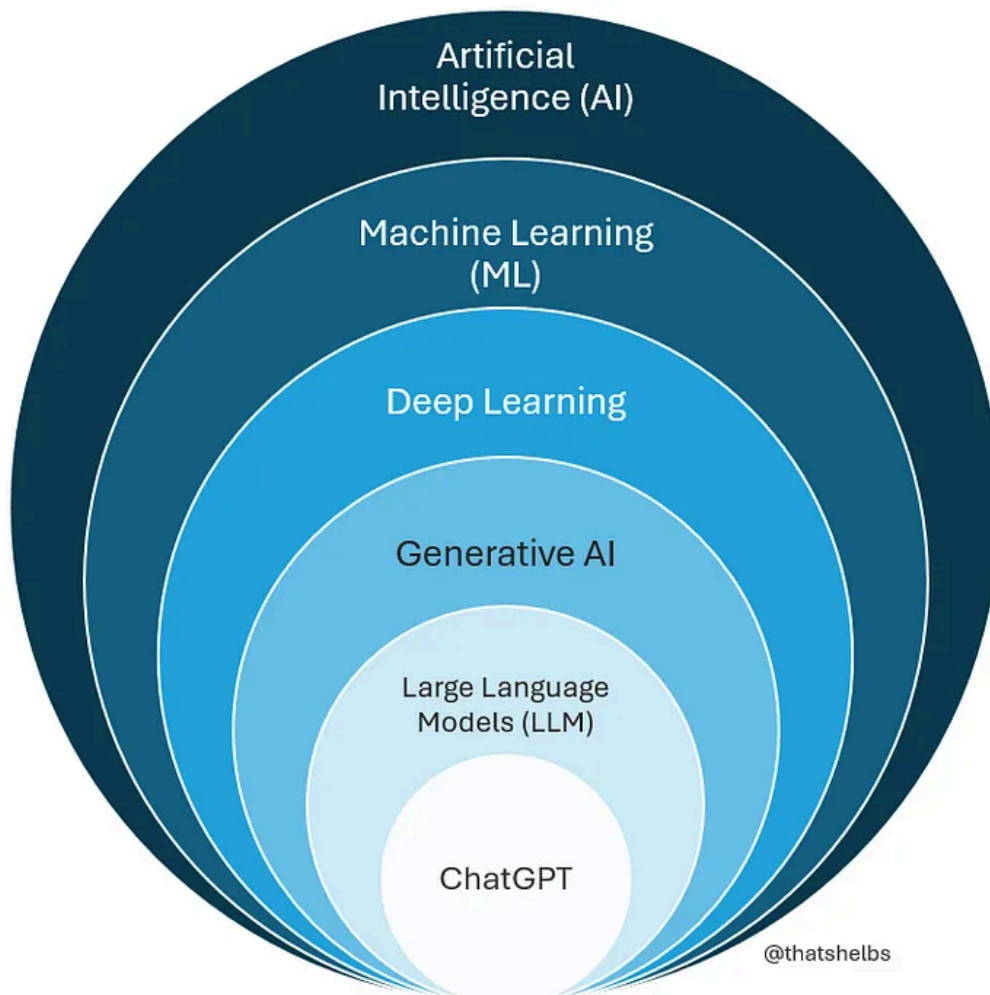
<https://developers.google.com/machine-learning/crash-course/overfitting/overfitting>

1. Teoreetiline ülevaade

Järgnevas peatükis antakse ülevaade keelemodelite olemusest ning nende tööpõhimõttest, mis aitab paremini mõista, kuidas keelemudelid teavet töötlevad. See tagab tausta, et seletada, kuidas on võimalik inimesel keelemudeli tööprotsessi uurida.

1.1 Keelemudel

Selleks, et paremini vastata küsimusele, mis on keelemudel, on mõistlik esmalt teada, mis tehnoloogiate alla see kuulub, sest keelemudelid on arendatud mitme eelneva tehnoloogia põhjal (vaata joonis 1).



Joonis 1. Venni diagramm tehisintellektist kuni ChatGPT-ni (Temple, 2024)

Esimene ning kõige laiem mõiste, mille alla keelemudelit kategoriseeritakse, on tehisintellekt, mis on arvuti võime läheneda ülesande lahendamisele inimesesarnase intelligentsiga. See hõlmab endas võimet õppida, arutleda, tajuda ning teha otsuseid (Temple, 2024).

Teiseks kuulub keelemudeli tehnoloogia masinõppe alla. Cole Stryker ja Eda Kavlakoglu kirjeldavad masinõpet kui mudelit, mis koosneb andmete peal treenitud algoritmidest, mis suudab etteantud andmetest lähtuvalt teha ennustusi (*What Is Artificial Intelligence (AI)?*, 2024). Näiteks kui on andmefail, kus on kirjas inimeste pikkus, kaal, vanus ning hinnang inimese tervisele (kujul “terve” või “mitteterve”) ning masinõppe mudelit treenitakse nende andmete põhjal, siis proovib mudel leida tunnuste (pikkus, kaal, vanus, tervise hinnang) vahelisi matemaatilisi seoseid nagu, kas pikkuse ja kaalu väärtus on omavahel seoses või, kas vanuse kasvades on tervisehinnang enamasti terve või mitteterve. See võimaldab ühte tunnust teiste tunnuste põhjal ennustada.

Kolmas tase on süvaõpe: masinõppe alamliik, mis kasutab kindlate algoritmide asemel mitmekihilisi närvivõrke, et andmetest õppida (Holdsworth & Scapicchio, 2024). Närvivõrkude kasutamine pakub mitmeid eeliseid, näiteks suudavad need kasutada struktureerimata andmeid (harilik masinõppe kasutab üldiselt andmeid, mis on struktureeritud sarnaselt tabeli formaadile), leida keerulisemaid tunnuste vahelisi seoseid kui harilikud algoritmid ning teostada järelvalveta õpet, mis tähendab, et suudetakse õppida sildistamata andmetest ja leida neist seoseid ilma, inimese suunamiseta. Tänu närvivõrkude eelistele on need suutelised täitma ülesandeid, millega masinõppe mudelid ei tule hästi toime nagu pildituvastus või keeletöötlus (*What Is Deep Learning?*, s.a.).

Neljandat taset tihti ei käsitleta eraldi ja järgmiseks loetaksegi keelemudeleid, kuid minnes spetsiifilisemaks on olemas ka generatiivne tehisintellekt, mis suudab luua kasutaja viipe põhjal mingi teose nagu pildi, teksti, video, heliteose või tarkvara koodi (Stryker & Scapicchio, 2024). Keelemudelid on spetsialiseerunud just keelega seonduvatele ülesannetele. Need on loodud mõistma ning looma teksti ning ka teisi sisuvorme inimese sarnaselt. Näiteks on keelemudelid võimelised järgnevateks tegevusteks: konteksti põhjal järelduste tegemine, küsimustele sidusalt ning kontekstikohaselt vastamine, teksti tõlkimine ühest keelest teise ning tekstist kokkuvõtete loomine (*What Are Large Language Models (LLMs)?*, 2023).

1.2 Keelemudelite tööpõhimõte

Keelemudelid kasutavad mitmekihilisi närvivõrke, et treenimisprotsessi käigus õppida ennustama järgmist sõnet lauses. Selleks vaatab mudel lause eelnevaid sõnesid ning omistab kõigile äraõpitud sõnedele tõenäosusskoori, määramaks kui võimalik on, et see tuleb valida järgnevas sõnaks (*What Are Large Language Models?*, s.a.; *What Are Large Language Models (LLMs)?*, 2023).

Arvutid ei kasuta oma töö jaoks inimkeelt, vaid numbreid, seega on teksti paremaks töötlemiseks mõttekas esitada sõnu arvulisel kujul.

Mudel saab sisendiks toore teksti, näiteks hulga lauseid. Esimene samm on seda sisendit tokeniseerida, ehk teisendada tekst sõnade või alamsõnade seeriaks (Wolfe, 2024). Neid sõnu ja alamsõnasid kutsutakse tokeniteks. Tokeniseerimine on tähtis, et jagada eesolevat tööd väiksemateks, kergemini käsitlevamateks osadeks. Tokenid võivad olla ühe tähe või terve sõna pikkused. Seejärel teisendatakse tokenid vektoriteks, mida kutsutakse vektorestituseks (ingl *embedding*), kus igale sõnale vastab tema arvuline esitus (Sügis et al., 2024).

Tulemuste täpsuse tagamiseks treenitakse keelemudeleid massiivsetel andmemahtudel, mida saadakse näiteks Wikipedia artikkelite põhjal loodud korpustest või Common Crawl⁸ korpustest, mis luuakse veebist massiliselt teksti kogumisest. See võimaldab keelemudelitel õppida grammatikat, semantikat ning ka faktiteadmisi (Radford et al., 2019).

1.3 Transformeri arhitektuur

Valdav enamus keelemudeleid kasutavad transformeri arhitektuuri, mis esitati 2017. aasta uurimuses “*Attention is All You Need*” ning millest on saanud kaasaegsete keelemudelite jaoks fundamentaalne osa (Markowitz, 2021). Dale Markowitz (Markowitz, 2021) toob välja, et transformeritel ehk keelemudelitel, mis kasutavad vastavat arhitektuuri, on mitmeid eeliseid varasemate (rekurrentsete) närvivõrkude ees:

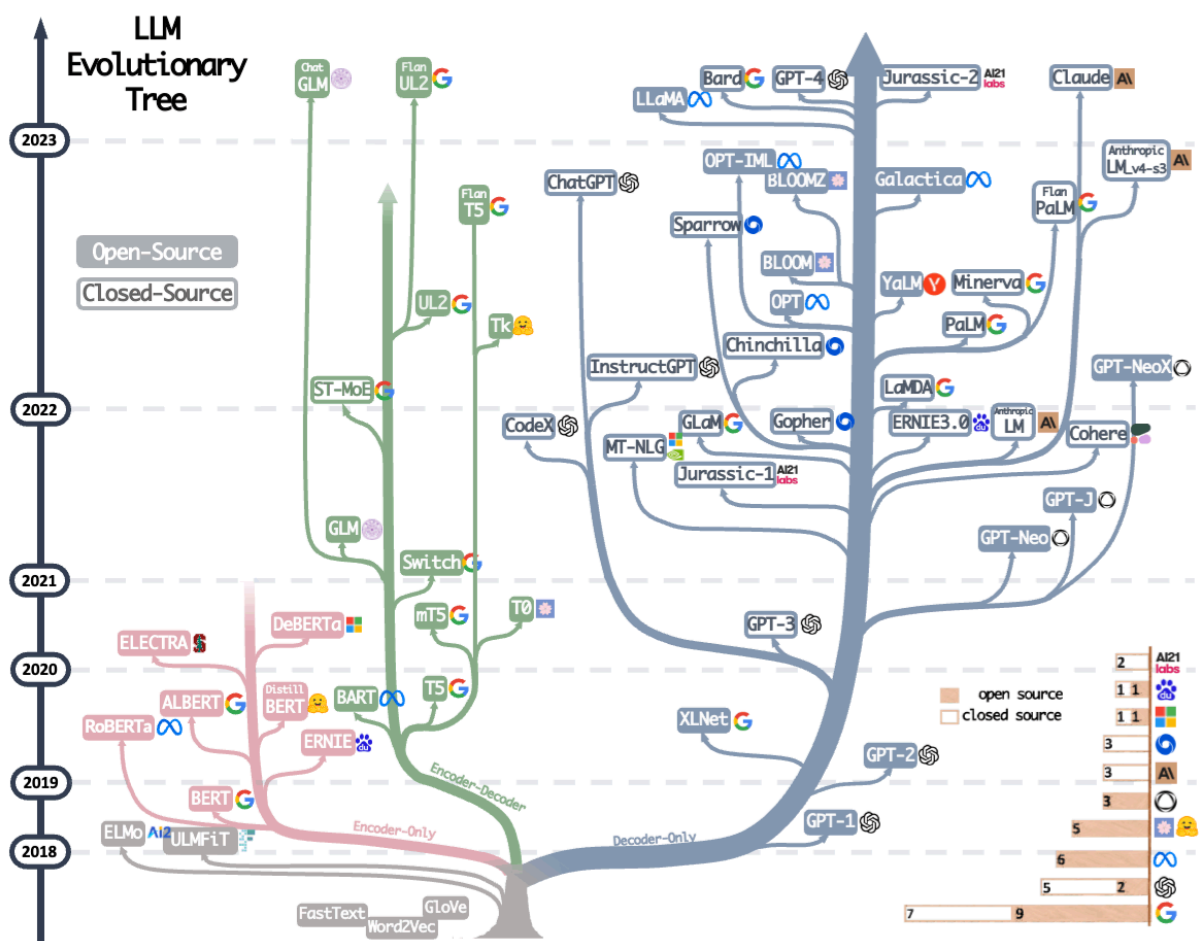
- Sisendlause sõnade töötlemine paralleelselt, mitte sekventsiaalselt ehk üks sõna järjest, mis muudab mudeli efektiivsemaks.
- Enesetähelepanuvõime, mis võimaldab mudelil lauses sõnu üksteisega seostada, nagu asesõna nimisõnaga.

⁸ (Buck et al., 2014)

- Hariliku tähelepanu kiht, mis aitab keskenduda lause asjakohastele sõnadele.
- Sõnade positsiooni kodeerimine, mis aitab õppida korrektset sõnajärge.

Enkooder on mudeli alamblokk, mis vastutab sisendtekstist asjakohase teabe mõistmise ja eraldamise eest. Seejärel väljastab see sisendteksti vektorsituse ja edastab selle mudelile, kus see lõpuks jõuab dekooderile, mis teisendab vektorsituse tagasi sõne kujule (Raschka, 2025)

Nagu joonisel 2 on näha, siis suur osa keelemudelitest on ainult dekooder tüüpi transformerid (Yang et al., 2024), kaasa arvatud selles töös kasutatav GPT-2 keelemudel. Seega ei keskenduta selles töös pikemalt enkooderi osale.



Joonis 2. Suurte keelemudelite evolutsiooni skeem. Punane haru on ainult enkooder, roheline enkooder-dekooder ja sinine ainult dekooder-ülesehitusega keelemudelid.⁹

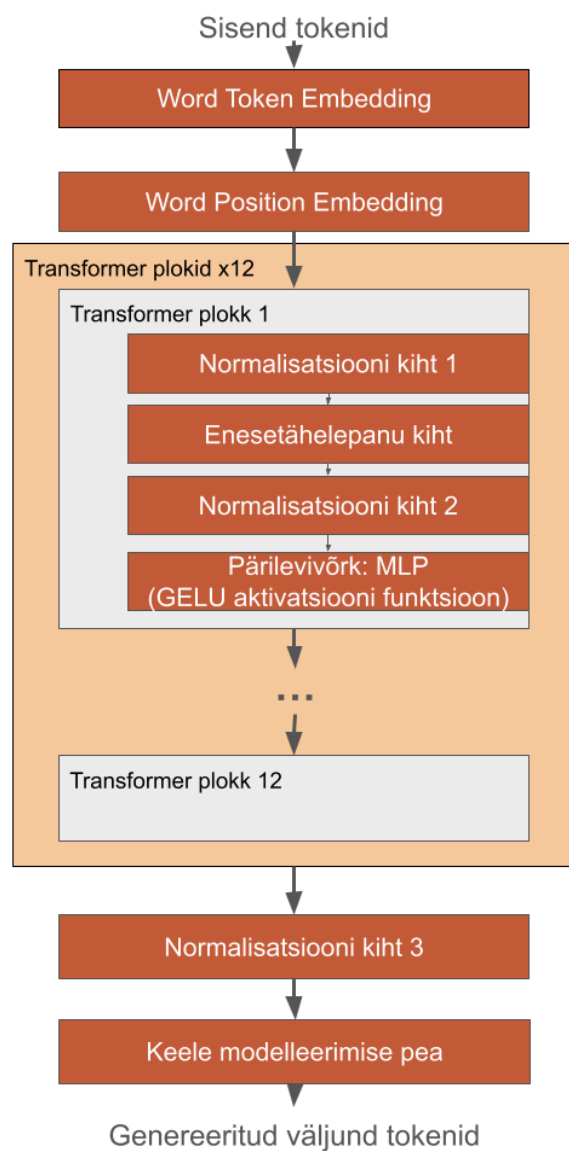
Wolfe kirjutab (Wolfe, 2024), et dekooder tüüpi transformerite on *embedding* kiht, mis teisendab kõik sisendiks saadud tokenid vastavale vektorkujule. Lisaks sisestab embedding

⁹ Kuvatõmmis, allikas: <https://dl.acm.org/doi/10.1145/3649506>

kiht infot tokeni positsiooni kohta tekstis, mis on oluline sest sõnajärg mõjutab lause tähendust.

1.4 Kasutatav keelemudel

Keelemudel, mida uurimistöös uuriti, on GPT-2 arhitektuuriga, mis avaldati 2019. aastal. Mudeli ülesehitus on hierarhiliselt välja toodud joonisel 3 (detailsem mudeli arhitektuur on lisas 2).



Joonis 3. EstMed-gpt2_fine_tuned4 mudeli kihid. Selguse eesmärgil pole sellel joonisel kujutatud nelja vahepealset dropout kihti.

Mudel koosneb järgnevatest kihtidest (sulgudes on lisas 2 kasutatud tähistused):

Word Token Embedding (wte) - teisendab 50258 õpitud tokenit 768 dimensionaalsele vektori kujule.

Word Position Embedding (wpe) - lisab tokeni vektoresitusele infot (kuni 1024 tokenit) positsiooni kohta.

Dropout (drop) - seab 10% neuronite aktivatsiooni väärtusteks 0. Reguleerimistehnika vältimaks ülesobitamist (ingl *overfitting*).

Mudeli moodulite nimekirjas on 12 transformeri plokki. Igas plokis on:

Normalisatsiooni kiht (ln_1) - normaliseerib sisendid vahemikku [0, 1], mis aitab treenimist stabiliseerida.

Enesetähelepanu plokk (attn) - arvutab iga tokeni puhul, kui palju peaks mudel sellele keskenduma. c_attn kiht jagab tokeni kolmeks osaks, millega arvutatakse tähelepanu skoori ning c_proj teisendab selle tagasi algsele kujule, et mudeli järgmisele kihile edastada.

Enesetähelepanu plokile järgneb veel kaks *dropout*-kihti ning normalisatsioonikiht.

Multi layer pertseptron (MLP) plokk on pärilevivõrk, mis toimib kui andmete töötleja, mis aitab õppida ja ennustada keelelisi mustreid. Selles on lineaarne konvolutsiooniline kiht (c_fc), tagasi algsele kujule teisendav kiht (c_proj), mitte-lineaarne aktivatsioonikiht (act) ja *dropout*-kiht.

Transformer plokkidele järgneb viimane normalisatsiooni kiht, pärast mida on *language modeling head* kiht, mis võtab vektoresituse ning saab sellest tõenäosused järgneva tokeni ennustamiseks, mis antakse kõige viimasele kihile — generaatorile, mis tagastab väljund tokeni.

2. Keelemudelitest kontseptsioonide eraldamine

Anthropicu rühm (Bricken et al., 2023) on kirjutanud, et mehhaanilise tõlgendatavuse eesmärk on mõista närvivõrke, jagades need komponentideks, mis on kergemini mõistetavad kui tervik. Mõistes iga komponendi funktsiooni ning nende omavahelist vastastikust toimet, loodetakse seletada kogu võrgu käitumist. Esimene samm sellele lähenemisele on tuvastada õige komponent, mida analüüsida. Närvivõrkude põhilisim komponent neuron, mille võiks analüüsi alla võtta, ei käitu aga inimesele enesestmõistetavalt.

Selleks, et närvivõrk saaks väljundit ennustada, kasutab see sisemisi neuroni aktiveerimisi kui muutujaid. Neuronite aktiveerimine loob hilisemates kihtides kõrgetasemelisi, semantiliselt rikkalikke kontseptsioone, kasutades varasemate kihtide madalama taseme kontseptsioone (Ayonrinde, 2023).

Anthropicu rühm (Bricken et al., 2023) seletab seda sellega, et neuronid on polüsementilised (ingl *polysemantic*): sama neuron võib reageerida täiesti erinevatele näiliselt mitteseotud sisenditele. Näiteks nende väljatoodud väikeses keelemudelis reageeris üks neuron nii akadeemilistele viidetele, inglise keelsele dialoogile, HTTP päringutele kui ka korea keelsele tekstile. Polüsementilisus muudab raskeks seletada närvivõrgu käitumist üksikute neuronite raames.

Üks pakutavatest polüsementilisuse esinemise põhjusteks pakutakse nähtust nimega superpositsioon (ingl *superposition*), kus närvivõrgud esitavad rohkem andmete “tunnuseid” kui sellel on neuroneid, määrates igale tunnusele lineaarse kombinatsiooni neuroneid. Niimoodi simuleerivad närvivõrgud palju suuremaid ning hõredamaid närvivõrke. Selle tulemusena võib eeldada, et tunnuse suunad moodustavad vektorruumis ületäieliku baasi (ingl *overcomplete basis*) (Bricken et al., 2023).

Seoses vektorruumidega on täielik baas vektorite kogum, nii et mis tahes vektorit vektorruumis saab esitada baasist pärinevate vektorite lineaarse kombinatsioonina. Baasi loetakse ületäielikuks, kui see on täielik ka pärast vektori eemaldamist baasist. (Krishna, 2014)

Uurimistöös “*Toy Models of Superposition*” (Elhage et al., 2022) toob Anthropicu rühm välja kolm viisi superpositsiooni fenomeni lahendamiseks:

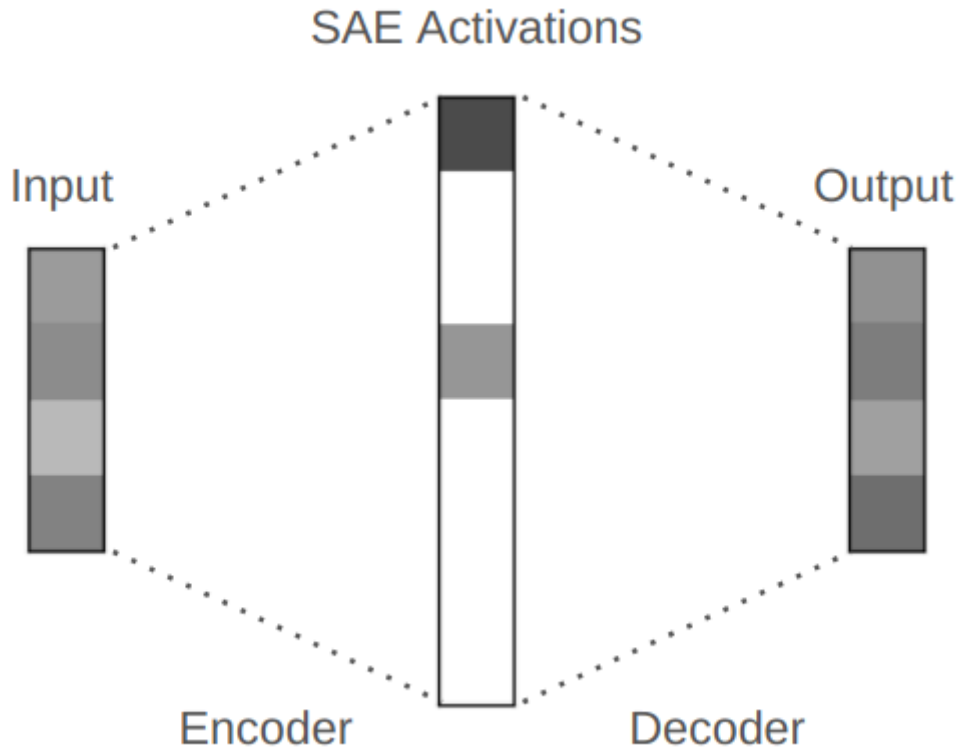
1. Luua keelemudeleid ilma superpositsioonita. Üliväikeste mängumudelite puhul on see võimalik. Hilisemas uurimistöös (Bricken et al., 2023) aga leiti, et see fundamentaalselt pole sobiv lahendus, sest neuronid võivad olla polüsementilised isegi ilma superpositsioonita. Seda seetõttu, et paljudel juhtudel saavutavad mudelid väiksema kao, esitades mitut tunnust mitmetähenduslikult (polüsementilises neuronis), kui esitades ühe tunnuse üheselt ja ignoreerides teisi.
2. Leida ületäielik baas (ingl *overcomplete basis*), mis kirjeldab, kuidas tunnuseid superpositsiooniga mudelites esitatakse. Algselt loodeti, et see on lahendatav hariliku hõreda kodeerimise või sõnastiku õppimise meetoditega ning et ei pea muretsema mudeli jõudluse vähendamise pärast. Järgnevas uurimistöös (Bricken et al., 2023) leiti aga, et hariliku sõnastiku õppimise meetodi rakendamine põhjustab lisaks muudele probleemidele liigset ülesobitamist (ingl *overfitting*).
3. Hübridne lähenemine, mille käigus muudetakse mudeleid nii, et ei eemaldata superpositsiooni täielikult. Kui aga mudelil esineb treenimise tulemusena vähem superpositsiooni, siis see hõlbustatakse analüüsi teises etapis ületäieliku aluse leidmist. Esimese viisi ebasobivuse tõttu seda valikut kasutusse ei võetud.

2023. aasta uurimistöös (Bricken et al., 2023) arendati edasi teise viisi metoodikat ning Anthropicu rühm kasutas “nõrka” sõnastiku õppimise algoritmi, mida kutsutakse hõredaks autoenkooderiks (ingl *sparse autoencoder. SAE*). Selle meetodiga genereeriti õpitud tunnuseid treenitud mudelist, mis pakkus ühetähenduslikumat analüüsi subjekti kui mudeli neuronid ise.

2.1 Hõre autoenkooder

Selleks, et jagada keelemudeli neuroni aktiveerimised arusaadavamateks osadeks, treenitakse hõredad autoenkooderid keelemudeli põhjal (Templeton et al., 2024).

Kola Ayonrinde (Ayonrinde, 2023) kirjeldab autoenkooderit kui mudelit, mis proovib rekonstrueerida andmeid, pärast seda, kui need on kitsaskohast läbi lastud. Traditsioonilistes autoenkooderites võib kitsaskohaks olla funktsioon, mis vähendab maatriksi dimensioone. Hõredates autoenkooderites aga on kitsaskohaks funktsioon, mis suurendab maatriksi dimensioone ning esitab nad hõredal kujul, nagu on kujutatud joonisel 4. Autoenkooderite eesmärk on kitsaskohast hoolimata originaalandmed võimalikult täpselt uuesti luua.



Joonis 4. Hõreda autoenkooderi diagramm. Värvitoonid näitavad aktivatsiooni väärtusi (Karvonen, 2024).

Hõredad autoenkooderid lisavad oma kaofunktsioonile hõredustrahvi (ingl *sparsity penalty*) L1-kaos kujul, pannes mudelit aktiveerima vaid väikese osa peitkihtides olevatest neuronitest. See sunnib mudelit looma võimalikult paljude nullväärtustega vahevektoreid (aktivatsiooni väärtusi) (Reberga, 2025).

L1-kadu arvutatakse kaalude absoluutväärtuste summana: $L1 = \lambda \sum |w|$, kus λ on regulariseerimisparameeter. Proovides saada kaalude absoluutväärtuste summat võimalikult väikeseks, seatakse osa kaale nulliks (Dey, 2024).

Hõreda autoenkooderi mudel töötab formaalselt järgmiselt: närvivõrk võtab sisendi $x \in \mathbb{R}^n$ ja rekonstrueerib selle väljundina \hat{x} kasutades võrrandit $\hat{x} = W' \sigma(Wx + b)$, kus $W \in \mathbb{R}^{h \times n}$ on enkooderi kaalude maatriks, b on vabaliige (ingl *bias term*), σ on mitte-lineaarne aktivatsiooni funktsioon ja W' on dekodeeri maatriks, mis tihti kasutab transponeeritud enkooderi maatriksit (Lan et al., 2025). Kaalude maatriksi ridade arv h on ettemääratav argument.

2.2 Sõnastiku õppimise meetod

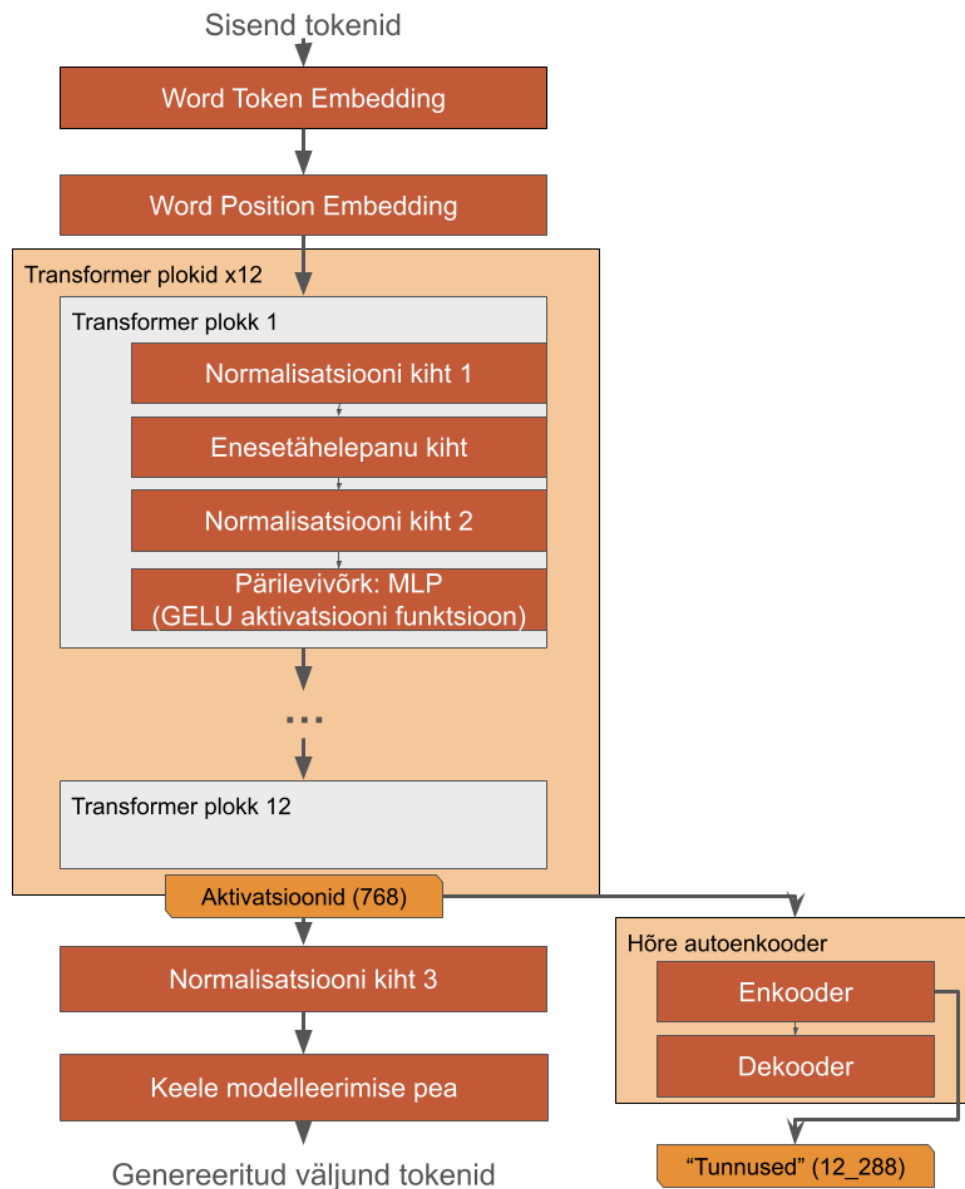
Sõnastiku õppimine (ingl *Dictionary Learning*) pärineb lineaarsest algebrast antud vektorruumi baasi leidmisest. Väljas (Väljas, s.a.) defineerib, et vektorruumi V lineaarselt sõltumatute vektorite süsteem $W = \{e_1, e_2, \dots, e_n\}$ moodustab baasi, kui ruumi V mistahes vektor on avaldatav süsteemi kuuluvate vektorite lineaarse kombinatsioonina, see tähendab

$$\forall x \in V \text{ korral } x = x_1 e_1 + x_2 e_2 + \dots + x_n e_n \text{ kus } x_i \in R (i = 1, 2, \dots, n)$$

Baasivektorite maatriksit kutsutakse ka sõnastiku maatriksiks ning algoritm, mis genereerib sõnastiku maatriksi annab sellele meetodile nime sõnastiku õppimine (Deelaka, 2024). Seega sõnastiku õppimine on selliste baasivektorite leidmine, et sisendandmeid saab esitada lineaarse kombinatsioonina nendest baasivektoritest, andes andmetele uue esituse.

Hõreda autoenkooderi mudeli enkooderi õpitud kaalude maatriksisse W saab suhtuda kui sõnastiku maatriksisse ning sellepärast loetakse autoenkooderi kasutamist sõnastiku õppimise meetodite hulka. Kaalude maatriksi W ridade arv (h) vastab ka õpitud tunnuste arvule.

Kui keelemudelile anda sisend, siis mudeli pärilevivõrgu neuronites tekivad aktivatsioonid. Tänapäevastes mudelites on pärilevikihte mitmeid, nagu selle töö mudelis on 12 MLP plokki. Iga eelnev kiht mõjutab järgmist, luues mudelis aina keerukamaid seoseid. Hõre autoenkooder võtab ühe kihi neuronite aktivatsioonid ning teisendab need hõredale kujule, kus seosed on nii-öelda eraldatud, muutes selle uue aktivatsioonide kuju ühetähenduslikuks (joonis 5). Neid saab siis vaadelda kui tunnuseid, mida keelemudel on ära õppinud.



Joonis 5. Ülevaade sõnastiku õppimise meetodi voost. Transformer plokkist tulevad aktivatsioonid teisendatakse hõreda autoenkooderiga tunnuste kujule.

Näiteks toob Anthropicu uurimisrühm (Templeton et al., 2024) välja tunnuse nende uuritavast mudelist Claude 3 Sonnet'st, mis aktiveerub tugevasti Los Angelese Kuldvärava silla mainimise ning soetud kontseptsioonide peale (joonis 6).

Feature #34M/31164353 Golden Gate Bridge feature example

The feature activates strongly on English descriptions and associated concepts

in the Presidio at the end (that's the huge park right next to the Golden Gate bridge), perfect. But not all people

repainted, roughly, every dozen years." "while across the country in san francisco, the golden gate bridge was

it is a suspension bridge and has similar coloring, it is often compared to the Golden Gate Bridge in San Francisco, US

They also activate in multiple other languages on the same concepts

ゴールデン・ゲート・ブリッジ、金門橋は、アメリカ西海岸のサンフランシスコ湾と太平洋が接続するゴールデンゲート海

골든게이트 교 또는 금문교는 미국 캘리포니아주 골든게이트 해협에 위치한 현수교이다. 골든게이트 교는 캘리포니아주 샌프란시

мост золотые ворота – висячий мост через пролив золотые ворота. Он соединяет город сан-фран

And on relevant images as well



Joonis 6. Visualiseeritud näide tunnusest Kuldvärava silla kohta. Tunnus aktiveerub nii inglise- kui ka muukeelse teksti puhul. Tumedamalt esiletõstetud sõnad aktiveerisid tugevamalt. Lisaks aktiveerib tunnus ka piltide korral.

3. Metoodika

Töö praktiline osa koosneb kolmest osast: Hõreda autoenkooderi treenimine, keelemudeli aktivatsioonidest tunnuste eraldamine ning tunnuste analüüsimine ning tõlgendamine. Töökeskkonnaks oli Jupyter notebook rakendus Miniconda keskkonnas. Töö valmis Tartu Ülikooli teadusarvutuste keskkuses.

3.1 Andmestik

Kasutatav keelemudel, EstMed-gpt2_fine_tuned4, on varasemalt peenhäälestatud Eesti terviseandmete peal. Selleks, et töö raames ei peaks tundlike terviseandmeid käsitlema, kasutati hõreda autoenkooderi treenimiseks sünteetilisi andmeid, mis on EstMed-gpt2_fine_tuned4 mudeli poolt genereeritud. Andmestik koosneb 100 000 haigusloo sissekandest, kus on olemid sissekande tüübi (näiteks epikriis, anamnees, uuring), vanusegrupi, soo, kategooria koodi ning sissekande teksti jaoks. Autoenkooderi treenimiseks kasutati sissekande teksti osa, sest ülejäänud olemite sisu kordus andmestikus sageli ning see oleksid seganud mudeli aktiveerimisi.

3.2 Autoenkooderi treenimine

Töös kasutati avatud lähtekoodiga teeki Dictionary_learning (Samuel Marks & Mueller, 2024), mis lubab luua hõreda autoenkooderi mudelit, seda treenida ning ka hinnata. Töö käigus valminud koodi ning kohandatud Dictionary_learning teegi leiab lisast 3.

Autoenkooderi mudeli arhitektuur loodi, põhinedes Anthropic uurimistöole (Bricken et al., 2023). Mudel koosneb ühest MLP plokist, mille ülesehitus on järgmine:

- 1) Sisendkiht, millele on rakendatud nihe.
- 2) Lineaarne kiht, millele on rakendatud nihe.
- 3) Enkooder, mis on ReLU aktiveerimisfunktsiooniga kiht.
- 4) Dekooder, milleks on veel üks nihkega lineaarne kiht.

Treenimisel antakse ette autoenkooderile keelemudeli mingi kiht, millest tulevate aktivatsioonide põhjal hakkab autoenkooder õppima. Kihi valik mõjutab õpitud tunnuste tõlgendatavust, sest mida rohkem on sisendit töödeldud, seda spetsiifilisemaks peaksid mudeli aktivatsioonid muutuma, millest saaks ka eraldada tähenduslikke tunnuseid, mis oleksid vähem abstraktsed, kui esmaste kihtide aktivatsioonid.

Õpitavate tunnuste arv ehk sõnastiku suurus sõltub keelemudeli väljundi suurusest, mis on konstantne (768 dimensionaalne vektor), ning etteantavast tegurist. Tunnuste arvu suurendamine võimaldab leida mudeli omadusi, mis ei pruugi väheste tunnuste puhul välja tulla, kuid toob kaasa ka jõudlusprobleemi, sest treenimiseks kuluv aeg ning ressurss on lineaarselt sõltuv tegurist.

Treenimiseks kasutati Tartu Ülikooli teadusarvutuste keskuse Tesla V100 graafikakaarti.

3.3 Keelemudeli aktivatsioonidest tunnuste eraldamine

Keelemudeli aktivatsioonidest tunnuste eraldamine toimub treenimisele sarnaselt - ühe keelemudeli kihi aktivatsioonid sõõdetakse hõredale autoenkooderile ning need teisendatakse õpitud tunnuste kujule. Nende eraldatud tunnuste talletamine toob kaasa mahutavuse probleemi, sest nagu närvivõrgu neuroni aktivatsioone on tunnuseid sisendi iga tokeni kohta. Kui näiteks sisendi pikkus on 512 tokenit, siis juba 12288 tunnuse korral on tulemuseks 512×12288 dimensionaalne maatriks. Elemendid on float32 andmetüüpi, mis on nelja baidi suurused. $512 \times 12288 = 6,291,456$ elemendiline maatriks võtaks $6,291,456 \times 4 = 25,165,824$ baiti ehk umbes 24 megabaiti ning sisendeid on andmestikus sada tuhat.

Kuna aga suvalise sisendi (tokeni) puhul võib eeldada, et aktiveeruvad ainult asjakohased tunnused ehk enamike tunnuste aktiveerimise väärtus on 0, siis maatriks on hõre. Hõredaid maatrikseid saab esitada efektiivsel kujul, kus on märgitud vaid nullist erinevate väärtuste rida, veerg ja väärtus ise.

Eraldatud tunnused salvestati kasutades *Hierarchical Data Format* (HDF5) failiformaati¹⁰, mis on mõeldud suurte andmehulkade hoiustamiseks ning organiseerimiseks. H5PY pythoni

¹⁰ <https://www.hdfgroup.org/solutions/hdf5/>

teegi¹¹ kasutamine võimaldab HDF5 failiformaati käsitleda pythoni andmestruktuuride süntaksiga, näiteks sarnaselt NumPy järjenditega.

Lisaks tunnuste aktiveerimise väärtustele on faili salvestatud ka tokenid selleks, et mugavalt seostada, mis tokenist pärinevad vastavad aktivatsioonid. H5 fail on binaarkujul, kuid praktikas võib ette kujutada, et pärast faili sisselugemist on andmed kujul, kus ühel real on tokeni numbriline esitus, millele järgneb kõigi tunnuste aktiveerimisväärtuste vektor selle tokeni puhul.

¹¹ <https://www.h5py.org/>

4. Tulemused

Töö käigus treeniti võrdlemiseks kolm erinevate kihtide valiku ning tunnuste arvuga hõredat autoenkooderit:

- X16L11 - Sõnastiku suuruse teguriga 16, ehk $768 \times 16 = 12,288$ tunnusega ning MLP plokk indeksil 11. (tegelikkuses 12. plokk. Indeksid algavad nullist) väljundkihiga.
- X16L5 - Teguriga 16 (12,288 tunnust) ning MLP plokk indeksil 5. väljundkihiga.
- X64L11 - Teguriga 64 (49,152 tunnust) ning MLP plokk indeksil 11. väljundkihiga.

4.1 Autoenkooderite hindamine

Anthropicu uurimisrühm kirjutab (Bricken et al., 2023), et harilikult on masinõppes küllaltki lihtne öelda olemasoleva meetriku põhjal, nagu testimisel kadu, kas rakendatav meetod töötab hästi või mitte. Tunnuste eraldamise puhul ühest selget meetrikut ei ole, vaid vaadati mitut meetrikut korraga:

1. Manuaalne inspeksioon: Kas tunnused tunduvad tõlgendatavad?
2. Tunnuste tihedus: Kui palju on “elusaid” tunnuseid.
3. Rekonstrueerimise kadu: Kui hästi autoenkooder suudab rekonstrueerida MLP aktivatsioone dekodeeriga.

Kolme mudeli tunnuseid manuaalselt analüüsid oli valdaval enamusel kindel kontseptsioon, mille peale tunnus aktiveeris. Siiski leidis tunnuseid, mis ei aktiveerunud kunagi tugevalt või aktiveerusid ainult keskmise tugevusega väga erinevate tokenite puhul, näiteks numbrite, kirjavahemärkide ja ka sõnade peale.

Hinnates oma treenitud hõredaid autoenkoodereid Dictionary Learning teegi (Samuel Marks & Mueller, 2024) evalveerimise funktsiooniga, leiti mudelite järgmised hindamismeetrikud:

- Elusate tunnuste protsent - murdosa sõnastiku funktsioonidest, mis on aktiivsed vähemalt ühel tokenil tunnuste arvu seast.
- L1 kadu - autoenkooderi hõreduse mõõt. Väiksem tähendab hõredamat.
- L2 kadu - keskmine L2 kadu rekonstrueerimise puhul. Väiksem tähendab paremat rekonstrueerimist dekodeeriga.

Tabel 1. Treenitud autoenkooderite evalveerimise tulemused.

Hõreda autoenkooderi mudel	Elus tunnuseid (%)	L1 kadu	L2 kadu
X16L11	66.1%	19.95	5.66
X16L5	51.7%	8.52	2.17
X64L11	28.5%	19.70	5.58

Nagu tabelis 1 on näha, siis autoenkooderi mudel X16L5 12,288 tunnusega ja 6. MLP ploki väljundiga on hõredam ning parema rekonstrueerimise tulemusega kui mudel sama tunnuste arvuga ning mudel X16L11, millel on 12. MLP ploki väljund, kuid X16L5-l on ka vähem elusaid tunnuseid. Tunnuseid analüüsidest olid X16L5 omad abstraktsemad ning halvemini tõlgendatavad. Selle põhjuseks võis olla, et keskmise kihi kasutamine, milles närvivõrgu aktivatsioonid polnud lõplikult arenenud. Tõlgendatavuse tõttu jäeti selle mudeli tunnused põhjalikust analüüsist välja.

Mudel X64L11 49,152 tunnusega ja 12. MLP ploki väljundiga oli kao meetrikute poolest praktiliselt võrdne, kuid elusaid tunnuseid oli märgatavalt vähem. Elusate tunnuste väärtus vahemikus 50-80% on hea, et leida haruldasi tunnuseid, kuid sellest madalam näit tähendab, et suurt osa sõnastikust kasutatakse ülimalt harva või mitte kunagi. Enamikel sisenditel põhjustab surnud tunnuste olemasolu suurenenud mälu vajadust ning arvutusressursside kulu. Kuna analüüsimist viidi läbi manuaalselt, siis oli ka rohkete tunnuste arv murekohaks.

Eelnevate puuduste tõttu keskenduti töös mudeli X16L11 analüüsimisele.

4.2 Tunnuste analüüsimine

Tunnuste analüüsimise ning tõlgendamine hõlmab endas tunnuste läbikäimist ning nende tähenduse või lühikese kontseptsiooni kirjelduse omastamist. Selleks vaadatakse, milliste tokenite puhul tunnus aktiveerus tugevasti ehk üle mingi määratud väärtuse lävendi.

Tunnuste automatiseeritud analüüsimiseks oli plaanis kasutada keelemudelit, mis antud tokenite põhjal annaks neid ühendava kontseptsiooni. Seda ideed prooviti tartnuNLP

Llammas¹² mudelil, kuid sellel esinesid mitmed probleemid. Esiteks on tokenite hulk liiga suur, et seda anda hariliku viiba kujul mudelile ning andes ainult väikese osa näidistokeneid, ei ilmne tunnuse kontseptsioon selgelt. Samuti ei suutnud Llammas mudel meditsiiniterminite puhul pakkuda üldistavat seletust, kui oli tegu erinevate kindla haiguse või seisundiga seonduvate terminitega, vaid pakkus kontseptsioonina nende samade terminite loetelu. Eelnevate probleemide põhjal tuleks kasutada peenhäälestatud keelemudelit selle ülesande jaoks. Prooviti kasutada ka sama terviseinformaatika uurimisrühma EstMed-gpt2_fine_tuned4 keelemudelit, kuid ka selle tulemused ei olnud piisavalt rahuldavad. Tunnuste osalist analüüsimist viidi läbi manuaalselt.

4.3 Näited tunnustest

Selleks, et eraldatud kontseptsioone visualiseerida, loodi meetodid näitamakstunnuse aktiveerumist etteantud sisendi iga tokeni puhul, kus tugevamalt aktiveeruvad tokenid on esile tõstetud tumedama värviga.

Tunnuse aktiveerimise tugevuse jaotuse visualiseerimiseks loodi funktsioon histogrammi kuvamiseks. See lubab analüüsida, kui sagedasti tunnus aktiveerib nõrgalt või tugevalt ning kas nõrgad või tugevad aktiveerimised on domineerivad.

Analüüsimise käigus tuli välja tunnuseid, mis vastasid nii konkreetsetele kontseptsioonidele, näiteks aktiveerides kõige tugevamini termini kindla kirjapildi peale, kui ka kontseptsioonidele, milles samaväärselt tugevaid aktivatsioone kutsusid esile erinevad mõisted, mis on teema poolest seotud, näiteks tunnus erinevate antibiootikumide ravimite nimede kohta või tunnus vigastusest/haigusest mõjutatud piirkonna kohta.

Järgnevalt on visualiseeritud joonistel 7 kuni 14 mõned tähelepanuväärsemad tunnused, toodud välja neile vastav kontseptsioon ning tunnuse aktivatsioonide jaotus. Tunnuste nimetus on kirja pandud kujul <auto enkooderi mudel>#<tunnuse indeks>

Tunnus X16_L11#1717: Positiivsed hinnangud ravile (täpsemalt dünaamikale või efektile).

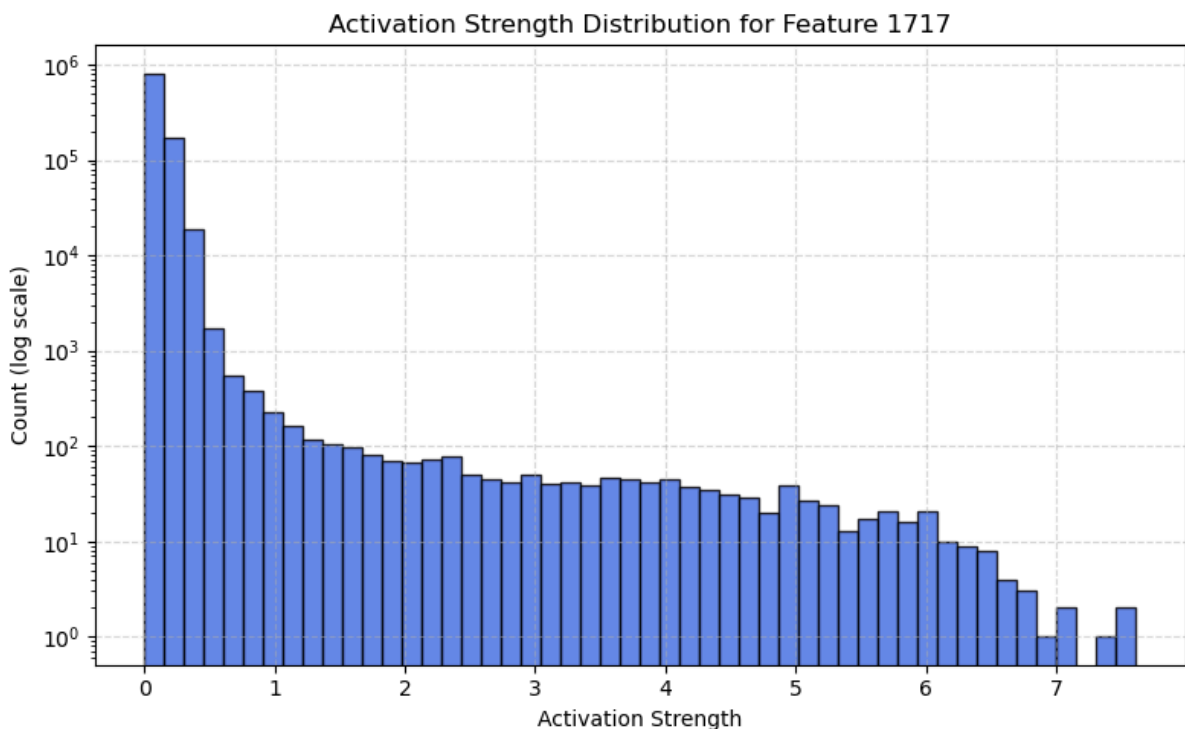
neuroloog (DATE)– aju - kt kontrastainega + re ko mend . eeg kontroll --> + ana eroob ne ja mittemed ikamentoosne ravi kerge **positiivse** dünaamikaga . retro dru aalne treemor kerge **positiivse** dünaamikaga , vasakul > paremal . rigiidsust kättes , jalgades ei ole . jätkata ravi (lev eti ra cetam 500 mg x 2) + flu anx oli 1 mg x 1 , cb z 150 mg x 2 perearsti juures soovitatud kontrolli kilpnäärme funktsiooni + vajadusel glükoosi toleran tsus test . neuroloog (DATE)– patsiendi sõnul **positiivse** dünaamikaga . ta treemor ei ole . peavalu hoog ide i esine . obj --> teadvusel , adekvaatne

¹² <https://huggingface.co/tartuNLP/Llammas>

psühholoogi konsultatsioon . kilpnäärme analüüsid olulise dünaamikata . tsh , ft 4 , ft 3 kontroll NAME : saanud varasemalt mitu korda escitaloprami tbl - d (DATE a sügisel) . esmane ravi sertr aliniga DATE - DATE a . kevadest , algselt otsusta ud ravi lõpetada . xxx DATE aprill a . paanikahoog udega , t . sertralini 15 - 20 - 40 mg x 1 . viimati käinud DATE a . detsembris . võtab praegu 25 mg x 1 , suvel - nov . alates 50 mg x 1 . ravi andnud hea efekti , kuid viimased paar päeva pole alprazolami võtnud . pt . - il laps 5 - aastane , sageli maga vad ja põhjustavad t ül li ja sega dust , mille tõttu

haiguse anamnees : DATE . xxx - c 20 pärasoole pahaloomuline kasvaja , tnm : c t 3 c n 0 c m 1 iv st . pt on saanud na kiiritusravi DATE - DATE rektumi kasvajale ja vaagna reg - le 54 gy , samaaeg ne keemiaravi s . cisplatini ga DATE : mrt : bra f dot sed . võrdluseks eelmine pet / kt DATE . presak raasel vasaku kusejuha suubumis kohal retsiidiivile / metasta asile viitav sõlm 1 , 9 cm . pildiliselt pigem soodsa dünaamikaga , mis eelnevatel uuringutel , fibroosne kude seal . paremal parailiakaasel 1 , 5 cm lümfisõlm dünaamikas vähenenud (oli 2 , 1

Joonis 7. tunnuse X16_L11#1717 näited andmestiku tekstilõikudest.



Joonis 8. tunnuse X16_L11#1717 histogramm aktivatsiooni tugevuste jaotusest. Tunnus aktiveerib väga tihti nullilähedase väärtusega. Üleminek keskmistest aktivatsioonidest tugevatele on lauge.

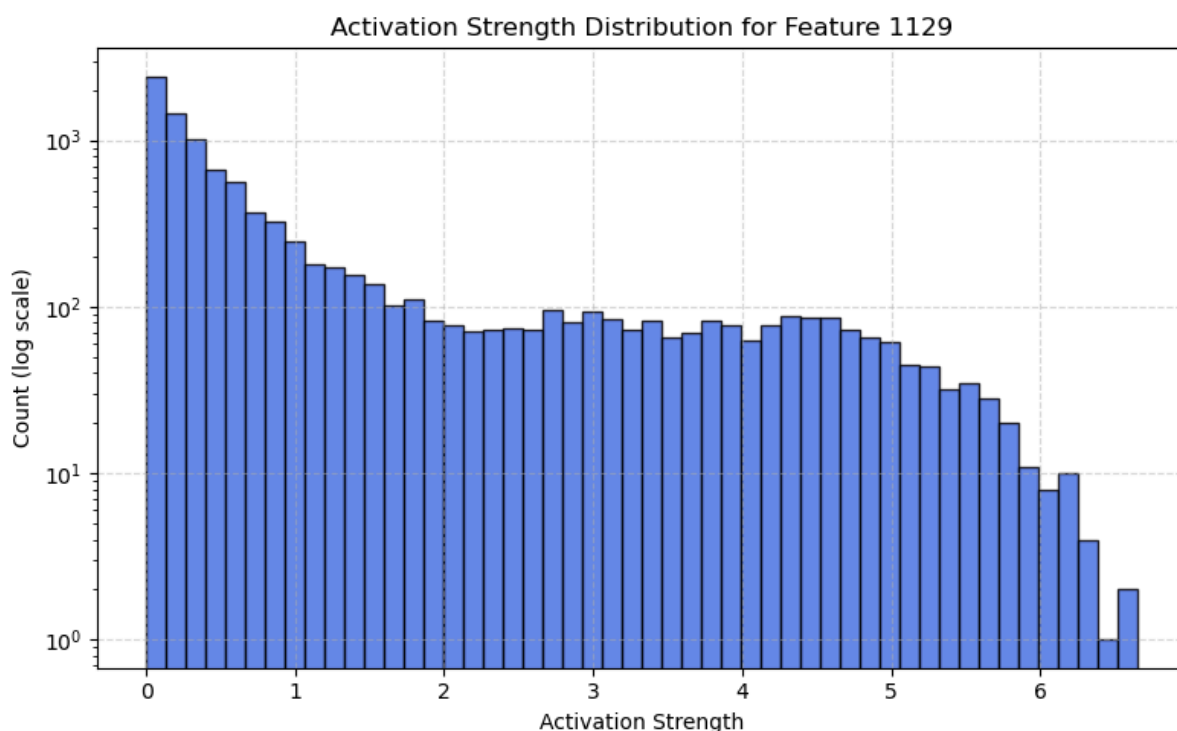
Tunnus X16_L11#1129: Kõrge vererõhu ravimid. Erinimelised ravimid, mis on kõik kõrge vererõhu ravimiseks.

anamnees DATE tiina tar kin : kardioloog oli teostanud ehho uuringu (sept umi seinä paks useks hinnatud arvestades + kaasnev ana i f / f kodade laperdus) . vererõhu ravimi foonil püsis vererõhk normis . kodune ravi : twynsta 80 / 10 mg x 1 metozoc 50 mg x 1 võtab tegelikult DATE aastast xarelto 20 mg x 1 diabeedi ravi 3 mg x 1 metformiin 850 mg x 2 crestor 10 mg x 1

11, NAME : ekg s siinusrütm, fr 56, n : tegi vasaku käe ja jala nõrkustunnet, kest is mõnda aega, kaebab urineerimishäireid, sage urin, põlet ustunne vas, küljes raviskeemis - telmis artan 40 mgx 1, amlodipini 10 mgx 1, rosuvastatin 20 mgx 1 rr 130 / 80 mmhg, südametöö rütmiline, fr 72 x', toonid puhtad, kopsudes vesikul, h / k DATE NAME : pereõe nõustamine hüpertoonia haige toitumisest - vähendada toidus soola sisaldust, rasva sed lih atooded ära ja vor sti minimaalselt, rohkesti tarbida köögivilju, ravimeid tarvitada iga päev, et hoida vererõhku

pigem vaskulaarne struktuur, suurenenud l / s - i kaelal ei näe, kilpnäärme maht NAME xxx : kodus süstoolne 145 ringis, diastoolne 80 ringis, vahel 66 vahel, eile õhtul oli nägu valge (mõõt is pulssi), eks raviks praegu valsacombi 160 / 12, 5 mg x 1, bisoprolol 10 mg x 1, amiodar on pole võtnud, aga kuna arütmia olnud kuu jooksul, hetkel ravimeid ei tarvita, ise seostab vererõhu probleemi glükoosi liiga kõrge olnud, m ära g selt jättis ravi ära, DATE öde xxx lõõ bas : teostatud ekg rr 140 / 100 mmhg pikkus 186 cm, kaal 113, 1 kg kmi 33, 1 v / s 7, 7 mmol

Joonis 9. tunnuse X16_L11#1129 näited andmestiku tekstilõikudest.



Joonis 10. tunnuse X16_L11#1129 histogramm aktivatsiooni tugevuste jaotusest. Tunnusel on tugevate aktivatsioonide arvu kohta harilikult vähem nõrkasid aktivatsioone.

Tunnus X16_L11#23: Füüsiline liikumine

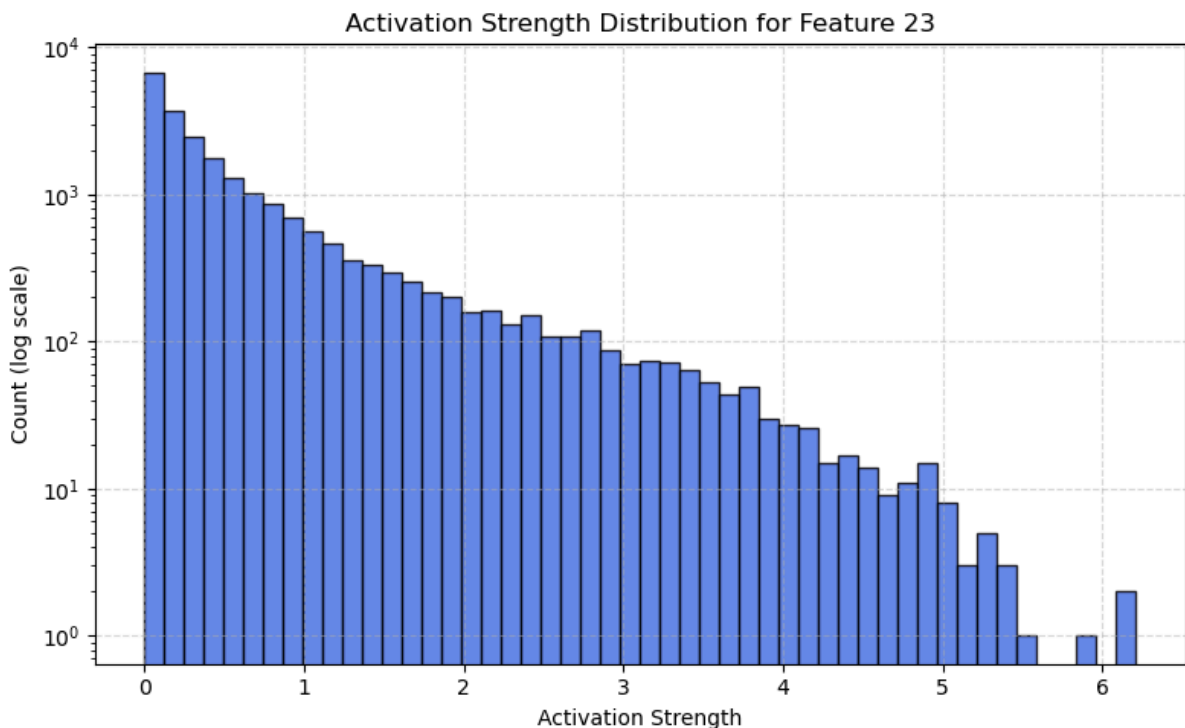
norm 37, 0 - 47, 0] DATE : leukotsüüdid 4, 8 e NAME : kaal 67 kg, pikkus 155 cm, rr 178 / 88 pulss 65, ei suitseta, ravimid võtab regulaarselt, teeb aiatööd, käib jalgsi, võetud vered, tehtud ekg, antropomeetria, eluanamnees, nad ekg - s siinusrütm, fr 52, repol, häiret ei esine, kaalutud, mõõdetud, kmi 24, 2, ta - 77 kg, pikkus - 153 cm, v, ü - 89 cm, puusa ümb - 104 cm, kmi - 32, 5, pt, helistab ja küsib analüüside vastuseid, analüüsid korras, nõustatud toitumise osas ja füüsilise aktiivsuse

NAME : pt tuli prof vv - le ja analüüsidele, rr 152 / 68 fr 65 x / min ravimid : t, roswera 5 mgx 1 ja co - prenessa 2 / 0, 625 mgx 1 + novorapid flex pen 2 - 8 tü vajadusel, pulmon ooloogil käis kevadel, midagi ei leitud, vaid siis, kui tõ sist kõha, töötab mü ma ra de müüjana, ei suitseta, pii mat ooteid pole 2 aastat kasutanud, kuna see ei sobi, füüsi lisi kaebusi hindab rahuldav aks, kon ts aga liigub päevas korra, kodus kõnnib mitu km 2 - 3 x

7 pt re al iop 13 / 13 mmhg jääb sama glaukoomi ravi NAME : kontrol lan al s kaebused sama vr vahel valud südame piirkonnas ja tor kav valu vasaku roidekaare all peale raskuse tõstmist (pa ele tugevat füüsilist pingutust) pt ülekaaluline lapsepõlvest , iga kuu 80 kg , käib jalutamas kuni 1 km . ravi tarvitab pidevalt regulaar selt . vr 170 / 90 ps 93 , südametoonid regulaarsed , tuhmid , kops ii , jalad turseta . konsulteeritud dr NAME (DATE) : pt dm ii - arteriaalne hüpertoonia , hüperkolesteroleemia , kardiostimulaator , hüpötüreos , eutüreoidne kilpnääre sõlmeline , uh - s üks sagar DATE mm x 0 , 76 mm x 2 , 99

NAME : teostatud kroonilise haige nõustamine elustiili osas , toitumise ja elustiili osas - 5 kg alla võtnud , füüsiline koormus igapäevaselt mõõdukas , 1 - 2 x nädalas sööb tootumise kava . ta mõõdukas tempos meeldib kõndida 20 - 30 min päevas . kahjulikud harjumised : ei suitseta , alkoholi tarbib kaks korda nädalas 1 - 2 klaasi veini , väga harva kanget alkoholi . viimati alkoholi tarvitamisel õi u 5 - 6 p eest . toitumine mitme kes ine , teadlik tervis li kest eluvii sidest .

Joonis 11. tunnuse X16_L11#23 näited andmestiku tekstilõikudest.



Joonis 12. tunnuse X16_L11#23 histogramm aktivatsiooni tugevuste jaotusest. Näha on tugevaimate aktivatsioonide ees olevat aktivatsioonide langust.

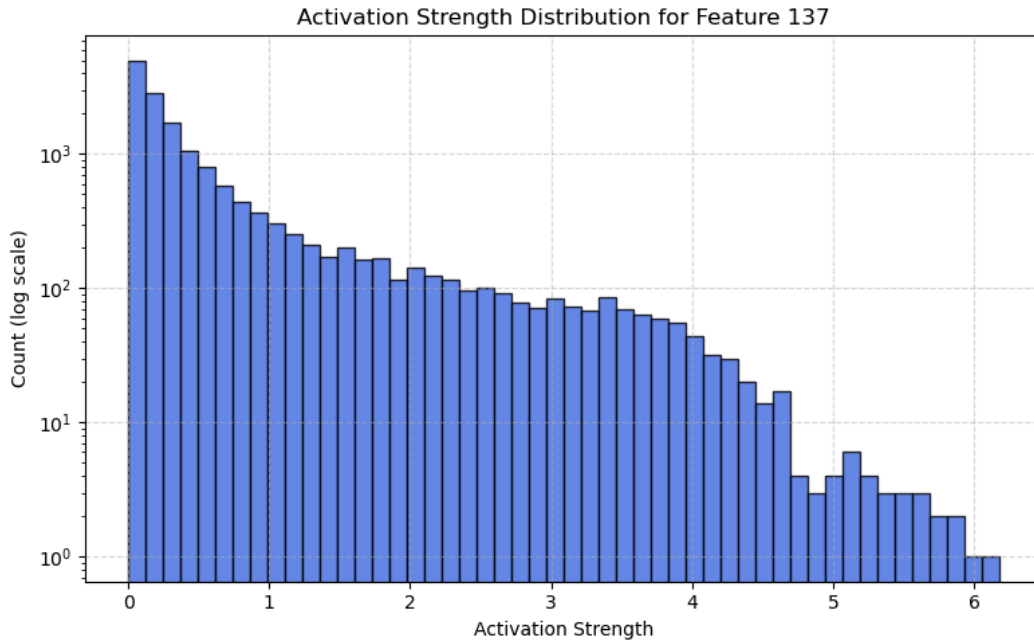
Tunnus X16_L11#137: Pereliikmete diagnoosid. Kõige tugevamalt aktiveerus tunnus pereliikmete endi peale ning keskmiselt nendega seostuva terviseprobleemi peale.

harjumused ? suitsetamine ? eipäriikkus : isa emal rinnavähk , emaalma maovähk - surmud , pahal _____ kuupäev hk hinnakirja kood monot . 0

- DATE - DATE menarhe : 14 a , 7 / 28 / regulaarne , verer o hk usega , mitte iga kord s = 2 a = 0 põetud haigused : angiinid , hüpertoonia -- saab ravi , põiepõletik ,, emakakaela haavand ,, ---- > kolposkoopia , kontrollib 2 x aastas suguhaigused : kla müdi oos ravitud . vereülekanded : ei allergia : ei suitsetamine : ei tbc - ei diabeet - ei kollatõbi - ei tbc - ei väh kkasvaja perekonnas - isa emal oli emaka ca , isal oli kopsu ca , emal rinna ca , onul prostata ca

NAME : tuli kontrolli . rvv jeanine , sobib . v . m . DATE , aga veritsus algas 1 , 5 näd tagasi . tuli kontrolli . r - 0 xxx
nilm (09 . 2015) stli - d : kiamüüdia aprill DATE . gün . haigused : eikr . haigused :- ravimid :- allergia ravimitele :-
operatsioonid :- pärlikkus : emal : rinna ca : vanaemal munasarja ca ei suitseta . st . gyn : välisgenitaalid ii . ta vaba ,
puhas . portio terve . emakas norm suurust , valutu , liikuv . adnekside pk vaba , valutu . uh : emakas 59 x 45 mm ,
endomeetrium

Joonis 13. tunnuse X16_L11#137 näited andmestiku tekstilõikudest.



Joonis 14. tunnuse X16_L11#137 histogramm aktivatsiooni tugevuste jaotusest. Tunnusel langeb aktivatsiooni väärtused lähenedes viiele, kuid seejärel veel tõusevad. Need tugevaimad on mõisted pereliikmete kohta.

Osa tunnuseid on veel lisades 4 ja 5.

Kokkuvõte

Töö eesmärk oli treenida GPT-2 arhitektuuril keelemudeli põhjal hõre autoenkooder ning eraldada selle mudeli neuroni aktivatsioonidest tunnused, millest saab tõlgendada vastavad kontseptsioonid. Tunnuse ning nende läbi kontseptsioonide eraldamine annab parema mõistmise keelemudeli käitumisse, mis suurendab mudeli usaldusväärsust ning avab võimaluse mudeli tööd parandada läbi tunnuste võimendamise.

Selle töö käigus siiski ei jõutud tunnuste võimendamist läbi viia. See osa seati alguses lõppeesmärgiks, kui kontseptsioonide eraldamine on saavutatud ning aega jääb üle, kuid sellele eelnevate sammude täitmine osutus piisavalt mahukaks ülesandeks. Tunnuste võimendamine ning mudeli käitumise parandamine jääb tulevaseks edasiarenduseks.

Töö raames treeniti kolm hõredat autoenkooderit, mis erinesid üksteisest tunnuste arvu ning mudeli kihi, millest autoenkooder sai mudeli neuronite aktivatsioonid, poolest. Treenimise andmetena kasutati sünteetilist andmestikku, mis koosnes eestikeelsetest haigusloo sissekannetest. Kahel autoenkooderil olid negatiivsed küljed, mis välistasid nad edasisest analüüsist: ühe tunnused olid raskesti tõlgendatavad ning teisel oli elavate tunnuste määr madal. Kolmest autoenkooderist valiti välja üks, millest saadud tunnuseid hakati põhjalikumalt analüüsima. Analüüsis uuriti tugevalt aktiveerivaid tokeneid, aktivatsiooni konteksti teksti lõigus ning aktivatsiooni jaotust üle kõigi tokenite. Mudelist leiti kontseptsioone, mis tugevalt sarnanesid algselt pakutud potentsiaalsetele kirjeldustele, nagu patsienti mõjutavad spetsiifilised terviseprobleemid, patsiendi füüsiline aktiivsus, ravi positiivne kulg.

Analüüsi viidi läbi manuaalselt, kuid edasiarendusena saaks peenhäälestada keelemudelit tunnuste analüüsimise jaoks ning rakendada seda automaatselt kontseptsiooni kirjeldusi pakkuma.

Töö viidi läbi vastavalt TÜ eetikakomitee ja Eesti bioetika ja inimuuringute nõukogu lubadele (load nr 300/T-23 ja 1.1-12/3088) ning projektide TEM-TA72 ja PRG1844 raames. Projekt TEM-TA72 on rahastatud Euroopa Liidu ja kaasrahastatud Haridus- ja Teadusministeeriumi poolt. Projekt PRG1844 on rahastatud Eesti Teadusagentuuri poolt.

Viidatud kirjandus

Ayonrinde, K. *Dictionary Learning with Sparse AutoEncoders* Kola Ayonrinde Blog. 2023

<http://www.kolaayonrinde.com/blog/2023/11/03/dictionary-learning.html>

(20.03.2025)

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C.,

Denison, C., Askeel, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T.,

Joseph, N., Hatfield-Dodds, Z., Tamkin, A., Nguyen, K., ... Olah, C. (2023). Towards

Monosemanticity: Decomposing Language Models With Dictionary Learning.

Transformer

Circuits

Thread.

<https://transformer-circuits.pub/2023/monosemantic-features/index.html> (13.05.2025)

Buck, C., Heafield, K., & van Ooyen, B. (2014). N-gram Counts and Language Models from

the Common Crawl. N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard,

J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Toim), *Proceedings of the Ninth*

International Conference on Language Resources and Evaluation (LREC'14) (lk

3579–3584). European Language Resources Association (ELRA).

<https://aclanthology.org/L14-1074/> (13.02.2025)

Deelaka, N. (2024, jaanuar 30). What “Dictionary Learning” actually is? *Analytics Vidhya*.

[https://medium.com/analytics-vidhya/what-dictionary-learning-actually-is-812d264e9](https://medium.com/analytics-vidhya/what-dictionary-learning-actually-is-812d264e9646)

646 (09.05.2025)

Dey, S. (2024, mai 31). *Deep Dive into Anthropic's Sparse Autoencoders by Hand*. Medium.

<https://medium.com/data-science/deep-dive-into-anthropics-sparse-autoencoders-by-h>

and-%EF%B8%8F-eebe0ef59709 (13.05.2025)

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z.,

Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D.,

Wattenberg, M., & Olah, C. (2022). Toy Models of Superposition. *Transformer*

- Circuits Thread*. https://transformer-circuits.pub/2022/toy_model/index.html
(26.03.2025)
- Holdsworth, J., & Scapicchio, M. (2024, juuni 17). *What Is Deep Learning? | IBM*.
<https://www.ibm.com/topics/deep-learning> (15.12.2024)
- Karvonen, A. (2024, juuni 11). An Intuitive Explanation of Sparse Autoencoders for LLM Interpretability [Blog]. *Adam Karvonen's Blog*.
https://adamkarvonen.github.io/machine_learning/2024/06/11/sae-intuitions.html
(24.04.2025)
- Krishna, S. (2014, juuli 11). *What is the difference between complete basis and overcomplete basis?*.
<https://www.quora.com/What-is-the-difference-between-complete-basis-and-overcomplete-basis> (28.03.2025)
- Lan, M., Torr, P., Meek, A., Khakzar, A., Krueger, D., & Barez, F. (2025). *Sparse Autoencoders Reveal Universal Feature Spaces Across Large Language Models* (No. arXiv:2410.06981). arXiv. <https://doi.org/10.48550/arXiv.2410.06981> (24.04.2025)
- Lindsey, J., Gurnee, W., Ameisen, E., Chen, B., Pearce, A., Turner, N. L., Citro, C., Abrahams, D., Carter, S., Hosmer, B., Marcus, J., Sklar, M., Templeton, A., Bricken, T., McDougall, C., Cunningham, H., Henighan, T., Jermyn, A., Jones, A., ... Batsion, J. (2025, märts 27). *On the Biology of a Large Language Model*. Transformer Circuits.
<https://transformer-circuits.pub/2025/attribution-graphs/biology.html#dives-medical>
(15.05.2025)
- Markowitz, D. (2021, mai 6). *Transformers, Explained: Understand the Model Behind GPT-3, BERT, and T5*. <https://daleonai.com/transformers-explained> (18.12.2024)
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language*

Models are Unsupervised Multitask Learners. (17.12.2024)

Raschka, S. (2025, veebruar 5). *Understanding Encoder And Decoder LLMs.*

<https://magazine.sebastianraschka.com/p/understanding-encoder-and-decoder>

(17.04.2025)

Reberga, L. (2025, märts 18). *Interpretability of LLMs: The Role of Sparse Autoencoders | Aqsome.*

<https://www.aqsome.com/en/blog/interpretability-of-llms-the-role-of-sparse-autoencoders> (27.03.2025)

Samuel Marks, Adam Karvonen, & Mueller, A. (2024). *Dictionary learning.*

https://github.com/saprmarks/dictionary_learning

Stryker, C., & Scapicchio, M. (2024, märts 22). *What is Generative AI? | IBM.*

<https://www.ibm.com/topics/generative-ai> (15.12.2024)

Sügis, E., Tampuu, A., Aljanki, A., Fišel, M., & Kull, M. (2024). *Praktiline andmeteadus. Kõrgkooliõpik.* Tartu Ülikooli arvutiteaduse instituut.

https://courses.cs.ut.ee/t/andmeteadus/Main/HomePage?action=download&upname=Praktiline_andmeteadus_ko%CC%83rgkoolio%CC%83pik.pdf

Temple, S. *Stop Confusing AI with Generative AI.* Medium. 2024

<https://generativeai.pub/stop-confusing-ai-with-generative-ai-understanding-the-key-differences-21da2b2d3374> (15.12.2024)

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N. L., McDougall, C., MacDiarmid, M., Freeman, C. D., Sumers, T. R., Rees, E., Batson, J., Jermyn, A., ... Henighan, T. (2024). Scaling Monosemanticity: Extracting Interpretable Features

from Claude 3 Sonnet. *Transformer Circuits Thread.*

<https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html>

(07.05.2025)

Väljas, M. (s.a.). *Vektorruumi baas ja mõõde.*

https://staff.ttu.ee/~mati.valjas/3710/loeng_06.pdf

What are Large Language Models? | A Comprehensive LLMs Guide. (s.a.).

<https://www.elastic.co/what-is/large-language-models> (15.12.2024)

What Are Large Language Models (LLMs)? | IBM. 2023.

<https://www.ibm.com/think/topics/large-language-models> (15.12.2024)

What Is Artificial Intelligence (AI)? | IBM. 2024.

<https://www.ibm.com/think/topics/artificial-intelligence> (15.12.2024)

What is Deep Learning? - Deep Learning AI Explained - AWS. (s.a.). Amazon Web Services,

Inc. Salvestatud, <https://aws.amazon.com/what-is/deep-learning/> (15.12.2024)

Wolfe, C. R. (2024, märts 4). Decoder-Only Transformers: The Workhorse of Generative

LLMs [Substack newsletter]. *Deep (Learning) Focus.*

<https://cameronrwolfe.substack.com/p/decoder-only-transformers-the-workhorse>

(17.04.2025)

Yang, J., Jin, H., Tang, R., Han, X., Feng, Q., Jiang, H., Zhong, S., Yin, B., & Hu, X. (2024).

Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond. *ACM*

Trans. Knowl. Discov. Data, 18(6), 160:1-160:32. <https://doi.org/10.1145/3649506>

Lisad

Lisa 1. Keelemudeli GPT-4o viip ning väljund termini mitmekihiline pertseptron jaoks.

Viip	Väljund	Link vestlusele
Palun defineeri termin mitmekihiline pertseptron eesti keeles kuni kahe lausega.	Mitmekihiline pertseptron (MLP) on tehisnärvivõrk, mis koosneb sisendkihist, ühest või enamast peidetud kihist ja väljundkihist, kus iga neuron on ühendatud järgmise kihi kõikide neuronitega. See suudab õppida keerukaid seoseid andmetes, kasutades mittelineaarseid aktivatsioonifunktsioone ja tagasi-levi algoritmi.	https://chatgpt.com/share/6812526c-7ad4-8006-8468-6527dd24eb8f

Lisa 2. Kasutatava keelemudeli detailne arhitektuur

```
GPT2LMHeadModel(  
  (transformer): GPT2Model(  
    (wte): Embedding(50258, 768)  
    (wpe): Embedding(1024, 768)  
    (drop): Dropout(p=0.1, inplace=False)  
    (h): ModuleList(  
      (0-11): 12 x GPT2Block(  
        (ln_1): LayerNorm((768,)), eps=1e-05, elementwise_affine=True)  
        (attn): GPT2Attention(  
          (c_attn): Conv1D()  
          (c_proj): Conv1D()  
          (attn_dropout): Dropout(p=0.1, inplace=False)  
          (resid_dropout): Dropout(p=0.1, inplace=False)  
        )  
        (ln_2): LayerNorm((768,)), eps=1e-05, elementwise_affine=True)  
        (mlp): GPT2MLP(  
          (c_fc): Conv1D()  
          (c_proj): Conv1D()  
          (act): NewGELUActivation()  
          (dropout): Dropout(p=0.1, inplace=False)  
        )  
      )  
    )  
    (ln_f): LayerNorm((768,)), eps=1e-05, elementwise_affine=True)  
  )  
  (lm_head): Linear(in_features=768, out_features=50258, bias=False)  
  (generator): WrapperModule()  
)
```

Lisa 3. Koodi repositoorium

https://github.com/JuhanPauklin/dictionary_learning

Lisa 4. Tunnus X16_L11#38

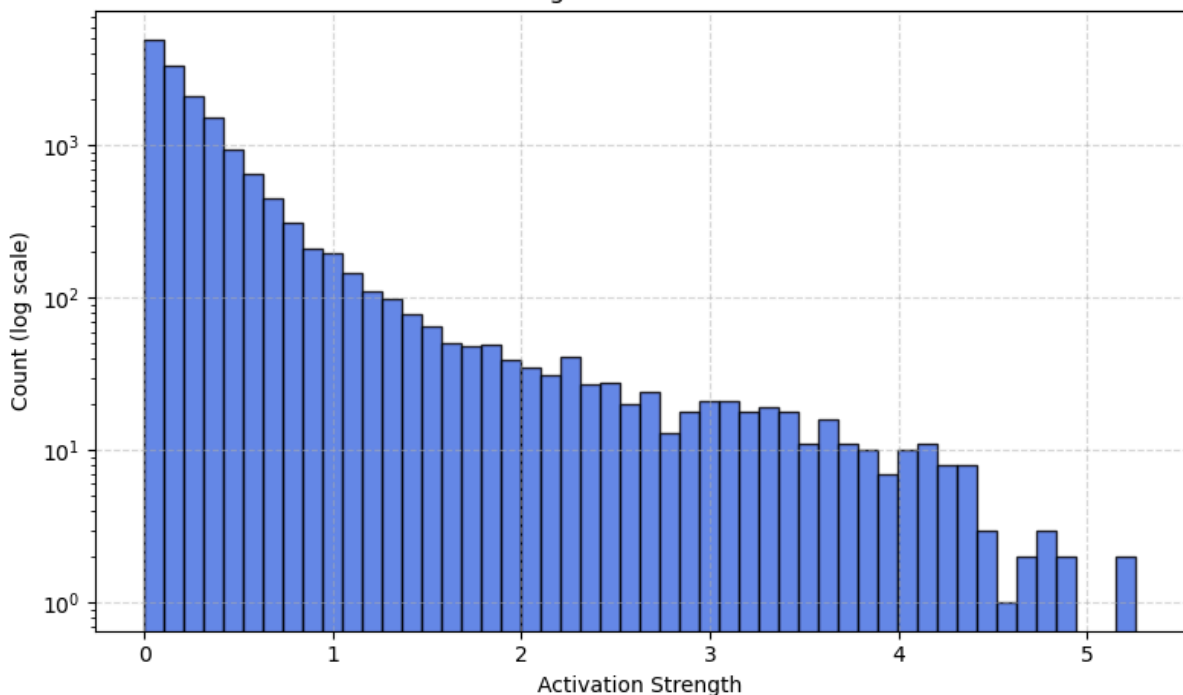
Tõved, mis on nimetatud isikute järgi

- probleemiks mäluhäired, mis on tekkinud viimase aasta jooksul. viibis uuringutel NAME haigla neuroloogia osakonnas: põhihaigused: parkinsonistlik sündroom, täpsustamata kõnehäire, mõõdukas depressioon somaatiliste sümptomitega, gastriit, kr. gastriit. patsiendil on varem diagnoositud ja paatiat dementsuse poolt, kuid varasemalt on diagnoositud alzheimeri tõbi, kus oli valdavalt alzheimeri tõve sensoorse komponendiga. kaasuvait parkinsoni tõbi, DATE. a. implanteeritud kardistimulaator. ei ole enam teadvuse kaotusi olnud. k. t. peast on DATE. a. haigusliku leiuta. ultraheli uuringul kaelaarterite hemodünaamiliselt olulisi stenose ei esine. mrt uuringul suuraju hemisfäärides f.

NAME: kaebuseks umbes 10 päeva jooksul valulikkus vasaku roidekaare all. pt sõnul valu kiirgub nii vasakule kui ka paremale roiete prk - da. ta tuim, pidev, intensiivistub liigutamisel, liikumisel. varasemalt sarnaseid valu episoodide esinenud ei ole. valuvaigistit kasutanud ei ole. iiveldust ja oksendamist eitab. seedetegevus regulaarne, väljaheite värvus tavapärane. urineerimine tavapärane. varasemalt diagnoositud maksas hemangioomid, crohni tõbi. obj: palpatoorselt valulikkus vasakul rindkeres roidevahe mik es, aksillaarselt, rinnaku kõrval. DATE: diafragma kaarjas, lateraalsiinused vabad, hiilused struktuursed. kopsuväljad transparentsed, koldelisi - infiltratiivseid muutusi ei sedasta. ta ristimõõdus laienenud.

ohu annab ulemise piiri. haige kassas on veidi viiruse esinenud. haiguse anamnees: xxx. DATE põdes viirusinfektsiooni, sellejärgselt jäi püsima kõha. anamneesis sjögreni sündroom, reumatoloogi jälgimisel. viimane konsultatsioon 11 / DATE: uriinis valg 1.0 g / l, kvantitatiivselt 5 g / l, kvantitatiivselt norm kuni 2 g / l. DATE kopsu rõkkorras. hetkel saab imuran 50 mg x 1. kaasuvana astma, raviks pulmicort th 200 mcg x 2 ja vajadusel ventolin aer. 07 / DATE usg - l maksasteatoos, kahtlus kivile. neerude kogumissüsteemis kivi nähtavale ei tule. eluanamnees:

Activation Strength Distribution for Feature 38

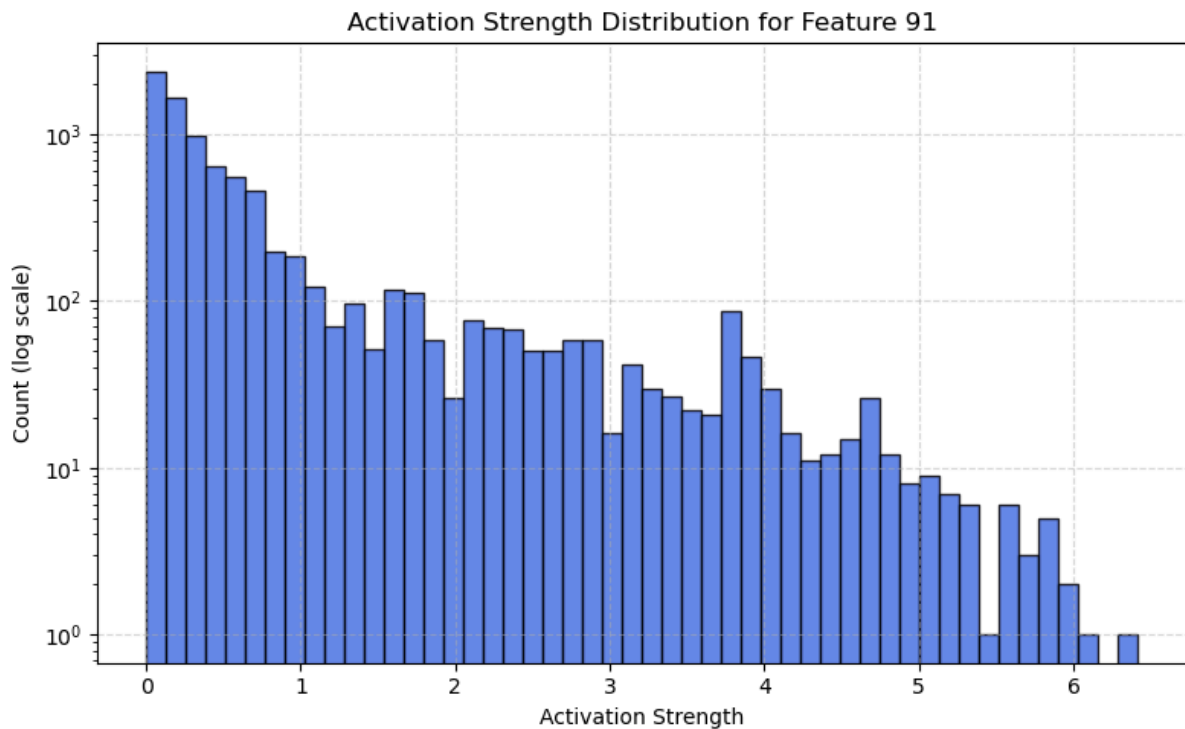


Lisa 5. Tunnus X16_L11#91:

Uuringud nagu kompuuteranalüüs, ehkardiograafia, koormustest, ultraheliuuring jne. Huvitaval kombel tunnuse ei sisalda elektrokardiograafiat.

hüpfüüs t 1 kõrge signaaliga - võimaliku mikro aden oomiga võimalik mujal patoloogilist signaalimuutust kuupäev : DATE kood ja nimetus : 6323 elektrokardiograafia koos kompuuteranalüüsiga põhjendus : int kirjeldus : siinustahhükardia 103 x min . madal voltaaz . kuupäev : DATE kood ja nimetus : 6340 täismahus ehkardiograafia põhjendus : palat 11 kirjeldus : vasak vatsake normaal mõõtmeline , normaalse summa arse ja segmentaarse süstoolse funktsiooniga , ef 65 % simpsoni j . parem vatsake normaal mõõtmeline , normaalse süstoolse funktsiooniga . kojad normaal mõõtmelised . aort sklerootiline . 3 - hõlmaline ak , hõlmad kaltsin ootilised , esineb kerge aordiklapi puudulikkus . mitraalklapi hõlmad kergelt fibrootilised , eesmine hõlm prolabe eruv . esineb kerge mitraalne ja trikuspidaalne regurgitatsioon . ph ei esine . ivc ei ole dilateerunud .

kuupäev : DATE kood ja nimetus : 6323 elektrokardiograafia koos kompuuteranalüüsiga põhjendus : prof . kirjeldus : siinusrütm fr . 76 xmin . horisontaalne ei positsioon . vas . vatsakese hüpertroofia . repolarisatsiooni häired külgsena ülaosas . kuupäev : DATE kood ja nimetus : 6324 elektrokardiograafia koormustest (veloergomeetril , koormusrajal) põhjendus : prof . kirjeldus : DATE , 12 : 05 : 04 , exercise test : NAME : total exercise NAME DATE NAME hr : 162 bpm 95 % of max predicted 165 bpm hr at rest : 82 NAME bp : 210 / 100 mmhg bp at rest : 130 / 70 NAME rpp : 27 260 mmhg bpm



Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Juhan Pauklin,

1. Annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose “Terviseandmetel treenitud keelemudelist kontseptsioonide eraldamine”, mille juhendaja on Raivo Kolde, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada Tartu Ülikooli digitaalarhiivi kuni autoriõiguse kehtivuse lõppemiseni;
2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi kaudu Creative Commons'i litsentsiga CC BY NC ND 4.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni;
3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile;
4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Juhan Pauklin

13.05.2025