



**TURUN
YLIOPISTO**

SUURET KIELIMALLIT LÄÄKETIETEELLISTEN LAUSUNTOJEN
RAKENTEELLISTAMISESSA

Ida-Maria Kaleva

Pro gradu -tutkielma
Kesäkuu 2025

Tarkastajat:
Prof. Ion Petre
Dos. Teemu D. Laajala

MATEMATIIKAN JA TILASTOTIETEEN LAITOS

Turun yliopiston laatujajestelmän mukaisesti tämän julkaisun alkuperäisyys on tarkastettu Turnitin OriginalityCheck-järjestelmällä

TURUN YLIOPISTO

Matematiikan ja tilastotieteen laitos

IDA-MARIA KALEVA: Suuret kielimallit lääketieteellisten lausuntojen rakenteellistamisessa

Pro gradu -tutkielma, 53 s., 8 liites.

Sovellettu matematiikka

Kesäkuu 2025

Tutkielma käsittelee lääketieteellisten lausuntojen rakenteellistamista suurten kielimallien avulla sekä teoriassa että käytännössä. Rakenteellistaminen on luonnollisen kielen käsittelyn tehtävä, josta voidaan myös käyttää nimitystä tiedonpoimintatehtävä. Tutkielmassa esitellään luonnollisen kielen käsittelyn perusperiaatteita ja tämän jälkeen erityisesti tiedonpoimintatehtävien piirteitä ja niiden erilaisia ratkaisutapoja. Koska tässä tutkielmassa ratkaistaan lääketieteen alan tutkimuskysymys, käsitellään tutkielmassa myös luonnollisen kielen erityispiirteitä lääketieteellisessä esimerkkisovelluskohteessa.

Tutkielmassa esitellään suurten kielimallien kehitystä ja historiaa, niiden rakennetta ja eri ominaisuuksien merkitystä suurten kielimallien toiminnassa. Tämän tutkielman implementointiosuudessa hyödynnetään kehotesuunnittelua. Tämän takia esitellään myös sen perusajatus ja juuri tämän tutkielman implementointiosuudessa hyödyllisiä tekniikoita. Lopuksi huomioidaan vielä suurten kielimallien haasteet, sekä käyttäjän näkökulmasta että yhteiskunnallisella tasolla.

Tutkielman implementointiosuudessa tarkastellaan neljän suuren kielimallin (GPT-3.5, GPT-4, Llama3-70b ja Gemini1.0-pro-001) suoritusta lääketieteellisten lausuntojen rakenteellistamisessa. Päähuomio on mallien suoritusten toistettavuudessa, oikeellisuudessa ja kehoitteiden vaikutuksessa mallien toimintaan. Lopuksi esitellään tutkielmassa saadut tulokset ja esitetään joitakin mallikohtaisia kehotesuunnitteluehdotuksia mallien suoritusten parantamiseksi.

Asiasanat: NLP, luonnollisen kielen käsittely, tiedonpoimintatehtävä, LLM, suuret kielimallit, kehotesuunnittelu, rakenteellistaminen, eturauhassyöpä, histopatologia, Gleason

Sisälllys

1 Johdanto	1
2 Luonnollisen kielen käsittely	4
2.1 Perusperiaatteet	4
2.2 Tiedonpoimintatehtävä	5
2.3 Tiedonpoimintatehtävän ratkaiseminen	7
2.4 Luonnollinen kieli elektronisissa potilastietokannoissa	9
3 Suuret kielimallit	12
3.1 Historia	12
3.2 Transformer-arkkitehtuuri	18
3.3 Enkooderi- ja dekooderipohjaiset mallit	23
3.4 Kehotesuunnittelu	25
3.5 Suurten kielimallien haasteet	27
4 Implementointi	30
4.1 Implementoinnin lähtökohdat	30
4.2 Mahdolliset lähestymistavat	33
4.3 Ratkaisuyritys	34
5 Tulokset ja tulkinta	37
5.1 Toistettavuus	38
5.2 Oikeellisuus	41
5.3 Mallien vertailu	48
6 Yhteenveto	52
Viitteet	54
Liitteet	61
A Tutkielman implementointiosuudessa käytetyt kehotteet	61
B Esimerkkejä saaduista vastauksista	67

1 Johdanto

Suuret kielimallit ovat tekoälyn osa-alue. Tekoäly ja suuret kielimallit ovat keskeisessä asemassa monella alalla niiden kehittyessä nopeasti ja samalla niiden mahdollistaessa uusia ratkaisutapoja haastaviin ongelmiin. Uusien ratkaisutapojen lisäksi ne tarjoavat ratkaisuja ongelmiin, joita ei aiemmin ollut mahdollista ratkaista resurssien saatavuuden takia, tai niiden ratkaiseminen oli erittäin haastavaa tai kallista. Keskeinen ero aikaisempiin tekniikoihin ja rakenteisiin on suurissa kielimalleissa käytettävä transformer-arkkitehtuuri, joka on myös eräänlainen keinotekoinen neuroverkko.

Rakenteellistaminen on luonnollisen kielen käsittelyn tehtävä, jossa epärakenteellisesta tiedosta pyritään saamaan rakenteellista dataa keräämällä alkuperäisestä aineistosta ennalta määrättyjä tietoja järjestelmällisesti. Tiedon rakenteellistaminen on keskeistä erityisesti raakadatan jalostamisessa. Lääketieteessä epärakenteellinen tieto voi tarkoittaa esimerkiksi lääketieteellisiä vapaatekstilausuntoja, joiden kirjoitus- ja merkintätavat saattavat erota suurestikin toisistaan. Tällöin niiden sisältämien tietojen yhtenäistäminen voi vaatia paljon manuaalista työskentelyä ja tietotaitoa, mikä ei usein ole mahdollista. Suurten kielimallien avulla tiedon rakenteellistaminen voidaan automatisoida ja näin lääketieteellisissä lausunnoissa esiintyvää tietoa voidaan hyödyntää edelleen visualisoinnissa ja mallintamisessa. Rakenteellistettu vapaatekstilausunto voisi olla esimerkiksi: "potilaan ikä: 55; potilaan PSA-arvo:10; potilaan diagnoosit:tyypin 2 diabetes, kohonnut verenpaine".

Perinteisessä ohjelmoinnissa haastavat tehtävät paloittellaan pienemmiksi, tarkemmin määritellyiksi, tehtäviksi ja koneelle annetaan ohjeet tehtävän ratkaisemiseksi. Keinotekoiset neuroverkot taas oppivat havainnoimastaan datasta ja rakentavat sen perusteella oman ratkaisunsa tehtävälle. Käytetään keinotekoisista neuroverkoista tästä eteenpäin vain nimitystä neuroverkot. Ihmisäivot pystyvät käsittelemään haastavia ja epäjohdonmukaisiakin aineistoja ja muodostamaan usein oikeita johtopäätöksiä moniulotteisen aineiston, esimerkiksi kuvien, perusteella. Tällaista toimintaa pyritään jäljittelemään myös koneille neuroverkkojen avulla. [33] [35] Neuroverkot koostuvat kerroksista, ja usean kerroksen neuroverkot voivat oppia monimutkaisiakin rakenteita sille annetusta syötteestä eli aineistosta. [2] Usean kerroksen neuroverkkoja hyödyntäviä menetelmiä kutsutaan syväoppimiseksi. Neuroverkot mahdollistavat monimutkaisten, kuten matemaattisesti vaikeasti määriteltävien, tehtävien ratkaisemisen ja ne ratkaisevatkin hyvin myös sellaisia tehtäviä, joiden ratkaisut eivät ole yksiselitteisiä. Tällaisia ovat esimerkiksi erilaiset tunnistamiseen, lajitteluun ja ennusteisiin liittyvät tehtävät.

Neuroverkkojen opettaminen on keskeisessä roolissa niiden toiminnassa. Opetus-

prosessissa tavoitteena on, että neuroverkkomalli oppii yleistettyjä periaatteita sille annetun aineiston pohjalta. [12] Neuroverkkomallin opettamisessa käytetään virhefunktioita, joka kuvaa saadun ulostulon ja halutun ulostulon välistä eroa. Tavoitteena on minimoida tätä virhefunktioita, jolloin saadut ulostulot ovat mahdollisimman lähellä oikeita vastauksia. [2] Opetustapa valitaan tilannekohtaisesti riippuen saatavilla olevista resursseista, kuten ajasta ja merkityn datan määrästä. Huomioitavaa on myös mallin käyttötarkoitus.

Neuroverkkkoa opettaessa aineisto jaetaan tavallisesti opetus-, validointi- ja testiaineistoon. Opetusaineistolla neuroverkkkoa opetetaan tietyn tehtävän ratkaisemiseen. Validointiaineiston avulla mallin toimintaa testataan ja sen parametrejä säädetään. Testiaineistolla taas voidaan tarkistaa neuroverkon kyky ratkaista samankaltaisia tehtäviä. Testiaineisto koostuu datasta, jota ei ole käytetty mallin opetuksessa. [2]

Ohjatussa oppimisessa (eng. supervised learning) nimensä mukaisesti voidaan ajatella, että jokin "ohjaa" mallia opetusprosessissa. Siinä tavoitteena on oppia kuvaus syötteeltä ulostulolle, kun oikeat vastaukset ovat tiedossa. Ohjatussa oppimisessa mallilla on opetusprosessissaan käytettävissään siis sekä syötteet että oikeat ulostulot ja koneoppimisalgoritmin tehtävänä on löytää sellainen funktio, joka arvioi kuvausta syötteeltä ulostulolle mahdollisimman hyvin. [12] Haasteena ohjattua oppimista käytettäessä on kuitenkin se, että koska malli tarvitsee oppiakseen myös oikeat ulostulot, on merkittävää dataa oltava riittävästi. Merkittävää dataa tarkoittaa aineistoa, joka on merkittävää tunnistetta. Usein tämä tehdään manuaalisesti ja datan merkitseminen taas vie paljon resursseja. [2]

Ohjaamattomassa oppimisessa (eng. unsupervised learning) malli pyrkii löytämään syötteestä säännöllisyyksiä. Käytettävissä ei ole oikeita ulostuloja, vaan pelkästään syöte. Tavoitteena on ymmärtää syötekerroksen rakennetta ja sen perusteella saada selville mitä yleensä tapahtuu tietyssä tilanteessa ja mitä taas toisessa. Ohjaamatonta ja ohjattua oppimista voidaan myös käyttää yhdessä niin, että ensin ohjaamatonta oppimista käytetään alkuarvon määrittämiseksi, jonka jälkeen voidaan käyttää ohjattua oppimista opetusprosessissa. Näin prosessi vaatii vähemmän aikaa ja merkittävää dataa kuin pelkkää ohjattua oppimista käytettäessä. [2]

Vahvistusoppimisessa (eng. reinforcement learning) mallille annetaan palautetta sen toiminnasta sen mukaan, kuinka hyvin se on suoriutunut. Toiminta nähdään tällöin yksittäisten päätösten tai toimintojen summana ja päätös tai toiminto on hyvä, jos se on osa hyvää toimintatapaa. Malli on toimija (eng. agent), joka tekee päätöksiä ympäristön kanssa käydyn vuorovaikutuksen perusteella siirtyen tilasta (eng. state) toiseen suorittamalla toimintoja (eng. action). Mallille annetaan positiivista

tai negatiivista palautetta sen toiminnan mukaan ja tämän palautteen perusteella malli muuttaa toimintaansa. Käytännössä palaute tarkoittaa matemaattisesti jotakin funktiota, joka määritellään tehtäväkohtaisesti. Näin pyritään lopulta tilanteeseen, jossa malli maksimoi positiivista palautetta ja näin löytää parhaan mahdollisen toimintatavan. Vahvistusoppimisessa mallille ei kerrota mitä sen tulee tehdä vaan kuinka hyvin se onnistuu tekemässään. [2]

Tässä tutkielmassa esitellään luonnollisen kielen käsittelyn perusteet ja tämän lisäksi käsitellään tarkemmin tiedonpoiminta- eli rakenteellistamistehtävää ja sen ratkaisemista. Tutkielmassa käsitellään suuria kielimalleja sekä teoriassa että käytännössä. Ensin käydään läpi suurille kielimalleille keskeisen transformer-arkkitehtuurin rakennetta ja toimintaa. Tämän jälkeen käsitellään suurten kielimallien toimintaa käyttäjän näkökulmasta esittelemällä enkooderi- ja dekooderipohjaisten mallien toimintaa, kehoitesuunnittelua ja suurten kielimallien haasteita.

Lopuksi tutkielmassa esitellään käytännön sovellus suurten kielimallien käytöstä lääketieteellisten lausuntojen rakenteellistamisessa. Tutkielman implementointiosuudessa käsitellään histopatologisia eturauhassyövästä kirjoitettuja potilaslausuntoja, joista pyritään löytämään ennalta määrättyjä tietoja neljän suuren kielimallin (GPT-3.5, GPT-4, Gemini1.0-pro-001 ja Llama3-70b) avulla kehoitesuunnittelua käyttäen. Tutkielmassa käytetyt potilaslausunnot ovat aidon kaltaisia synteettisiä lausuntoja. Tavoitteena on vertailla suurten kielimallien toimintaa rakenteellistamistehtävässä ja tämän lisäksi tarkastella kehoitesuunnittelun merkitystä suurten kielimallien suorituksissa.

2 Luonnollisen kielen käsittely

Luonnollisen kielen käsittelyssä yhdistyvät tekoäly ja kielitieteet. Kieli voidaan määrittellä joukkona sääntöjä tai symboleita, joita yhdistelemällä saadaan ilmaistua tietoa. Luonnollisen kielen käsittelyssä tietokone pyrkii ymmärtämään sanoja, niiden yhdistelmiä ja merkityksiä ihmisen käyttämässä kielessä. [23] Tässä tutkielmassa pyritään ratkaisemaan luonnollisen kielen käsittelyn tehtävä, tarkemmin ottaen tiedonpoimintatehtävä, suurten kielimallien avulla. Tutkielmassa ratkaistavassa tiedonpoimintatehtävässä tarkastellaan suurten kielimallien kykyä käsitellä lääketieteellisiä lausuntoja ja poimia niistä ennalta määrättyjä tietoja. Käytetyt lausunnot ovat vapaatekstimuodossa. Tässä luvussa käydään ensin läpi luonnollisen kielen käsittelyn peruseriaatteita, tämän jälkeen esitellään tarkemmin tiedonpoimintatehtävä luonnollisen käsittelyn tehtävänä, sen piirteitä ja tapoja ratkaista tätä tehtävää. Lopuksi käydään vielä läpi luonnollisen kielen käsittelyä lääketieteessä ja vapaatekstimuodossa olevien luonnollisen kielen aineistojen haasteita.

2.1 Peruseriaatteet

Tässä aliluvussa esitellään luonnollisen kielen käsittelyn peruseriaatteita. Ensin määritellään luonnollisen kielen käsittely alana ja sen jälkeen tarkastellaan sen peruseriaatteita. Aliluku perustuu pääosin lähteeseen [20], joka on julkaistu vuonna 2009. Luonnollisen kielen käsittelyn peruseriaatteet ovat pysyneet samoina, vaikka näiden tehtävien ratkaisutavat ovat kehittyneet huomattavaa vauhtia tämän jälkeen.

Luonnollinen kieli tarkoittaa ihmisten laajasti käyttämää kieltä. Se voi olla eri muodoissa, esimerkiksi puheena tai tekstinä. Luonnollisen kielen käsittelyn ensisijainen tavoite on käsitellä puhetta tai tekstiä syötteenä tai ulostulona niin, että kone ymmärtää kieltä kokonaisuutena eikä vain merkkijonoina. Tällöin koneet pystyvät löytämään annetusta luonnollisen kielen aineistosta laajempia käsitteitä, kokonaisuuksia ja merkityksiä. On kuitenkin huomioitava, että koneet eivät ymmärrä luonnollista kieltä samalla tavalla kuin ihmiset vaan ne käsittelevät sitä muun muassa erilaisten mallien ja algoritmien ja metodien avulla käyttämällä tietojenkäsittelytieteen, tilastotieteen ja matematiikan tarjoamia työkaluja. Ihminen pystyy usein analysoimaan luonnollista kieltä automaattisesti ja löytämään sanojen välisiä ja niiden muodostamia merkityksiä helposti. Kun samaa tehtävää suoritetaan tekoälyn avulla, yksi etu on, että kone pystyy analysoimaan huomattavasti suurempia datamääriä paljon lyhyemmässä ajassa. Tämän takia luonnollisen kielen käsittelemiseksi kehitetään jatkuvasti uusia tekniikoita ja malleja, jotta koneet pystyisivät käsittelemään kieltä toimivasti ja tehokkaasti. [28]

Luonnollisen kielen aineistojen käsittely koneen avulla saattaa kuitenkin osoittautua haastavaksi useista syistä. Jotkin luonnollisen kielen piirteet voivat olla haastavia koneelle. Esimerkiksi tietyt konteksteista riippuvat sanat, synonyymit ja idiomit voivat muodostua ongelmiksi. Samoja sanoja voidaan käyttää tarkoittamaan eri asioita eri konteksteissa. Näissä tilanteissa koneen tulisi pystyä erottamaan juuri siinä kontekstissa käytetty sanan merkitys. Synonyymit taas ovat haastavia moninaisuutensa takia. Myös niiden sävyerot voivat olla koneelle haastavia ymmärtää ja eri luonnollisen kielen käyttäjät saattavat käyttää samankaltaisia sanoja eri tavoilla. Esimerkiksi suuri, iso ja valtava saattavat eri ihmisten kielenkäytössä tarkoittaa samaa asiaa. Idiomit taas eivät ole vain osiensa summia, eli niitä ei voi ymmärtää pelkästään ymmärtämällä niissä esiintyviä yksittäisiä sanoja, vaan ne pitää käsitellä kokonaisuutena. Luonnollisen kielen moninaisuus ja moniselitteisyys on muun muassa näistä syistä otettava huomioon, kun ratkaistaan sen tehtäviä koneiden avulla. [23]

2.2 Tiedonpoimintatehtävä

Tämän tutkielman implementointiosuudessa pyritään rakenteellistamaan lääketieteellisiä, vapaatekstimuodossa olevia, lausuntoja suurten kielimallien avulla. Tällaista tehtävää voidaan myös kutsua tiedonpoimintatehtäväksi. Tässä aliluvussa keskitytäänkin tästä syystä juuri tiedonpoimintatehtäviin. Ensin aliluvussa esitellään pintapuolisesti luonnollisen kielen käsittelyn tehtäviä laajemmin. Tämän jälkeen käydään läpi nimenomaan tiedonpoimintatehtävän piirteitä ja sen osa-alueita. Aliluku perustuu pääosin lähteeseen [20].

Luonnollisen kielen käsittelyn tehtävät voidaan jakaa kahteen osaan: kielen analysointiin liittyviin tehtäviin ja kielen generointiin liittyviin tehtäviin. Kielen analysointiin liittyvissä tehtävissä pyritään analysoimaan ja käsittelemään olemassaolevaa tekstiä halutulla tavalla. Kielen generointiin liittyvissä tehtävissä taas luodaan uutta tekstiä annettujen ohjeiden perusteella. Tässä tutkielmassa ratkaistava tiedonpoimintatehtävä on kielen analysointiin liittyvä tehtävä. Kielen analysoinnissa on tärkeää, että kone ymmärtää jollain tasolla käsittelemänsä tekstin. Tämä tarkoittaa esimerkiksi kielen yksittäisiä sanoja tai lauseita, kielioppirakenteita ja sanojen merkityksiä yksin ja yhdessä sekä tietyn kontekstin merkitystä lauseita käsitellessä. Tällöin kone voi poimia annetusta tekstistä erilaisia ennalta määrättyjä tietoja. [4]

Joitakin tiedonpoimintatehtäviä voidaan myös ratkaista yksinkertaisemmilla tekniikoilla, jotka eivät vaadi kielen laajempaa ymmärtämistä. Eräs tunnetuimmista yksinkertaisista tekniikoista on säännöllinen lauseke (eng. regular expression), jonka avulla voidaan poimia tekstistä ennalta määritettyjä merkkijonoja. Vapaateksti-

lausuntoja käsitellessä säännöllinen lauseke ei kuitenkaan usein ole riittävä työkalu, sillä vapaateksti on usein muodoltaan haastavaa ja vaihtelevaa, kun taas säännöllinen lauseke toimii rajoitetusti vai määrätetyn merkkijonon kautta.

Tiedonpoimintatehtävissä tekstissä olevasta epärakenteellisesta tiedosta pyritään muodostamaan rakenteellista tai semirakenteellista (eng. semi-structured) dataa. Epärakenteellinen tieto ei ole järjestelmällistä ja sitä on tämän takia haastavaa käsitellä. Käytännössä tiedon epärakenteellisuus tarkoittaa lääketieteellisten lausuntojen yhteydessä esimerkiksi käytettyjen formaattien ja tekstin pituuden vaihtelua, mahdollisia kirjoitusvirheitä ja lyhenteitä ja erityistä lääketieteellistä sanastoa. [1] Rakenteellinen tieto taas tarkoittaa tilannetta, jossa tieto on järjestetty sellaiseen yhtenäiseen muotoon, että sitä voidaan käsitellä tietokantajärjestelmien, tilastollisten analyysien ja muiden työkalujen avulla. Näin rakenteellistettua tietoa voidaan hyödyntää jatkotehtävissä. Eräs tunnetuimmista semirakenteellisista tietomuodoista on JSON-formaatti, jota käytetään tässäkin tutkielmassa. Semirakenteellinen tieto on muodoltaan joustavampi rakenteelliseen tietoon verrattuna, mutta samalla mahdollistaa datan laajan jatkokäytön.

Tiedonpoimintatehtävät voidaan edelleen jakaa pienempiin tehtäviin kuten nimettyjen entiteettien tunnistukseen (eng. named entity recognition), yhteyksien tunnistamiseen ja lajitteluun (eng. relation detection and classification), ajallisen ilmaisen tunnistamiseen ja analyysiin (eng. temporal expression recognition and analysis) ja templaatin täyttöön (eng. template filling). Seuraavaksi käydään lyhyesti läpi joidakin tiedonpoimintatehtävän osatehtäviä, jotta saataisiin kokonaiskuva yksittäisen luonnollisen kielen tehtävän ratkaisemiseen vaadittavasta työstä.

Nimettyjen entiteettien tunnistuksessa nimetyllä entiteetillä tarkoitetaan yksinkertaisesti jotakin kokonaisuutta, johon voidaan viitata tietyllä nimellä. Tällaisia voivat olla esimerkiksi ihmisten, paikkojen tai yritysten nimet. Lääketieteen soveluksissa nimetyt entiteetit voivat olla esimerkiksi sairauksien tai lääkkeiden nimiä. Nimettyjen entiteettien tunnistuksessa tunnistetaan ensin nimet ja sen jälkeen luokitellaan sisältö ja yhdistetään ne oikeisiin ryhmiin tyyppinsä mukaisesti. Tutkielman implementointiosuudessa tällaisia nimettyjä entiteettejä ovat esimerkiksi tietyt syöpätutkimuksessa keskeiset luokitukset ja piirteet, joiden avulla hoidontarvetta voidaan arvioida ja sairauden kulkua ennustaa.

Yhteyksien tunnistamisessa ja luokittelussa tehtävänä on löytää ja luokitella tekstistä löydettyjen entiteettien välisiä semanttisia yhteyksiä ja luokitella niitä sopiviin kategorioihin. Käytännössä tällaisia yhteyksiä ovat esimerkiksi toisiinsa liittyvät syöpätutkimuksessa esiin tulevat tekijät ja niiden merkitys sairauden vakavuutta arvioiessa. Jokin tekijä voi esimerkiksi olla merkki huonosta ennusteesta potilaalle.

Myös ajallinen ulottuvuus on tärkeää yhteyksien ymmärtämisessä. Ajallisen ilmaisun tunnistamisessa ja analyysissä tarkoituksena on määrittää tapahtumien ajallista yhteyttä keskenään. Lääketieteessä tällaisia ajallisia yhteyksiä voivat esimerkiksi olla sairauden ajallinen kulku ja milloin potilaalla on havaittu tiettyjä sairaudelle keskeisiä oireita.

Templaatin täyttö on tiedonpoiminnan osatehtävä, jossa tarkoituksena on muodostaa eräänlainen tiivistelmä tietystä tekstistä täyttämällä aiheeseen sopiva templaatti. Templaatti voidaan täyttää yksittäisillä sanoilla tai tekstistä otetuilla pidemmillä segmenteillä. Tämän tutkielman implementointiosuudessa templaattina toimii lista malleille esitetyistä kysymyksistä ja se täytetään lausunnoista löydettyillä tiedoilla annettujen ohjeiden mukaisesti.

Tiedonpoimintatehtävä koostuu siis useammasta pienemmästä tehtävästä, joiden avulla voidaan ratkaista alkuperäinen, laajempi, tehtävä. Usein oikean elämän tehtäviä ratkaistaessa suoritetaan useaa luonnollisen kielen käsittelyn tehtävää saman tehtävän yhteydessä.

2.3 Tiedonpoimintatehtävän ratkaiseminen

Tässä tutkielmassa käytetään suuria kielimalleja tiedonpoimintatehtävän ratkaisemisessa. Usein luonnollisen kielen tehtäviä voi ratkaista usealla tavalla ja ratkaisumallin tai -tekniikan valitseminen riippuu muun muassa käytettävissä olevasta ajasta ja tietoteknisistä resursseista, tehtävän luonteesta ja saatavilla olevasta tietotaidosta. Tässä aliluvussa esitellään ensin kevyesti niitä piirteitä, jotka tekevät tiedonpoimintatehtävästä haastavan. Tämän jälkeen käydään läpi joitakin tekniikoita, joilla tehtävää on ratkaistu. Lopuksi esitellään vielä joitakin keskeisimpiä arviointimetriikoita tiedonpoimintatehtävän ratkaisujen arvioinnissa.

Tiedonpoimintatehtävä on monimutkainen luonnollisen kielen käsittelyn tehtävä. Tehtävän ratkaisun laatu riippuu monista asioista kuten poimittavien tietojen loogisten rakenteiden haastavuudesta, syötetekstin luonteesta, syötteen kielestä ja siitä, minkä alan tekstiä käsitellään. Tämän tutkielman implementointiosuudessa käytetään syötteenä suomenkielistä tekstiä. Suomen kieli tuo lisähaasteita tiedonpoimintatehtävän ratkaisemiseen, sillä jotkin sen piirteet ovat monimutkaisempia tiedonpoimintatehtävän näkökulmasta esimerkiksi englantiin verrattuna. Suomen kielestä haastavan tekee muun muassa suuri määrä sijamuotoja, jotka lisätään suoraan sanaan ja ne pitäisi tehtävää ratkaistaessa pystyä käsittelemään oikealla tavalla. Myös esimerkiksi yhdyssanat saattavat aiheuttaa samankaltaisia haasteita. [38]

Laskennallisesti yksinkertaisin tapa ratkaista tiedonpoimintatehtävä on manuaalinen tiedonpoiminta, missä asiantuntijat poimivat teksteistä tietoja ennalta määri-

tettyjen sääntöjen mukaisesti. Tämä tapa vie kuitenkin paljon aikaa ja vaatii luonnollisesti paljon manuaalista työtä verrattuna saatuun tulokseen. Tämän takia tätä tapaa ei voikaan yleistää suurille aineistoille, vaikka pienillä aineistolla sillä saadaan usein hyvin tarkkoja tuloksia. Tilastolliseen koneoppimiseen perustuvissa tekniikoissa käytetään ohjattua oppimista ja merkittyä dataa opetusprosessissa ja tehtävä ratkaistaan näin opetetun mallin avulla. Tässäkin tavassa tarvitaan kuitenkin paljon alan asiantuntijoiden manuaalista työtä datan merkitsemiseen ja tällainen työ on aikaavievää ja kallista. Toisaalta se on usein sopiva tapa arvioida esimerkiksi automaattisten kielimallien tarkkuutta tehtävässä, sillä alan asiantuntijoilla on paljon tietotaitoa ja näin mallien vastauksia voidaan verrata absoluuttiseen totuuteen.

Syväoppimiseen ja neuroverkkoihin perustuvat tekniikat sopivat yleensä paremmin suuren aineiston käsittelemiseen kahteen muuhun esiteltyyn tekniikkaan verrattuna, sillä nämä tekniikat eivät vaadi monimutkaista ominaisuussuunnittelua (eng. feature engineering). Ominaisuussuunnittelussa raakadataa muunnetaan tehokkaammaksi syötteenä käsittelemällä sitä manuaalisesti ennen varsinaista tehtävän ratkaisemista. [65] Muista tiedonpoimintatehtävän ratkaisemisessa käytetyistä syväoppimiseen ja neuroverkkoihin perustuvista tekniikoista voi lukea lisää esimerkiksi lähteistä [37], [65] ja [68].

Tiedonpoimintatehtävän ratkaisua arvioidessa tärkeää on erityisesti ratkaisujen tarkkuus ja oikeellisuus. Tarkkuus (eng. precision) ja oikeellisuus (eng. recall) mittavat ratkaisun tehokkuutta käyttäjän näkökulmasta. Templaatti tarkoittaa tiedonpoimintatehtävässä ennalta määritettyä muotoa, jossa poimittu tieto tulisi esittää. Lääketieteen tehtävissä tämä voisi tarkoittaa esimerkiksi templaattia:

Potilaan nimi: “potilaan nimi”

Potilaan ikä: “potilaan ikä”

missä kohdat “potilaan nimi” ja “potilaan ikä” täytettäisiin aineistosta löydettävillä tiedoilla. Kaikkia templaatin kohtia, joita odotetaan täytettävän voidaan merkitä #avain -merkinnällä. Samalla merkinnöillä #oikea (#väärä) merkitään oikein (väärin) täytettyjä kohtia. Vastaus voidaan määritellä vääräksi, jos se ei noudata ennalta määritettyjä oikean vastauksen kriteerejä. Tarkkuus ja oikeellisuus voidaan tällöin määritellä seuraavasti:

$$\text{tarkkuus} = \frac{\#oikea}{\#oikea + \#väärä} \quad \text{oikeellisuus} = \frac{\#oikea}{\#avain} \quad (1)$$

Tarkkuus- ja oikeellisuusmetriikoita voidaan tarvittaessa käyttää erikseen jokaisen templaatin kohdan tarkasteluun, jotta saataisiin tarkempi kuva vastausten tarkkuudesta ja oikeellisuudesta.

Tarkkuuden ja oikeellisuuden lisäksi usein käytetty metriikka on F-mitta (eng. F-measure/F-score), joka on tarkkuuden ja oikeellisuuden painotettu harmoninen keskiarvo. Sen avulla voidaan arvioida käytetyn mallin tai tekniikan suoritusta tehtävän ratkaisemisessa. F-mitta voidaan määritellä seuraavasti:

$$F = \frac{(\beta^2 + 1) \times \text{tarkkuus} \times \text{oikeellisuus}}{(\beta^2 \times \text{tarkkuus}) + \text{oikeellisuus}}, \quad (2)$$

missä β on ei-negatiivinen arvo, jolla voidaan vaikuttaa tarkkuuden ja oikeellisuuden painotuksiin metriikoissa. [38]

2.4 Luonnollinen kieli elektronisissa potilastietokannoissa

Ympäri maailmaa, myös Suomessa, on siirrytty sähköisten potilastietojärjestelmien (Electronic Health Records, EHRs) käyttöön. Sähköisten potilastietojärjestelmien avulla terveystietojen tallentaminen ja jakaminen eri terveydenhuollon tarjoajien kesken on helpompaa paperipohjaisiin järjestelmiin verrattuna. Sähköisten potilastietojärjestelmien ensisijainen tarkoitus onkin juuri se, että terveystiedot on tallennettu johonkin järjestelmään, jonka kautta niihin pääsevät käsiksi potilaan hoidossa keskeisenä toimivat tahot. Sähköiset potilastietojärjestelmät mahdollistavat myös terveystietojen analyysin. Analyysin tarkoituksena on tukea potilaanhoitoa, lääketieteellistä tutkimusta ja tarjota sellaisia mahdollisuuksia, jotka eivät pelkän manuaalisen tarkastelun tuloksena olisi mahdollisia.

Merkittävä osa sähköisiin potilastietojärjestelmiin tuotetusta tiedosta ei ole rakenteellisessa muodossa, mikä estää niiden suoran käytön esimerkiksi tilastollisessa tarkastelussa. Vaikka tämä on ilmaisun joustavuutensa takia hyvä käyttäjälleen, aiheuttaa se kuitenkin usein ongelmia vapaatekstimuodossa kerätyn tiedon käytön kanssa. Tekstin epärakenteellinen luonne tekee vapaatekstimuodossa olevien terveystietojen käytöstä usein haastavaa. Vapaatekstin epärakenteellisuuden vuoksi tekstissä oleva tieto voi olla hyvinkin subjektiivista ja sen löytäminen vaikeaa. Tieto voi riippua käytetystä kontekstista mitä ei erikseen ole mainittu ja näin sitä voi olla haastavaa poimia kirjoitetusta lausunnosta. [1] Luonnollisen kielen käsittelyn avulla epärakenteellisia vapaatekstejä pystytään kuitenkin rakenteellistamaan ja vapaateksteissä olevan tiedon käyttömahdollisuudet realisoituvat. Vapaatekstit sisältävät paljon sellaista tietoa, mitä ei muualla terveystiedoissa ole, joten niiden avulla voidaan saada parempi kokonaiskuva käsiteltävästä asiasta.

Esimerkiksi lääkärin kirjoittamat lausunnot ovat usein vapaatekstimuodossa. Niiden analysointi voi auttaa ennustamaan potilaiden hoitojen toimivuutta ja sillä voidaan parantaa potilaiden hoitotarpeen arviointia. Analysoinnin avulla voidaan

myös luoda matemaattisia malleja, joiden ansiosta voidaan esimerkiksi havaita tauti varhain. Vapaatekstien analysoinnin ja prosessoinnin lisäksi luonnollisen kielen generoinnin avulla voidaan luoda esimerkiksi keskustelubotteja, joiden avulla potilaat voivat kysyä kysymyksiä ja saada tietoa etsimästään asiasta. Luonnollisen kielen käsittelyä käytetään lääketieteessä erityisesti toimivuuden parantamiseen, epärateellisen tekstin hyödyntämiseen sekä potilaiden kanssa kommunikointiin. [28]

Seuraavaksi käydään vielä läpi joitakin oikean elämän lääketieteen sovelluksia, joita on ratkaistu luonnollisen kielen käsittelyn avulla. Koska tässä tutkielmassa ratkaistaan tiedonpoimintatehtävää, esitellään tässä myös samankaltaisia sovelluksia. Artikkelissa [51] pyrittiin tunnistamaan yleisiä riskitekijöitä kotihoidon lääketieteellisistä lausunnoista ja arvioimaan niiden yhteyksiä sairaala- ja päivystyskäynteihin. Tavoitteena oli riskitekijöiden tunnistamisen avulla ennustaa potilaan tulevaa hoitotarvetta ja näin parantaa potilaille annettavaa hoitoa. Metodeina käytettiin aikasarja-analyysiä ja hierarkkista klusterointia. Artikkelissa [1] suurta määrää sähköisiä terveystietoja aliravitsemukseen liittyen pyrittiin rakenteellistamaan suuren kielimallin, Llama 2 -mallin, avulla käyttämällä nollaoppimista. Rakenteellistettua tietoa voidaan käyttää ongelman analysointiin. Artikkelissa [58] tavoitteena oli luoda sähköisten psykiatristen terveystietojen tiedonpoimintaan sopiva työkalu luonnollisen kielen käsittelyn avulla. Erityisesti haluttiin poimia tietoja liittyen psykiatrisen hoidon tuloksiin ja toimivuuteen. Näiden avulla haluttiin luoda ennustavia malleja psykiatrisen hoidon arviointiin. Nämä mallit voivat auttaa tulevassa hoidontarpeen arvioinnissa. Tiedonpoimintatehtävän ratkaiseminen on siis laaja sovelluskohde lääketieteessä, jonka avulla saadaan paljon tärkeää tietoa. Tätä tietoa voidaan käyttää edelleen jatkotehtävissä.

Lääketieteen alalla oman haasteensa aineiston käsittelyyn tuo sen arkaluontoisuus ja aineiston käytön eettisyyteen ja yksityisyyteen liittyvät ongelmat. Jotta sähköisiä terveystietoja voidaan käyttää esimerkiksi tutkimuksessa, tarvitaan siihen viranomaiselta lupa. [9] Suomessa tietolupaviranomaisena sosiaali- ja terveysalalla toimii Findata. Lupa tarvitaan, koska terveystiedot sisältävät paljon yksityisten ihmisten arkaluonteista tietoa. Sen lisäksi, että aineistoja voi olla haastavaa saada tutkimuskäyttöön, myös aineistojen siirtäminen esimerkiksi maiden rajojen ja eri yritysten välillä on haastavaa. Tämän takia terveystietoihin liittyvät tutkimustulokset ovat usein rajattuja ja kattavia kansainvälisiä tutkimusaineistoja on erittäin rajallisesti saatavilla. [52]

Eettisyyteen ja yksityisyyteen liittyvät haasteet terveystietojen jakamisessa pätevät myös, kun sovelluksia ratkaistaan suurten kielimallien avulla. Suuret kielimallit pystyvät kuitenkin useissa tapauksissa ratkaisemaan sovelluksia muita luonnollisen

kielen käsittelyn tekniikoita paremmin. Keskeisenä erona on suurten kielimallien kyky ymmärtää kieltä eri tavalla kuin aiemmat tekniikat. Kuten aiemmin on mainittu, suuret kielimallit eivät myöskään usein vaadi yhtä paljon merkittävää dataa ja ihmisen manuaalista työtä tehtävien ratkaisemisessa. Tällöin suurten kielimallien käyttäminen tehtävien ratkaisemisessa on usein resursseiltaan tehokkaampaa ja niiden käyttömahdollisuudet ovat laajemmat kuin aikaisempien tekniikoiden. Eettisyyteen ja yksityisyyteen liittyviä lisähaasteita suurten kielimallien käytössä tuo se, että useimmat suurista toimijoista ovat kaupallisia. Tällöin niiden kanssa jaettujen tietojen jatkokäytön oikeudet jäävät toimijoiden käsiin, eikä niiden kautta voida käsitellä arkaluonteisia tietoja. Vaihtoehtona ovat avoimet suuret kielimallit, jotka tarjoavat mahdollisuuden niiden käyttöön esimerkiksi sairaaloiden sisäverkoissa, jolloin arkaluonteisia tietoja ei tarvitse jakaa ulkopuolisille toimijoille. Seuraavassa luvussa esitellään tarkemmin suurten kielimallien toimintaa ja niiden vahvuuksia.

3 Suuret kielimallit

Suuret kielimallit (eng. Large Language Models, LLM) ovat kasvattaneet suosiotaan eri käyttötarkoituksissa arkielämän avusta erilaisten haastavien tehtävien ratkaisuun. Niitä voidaan käyttää laajasti erilaisissa sovelluskohteissa ja niiden nopea ja jatkuva kehitys mahdollistaa myös uusia tulevaisuuden käyttötarkoituksia. Tässä luvussa käydään ensin läpi suurten kielimallien lähihistoriaa ja niiden kehitystä. Tämän jälkeen esitellään suurten kielimallien toiminnassa keskeinen transformer-arkkitehtuuri ja enkooderi- ja dekodeeripohjaisten suurten kielimallien toimintaa ja niiden eroja. Lopuksi esitellään keskeisimpiä suurten kielimallien käytön haasteita ja rajoitteita.

3.1 Historia

Generatiivinen tekoäly tarkoittaa laskennallisia tekniikoita, jotka voivat tuottaa ainakin näennäisesti uutta ja merkityksellistä sisältöä, kuten tekstiä, kuvia tai ääntä, opetusaineiston avulla. [55] Generatiivinen tekoäly on laaja käsite, joka sisältää monia eri työkaluja. Eräitä näistä ovat suuret kielimallit, jotka keskittyvät erityisesti käsittelemään erilaisia luonnolliseen kieleen liittyviä tehtäviä.

Kielimallit ovat matemaattisia malleja, joilla on kapasiteettia ymmärtää ja generoida luonnollista kieltä. Suuret kielimallit taas ovat edistyneitä kielimalleja, joiden parametrimäärät ovat erittäin suuria ja ne pystyvät oppimaan monimutkaisiakin rakenteita. Keskeisenä osana suurissa kielimalleissa on tässä tutkielmassakin esitelty transformer-arkkitehtuuri (eng. transformer architecture) ja siinä toimiva omahuomiomekanismi (eng. self-attention mechanism). Esiopetusvaiheessa suuri kielimalli kehittää laajan taitorepertuaarin luonnollisen kielen käsittelyyn ja hahmontunnistukseen. Erilaisia tehtäviä ratkaistaessa suuri kielimalli osaa käyttää monipuolisesti näitä taitoja tehtävän tunnistamisessa ja sitä ratkaistaessa. [7] Tämä tarkoittaa, että mallit oppivat asioita, joita niille ei suoranaisesti ole opetettu. Tällöin suuret kielimallit pystyvät antamaan yhtenäisempiä ja asiayhteyden keskeisemmin liittyviä vastauksia tehtäviä ratkaistaessa. Tätä kykyä kutsutaan kontekstioppimiseksi ja juuri sen takia suuret kielimallit sopivat hyvin esimerkiksi interaktiivisiin ja keskustelukäyttötarkoituksiin. [8] Haasteena suurien kielimallien käytössä on, että käytännössä suuret kielimallit ovat kuitenkin niin sanottuja “black box” -malleja, eli niiden päätöksentekoprosessi ja toiminta eivät ole suoraan tulkittavissa.

Suurten kielimallien kehitys on mennyt nopeasti eteenpäin viimeisen vajaan vuosikymmenen aikana. Mielenkiinto suurista kielimalleista kohtaan on suurta, koska mahdollisten sovellusten määrä on valtava ja niitä voidaan hyödyntää monella alalla.

Tässä aliluvussa esitellään suurten kielimallien lähistoriaa ja kehitystä. Tämän tutkielman implementointiosuudessa käytetään neljää suurta kielimallia: GPT-3.5 -mallia, GPT-4 -mallia, Gemini-1.0-pro-001 -mallia ja Llama3-70b -mallia. Tämän takia keskitytään erityisesti näiden mallien kehitykseen. Tämän lisäksi esitellään myös muita keskeisimpiä trendejä ja kehityssuuntia suurten kielimallien historiassa.

Varsinaisesti suurten kielimallien historian voidaan nähdä alkaneen vuonna 2017, kun artikkeli “Attention is all you need” [59] julkaistiin. Artikkelissa esiteltiin Transformer-arkkitehtuuri, jonka ansiosta suuret kielimallit pystyvät käsittelemään luonnollista kieltä tavalla, joka ei aiemmin käytetyille malleille tai tekniikoille ollut mahdollista. Monet aiemmat syväoppimisen tekniikat luonnollisen kielen käsittelyn tehtäviin vaativat paljon merkittävää dataa, jonka saaminen vaatii paljon resursseja. Tämän takia ratkaisutekniikat, jotka toimivat ilman merkittävää dataa voivat usein olla tehokkaampia. Merkitsemätöntä dataa on saatavilla suuria määriä ja transformer-arkkitehtuurilla toimivat suuret kielimallit voivatkin käyttää tällaista dataa luonnollisen kielen käsittelyn tehtävien ratkaisemiseksi hyvinkin tehokkaasti ja onnistuneesti. Transformer-arkkitehtuurin kehittämisen jälkeen sitä on hyödynnetty useissa erilaisissa suurissa kielimalleissa.

OpenAI julkaisi ensimmäisen GPT-mallin (Generative Pre-trained Transformers), GPT-1:n [42], vuonna 2018. GPT-1 -malli on esiopetettu laajalla aineistolla, joka on kerätty internetistä eri muodoissa. Pian tämän jälkeen Google esitteli oman transformer-arkkitehtuurilla toimivan esiopetetun suuren kielimallinsa, BERT:in [10]. BERT (Bidirectional Encoder Representations from Transformers) eroaa GPT-1 -mallista, sillä se on enkooderipohjainen malli, kun taas GPT-mallit ovat dekodeeripohjaisia. Tämä tarkoittaa, että mallit käsittelevät syötteenä tulevaa kieltä eri tavoilla. Enkooderi- ja dekodeeripohjaisten mallien eroihin palataan myöhemmin tässä luvussa.

Jotta suuria kielimalleja voitaisiin käyttää erilaisten tehtävien ratkaisemisessa, tulee niitä ensin opettaa opetusaineistolla. Tätä kutsutaan esiopetusvaiheeksi ja suuret kielimallit ovatkin yleensä esiopetettuja. Tämä tarkoittaa, että niitä on opetettu laajalla korpuksella. Korpus tarkoittaa suurta määrää teksti- tai puhedataa, jonka tietokone voi lukea. [20] Se voi koostua yhdestä tai useasta asiakirjasta ja sen lähteenä voivat toimia esimerkiksi kirjat, artikkelit tai keskustelufoorumit. Korpus voi olla yksi- tai monikielinen. Esimerkiksi käännöstehtäviä varten mallia tulisi opettaa monikielisellä korpuksella. [4] Käytetty opetusaineisto ja sen suuruus ovat merkittäviä, sillä opetusaineiston avulla suuret kielimallit oppivat kieltä ja sen rakenteita.

Esiopetusvaiheessa opetetaan ensin tokenisoija ja sen jälkeen suuri kielimalli. Tokenisoinnissa teksti jaetaan pienempiin pätkiin, tokeneihin, joita sitten käytetään

syötteenä suurille kielimalleille. Tokenit ovat joko sanoja tai sanojen osia. Tokenisointi voidaan toteuttaa usealla eri tavalla ja suuret kielimallit käyttävät eri tapoja toiminnassaan. Esimerkiksi BERT [10] käyttää WordPiece-tokenisointia [63]. GPT-2 [43] taas käyttää BPE-tokenisointia (Byte-Pair Encoding) [50]. [44] Esiopetuksen tarkoituksena on opettaa mallille kielen rakennetta ja yhteyksiä sanojen välillä. Tavoitteena on, että näin malli oppisi opetettavan kielen piirteitä riittävän hyvin, jotta se osaisi ratkaista sille esitettyjä luonnollisen kielen tehtäviä. Konteksti tai kontekstipituus (eng. context length) tarkoittaa suurinta tokenien määrää, minkä suuri kielimalli voi käsitellä yksittäisen syötteen yhteydessä.

GPT-1:n seuraaja GPT-2 [43] esiteltiin vuonna 2019. GPT-2 -malli on edeltäjänsä suurempi parametrikooltaan, 1,5 miljardia parametriä verrattuna GPT-1 -mallin 117 miljoonaan parametriin. Varsinkin suurten kielimallien kehityksen alkuvaiheessa mallien parametrimäärät kasvoivat jokaisella mallin iteraatiolla sekä uusien mallien myötä. Suurempi parametrimäärä implikoi tällöin tehokkaampaa mallin toimintaa. Vuonna 2020 julkaistussa GPT-3 -mallissa [7] parametrimäärä kasvoi edelleen edeltäjänsä verrattuna. GPT-3 oli myös OpenAI:n ensimmäinen suljettu (eng. closed-source) malli, mikä tarkoittaa, että mallin neuroverkkoarkkitehtuuri ja parametriarvot eivät ole julkisesti saatavilla. Tämän jälkeen suurin osa markkinoiden johtavista malleista on ollut suljettuja malleja.

Suurten kielimallien parametrimäärä on kasvanut reilusta sadasta miljoonasta satoihin miljardeihin. Tämä tarkoittaa myös suurten kielimallien kapasiteetin kasvua, sillä parametrien määrä vaikuttaa mallien suoritukseen. [7] Suurissa kielimalleissa parametrit ovat samoja kuin neuroverkoissa, eli painoja ja vakiotermejä. Pelkkä suuri parametrimäärä ei kuitenkaan lupaa suuren kielimallin tehokkaampaa toimintaa. Artikkelissa [17] todetaan, että mallin parametrimäärän kasvaessa myös opetustokeneiden määrän tulisi kasvaa samassa suhteessa. Näin saataisiin kehitettyä mahdollisimman tehokkaita suuria kielimalleja. Mallien suoritukseen vaikuttavat mallin koko eli parametrien määrä, opetusaineiston määrä ja opettamiseen käytetyt tietotekniset resurssit. [22] Tietotekniset resurssit tarkoittavat saatavilla olevaa prosessointitehoa, muistia ja tallennustilaa. [55] Asiantuntijoiden yhdistelmä (eng. Mixture of Experts, MoE) on suurten kielimallien rakennetyyli, missä on useampia tiettyyn tarkoitukseen koulutettuja alineuroverkkoja eli niin sanottuja asiantuntijoita. Näistä vain olennaisimmat aktivoidaan tehtävää ratkaistaessa. Asiantuntijoiden yhdistelmätekniikkaa käyttämällä samalla tietoteknisten resurssien määrällä voidaan kasvattaa mallin kokoa tai opetusaineiston määrää huomattavasti verrattuna perinteisempään tiheään malliin (eng. dense model), jossa kaikki parametrit aktivoituvat jokaisella syötteellä. [48]

Google DeepMind julkaisi Geminin ensimmäisen version, Gemini 1.0:n [53], loppuvuodesta 2023. Gemini 1.0 -versiosta esiteltiin kolme eri kokoa: Ultra, Pro ja Nano. Ultra on ensisijaisesti tarkoitettu erittäin monimutkaisten tehtävien ratkaisuun, Pro tarjoaa tehostettua suorituskykyä ja laajempaa käyttökelpoisuutta ja Nano taas on tarkoitettu käytettäväksi suoraan käyttäjän laitteissa. Näiden lisäksi Gemini-malleja on hienosäädetty ja näin on kehitetty kaksi Gemini Apps -mallia, joiden avulla toimivat Googlen keskustelukumppanit, Gemini ja Gemini Advanced. Aiemmin Googlen keskustelukumppanina toimi Bard, joka toimi PaLM2 -mallilla. PaLM2-malli [3] julkaistiin vuonna 2023. Alkuvuodesta 2024 julkaistiin Gemini 1.5 -mallit, jotka on kehitetty erityisesti pitkien kontekstien ymmärtämiseen ja säilyttämiseen luonnollisen kielen käsittelyssä. [54] Tämän julkaisun myötä myös keskustelukumppani Gemini on päivitetty toimimaan Gemini 1.5 Flash -mallin avulla.

Alkuvuodesta 2023 Meta AI julkaisi ensimmäiset Llama-mallit [56]. Näihin malleihin kuului neljä perusmallia, joiden parametrimäärät vaihtelivat 7 - 65 miljardin parametrin välillä. Perusmalli tarkoittaa, että mallia on esiopetettu ratkaisemaan laajasti erilaisia tehtäviä. Llama-malleja opetettiin vain julkisesti saatavilla olevilla aineistoilla, tällöin ne sopivat avoimen lähdekoodin käyttötarkoituksiin. Heinäkuussa 2023 Meta AI julkaisi Llama 2 -mallit [57] Microsoft ensisijaisena kumppaninaan. Llama2-malleja julkaistiin kolme kappaletta ja niiden parametrimäärät vaihtelivat 7 - 70 miljardin välillä. Näihin malleihin kuului myös hienosäädetty Llama 2-chat -malli, joka on tarkoitettu erityisesti keskustelukäyttötarkoituksiin.

Hienosäätäminen tarkoittaa sellaista mallin lisäopettamista, jossa mallille esitellään lisää juuri käsiteltävän tehtävän ratkaisemiseen liittyvää aineistoa. Hienosäätövaiheessa mallia siis muunnetaan sopivalla tavalla, jotta se sopisi juuri tietyn tehtävän ratkaisemiseen. Tämä tapa vaatii kuitenkin paljon resursseja, kuten aikaa, tietotaitoa, aineistoa ja teknisiä resursseja. Kesällä 2024 julkaistiin Llama 3 -mallit [11]. Nämä ovat perusmalleja, jotka tukevat muun muassa monikielisyyttä ja koodin kirjoitusta. Aiempiin Llama-malleihin verrattuna Llama 3 -malleissa keskityttiin esimerkiksi esiopetuksessa käytetyn aineiston laatuun ja määrään. Llama 3 -malleja on kolme kappaletta: 8, 70 ja 405 miljardin parametrin versiot.

Llama-mallit ovat esimerkkejä avoimista suurista kielimalleista. Useimmat GPT- ja Gemini-mallit taas ovat yksityisomistuksessa olevia suljettuja suuria kielimalleja. Yksinkertaisesti, avoimet suuret kielimallit tähtäävät läpinäkyvyyteen mallin käytössä. Opetettujen mallien parametriarvot ovat saatavilla, jolloin käyttäjät voivat muokata malleja omiin käyttötarkoituksiinsa sopiviksi esimerkiksi lisäopettamalla tai hienosäätämällä malleja. Avoimet mallit ovat usein ilmaisia, tai ainakin edullisia, ja saatavilla vapaasti sekä yksityis- että yrityskäyttöön. Yksityisomistuksessa

olevat mallit taas vaativat yleensä jonkin mallin julkaisijan määrittämän käyttöoikeuden, esimerkiksi API-avaimen ja mallin käyttö on harvoin ilmaista. Keskeinen ero avoimissa ja yksityisomistuksen malleissa on käyttäjäkokemus. Yksityisomistuksessa olevat mallit vaativat myös niitä käyttäessä aktiivisen internetyhteyden ja ne varaavat käyttöoikeudet niiden kanssa jaettuun informaatioon. Tällöin arkaluontoisien tiedon, kuten potilaslausuntojen, käsittely on mahdotonta.

Esiopetusvaihe vaatii paljon aikaa, vaivaa ja tietoteknisiä resursseja. Tämän takia kielikohtaisia malleja kehitetäänkin usein vain pienelle määrälle suurten resurssien kieliä, kuten englannille. Kielikohtaisten mallien lisäksi on olemassa monikielisiä malleja, kuten tässäkin tutkielmassa käytettävät GPT-, Llama- ja Gemini-mallit. Mallin yksikielinen opetus tarjoaa usein parempia ratkaisuja, koska sen opetus on tarkempaa kielikohtaisesti. Tämä voikin osoittautua ongelmaksi suurten kielimallien käyttämisessä matalan resurssin kielille, kuten suomen kielelle, koska niillä opetettuja kielikohtaisia malleja on saatavilla vähemmän. [40]

Yksikielisiä malleja on kuitenkin kehitetty myös pienempien resurssien kielille. Suomen kielellä on muun muassa opetettu seitsemän yksikielistä suurta kielimallia, joiden parametrimäärät vaihtelevat 186 miljoonasta 13 miljardiin. Näitä malleja kutsutaan FinGPT-malleiksi. [32] Tämän lisäksi suomen kielellä on opetettu monikielisiä suuria kielimalleja, kuten 176 miljoonan parametrin BLUUMI [32] ja 34 miljardin parametrin PORO [31], jota on opetettu suomen ja englannin kielillä ja useilla ohjelmointikielillä. Näiden lisäksi vuonna 2025 julkaistiin Viking 33B [30], jota on opetettu kaikilla pohjoismaisilla kielillä, myös suomen kielellä.

Suuria kielimalleja kehitetään myös tiettyihin tarkoituksiin esiopettamalla niitä tietyn alan aineistolla. Biolääketieteen alalle on kehitetty esimerkiksi BioBERT [25] ja BioGPT [29], joiden on tarkoitus toimia paremmin erityisesti biolääketieteellisten tekstien tiedonetsinnässä ja tekstin generoinnissa. Suuria kielimalleja esiopetetaan yleensä aineistoilla, jotka sisältävät yleisiä verkosta löydettäviä tekstejä, esimerkiksi Wikipedia-artikkeleilla. Tiettyihin tarkoituksiin esiopetettuja suuria kielimalleja taas on opetettu tehtävän ratkaisuun relevantilla aineistolla. Tällöin näiden mallien tavoitteena on antaa tarkempia ja faktapohjaisempia vastauksia verrattuna yleisempiin malleihin esiopetuksessa käytetyn aineiston luonteen ja rakenteen takia.

Tiettyyn tarkoitukseen kehitettyjen suurten kielimallien lisäksi on kehitetty myös pieniä kielimalleja (eng. small language model), joiden tarkoituksena on toimia tehokkaasti jossakin tietyssä käyttötarkoituksessa. Kuten aiemmin todettiin, suurten kielimallien parametrimäärät ovat kasvaneet satoihin miljardeihin. Tällöin niiden opettaminen ja käyttö vaatii suuria määriä tietoteknisiä resursseja, mikä taas aiheuttaa suuria hiilijalanjälkiä ja saattaa tehdä niiden käytöstä haastavaa. [49] Esi-

merkiksi yrityskäyttöön pienet kielimallit saattavat usein olla parempi vaihtoehto niiden matalampien resurssivaatimusten vuoksi.

Suurten kielimallien ja hakukoneiden yhdistäminen tarjoaa uusia mahdollisuuksia muun muassa tiedon etsimiselle, sisällön ymmärtämiselle ja internetpalveluiden kanssa käydylle kanssakäymiselle. [64] Hakukoneet voivat parantaa suurten kielimallien antamien vastausten tarkkuutta ja informatiivisuutta, kun hakukoneiden avulla voidaan muun muassa etsiä suurille kielimalleille mahdollisimman hyviä opetusaineistoja mallien esiopetukseen. Suurten kielimallien avulla taas voidaan muun muassa parantaa hakukoneiden antamia vastauksia ja hakutulosten järjestystä. [64] GPT-3.5 -mallin avulla toimiva ChatGPT oli ensimmäinen suuri kielimalli, joka yhdistettiin Bing-hakukoneeseen vuonna 2023. Hakukoneisiin yhdistetyt suuret kielimallit, tai niiden avulla toimivat keskustelubotit, ovat yleistyneet ja muun muassa heinäkuussa 2024 OpenAI ilmoitti julkaisevansa prototyypin uudentyyppistä tekoälyllä toimivista hakuominaisuuksista, SearchGPT:n [36].

Vuonna 2021 OpenAI julkaisi DALL-E -mallin, joka pystyy luomaan kuvia sille annettujen luonnollisen kielen ohjeiden perusteella. Tällaista mallia, joka pystyy vastaanottamaan ja/tai tuottamaan sisältöä usealla eri tavalla kuten tekstinä, kuva- tai äänimateriaalina, kutsutaan multimodaaliseksi. Multimodaaliset mallit voivat käyttää syötteinä tai antaa ulostuloina multimodaalisia kokonaisuuksia. Tämän lisäksi mallien syötteet ja ulostulot voivat olla eri muodoissa, esimerkiksi tekstinä ja kuvana. Tällainen suuri kielimalli on esimerkiksi OpenAI:n kehittämä GPT-4o [19]. Multimodaalisuus mahdollistaa laajempia, uusia kykyjä suurille kielimalleille. Tämän takia se onkin erityisen mielenkiinnon aiheena suurten kielimallien kehittämisessä.

RAG (Retrieval-Augmented Generation) [27] on tekniikka, jonka avulla suurten kielimallien antamien vastausten ajankohtaisuutta ja laatua voidaan parantaa. Sen avulla suurille kielimalleille voidaan syöttää kohdistettua dataa esimerkiksi yrityksen tietokannasta, mikä auttaa mallia parempien vastausten muodostamisessa. Keskeistä on, että mallille syötetty lisäaineisto on mallin opetusaineiston ulkopuolista. Usein tämä tarkoittaa, että lisäaineisto on julkaistu mallin opettamisen jälkeen. Jos esimerkiksi haluttaisiin pyytää suuria kielimalleja ehdottamaan sopivia hoitotapoja johonkin sairauteen, voitaisiin mallille antaa lisämateriaalina suomenkielisiä artikkeleita viimeisimmistä hoitotavoista.

Viimeisimpiä suurten kielimallien kehityksen trendejä ovat olleet muun muassa multimodaalisuus, tiettyyn käyttöön tarkoitettut mallit, mallien käytön kohdistaminen ja erilaiset arjessa käytettävät keskustelubotit ja hakukoneisiin yhdistetyt suuret kielimallit. Tekoäly on alana nopeasti kehittyvä ja kilpailtu. Tämän takia uusia

innovaatioita kehitetäänkin jatkuvasti. Tekoäly on myös osana monien arkielämää töissä ja vapaa-ajalla, ja monet innovaatiot liittyvätkin suurten kielimallien parempaan ja tehokkaampaan käytettävyyteen. Esimerkiksi monet yritykset ovat ottaneet yrityksen sisäisesti käyttöön suuria kielimalleja, jotka auttavat juuri sen yrityksen toiminnassa.

3.2 Transformer-arkkitehtuuri

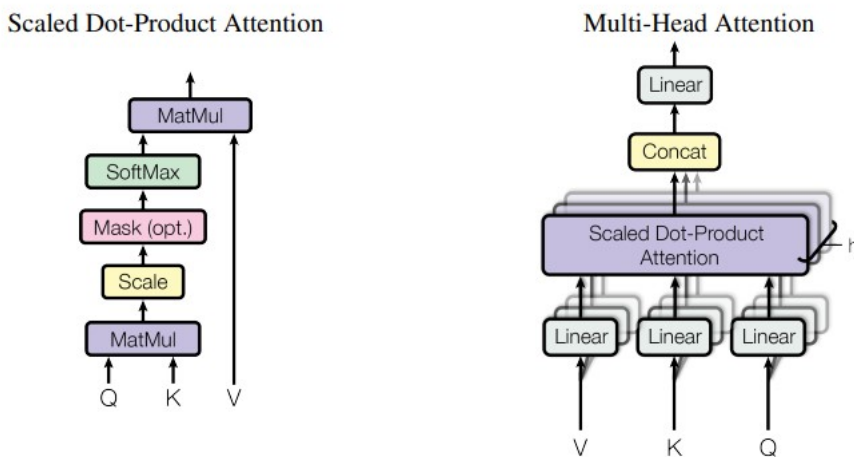
Kuten aiemmin mainittiin, transformer-arkkitehtuuri (eng. transformer architecture) on erittäin keskeinen suurten kielimallien toiminnassa. Arkkitehtuurissa käytetty huomiomekanismi (eng. attention mechanism) mahdollistaa erittäin pitkäaikaisen muistin, joka ei aiemmissa rakenteissa ollut mahdollista. [44] Periaatteessa resurssien ollessa rajattomat, myös pitkäaikainen muisti voi olla rajaton. Tässä aliluvussa käsitellään ensin huomiomekanismin toimintaa ja tämän jälkeen laajemmin transformer-arkkitehtuurin rakennetta. Aliluku pohjautuu pääosin lähteeseen [59].

Transformer-arkkitehtuuri on ensimmäinen neuroverkkorakenne, jossa syötteen ja ulostulon välisiä yhteyksiä tarkastellaan pelkästään huomiomekanismin avulla. Huomiomekanismissa malli pyrkii löytämään syötesarjasta keskeiset kohdat, joihin sen tulisi kiinnittää huomiota. Tämä tarkoittaa esimerkiksi tärkeitä sanoja tai yhteyksiä sanojen välillä. Huomiomekanismin avulla malli pystyy paremmin ymmärtämään sanojen välisiä yhteyksiä ja merkityksiä. Erityisesti luonnollisen kielen tehtäviä, kuten tiedonpoimintaa tai kääntämistä, ratkaistaessa onkin tärkeää, että malli ymmärtää sanojen ja lauseiden väliset yhteydet ja pystyy näin paremmin käsittelemään luonnollista kieltä.

Transformer-arkkitehtuurissa käytetään omahuomiomekanismia (eng. self-attention mechanism). Tämä eroaa huomiomekanismista siten, että omahuomiomekanismissa malli pyrkii tarkastelemaan syötesarjaa suhteessa itseensä, kun taas huomiomekanismissa malli tarkastelee syötesarjaa suhteessa muihin sarjoihin. Tässä yhteydessä sarja tarkoittaa esimerkiksi merkkijonoa. Syöte voi olla myös jotain muuta. Tässä tutkielmassa käsitellään kuitenkin luonnollista kieltä, joka voidaan useimmiten jakaa merkkijonoihin. Omahuomiomekanismi mahdollistaa lauseen sisäisten yhteyksien ymmärtämisen, koska se käsittelee syötesarjan osia suhteessa toisiinsa ja näin niiden merkitys muodostuu nimenomaan tässä asiayhteydessä. Näin omahuomiomekanismi toimii laajemmin luonnollisen kielen käsittelyn tehtävien ratkaisemisessa muun muassa kielen generoinnissa ja tiedonpoimintatehtävissä. Huomiomekanismia käytetään usein suoraviivaisempiin tehtäviin, jotka eivät vaadi monimutkaisten rakenteiden ja yhteyksien ymmärtämistä, vaan syötettä verrataan johonkin ulkopuoliseen sarjaan. Esimerkki tästä on käännöstehtävä, jossa pyritään kääntämään syö-

tesarja ulostulosarjaksi kääntämällä sana kerrallaan kohdekielelle.

Tässä tutkielmassa esitellään tarkemmin nimenomaan artikkelissa [59] transformer-arkkitehtuurissa käytetty skaalattu pistetulohuomiomekanismi (eng. scaled dot-product attention). Muita yleisiä huomiomekanismeja ovat Bahdanau-huomiomekanismi [5] ja pistetulohuomiomekanismi, joka on skaalausta lukuunottamatta samanlainen tässä tutkielmassa esitetyn huomiomekanismin kanssa.



Kuva 1: Huomiomekanismi. Vasemmalla skaalattu pistetulohuomiomekanismi ja oikealla monipäinen huomiomekanismi, joka hyödyntää skaalattua pistetulohuomiomekanismia [59]. Otettu muokkaamattomana artikkelista.

Ennen kuin suuri kielimalli voi suorittaa luonnollisen kielen tehtäviä, on luonnollinen kieli muutettava sellaiseen muotoon, että kielimalli voi käsitellä sitä. Tämä tapahtuu upotusten (eng. embeddings) ja paikkaenkoodauksen (eng. positional encoding) avulla muuttamalla luonnollista kieltä diskreeteistä kielen symboleista vektoriesitykseksi. [41] Upotukset voivat olla konteksteista riippuvaisia tai niistä riippumattomia. Konteksteista riippuvaisissa upotuksissa sanan upotus muuttuu dynaamisesti riippuen yhteydestä, jossa sitä käytetään. Tällaisia kontekstiriippuvaisia upotuksia käytetään myös transformer-arkkitehtuurissa. Tekstissä x_1, x_2, \dots, x_T , missä jokainen tokeni $x_t \in V$ merkitsee yksittäistä sanaa tai sanan osaa, tokenin kontekstiesitys riippuu koko tekstistä. [41] Merkintä V kuvaa sanastoa. Transformer-arkkitehtuurissa omahuomiomekanismi laskee jokaisen kahden sanan väliset yhteyspainot dynaamisesti. Upotusta käytetään muuttamaan syöte- ja ulostulotokenit vektorimuotoon dimensiossa d_{model} . Mallille on annettava tietoa tokeneiden suhteellisesta tai absoluuttisesta paikasta sarjoissa. Tämä voidaan saavuttaa paikkaenkoodauksen avulla. Enkooderin (eng. encoder) ja dekooderin (eng. decoder) alapuolella ole-

viin syöteupotuksiin lisätään paikkaenkoodaukset transformer-arkkitehtuurissa (2). Paikkaenkoodaukset ovat myös dimensiossa d_{model} , jotta ne voidaan laskea upotusten kanssa yhteen. Paikkaenkoodauksen avulla suuri kielimalli voi ymmärtää sanojen järjestystä lauseissa. Artikkelissa [59] käytettiin esimerkiksi eri taajuuksien sini- ja kosinifunktioita paikkaenkoodauksen suorittamiseen:

$$PE_{(pos,2i)} = \sin(pos/10000^{2i/d_{model}}) \quad (3)$$

$$PE_{(pos,2i+1)} = \cos(pos/10000^{2i/d_{model}}) \quad (4)$$

Suuret kielimallit käsittelevät luonnollista kieltä siis siitä muodostettujen vektorikuvausten kautta. Huomiomekanismia voidaan kuvailla kyselyjen ja avain-arvo-parien kuvauksena ulostulolle. Kyselyt Q , avaimet K , arvot V ja ulostulo esitetään kaikki vektoreina. Syöte koostuu kyselyistä ja avaimista dimensiossa d_k ja arvoista dimensiossa d_v . Käydään seuraavaksi läpi kuvassa (1) vasemmalla esitetty skaalattu pistetulohuomiomekanismi (eng. scaled dot-product attention) askel askeleelta. Kertomalla jokainen kysely Q kaikilla avaimilla K saadaan pistetulot, jotka kertovat jokaisen sanan merkityksen jokaiseen toiseen sanaan verrattuna eli kuinka todennäköisesti nämä sanat esiintyvät yhdessä (eng. MatMul, matrix multiplication). Tämän jälkeen tulot skaalataan (eng. scale) jakamalla ne kyselyjen ja avainten dimension d_k neliöjuurella. Näin tasataan tulojen gradientteja. Tämän jälkeen käytetään softmax-funktiota skaalatuille tuloille huomiopainojen saamiseksi. Nämä huomiopainot ovat todennäköisyysarvoja välillä $[0, 1]$. Softmax-funktio korostaa suurempia tuloja ja vastaavasti antaa vähemmän arvoa pienille tuloille. Huomiopainot auttavat mallia keskittymään tärkeisiin kohtiin. Huomiopainomatriisit kerrotaan vielä arvektorin kanssa ja näin saadaan ulostulomatriisit. Alla on esitetty pistetulohuomiomekanisimifunktio kokonaisuudessaan:

$$Attention(Q, K, V) = softmax\left(\frac{QK^\top}{\sqrt{d_k}}\right)V. \quad (5)$$

Kuvassa (1) oikealla esitetty monipäinen huomiomekanismi (eng. multi-head attention) koostuu useista rinnakkaisista huomiokerroksista. Kyselyille, avaimille ja arvoille muodostetaan h määrä lineaarikuvauksia dimensioissa d_k, d_k ja d_v . Jokainen kuvaus menee huomiokerroksen läpi, jolloin saadaan d_v -dimensioiset ulostulot. Teoriassa jokainen pää oppii eri asian, jolloin mallilla on enemmän esityskykyä. Esiityskyky tarkoittaa mallin kykyä esittää datan rakenteita ja siinä olevia suhteita

onnistuneesti. Näin saadut ulostulot yhdistetään yhdeksi (eng. *concat*) ja saadusta ulostulosta muodostetaan lineaarikuvaus (eng. *linear*). Monipäinen huomiomekanismifunktio esitetään kokonaisuudessaan seuraavasti:

$$MultiHead(Q, K, V) = Concat(head_1, \dots, head_h)W^O, \quad (6)$$

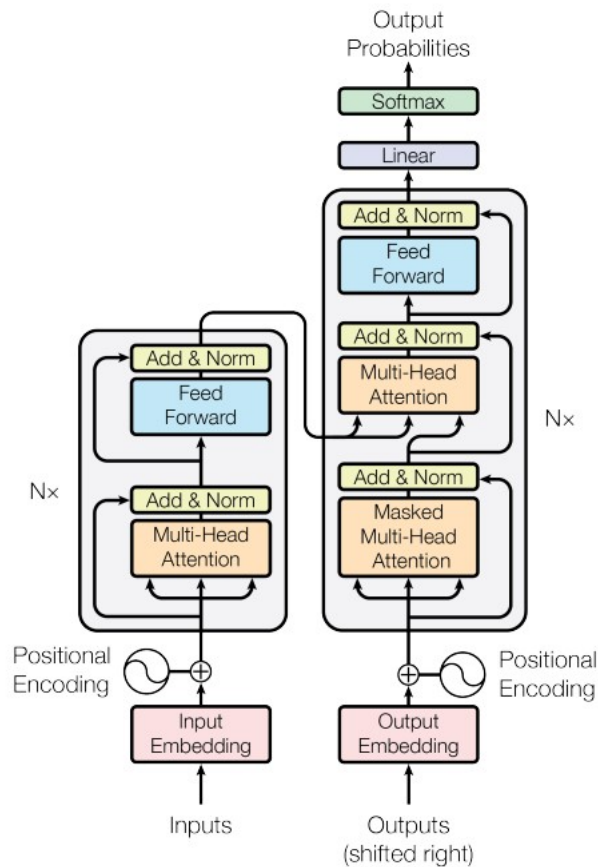
missä

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \quad (7)$$

Kuvaukset ovat parametrimatriiseja $W_i^Q \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{model} \times d_v}$ ja $W^O \in \mathbb{R}^{hd_v \times d_{model}}$. Monipäinen huomiomekanismi on keskeinen osa transformer-arkkitehtuuria ja se mahdollistaa rakenteen tehokkaan toiminnan erityisesti luonnollisen kielen tehtäviä ratkaistaessa. Transformer-arkkitehtuuri voi säilöä tietoa pitkäaikaisessa muistissa avain-arvo -pareina ja myöhemmin palauttaa tätä tietoa luomalla kyselyn, joka ottaa säilötyn tiedon huomioon. [44] Tämä mahdollistaa transformer-arkkitehtuurille tyypillisen pitkäaikaisen muistin.

Kuvan (2) transformer-arkkitehtuuri koostuu enkooderi- ja dekooderipinoista. Enkooderia ja dekooderia käytetään suurten kielimallien lisäksi myös muissa neuroverkkoihin perustuvissa rakenteissa. Enkooderi koodaa symbolien muodostaman syötesarjan (x_1, \dots, x_n) sarjaksi jatkuvia esityksiä (z_1, \dots, z_n) . Dekooderi taas purkaa esitykset z ja muodostaa niiden avulla ulostulon (y_1, \dots, y_n) yksi elementti kerrallaan. Malli on autoregressiivinen jokaisessa askeleessa eli se käyttää jo-muodostettuja ulostulon elementtejä lisäsyötteenä muodostaessaan seuraavia ulostuloelementtejä.

Enkooderipino on kuvassa (2) vasemmalla. Ennen varsinaista enkooderipinoa syötteen (eng. *inputs*) muutetaan aiemmin selitetyllä tavalla vektoriesitykseksi upotusten (eng. *input embedding*) ja paikkaenkoodauksen (eng. *positional encoding*) avulla. Enkooderipino koostuu kuudesta identtisestä kerroksesta, jotka kukin muodostuvat kahdesta alikerroksesta. Alikerrokset on merkitty kuvassa erikseen laatikoilla ja nuolilla, itse kerrokset ovat keskenään samanlaisia. Niitä merkitään sivussa olevalla merkinnällä Nx , tässä tapauksessa $N = 6$. Ensimmäinen alikerros on monipäinen omahuomiomekanismi (eng. *multi-head attention*) ja toinen alikerros eteenpäinkytketty neuroverkko (eng. *feed forward*). Eteenpäinkytketyssä neuroverkkossa edellisen neuroverkkokerroksen ulostuloa käytetään seuraavan neuroverkkokerroksen syötteenä eikä rakenteessa tapahtuvia silmukoita sallita. [35] Molempien



Kuva 2: Transformer-arkkitehtuuri. Kuvassa enkooderipino on vasemmalla, ja se toistuu rakenteessa kuusi kertaa. Dekooderipino on kuvassa oikealla ja toistuu myös rakenteessa kuusi kertaa. Enkooderi- ja dekodeeripinot muodostuvat kerroksista, jotka koostuvat alikerroksista. Jokaisen alikerroksen jälkeen käytetään jäännösyhteyttä ja kerroksen normalisointia (eng. Add & Norm). Otettu muokkaamattomana artikkelista [59]

alikerrosten jälkeen käytetään jäännösyhteyttä (eng. add) ja kerroksen normalisointia (eng. norm). Syvempiä neuroverkkoja on haastavampaa opettaa. Jäännösyhteydessä kerroksen ulostuloon lisätään suoraan sen syöte. Näin jäännösyhteyksiä käyttävät neuroverkkomallit oppivat paremmin suhteessa syötteeseen. Jäännösyhteydet helpottavat neuroverkkomallin optimoinnissa ja auttavat usein mallia saavuttamaan paremman tarkkuuden neuroverkkokerrosten määrän noustessa. [16]

Syvien neuroverkkojen opettaminen on tietoteknisesti kallista. Kerrosten normalisoinnin tavoitteena on standardisoida syötteitä, jolloin malli voi käsitellä tietoa johdonmukaisemmin. Tämä vähentää syvien neuroverkkojen opetusaikaa. [26] Kun käytetään jäännösyhteyksiä ja kerrosten normalisointia, jokaisen alikerroksen ulos-

tulo on

$$\text{LayerNorm}(x + \text{SubLayer}(x)), \quad (8)$$

missä $\text{SubLayer}(x)$ on itse alikerroksessa toteutettu funktio ja x on alikerroksen käytämä syöte.

Dekooderipino on kuvassa (2) oikealla. Myös ennen varsinaista dekooderipinoa suoritetaan upotukset (eng. output embeddings) ja paikkaenkoodaus (eng. positional encoding) enkooderista saadulle ulostulolle. Dekooderipino koostuu myös kuudesta identtisestä kerroksesta ja nämä kerrokset muodostuvat kukin kolmesta alikerroksesta, joista kaksi ovat samoja kuin enkooderipinossa. Kolmas alikerros on maskattu monipäinen omahuomiokerros (eng. masked multi-head attention). Myös dekoodeerissa käytetään jokaisen alikerroksen jälkeen jäännösyhteyttä ja kerrosten normalisointia.

Transformer-arkkitehtuurissa käytetään monipäistä huomiomekanismia kolmella tavalla. Dekooderipinossa alimmaisena kuvassa (2) oleva monipäinen omahuomiokerros on maskattu (eng. masked multi-head attention). Tämä tarkoittaa, että kaikkiin skaalatuun pistetulon (1) softmax-funktion syötteessä oleviin ei-sallittuihin yhteyksiin liitettävät arvot on asetettu arvoiksi $-\infty$. Näin voidaan estää tiedon liikkuminen oikealta vasemmalle, eli malli ei voi käyttää tietoa tulevista sanoista nykyistä tilaa määrittäessä. Tämä on tärkeää mallin autoregressiivisyyden säilyttämisessä. Dekooderipinon toinen huomiokerros toimii enkooderi-dekooderi -huomiokerroksena. Siinä kyselyt saadaan edellisestä dekooderikerroksesta ja avaimet ja arvot saadaan enkooderin ulostulosta. Näin dekooderi voi kiinnittää huomiota jokaiseen syötteen kohtaan ulostuloa muodostettaessa. Enkooderipinon omahuomiokerroksissa kyselyt, arvot ja avaimet saadaan enkooderin edellisen kerroksen ulostulosta. Enkooderipinon omahuomiokerros kiinnittää huomiota jokaiseen edellisen enkooderikerroksen kohtaan ulostuloa muodostettaessa.

3.3 Enkooderi- ja dekooderipohjaiset mallit

Edellisessä aliluvussa esiteltiin alkuperäinen transformer-arkkitehtuuri, joka sisältää sekä enkooderin että dekooderin. Artikkelissa [59] alkuperäistä transformer-arkkitehtuuria käytettiin kielen kääntämisessä englannista ranskaksi ja saksaksi. Tätä tehtävää ratkaistaessa enkooderin tehtävä oli oppia ymmärtämään syötekieltä ja poimia syötteestä keskeiset tiedot. Dekooderin tehtävä taas oli generoida tekstiä ulostulokielellä enkooderin antaman tiedon perusteella. Toisenlaisia tehtäviä ratkaistaessa on usein riittävää käyttää suuria kielimalleja, jotka hyödyntävät vain enkooderia

tai dekooderia rakenteessaan. Tässä aliluvussa käydään läpi enkooderi- ja dekooderipohjaisten suurten kielimallien eroja ja niiden käyttötarkoituksia. Tämän tutkielman implementointiosuudessa käytetään neljää dekooderipohjaista suurta kieli-mallia: GPT-3.5 -mallia, GPT-4 -mallia, Gemini-1.0-pro-001 -mallia ja Llama3-70b -mallia. Tämän takia keskitytään erityisesti dekooderipohjaisten kielimallien toimintaan ja vahvuuksiin.

Enkooderipohjaisia malleja kutsutaan AET-malleiksi (Auto-Encoding Transformers). AET-malleissa tavoitteena on luoda kaksisuuntainen enkooderirepresentaatio kokonaisista lauseista. Tämä tarkoittaa, että malleilla on rajoittamaton pääsy kaikkiin syötteisiin ulostuloa muodostettaessa. Tällöin mallit pystyvät käyttämään kaksisuuntaisesti sekä menneitä että tulevia lauseen osia nykyistä ulostuloa muodostaessaan. Esimerkki AET-malleista on Googlen kehittämä BERT[10]. AET-mallit sopivat yleensä erityisesti luonnollisen kielen ymmärtämisen tehtäviin, kuten lauseiden lajitteluun niiden merkityksen mukaisesti. Lisäopetuksen ja hienosäädön avulla AET-mallit saavuttavat hyviä tuloksia myös muunlaisissa tehtävissä, kuten tekstin generoinnissa. [44]

Dekooderipohjaisia malleja kutsutaan ART-malleiksi (Auto-Regressive Transformers). ART-malleissa ensisijainen tavoite on ennustaa seuraava sana jo-luetun tekstin perusteella ja mallien yleisin käyttötarkoitus on tekstin generointi. Tämä ei kuitenkaan tarkoita, etteivät mallit sopisi myös toisenlaisten tehtävien ratkaisemiseen. ART-mallien keskeisin ero AET-malleihin verrattuna on se, että niillä ei ole pääsyä tuleviin sanoihin nykyistä ulostuloa muodostettaessa. [44] Tämän tutkielman implementointiosuudessa ratkaistaan tiedonpoimintatehtävää ART-malleilla.

Hienosäätö ja lisäopetus vaativat usein paljon tehtäväkohtaista aineistoa ja laajoja resursseja, mitä ei välttämättä ole saatavilla. Tällaisia tehtäviä ratkaistaessa ART-mallien vahvuudeksi osoittautuu niiden valmius ratkaista erilaisia tehtäviä jo saatavilla olevien mallien avulla ilman erillistä lisäopetusta tai hienosäätöä. Tässäkin tutkielmassa päädyttiin käyttämään ART-malleja muun muassa tästä syystä.

Tässä tutkielmassa ratkaistaan lääketieteellisten tekstien rakenteellistamisteh-tävää. ART-malleja on käytetty samankaltaisten tehtävien ratkaisemiseksi esimerkiksi artikkeleissa [1], [13], [18] ja [66]. Yleisimmin käytetty ART-malli on jokin GPT-malleista. On kuitenkin otettava huomioon, että GPT-1 on julkaistu jo vuonna 2018, kun taas Geminin ja Llamen ensimmäiset versiot julkaistiin vasta vuonna 2023. Myös jälkimmäiset ovat lupaavia tällaisten tehtävien ratkaisemisessa. Kun käytetään ART-malleja lääketieteellisten tekstien rakenteellistamisessa, on parempien ratkaisujen saamiseksi yleensä käytetty myös kehotesuunnittelua.

3.4 Kehotesuunnittelu

Kehote (eng. prompt) tarkoittaa suurelle kielimallille annettavia ohjeita, joiden avulla voidaan vahvistaa sääntöjä, automatisoida prosesseja tai varmistaa tiettyjä piirteitä mallien antamissa vastauksissa. [62] Kehotesuunnittelu (eng. prompt engineering) taas on kehotteiden suunnittelua, kehittämistä ja implementointia. Sen avulla voidaan ohjata suurten kielimallien ulostuloa halutun kaltaiseksi. [34] Kehotesuunnittelulla voidaan parantaa suurten kielimallien toimintaa ja sen merkitys lääketieteessä kasvaa suurten kielimallien yleistyessä erilaisissa tehtävissä. Tässä aliluvussa käydään läpi kehotesuunnittelun peruseriaatteita ja joitakin kehotesuunnittelun tekniikoita

Kuten edellisessä aliluvussa todettiin, esiopettujen suurten kielimallien hienosäätäminen vaatii paljon aikaa, rahaa, dataa ja tietoteknisiä resursseja. Kehotesuunnittelu vaatii huomattavasti vähemmän resursseja, ja tämän takia se onkin herättänyt laajaa mielenkiintoa. Kehotesuunnittelun avulla suuret kielimallit voivat ratkaista erilaisia tehtäviä hyvinkin pienellä määrällä merkittävää dataa. [60]

Kehotesuunnittelu tarkoittaa prosessia, jossa luodaan kehoite, joka ohjaa mallia tuottamaan tehokkaan ratkaisun jatkotehtävään. Jatkotehtävä tarkoittaa sellaista tehtävää, jonka ratkaisemiseen suurta kielimallia ei varsinaisesti ole opetettu. Sen ratkaisemisessa malli voi kuitenkin käyttää muuten oppimiaan taitoja. Kehotteet voidaan jakaa kahteen ryhmään: täydennettäviin kehotteisiin (eng. cloze prompt) ja etuliitekehotteisiin (eng. prefix prompt). Täydennettävissä kehotteissa mallia pyydetään täydentämään jokin tekstiin jätetty aukko kehotteen antamien ohjeiden mukaisesti. Etuliitekehotteissa syöteteksti taas on esitetty kokonaisuudessaan ennen mallin generoimaa vastausta. [60] Tämän tutkielman implementointiosuudessa käytetään etuliitekehotteita, mallille annetaan kaikki ohjeet ensin ja sen jälkeen mallilta odotetaan ohjeiden mukaisesti generoitua vastausta.

Kehotteet voidaan jakaa kahteen osaan myös manuaalisiin ja automaattisiin kehotteisiin. Ihmiset kirjoittavat manuaalisia kehotteita, ja niissä malleille annetaan tarkat ohjeet siitä, mihin osiin dataa mallin tulee keskittyä ja miten annettua tehtävää tulee lähestyä. Automaattiset kehotteet taas luodaan algoritmien ja tekniikoiden avulla, eikä niiden tekemisessä vaadita ihmisen puuttumista kehotteen käyttämisessä. Esimerkiksi kehotelouhinnassa (eng. prompt mining) suuresta korpuksesta etsitään usein esiintyviä yhdistäviä sanoja tai riippuvuuksia syöte-ulostuloparien välillä ja näin muodostetaan toimivia kehotteita. Näitä yhteyksiä voidaan etsiä jonkin algoritmin avulla. Tämän tutkielman implementointiosuudessa käytetään manuaalisia kehotteita. Tällainen kehoite toimii parhaiten silloin, kun aineisto on hyvin määriteltyä ja ulostulon tulee olla tietyssä muodossa. Manuaalisia kehotteita kirjoittaessa

tulee kuitenkin ottaa huomioon niiden vaatima aika ja asiantuntemus. Mallit saattavat vastauksissaan reagoida pieniinkin muutoksiin kehoitteissa, joten kehoitteiden onnistunut muotoilu on erittäin tärkeää tehtävän onnistumisen kannalta. [60]

Kehotesuunnittelutekniikat auttavat malleja ratkaisemaan monimutkaisia ongelmia, välttämään hallusinaatioita ja antamaan tarkempia vastauksia. [46] Eräitä yleisimmistä tekniikoista ovat alan asiantuntija - ja ajatusketjutekniikat. Alan asiantuntija -tekniikassa suurta kielimallia pyydetään toimimaan tietyn alan tai aiheen asiantuntijana vastatessaan kysymyksiin. Suurille kielimalleille voidaan antaa jokin rooli, minkä perusteella niiden tulisi antaa vastaukset. Tällöin tavoitteena on, että tekniikka auttaisi suurta kielimallia valitsemaan oikeanlaisen ulostulon generoidessaan vastausta ja se keskittyisi tehtävälle keskeisiin yksityiskohtiin. [62] Ajatusketjutekniikassa mallia taas pyydetään ajattelemaan kysymystä askel askeleelta. Tämän tekniikan tavoitteena on, että malli tehtävää ratkaistessaan paloittelisi monimutkaisen tehtävän useisiin pienempiin ja helpompiin tehtäviin. [24][61] Eräs usein käytetty tekniikka on myös yhdistellä useita kehoitesuunnittelutekniikoita sopivalla tavalla tehtävän ratkaisemiseksi.

Luonnollisen kielen muotoilun lisäksi mallien antamiin vastauksiin vaikuttavat niille annetut esimerkit toivotuista vastauksista. Annettujen esimerkkien määrä voi olla nollan ja usean kymmenen välillä ja esimerkkien antaminen voi vaikuttaa mallien suorituksiin.

Harvaoppimisessa (eng. few-shot learning) mallille annetaan joitakin esimerkkejä tehtävän suorittamisesta. Mallin toivotaan oppivan tehtävän ratkaiseminen näiden esimerkkien perusteella. Tyypillisesti annetuissa esimerkeissä on tehtävään tarvittava konteksti ja toivottu tulos. Esimerkiksi käännöstehtävässä mallille voitaisiin antaa esimerkeiksi jokin määrä jo-käännettyjä sanapareja ja tämän lisäksi ohje kääntää annettu sana halutulle kielelle.

Yksioppiminen (eng. one-shot learning) toimii muuten samalla tavoin kuin harvaoppiminen, mutta siinä mallille annetaan vain yksi esimerkki suoritettavan tehtävän ratkaisusta. Yksioppimista käytetään, sillä se on usein ihmisen tapa suorittaa tehtävää. Ihminen saa usein vain yhden esimerkin suoritettavasta tehtävästä ennen sen tekemistä.

Nollaoppimisessa (eng. zero-shot learning) mallille ei anneta lainkaan esimerkkejä vaan mallin tulee suorittaa tehtävä pelkästään annettujen luonnollisen kielen ohjeiden perusteella. Tämä lähestymistapa on kaikista haastavin, mutta onnistuessaan sen hyviä puolia ovat vastausten robustisuus ja esimerkeistä saatujen väärin yhteyksien puuttuminen. Nollaoppimisessa on tärkeää muotoilla luonnollisen kielen ohjeet riittävän selkeästi, jotta malli osaa suorittaa tehtävän ilman esimerkkejä.

Kehotesuunnittelun avulla voidaan ratkaista tehokkaasti sellaisiakin lääketieteellisiä tehtäviä, joiden ratkaisu olisi aiemmin vaatinut suuria määriä aineistoja ja paljon resursseja. Esimerkki tällaisesta tehtävästä on tässäkin tutkielmassa ratkaistava rakenteellistamistehtävä. Muita kehoite-suunnittelulla ratkaistavia lääketieteellisiä tehtäviä ovat esimerkiksi lajittelutehtävät, kysymyksiin vastaaminen lääketieteellisen tiedon perusteella ja päättelytehtävät, joissa mallin tavoitteena on perustella annetun skenaarion diagnostinen prosessi tai kliininen päättely. [60]

3.5 Suurten kielimallien haasteet

Suuret kielimallit tuovat paljon uusia mahdollisuuksia erilaisten ongelmien ratkaisemiseen. Niiden suhteellisen uutuuden ja nopean kehityksen takia niitä käyttäessä tulee kuitenkin muistaa niiden rajoitteet ja haasteet. Tässä aliluvussa esitellään keskeisimpiä suurten kielimallien haasteita ja niiden kehitysmahdollisuuksia tulevaisuudessa. Ensin käydään läpi suurten kielimallien käytön rajoitteita ja sen jälkeen eettisiä ongelmia, joita suurten kielimallien käytöstä voi seurata. Lopuksi annetaan joitakin esimerkkejä suurten kielimallien haitallisesta käytöstä. Tämä aliluku perustuu pääosin lähteeseen [55].

Kuten edellä todettiin, suuria kielimalleja opetetaan erilaisilla teksteillä ja aineistoilla. Opetus on kuitenkin lopetettava jossain kohtaa, kun malli otetaan käyttöön. Tämä tarkoittaa, että mallia on opetettu aineistolla tiettyyn päivään asti. Sen jälkeisestä kehityksestä malli ei tiedä mitään. Tieteessä kehitys saattaa olla hyvinkin nopeaa, ja tällöin mallien vastaukset saattavat olla vääriä tai niiden tieto vanhentunutta. Suurilta kielimalleilta puuttuu viimeaikaaisuutta (eng. recency), jonka avulla niiden tieto olisi ajan tasalla. Tätä voitaisiin korjata opettamalla malleja uudemmalla aineistolla tai vaihtoehtoisesti ajankohtaista opetusaineistoa voitaisiin saada internetistä reaaliajassa.

Toinen suurten kielimallien rajoitteista on niiden luotettavuus (eng. accuracy). Opetusaineistona käytetään laajaa valikoimaa erilaisia tekstejä ja näiden tekstien todenmukaisuuden tarkistamiseen tai tiedon validointiin ei vielä ole kehitetty sopivaa tapaa. Tällöin siis suuria kielimalleja saatetaan opettaa tiedolla, joka ei ole todenmukaista ja malli antaa näin vastauksensa tämän tiedon mukaisesti myös virheellisesti. Mahdollisia ratkaisuja näihin ongelmiin voisivat olla opetusaineiston validointi ja sen todenmukaisuuden tarkistaminen sekä hienosäätö tietyn alan tekstin ymmärtämisen helpottamiseksi.

Suuret kielimallit eivät ymmärrä kielen semanttista merkitystä. Sen sijaan ulostulot perustuvat opituille sanojen välisille yhteyksille. Koska suuret kielimallit eivät ymmärrä luonnollista kieltä samalla tavalla kuin ihmiset, eivät ne osaa arvioida an-

tamiensa vastausten todenmukaisuutta ja loogisuutta samalla tavalla. Suuret kielimallit saattavatkin antaa vastauksena hallusinaatiota. Hallusinaatioista puhutaan silloin, kun suuri kielimalli generoi sisältöä, joka eroaa käyttäjän antamasta syötteestä, on ristiriidassa aiemmin generoidun sisällön tai vakiintuneen tiedon kanssa. Hallusinaatiot vähentävät suurien kielimallien luotettavuutta oikean elämän sovelluksissa, sillä usein kielimallit esittävät hallusinaatiot uskottavina osina vastauksia ja joskus käyttäjän voi olla vaikeaa erottaa oikeaa vastausta hallusinaatiosta. [67] Joissakin suurissa kielimalleissa on käytetty suojausmekanismia, jotka estävät mallien vastaamisen kliinisiin kysymyksiin mahdollisten vastausten haitallisuuden tai aineiston arkaluonteisuuden vuoksi. Tämä saattaa vaikeuttaa suurten kielimallien käyttöä lääketieteellisissä sovelluksissa.

Usein on myös epäselvää, miten suuri kielimalli on saanut antamansa vastauksen. Ihmisen on vaikea selvittää, miten suuri kielimalli prosessoi saamaansa syötettä ja miten se muodostaa antamansa vastauksen tai tekee päätöksiä. Tällaisia ongelmia kutsutaan ”black box” -ongelmiksi. Jos ei tiedetä, miten vastauksia muodostetaan, on mahdollisia virheitä myös haastavaa korjata. Tätä ongelmaa voitaisiin ratkaista asettamalla vaatimuksia siitä, että suurten kielimallien tulisi ulostulossa esittää myös koko päätöksentekoprosessi. Kehitteillä on myös selitettävää tekoälyä (eng. explainable AI), joka voisi parantaa suurten kielimallien tulkittavuutta. Tämä teknologia luonnollisen kielen käsittelyn yhteydessä on kuitenkin vielä uutta ja puutteellista.

Suurten kielimallien eettiset ongelmat ovat olleet laajasti esillä ja maaliskuussa 2023 avoin kirje niiden kehittämisen väliaikaiselle pysäyttämiseksi [45] keräsi laajaa huomiota maailmanlaajuisesti. Suuret kielimallit voivat tuottaa tekstiä, jota ei usein voi erottaa ihmisen kirjoittamasta tekstistä. Tämä voi johtaa tahallisiin ja tahattomiin väärinkäytöksiin ja suuri osa kielimallien eettisistä ongelmista johtuu juuri tästä.

Suurten kielimallien oppimat asiat riippuvat opetusaineistosta. Kuten edellä todettiin, nämä opitut asiat eivät välttämättä aina ole todenmukaisia tai ne saattavat olla vihamielisiä tai edustaa jotakin väestöryhmää liiallisesti tai vajavaisesti. Tällä voi olla seurauksia suurten kielimallien toiminnassa. Suuret kielimallit saattavat pitää yllä haitallisia stereotyyppioita tai ennakoasenteita. Ne saattavat myös käyttää vihamielistä kieltä tai toimia epäsuhtaisesti eri väestöryhmille tai aiheuttaa harmia huonommista lähtökohdista tuleville ihmisille. Jos suuria kielimalleja on opetettu arkaluonteisilla aineistoilla, saattavat ne aiheuttaa tietoturvariskejä.

Suurten kielimallien tuottaman misinformaation haitat voivat olla sekä yksilökohtaisia että koko yhteiskunnan laajuisia. Yksilötasolla kielimalli saattaa antaa

epäluotettavaa tietoa esimerkiksi lääketieteellisiin ongelmiin liittyen tai yleisesti edesauttaa väärän tiedon leviämistä ja vähentää yleistä tietoon luottamista. Yhteiskunnallisella tasolla suuria kielimalleja saatetaan käyttää lisäämään disinformaatiokampanjoiden tehokkuutta, luomaan huijauksia tai kirjoittamaan koodia esimerkiksi viruksia tai aseita varten. Esimerkiksi artikkelissa [15] on esitetty, kuinka lääketieteelliset suuret kielimallit ovat alttiita vahingollisille kohdistetuille misinformaatiohyökkäyksille.

Suurten kielimallien avulla luodut keskustelukumppanit (eng. conversational agent) ovat hyvin ihmismäisiä ja tällöin ihmisen voi olla vaikeaa ymmärtää tämän keskustelukumppanin rajoitteet ja vaarat. Ihmiset saattavat yliarvioida keskustelukumppanin taitoja tai jakaa hyvinkin henkilökohtaisia asioita sen kanssa. Suurten kielimallien toiminta on uutuutensa vuoksi vielä melko vierasta suurelle osalle ihmisistä ja tällöin ei välttämättä ymmärretä millaista tietoa niiden kanssa kannattaa jakaa tai millaisissa asioissa keskustelukumppaniin kannattaa luottaa.

4 Implementointi

Seuraavaksi esitetään käytännön sovellus lääketieteellisten lausuntojen rakenteellistamisesta suurten kielimallien avulla. Tehtävän tarkoituksena oli löytää suomenkielisistä histopatologisista potilaslausunnoista ennalta määrättyjä tietoja suurten kielimallien avulla. Tutkielmassa esitelty tulokset ja näiden taustalla olevat synteettiset lausunnot on esitelty aiemmassa muodossa myös konferenssijulkaisussa [21]. Implementoinnissa haluttiin vertailla erilaisten eniten huomiota saaneiden suurten kielimallien toimintaa ja niille annettujen kehoitteiden vaikutusta saatuihin vastauksiin. Tässä luvussa käydään ensin läpi tehtävän lähtökohdia ja käytettävää aineistoa sekä tehtävän rajoituksia. Tämän jälkeen esitellään mahdollisia lähestymistapoja ja valitaan niistä sopivimmat tehtävän ratkaisemiseksi. Lopuksi esitellään tehtävän ratkaisuyritys.

4.1 Implementoinnin lähtökohdat

Implementoinnissa käsitellään eturauhassyöpäpotilaiden histopatologisia lausuntoja. Tällaiset lausunnot sisältävät paljon keskeistä tietoa eturauhassyövästä ja tätä tietoa voidaan käyttää edelleen monella tavalla esimerkiksi tutkimuksessa, potilaiden hoidontarpeen arvioinnissa ja erilaisissa ennusteissa. Haasteena on kuitenkin lausuntojen vapaatekstimuotoisuus. Tämän takia ne onkin ensin rakenteellistettava, eli niiden muotoilu ja sisältö on yhtenäistettävä, jotta niissä olevaa tietoa voidaan käyttää tehokkaasti jatkotehtävissä. Tässä aliluvussa esitellään ensin yleisesti eturauhassyövän piirteitä, tämän jälkeen käydään läpi tässä tutkielmassa käytettyä aineistoa ja sen haasteita. Lopuksi esitellään vielä implementoinnin päätavoitteet ja erityiset mielenkiinnon kohteet suorituksia tarkastellessa.

Syöpä on laaja kansanterveydellinen, yhteiskunnallinen ja taloudellinen haaste. Vuoden 2022 eturauhassyöpätilastojen mukaan se aiheuttaa maailmanlaajuisesti lähes joka kuudennen kaikista kuolemista (16,8%) ja joka neljännen tarttumattomista sairauksista johtuvista kuolemista (22,8%). Eturauhassyöpä on toiseksi yleisin syöpädiagnoosi ja syöpäkuoleman aiheuttaja miehillä maailmanlaajuisesti keuhkasyövän jälkeen. Yleisimpiä eturauhassyöpädiagnoosit ovat yli 65-vuotiailla miehillä. Väestön ikääntyminen onkin keskeinen tekijä eturauhassyövän esiintyvyyden nousussa. Vuonna 2022 eturauhassyöpään sairastuneita oli 7,3% kaikista syöpään sairastuneista maailmanlaajuisesti.[6] Suomessa eturauhassyöpä on miesten yleisin syöpä. Vuonna 2022 diagnosoitiin 5514 uutta tapausta. Elämänsä aikana eturauhassyöpään sairastuu 13,9 % ja siihen kuolee 3,9 % suomalaisista miehistä. [39]

Eturauhassyöpädiagnoositiikassa yksi tavallisimpia aggressiivisuuden arviointikri-

teerejä on ns. Gleason-luokitus, joka suoritetaan neulabiopsiasta otetuista kudoksenäytteistä. Gleason kuvaa solujen erilaistumisen piirteitä, jotka histopatologi arvioi mikroskoopin avulla. Korkeammat arvot kuvaavat näytteessä näkyviä huonoennusteisempia solupiirteitä. Syöpälöydöksen Gleason-luokitus annetaan kahtena lukuna ja niiden summana. Ensimmäinen luku kuvaa yleisintä näytteessä nähtävää kasvutapaa ja toinen luku toiseksi yleisintä. Gleason-luokitus perustuu syöpäsolujen rauhasmuodostukseen. Luvut annetaan asteikolla 1-5, jossa 1 tarkoittaa hyvin erilaistunutta ja 5 huonosti erilaistunutta syöpäfokusta. Näin esimerkiksi Gleason-luokitukset $3+4 = 7$ ja $4+3 = 7$ eroavat ennusteellisesti toisistaan. Gleason-luokitus on diagnoosivaiheessa eturauhassyövän tärkein ennustetekijä. [47] Eturauhassyövän ensioireet ovat pääasiassa samanlaisia kuin hyvälaatuisen eturauhasen liikakasvun eli BPH:n (Benign Prostatic Hyperplasia) oireet. BPH ei kuitenkaan itsessään ole eturauhassyövän oire. Näytteessä havaittu atypia tarkoittaa solujen tai solutumien lievää epänormaaliutta, jonka syynä voi olla esimerkiksi kasvainprosessi. Perineuraalinen kasvu tarkoittaa syövän metastaasia eli leviämistä sen pääkasvaimen ulkopuoliseen ympäristöön, nimenomaan hermon läheisyyteen. Perineuraalinen kasvu antaa yleensä huonon ennusteen syöpään sairastuneelle potilaalle. Gleason-luokitus, BPH, atypia ja perineuraalinen kasvu ovat kaikki tekijöitä, joita tarkkailemalla voidaan ennustaa eturauhassyövän etenemistä, sen hoidontarvetta ja vakavuutta.

Histopatologiset lausunnot ovat yleensä patologiaan erikoistuneen lääkärin kirjoittamia. Patologiaan erikoistuneet lääkärit tutkivat saamiaan näytteitä ja kuvailevat niitä vapaatekstimuodossa patologisissa lausunnoissa. Patologisissa lausunnoissa voidaan kertoa muun muassa se, miltä saatu näyte näyttää verrattuna normaaliin kudokseen ja soluihin, mikä on diagnoosi ja mahdolliset tehdyt diagnostiset testit, kuten kasvaimen koko tai sen luokitus. Patologisten lausuntojen avulla voidaan esimerkiksi tehdä päätöksiä tarvittavasta hoidosta. [9] Tässä tutkielmassa käsiteltävät synteettiset histopatologiset lausunnot olivat eturauhassyöpänäytteistä kirjoitettuja.

Tässä tutkielmassa käytetyt histopatologiset potilaslausunnot olivat vapaatekstimuotoisia lääketieteellisiä lausuntoja, jotka oli kirjoitettu erikseen tätä käytännön sovellusta varten. Lausunnot oli manuaalisesti kuratoitu oikeaa dataa muistuttaviksi, mutta niissä kuvattuja näytteitä tai potilaita ei ole koskaan ollut olemassa. Alkuperäinen aineisto koostui tuhansista lääketieteen ammattilaisten kirjoittamista aidoista potilaslausunnoista, joissa oli luonnollisesti mainittu myös potilaiden henkilökohtaisia tietoja. Alkuperäisen aineiston arkaluonteisuuden vuoksi sitä ei siis voinut suoraan käyttää tämän tutkielman implementointiosuudessa, mutta tehtävässä käytetty aineisto on muodoltaan ja sisällöltään hyvin samankaltaista alkuperäisen

aineiston kanssa. Näin tässä tutkielmassa saatujen tulosten perusteella voidaan tehdä päätelmiä, joiden perusideat pätevät myös alkuperäiseen aineistoon.

Tätä tutkielmaa varten kirjoitettuja histopatologisia lausuntoja oli 30 kappaletta. Niistä haluttiin löytää suurten kielimallien avulla ennalta määritettyjä tietoja, jotka ovat keskeisiä eturauhassyöpätutkimuksen kannalta. Näitä tietoja olivat:

Lista 4.1 Suurilta kielimalleilta kysytyt asiat.

1. Otettujen näytteiden määrä (numeerinen vastaus)
2. Otettujen näytteiden koko (numeerinen vastaus)
3. Näytteiden Gleason-luokitus (numeerinen vastaus)
4. Mahdollinen eturauhasen hyvälaatuinen liikakasvu (kyllä/ei)
5. Mahdollinen perineuraalinen kasvu (kyllä/ei)
6. Mahdollinen atypia (kyllä/ei)

Suurilta kielimalleilta kysyttiin siis kuusi kysymystä. Suluissa on esitetty kysymyksiin odotettujen vastausten muoto. Tämän lisäksi kielimalleille annettiin vastausvaihtoehto tilanteita varten, joissa vastausta kysymykseen ei voitu suoraan löytää annetusta histopatologisesta lausunnosta. Histopatologisten lausuntojen lisäksi kirjoitettiin 10 kappaletta ns. puppulausuntoja, jotka eivät liittyneet eturauhassyöpään eivätkä olleet potilaslausuntoja. Puppulausunnot olivat muun muassa runoja, merkityksettömiä merkkijonoja ja lainauksia kirjoista. Näiden avulla haluttiin tarkkailla suurten kielimallien toimintaa tilanteissa, joissa ennalta määritettyjä tietoja ei aineistosta ollut saatavilla. Koska synteettiset lausunnot oli luotu manuaalisesti itse, olivat oikeat vastaukset kysymyksiin saatavilla.

Oman haasteensa implementointiin toi lausuntojen suomenkielisyys. Suomen kieli on pienten resurssien kieli, toisin kuin esimerkiksi englanti. Suurten kielimallien sovellukset keskittyvät usein suurempien resurssien kieliin. Resurssit tarkoittavat tässä yhteydessä saatavilla olevaa aineistoa esiopetuksessa. Suurten resurssien kielissä, kuten englannissa, tätä aineistoa on saatavilla lähes rajattomasti. Suurten kielimallien opettaminen on erittäin kallista eikä pienempien resurssien kielillä ole saatavilla samanlaista määrää aineistoa. Suomen kieltä puhuu alle 0,1% maailman väestöstä. [32] Suomen kieli on edustettuna joissakin transformer-arkkitehtuuria käyttävissä malleissa ja tämän lisäksi artikkelissa [32] esitellään joukko GPT-malleja ja BLUUMI, joita on opetettu nimenomaan suomenkielisellä aineistolla. Näiden lisäksi

TurkuNLP-ryhmä on luonut yhteistyössä SiloAI:n kanssa mallit Poro [31] ja Viking 33B [30]. Näiden mallien tarkoituksena on tarjota toimivia suuria kielimalleja myös eurooppalaisille kielille englannin lisäksi ja niitä on opetettu erityisesti suomen kielellä. Tässä tutkielmassa käytetään sellaisia suuria kielimalleja, joita ei ole erikseen esiopetettu suomenkielisellä aineistolla. Näin halutaan tarkastella suurten kielimallien valmiutta toimia suomen kielellä ilman erillistä suuria resursseja vaativaa esiopetusta. Suurimpien yritysten mallit ovat perusmalleja (eng. foundation model), joiden on tarkoitus soveltua kaikkiin tilanteisiin ainakin kohtuullisesti. Tässä tutkielmassa erityisen tilanteen luo suomen kielen lisäksi potilaslausuntojen lääketieteellinen sisältö.

Implementoinnissa saatuja suurten kielimallien suorituksia tarkastellessa tärkeää on erityisesti niiden oikeellisuus ja toistettavuus. Tärkeää oli myös se, miten mallit reagoivat erilaisiin niille annettuihin kehoitteisiin. Suorituksissa mielenkiintoista oli myös se, miten mallit toimivat tilanteissa, joissa lausunnosta löydettävä vastaus oli epävarma tai sitä ei ollut ollenkaan löydettävissä annetuista syötteistä. Erityisen tärkeää on, että mallit eivät antaisi vääriä tai harhaanjohtavia vastauksia eli hallusinaatioita. Rakenteellistamisessa saatuja vastauksia voidaan parhaassa tapauksessa käyttää suoraan jatkotehtävissä, jolloin on tärkeää, että niiden antamaan tietoon voidaan luottaa.

4.2 Mahdolliset lähestymistavat

Tutkielmassa haluttiin vertailla usean mallin toimintaa saman tehtävän ratkaisemisessa. Tehtävään haluttiin löytää sellaiset suuret kielimallit, jotka toimisivat vähintään kohtuullisesti suomen kielellä. Oikean elämän käyttöä ajatellen tärkeää olisi myös se, että malleja voitaisiin soveltaa yksityiseen dataan jakamatta sitä ulkopuolisille osapuolille. Malleja ei ollut tarkoitus hienosäätää edelleen juuri tämän tehtävän ratkaisemiseen, vaan haluttiin tarkastella nimenomaan mallien toimintaa ilman tehtäväkohtaista lisäopetusta. Mallien valinnan haasteina olivat mallien saatavuus, lääketieteellisten lausuntojen suomenkielinen sisältö ja lääketieteellinen sanasto.

Suuria kielimalleja käyttämällä on saatu hyviä tuloksia luonnollisen kielen tehtävien ratkaisemisessa ensin esiopettamalla mallia suurella korpuksella ja tämän jälkeen hienosäätämällä mallin toimintaa tietyn tehtävän ratkaisemiseksi. [7] Hienosäätämiseen vaadittaisiin kuitenkin tuhansia tai kymmeniätuhansia esimerkkejä sisältäviä tehtäväkohtaisia hienosäätödatasettejä, tämä vaatii siis paljon aineistoa ja resursseja. [7] Tässä tutkielmassa käsiteltävässä tehtävässä erillinen esiopettaminen ja hienosäätäminen eivät resurssien rajallisuuden vuoksi olleet mahdollisia.

Mallien valinnassa keskityttiin siihen, millaisten tehtävien ratkaisemiseen mallit

parhaiten sopivat. Tämän jälkeen malleja testattiin alustavasti antamalla niille ratkaistavaksi yksittäisiä rakenteellistamistehtäviä. Tällöin keskityttiin erityisesti mallien suomen kielen ymmärrykseen ja kielen tuottamiseen. Tämän jälkeen pohdittiin vielä mallien saatavuutta. Lopulta päädyttiin valitsemaan neljä suurta kielimallia: GPT-3.5, GPT-4, Gemini-1.0-pro-001 ja Llama3-70b.

Tehtävässä oli huomioitava hyperparametrien merkitys mallien suorituksiin. Suurelle kielimallille asetettua lämpötilahyperparametriä käytetään säätämään mallien antamien vastausten satunnaisuutta ja luovuutta. Lämpötila säätelee ennustettavien sanojen todennäköisyyttä mallin softmax-ulostulokerroksessa. Yleensä lämpötila voidaan säätää välille $[0, 1]$. Jos hyperparametri on säädetty lähemmäs arvoa 0, lämpötila terävöittää todennäköisyyksiä niin, että korkeimman todennäköisyyden omaava sana tulee todennäköisemmin valituksi. Näin suuri kielimalli antaa konservatiivisempia ja ennalta-arvattavampia vastauksia. Jos hyperparametri taas on säädetty lähemmäs arvoa 1, malli valitsee vapaammin käytettävien sanojen joukosta. Näin suuri kielimalli käyttää enemmän luovuutta antamissaan vastauksissa. [66]

Lääketieteellisten lausuntojen rakenteellistaminen vaatii tehtävänä faktapohjaisia vastauksia, sillä on tärkeää, että mallit antavat vastauksina vain tosiasiallisia tietoja ja tarkasti rakenteellistettua dataa. Vaikka esimerkiksi artikkeleissa [46] ja [66] on todettu, että eri lämpötilojen käytöllä ei ole ollut huomattavaa merkitystä mallien antamissa tuloksissa, haluttiin implementoinnissa kuitenkin saada mahdollisimman deterministisiä vastauksia tehtävän luonteen vuoksi. Lopulta päädyttiin tämän takia käyttämään valittuja suuria kielimalleja API-avainten kautta. API:n (eng. Application Programming Interface) eli ohjelmointirajapinnan avulla voidaan käyttää suuria kielimalleja. API-avain taas on yksilöllinen tunniste, jonka avulla käyttäjä tai sovellus voidaan tunnistaa ja todentaa. API-avaimia käyttämällä pystyttiin erillisillä käskyillä säätämään haluttuja parametreja tehtävälle sopivalla tavalla. Tällöin myös lämpötilahyperparametri asetettiin arvoon 0 ja näin pyrittiin saamaan mahdollisimman deterministisiä vastauksia.

4.3 Ratkaisuyritys

Neljä valittua suurta kielimallia ovat kolmen eri yrityksen tarjoamia. GPT-3.5, GPT-4 ja Gemini1.0-pro-001 ovat yksityisomistukseen perustuvia suljettuja malleja, kun taas Llama3-70b on avoin suuri kielimalli. Tässä tutkielmassa haluttiin tasavertaisuuden vuoksi vertailla valittujen kielimallien uusimpia saatavilla olevia versioita tutkielman kirjoittamisen aikaan. Lopulliset ajot tehtiin heinäkuussa 2024, jolloin tutkielmaan valitut kielimallien versiot olivat viimeisimpiä. Tämän takia implemen-

tointiosuuteen valittiinkin GPT-4, Gemini-1.0-pro-001 ja Llama3-70b. Tämän lisäksi haluttiin myös tarkastella GPT-3.5-mallin suoritusta erityisesti verrattuna seuraajaansa, sillä GPT-3.5-mallia on käytetty laajasti oikean elämän, myös lääketieteen, sovelluksissa.

Aluksi määriteltiin sopiva kehote, jonka muotoiluun käytettiin inspiraationa lähdettä Fink et al. [13], jossa alkuperäinen kehote oli samankaltaisessa sovelluskohhteessa keuhkosyövästä kirjoitettujen vapaatekstimuodossa olevien potilaslausuntojen rakenteellistamistehtävässä. Tällä kehotteella ajettiin erilaisia syötteitä, joissa yhdisteltiin pieniä muutoksia kehotteessa ja erilaisia lausuntoja. Näin pyrittiin ymmärtämään kokeilun kautta keskeisimpiä ongelmia ja haasteita kehotteen muotoilussa eri malleille.

Aluksi käytettiin kuutta histopatologista potilaslausuntoa. Kokeilemalla erilaisia kehotteita kaikille valituille malleille näiden kuuden lausunnon kanssa pyrittiin luomaan sellainen kehote, joka toimisi riittävällä tavalla kaikkia malleja käytettäessä. Haasteena oli luoda kehote, joka olisi tasa-arvoinen kaikille malleille ja joka ottaisi huomioon kaikkien mallien heikkoudet ja vahvuudet. Yhden ainoan kehotteen tekeminen osottautui erittäin haastavaksi. Tämän takia tehtiinkin neljä erilaista kehotetta, jotta mallien toimintaa voitaisiin vertailla tasapuolisemmin. Kaikki kehotteet muodostuvat kolmesta osasta: ennalta määräytyistä kysymyksistä, kielimalleille annetuista ohjeista ja synteettisestä histopatologisesta lausunnosta. Kehotteita kirjoitettaessa käytettiin aiemmin tutkielmassa mainittuja tekniikoita kuten alan asiantuntija - ja ajatusketjutekniikoita. Tutkielmassa käytetyt kehotteet on esitetty tutkielman liitteessä A.

Ensimmäinen kehote perustuu lähteeseen Fink et al. [13]. Siinä kysymykset (4.1) jaettiin kahteen osaan odotetun vastauksen perusteella. Kysymyksiin 1.-3. odotettiin monipuolisempia, usein numeerisia, vastauksia ja kysymyksiin 4.-6. kyllä/ei -vastauksia. Kehotteessa pyrittiin antamaan mallille laajemmat yleisohjeet ja pitämään itse kysymykset mahdollisimman lyhyinä. Toinen kehote perustuu lähteen [66] kehotteeseen 4. Siinä yleisohjeet pidettiin mahdollisimman lyhyinä ja malleille annettiin laajemmat ohjeet kysymyskohtaisesti. Kolmas kehote pidettiin lyhyenä sekä yleisten ohjeiden että kysymysten osalta. Sen tarkoituksena oli tarkastella, millaisia vastauksia mallit antavat minimaalisilla ohjeilla.

Alustava kokeilu tehtiin kaikilla malleilla. Tällöin Llamasta käytettiin kokeiluaihana uusinta Llama2-versiota. Tämän käytössä huomattiin haasteita nimenomaan suomen kielen ymmärtämisessä. Tämän takia kirjoitettiin vielä neljäs kehote, joka oli englanniksi. Tämä kehote pidettiin sisällöltään samankaltaisena kolmannen kehotteen kanssa ja tarkoituksena oli tarkkailla nimenomaan kielen vaikutuksia mal-

lien suoriutumisessa. Llama3-mallien ilmestyessä vaihdettiin vanhempi Llama2-malli uudempaan. Tästä huolimatta neljäs kehoite pidettiin mukana, jotta voitiin edelleen seurata mahdollisia eroja kehotteen ollessa englanniksi tai suomeksi.

Kaikista neljästä kehotteesta tehtiin vielä perusteltuja variaatioita, jotta voitaisiin seurata myös pienien kehoitteissa olevien erojen merkitystä saaduissa vastauksissa. Variaatiot tehtiin alustavien kokeilujen perusteella ja niiden ideana oli muuttaa jotakin pientä, mutta merkityksellistä osaa kehotteesta. Malleille annettiin muun kehotteen lisäksi käsky toimia erittäin kokeneena histopatologisena lääkärimäisenä alan asiantuntijana -tekniikan mukaisesti. Sama variaatio tehtiin kaikista neljästä kehotteesta. Näin tehtävässä käytettyjä kehoitteita oli yhteensä kahdeksan kappaletta.

Lopulta histopatologisten potilaslausuntojen rakenteellistamistehtävässä käytettyjä synteettisiä histopatologisia lausuntoja oli yhteensä 40, joista 30 sisälsivät oikeaa lääketieteellistä informaatiota ja 10 taas epärelevanttia tekstiä eli ns. puppua. Kehotteina käytettiin neljää eri tavalla rakennettua kehotetta, joiden tarkoituksena oli tarkastella annettujen ohjeiden, niiden muotoilun ja laajuuden, merkitystä mallien antamiin vastauksiin. Tämän lisäksi kaikista neljästä kehotteesta käytettiin alan asiantuntijana -tekniikalla tehtyä variaatiota. Näin kehoitteita oli yhteensä kahdeksan. Kaikki kehotteet käyttivät nollaoppimista. Kehotteet on esitetty tutkielman liitteessä A. Mallien antamia vastauksia saatiin yhteensä 3840 kappaletta. Malleja pyydettiin antamaan vastaukset JSON-formaatissa ja näin saadut lausunnot kerättiin yksitellen tiedostoihin ja niitä tarkasteltiin manuaalisesti keräämällä ne Excel-taulukoihin.

5 Tulokset ja tulkinta

Tässä aliluvussa käydään läpi tutkielman implementointiosuudessa saatuja tuloksia matemaattisten metriikoiden avulla. Luku on jaettu kolmeen eri tarkasteluosioon. Ensin tarkastellaan mallien vastausten toistettavuutta. Tämän jälkeen tarkastellaan mallien antamien vastausten oikeellisuutta sekä puppulausuntojen että histopatologisten potilaslausuntojen rakenteellistamistehtävissä. Lopuksi vertaillaan mallien suorituksia ja käydään läpi mallien vahvuuksia ja heikkouksia käytännön sovelluksessa. Tämän lisäksi pohditaan käytettyjen kehoitteiden heikkouksia ja joitakin kehotesuunnittelutekniikoita, joiden avulla mallien suorituksia voitaisiin mahdollisesti parantaa.

Jokaisessa kehoitteessa malleille annettiin erikseen ohjeet antaa vastaukset JSON-formaatissa. Ohjeista huolimatta mallit antoivat satunnaisesti vastauksia myös siten, että ne eivät täyttäneet JSON-formaatin vaatimuksia. Yleisimpiä mallien tekemiä muotoiluvirheitä vastausten antamisessa olivat ylimääräiset, JSON-vastauksen ulkopuoliset, kuvailevat sanat tai lauseet ja muuten JSON-formaatista eroava muotoilu, esimerkiksi merkkijonojen antaminen lainausmerkkien ulkopuolella. Jos samaa tehtävää haluttaisiin soveltaa oikean elämän aineistolle, olisi lausuntoja vähintään tuhansia ja tarkoituksena olisi automatisoida lääketieteellisten lausuntojen rakenteellistaminen ja saatujen tietojen lukeminen ja käyttäminen ilman ihmisen työpanosta jokaisen vastauksen kohdalla. Tämän takia malleja pyydettiin noudattamaan tiettyä formaattia, jotta vastaukset olisivat yhtenäisiä ja niitä pystyttäisiin helpommin käsittelemään. Jos vastaukset eivät noudata samaa formaattia, on niitä erittäin haastavaa käsitellä ilman manuaalista työpanosta.

Mallien odotettiin siis noudattavan JSON-formaattia annettujen kehoitteiden mukaisesti. Tässä tehtiin kuitenkin kaksi poikkeusta kohtuullisuuden rajoissa. Gemini-1.0-pro-001 -malli antoi kaikissa vastauksissa merkkijonoja ennen itse JSON-vastausta. Nämä olivat kuitenkin jokaisen tiedoston kohdalla täysin samat: ennen JSON-vastausta merkkijono “````json`” ja tiedoston jälkeen merkkijono “`````”. LLama3-70b -malli taas antoi toistuvasti JSON-formaatin lisäksi ylimääräistä tietoa, joko esitelläkseen tulevan JSON-formaatissa olevan vastauksen tai antaakseen lisätietoa annetuista vastauksista. Nämäkin ylimääräiset merkkijonot oli vastauksissa eroteltu merkein “`````” ja “`'''`”. Tulosten tarkastelussa tehtiinkin päätös poistaa näiden merkkijonojen ulkopuolella oleva materiaali. Näin annetut vastaukset käsiteltiin näistä toistuvista ylimääräisistä merkkijonoista huolimatta. Nämä poikkeukset hyväksyttiin tulosten tarkastelussa, jotta kaikilta malleilta saatiin tarpeeksi vastauksia tarkasteluun.

Kehotteissa malleja pyydettiin rakenteellistamaan annettuja lääketieteellisiä lausuntoja kysymällä niiltä kuusi eri kysymystä (4.1) ja pyytämällä malleja kokoamaan näi-

den kysymysten vastaukset yhteen JSON-vastaukseen. Esimerkit oikeasta, väärästä ja JSON-syntaksivirheen sisältävästä vastauksesta löytyvät tutkielman liitteestä B. Työn edetessä mielenkiinto siirtyi kuitenkin erityisesti kahteen kysymykseen. Tässä tehtävässä tärkeimpiä kysymyksiä olivat näytteistä löydetty Gleason-luokitukset (kysymys kolme) ja mahdollinen hyvälaatuinen eturauhasen liikakasvu (kysymys neljä). Näiden kahden kysymyksen keskeisyyden takia ne valittiin lopulliseen tarkasteluun. Yleensä jos potilaslausunnosta on löydettävissä Gleason-luokitus, hyvälaatuiseen eturauhasen liikakasvuun ei oteta sen enempää kantaa, sillä Gleason-luokitus määrittää syövän vakavuutta. Tuloksia tarkastellessa on otettava huomioon, että malleilta on kysytty myös muita kysymyksiä. Tämä saattaa vaikuttaa vastausten laatuun verrattuna tilanteeseen, jossa malleilta olisi kysytty alun perinkin vain mainitut kaksi kysymystä. Yleisesti rakenteellistamistehtävässä mahdollisimman hyvien vastausten saamiseksi malleille esitettävien kysymysten määrää tulisi rajata niiden merkityksellisyyden mukaan.

5.1 Toistettavuus

Jokainen kehote jokaisella histopatologisella potilaslausunnolla ajettiin täysin samalla tavalla kolme kertaa, eli jokaisella ajolla oli kolme replikaattia. Lämpötilaparametri asetettiin arvoon 0, jolloin mallien tulisi antaa kaikkein deterministisempiä vastauksia eli periaatteessa mallien olisi tullut antaa jokaisella saman ajon replikaattilla täysin sama vastaus. Replikaatteja vertailemalla pyritään saamaan kokonaiskuva tulosten toistettavuudesta ja mallien luotettavuudesta. Tässä tarkastelussa otettiin huomioon vain histopatologiset potilaslausunnot, joita oli 30 kappaletta.

Eroja saman syötteen replikaattien välillä voi olla kahdenlaisia: merkintätavassa ja sisällössä. Edellä mainitut ovat vähemmän haitallisia mallin luotettavuutta arvioi-
dessa, sillä vastausten ollessa sisällöltään samat, mutta merkintätavaltaan erilaiset, voidaan silti katsoa vastausten olleen samat useissa tapauksissa. Poikkeuksena tästä tapaukset, joissa merkintätapa eroaa siten, että vastauksessa on ylimääräisiä merkkejä tai merkkijonoja, kuten sulkeita tai lainausmerkkejä. Vastausten oikeellisuutta arvioitiin unohtaen ero pienten ja isojen kirjainten välillä, koska käytännön rakenteellistamisessa esimerkiksi vastausten “Ei” ja “ei” yhdistäminen on triviaalia.

Kuvissa (3a), (3b) ja (3c) on eroteltu toistettavuustarkastelun samat replikaatit, eroavat replikaatit ja JSON-syntaksivirheitä sisältävät vastaukset. Samat replikaatit tarkoittavat tässä yhteydessä sitä, että täysin samalla syötteellä ajatut kolme replikaattia ovat samat. Eroavat replikaatit taas sitä, että yksi kolmesta eroaa muis-

	GPT-3.5	GPT-4	Gemini1.0-pro-001	Llama3-70b
samoja (lkm)	147	156	228	158
samoja (%)	61.3	65.0	95.0	65.8

(a) Samojen replikaattien lukumäärä ja prosenttiosuus kaikista replikaateista malleittain. Tässä tarkastelussa mitä suurempi arvo sitä parempi.

	GPT-3.5	GPT-4	Gemini1.0-pro-001	Llama3-70b
eroavia (lkm)	93	82	3	2
eroavia (%)	38.8	34.2	1.3	0.8

(b) Eroavien replikaattien lukumäärä ja prosenttiosuus kaikista replikaateista malleittain. Tässä tarkastelussa mitä pienempi arvo sitä parempi.

	GPT-3.5	GPT-4	Gemini1.0-pro-001	Llama3-70b
ei arvioitavissa (lkm)	0	2	9	80
ei arvioitavissa(%)	0	0.8	3.8	33.3

(c) JSON-syntaksivirheitä sisältävien vastausten lukumäärä ja prosenttiosuus kaikista replikaateista malleittain. Tässä tarkastelussa mitä pienempi arvo sitä parempi.

Kuva 3: Toistettavuustarkastelu. Kaikki vastaukset oli ensin muutettu pienaakkoisiin. Histopatologisia potilaslausuntoja oli 30 kappaletta ja kehoitteita kahdeksan kappaletta. Kolmen replikaatin ajoja oli näin yhteensä 240 kappaletta per malli.

ta kahdesta tai kaikki kolme samalla syötteellä ajettua replikaattia ovat keskenään eroavia. JSON-syntaksivirheitä sisältävät vastaukset tarkoittavat käytännössä sellaisia vastauksia, jotka eivät noudata JSON-formaattia ja näin niiden sisältöä ei voida arvioida sen pidemmälle. Tässä toistettavuustarkastelussa ei oteta kantaa siihen ovatko vastaukset oikein, vaan tarkastellaan ainoastaan mallien vastausten toistettavuutta.

Kuvasta (3a) nähdään, että toistettavuus on Gemini1.0-pro-001 mallilla erittäin hyvä. Muiden mallien toistettavuus on keskenään samalla tasolla. Kuvasta (3b) taas nähdään, että GPT-3.5 - ja GPT-4 -mallilla on huomattavasti enemmän eroavia replikaatteja Gemini1.0-pro-001 - ja Llama3-70b -malleihin verrattuna. Samoin Gemini1.0-pro-001 ja Llama3-70b -mallien suoritus eroavien replikaattien määrässä on hyvin samankaltainen keskenään. Kuvasta (3c) nähdään, että Llama3-70b -mallilla JSON-syntaksivirheitä sisältäviä vastauksia oli huomattavasti enemmän muihin malleihin verattuna. Tarkastelussa on otettava huomioon, että sekä Gemini1.0-pro-001 - että Llama3-70b -malleilla oli jo valmiiksi otettu huomioon toistuvat virheet vastausten muotoilussa.

Kuvassa (4) on esitetty eroavien replikaattien määrä malli- ja kehotekohtaisesti. Kuvassa on myös esitetty eroavien replikaattien lukumäärien summa jokaiselle mallille ja kehotteelle erikseen. Käydään seuraavaksi läpi kehoitteiden vaikutusta vas-

	GPT-3.5	GPT-4	Gemini1.0-pro-001	Llama3-70b	yht.
P0	12	9	0	1	22
P1	11	13	1		25
P2	9	19	0	0	28
P3	8	14	0	0	22
P4	13	10	0		23
P5	9	4	0	0	13
P6	13	6	0	1	20
P7	18	7	0	0	25
yht.	93	82	1	2	

Kuva 4: Toistettavuustarkastelussa eroavien replikaattien lukumäärä malli- ja kehotekohtaisesti. Vastauksia oli yhteensä jokaiselle mallille 240 kappaletta ja jokaiselle kehotteelle 120 kappaletta. Harmaat ruudut merkitsevät, että kyseisellä malli-kehoteparilla kaikki vastaukset sisälsivät JSON-syntaksivirheitä.

tauksiin toistettavuustarkastelussa. Tämän tutkielman kokeiluosuudessa eräs mielenkiinnon kohde on eri kehotteiden vaikutus mallien suorituksiin ja se, millaiset kehotteet toimivat parhaiten eri malleilla.

Kuvasta (4) nähdään, että kehotteiden välillä ei ollut suurta eroa mallien suorituksissa. Tämä saattaa johtua siitä, että kaikki kehotteet olivat usean iteraation tulosta, ja ne olivat jo alun perin niin hyviä, etteivät ne vaikuttaneet toistettavuustarkastelussa mallien vastauksiin merkittävästi. Kehotteet P4 - P7 ovat alan asiantuntija - tekniikalla tehtyjä variaatioita kehotteista P0 - P3 järjestyksensä mukaisesti (P4 on variaatio kehotteesta P0 ja niin edelleen). Kehotteiden variaatioilla ei ole huomattavaa merkitystä mallien suorituksiin toistettavuustarkastelussa. Kuvasta nähdään, että kehote P5 oli universaalisti hieman parempi muihin kehotteisiin verrattuna. Tämä kehote oli alan asiantuntija -tekniikalla tehty kehote, jossa malleja kannustettiin ajattelemaan kysymyksiä askel askeleelta. Kehotteen P5 tulokset olivat myös hieman parempia verrattuna kehotepariinsa P0, joka oli muuten samanlainen, mutta siinä malleja ei ohjeistettu toimimaan histopatologisenä erikoislääkärinä.

Mahdollisessa jatkotarkastelussa olisi mielenkiintoista selvittää, millaisia GPT-3.5 - ja GPT-4 -mallien toistettavuustarkastelun keskenään eroavat replikaatit olivat ja kuinka usein kaikki kolme replikaattia olivat kaikki keskenään erilaisia. Jos yleisimmässä tapauksessa kaksi kolmesta replikaatista on ollut samoja ja vain yksi kolmesta on ollut eroava, voitaisiin tämä ottaa huomioon vastausten tarkastelussa ja käytössä. Tämä vaatisi kuitenkin tarkempaa ja yksityiskohtaisempaa tarkastelua. Tämän tutkielman tavoitteena on antaa yleiskuva mallien suorituksista rakenteellistamistehtävässä ja erilaisten kehotteiden vaikutuksesta mallien toimintaan. Tämän tutkielman tuloksista huomataan, että edes valituista malleista parhaiten toistetta-

vuustarkastelussa suoriutunut Gemini1.0-pro-001 -malli ei ollut 100 % toistettava lämpötilaparametrilla 0. Tällöin voidaan yleisesti ajatella, että stabiileimmatkaan suuret kielimallit eivät ole vielä ideaalin täydellisen toistettavuuden tasolla.

5.2 Oikeellisuus

Seuraavaksi tarkastellaan mallien vastausten oikeellisuutta. Tarkastelu on jaoteltu kahteen osaan: puppulausuntosyötteiden rakenteellistamistehtävän ja histopatologisten potilaslausuntojen rakenteellistamistehtävän tarkasteluihin. Ensin tarkastellaan puppulausuntosyötteiden rakenteellistamistehtävän vastauksia. Puppulausuntojen oikeellisuutta tarkastelemalla pyritään saamaan käsitys mallien suorituksista tilanteissa, joissa syötteet eivät sisällä kehoitteiden kuvailemia tietoja. Tällaisissa tilanteissa mallit saattavat esimerkiksi alkaa hallusinoimaan, mikä on erittäin haitallista erityisesti jatkotehtäviä ajatellen. Tällaisissa tilanteissa toivottaisiinkin, että mallit antaisivat ennemmin liian varovaisia vastauksia yli-itsevarmojen vastausten sijaan.

Puppulausuntosyötteiden rakenteellistamistehtävän vastauksia tarkastellaan jakamalla saadut vastaukset kolmeen osaan: oikeisiin, väriin ja JSON-syntaksivirheitä sisältäviin vastauksiin. Oikeaksi vastaukseksi hyväksyttiin vain vastaus “na” (eng. not available), mikä on yleinen tapa merkitä puuttuvaa arvoa. JSON-syntaksivirheitä sisältäviä vastauksia ei oltu annettu JSON-formaatissa. Vääriksi vastauksiksi luokiteltiin kaikki jäljelle jäävät vastaukset eli JSON-formaatissa annetut, mutta sisällöltään väärät vastaukset. Puppulausuntoja tarkastellessa replikaatteja käsiteltiin erillisinä vastauksina. Kehotteita oli kahdeksan, erilaisia puppulausuntoja kymmenen ja jokaisella lausunto-kehoteparilla ajettuja replikaatteja oli kolme. Tällöin tarkasteltavia vastauksia oli yhteensä jokaiselle mallille 240 kappaletta.

Kuvasta (5a) nähdään, että oikeiden vastausten määrät vaihtelevat suuresti mallien välillä. Puppulausuntosyötteiden rakenteellistamistehtävässä ainoa oikea vastaus oli kieltäytyä antamasta vastausta eli palauttaa ohjeiden mukaisesti arvo “na”. GPT-4 -malli antoi eniten oikeita vastauksia, GPT-3.5 -malli taas antoi vähiten oikeita vastauksia. Kuvista (5b) ja (5c) nähdään, että oikeita vastauksia lukuun ottamatta vastaukset jakautuivat malleittain niin, että GPT-3.5 - ja GPT-4 -mallit ovat antaneet enemmän väärää vastauksia ja Gemini1.0-pro-001 ja Llama3-70b -mallit taas ovat antaneet enemmän JSON-syntaksivirheitä sisältäviä vastauksia.

Tarkastelussa on otettava huomioon JSON-syntaksivirheitä sisältävien ja väärin vastausten laatu. Pintapuolisen tarkastelun perusteella mallit antoivat esimerkiksi joissain tilanteissa vastaukseksi toteamuksen eturauhassyöpätietojen puuttumisesta

	GPT-3.5	GPT-4	Gemini1.0-pro-001	Llama3-70b
oikein (lkm)	51	201	144	95
oikein (%)	21.3	83.8	60.0	39.6

(a) Oikeat vastaukset lukumäärinä ja prosenttiosuuksina malleittain. Ainoa oikeaksi hyväksytty vastaus oli “na”. Tässä tarkastelussa mitä suurempi arvo sitä parempi.

	GPT-3.5	GPT-4	Gemini1.0-pro-001	Llama3-70b
väärin (lkm)	189	38	12	15
väärin (%)	78.8	15.8	5.0	6.3

(b) Väärät vastaukset lukumäärinä ja prosenttiosuuksina malleittain. Tässä tarkastelussa mitä pienempi arvo sitä parempi.

	GPT-3.5	GPT-4	Gemini1.0-pro-001	Llama3-70b
ei arvioitavissa (lkm)	0	1	84	130
ei arvioitavissa (%)	0.0	0.4	35.0	54.2

(c) JSON-syntaksivirheitä sisältävät vastaukset lukumäärinä ja prosenttiosuuksina malleittain. Tässä tarkastelussa mitä pienempi arvo sitä parempi.

Kuva 5: Puppulausuntosyötteiden rakenteellistamistehtävän vastaukset. Puppulausuntoja oli 10 kappaletta, kehoitteita kahdeksan ja jokaisella ajolla oli kolme replikaattia. Vastauksia oli näin yhteensä 240 kappaletta per malli.

ja näin vastauksen antamisen mahdottomuudesta. Tällainen vastaus on luokiteltu JSON-syntaksivirheeksi, koska sitä ei ole esitetty JSON-formaatissa. Esimerkiksi tällaisissa tilanteissa mallien antamat vastaukset eivät välttämättä kuvaa parhaiten mallien toimintaa sellaisissa tilanteissa, joissa käytetään aitoja lääketieteellisiä lausuntoja. Puppulausuntosyötteiden rakenteellistamistehtävän vastausten tarkastelun oli tässä tutkielmassa tarkoitus avata mallien toimintaa tilanteissa, joissa annetut syötteet eivät vastaa kehoitteiden perusteella odotettavissa olevia syötteitä.

Mahdollisessa jatkotarkastelussa voisi olla mielenkiintoista selvittää tarkemmin, millaisia vastauksia mallit ovat antaneet. Kaikista haastavimpia ovat tilanteet, joissa mallit ovat antaneet harhaanjohtavan vastauksen. Tällaisia olisivat esimerkiksi puppulausuntosyötteillä annetut Gleason-luokitukset, joita ei tosiasiallisesti löydy syötteistä ollenkaan. Tällaiset vastaukset ovat haitallisia rakenteellistettujen lausuntojen jatkokäytössä.

Seuraavaksi käsitellään histopatologisten potilaslausuntojen rakenteellistamistehtävän vastausten oikeellisuutta. Kuten muissakin tarkasteluissa, vastausten oikeellisuutta arvioidessa tarkasteltiin kahta alkuperäisistä kysymyksistä: Gleason-luokitusta (kysymys kolme) ja BPH:ta (kysymys neljä). Tässä tarkastelussa kysymyksiä on käsitelty erikseen. Kysymykset on taulukoissa merkitty lyhenteillä Q3 (Gleason-luokitus) ja Q4 (BPH).

Gleason-luokituksen vastauksia tarkastellessa hyväksyttiin neljä merkintätapaa:

“major + minor = sum”, “major + minor”, “sum” ja “sum (major + minor)”. Vastausten muotoilu on keskeistä oikeita vastauksia tarkastellessa, sillä kuten aiemmin on todettu, jatkotehtävissä on tarkoitus käyttää rakenteellistettuja lausuntoja, joita voidaan käsitellä automaattisesti. Tällöin ei ole mahdollista hyväksyä kaikkia mahdollisia muotoiluja, vaan pyritään siihen, että suuret kielimallit antaisivat vastaukseksi vakiintuneita muotoiluja.

Eturauhasen hyvälaatuista liikakasvua (BPH) tarkastellessa hyväksyttiin vastaukset “kyllä”, “ei” ja “na” ja niiden englanninkieliset vastineet riippuen annetusta lausunnosta. Tässäkin tärkeää on vastausten yhteneväisyys ja se, että vastauksia voidaan käsitellä automaattisesti jatkotehtävissä. Tällöin ei hyväksytty esimerkiksi vastauksia “ehkä” ja “ei mainittu”, vaikka periaatteessa ne olisivat olleet sisällöltään oikean suuntaisia, mutta eivät ohjeiden mukaisesti rakenteellistettuja.

Koska kaikissa kehoitteissa suuria kielimalleja on ohjeistettu antamaan vastaukset tietyssä muodossa, on myös kohtuullista arvioida oikeita vastauksia näiden annettujen ohjeiden mukaisesti. Vastaukset käytiin läpi manuaalisesti. Manuaalisessa vastausten tarkastelussa pystyttiin paremmin arvioimaan yksittäisten vastausten oikeellisuutta, kuin mikä koodin avulla olisi ollut mahdollista.

Oikeita vastauksia tarkastellessa käsiteltiin kunkin mallin kohdalla 720 vastausta. Histopatologisia potilaslausuntoja oli 30 kappaletta, nämä kaikki ajettiin kahdeksalla kehoitteella ja täysin sama ajo toistettiin kolme kertaa ja näin saatiin kaikista vastauksista kolme replikaattia. Oikeita vastauksia tarkastellessa replikaatit käsiteltiin erillisinä vastauksina. Mallien antamat vastaukset on luokiteltu kolmeen ryhmään: oikeisiin, väärisiin ja JSON-syntaksivirheitä sisältäviin vastauksiin.

Kysymysten kolme ja neljä tuloksia (kuva 6) ei voida täysin suoraviivaisesti vertailla keskenään sillä kysymysten vastausvarauudet ovat erilaiset. Kysymyksessä kolme (Q3, Gleason) hyväksytyt vastausvaihtoehdot oli neljä kappaletta ja mahdolliset Gleason-luokitukset sisälsivät sekä major- että minor-luokitukset, jotka voivat saada kokonaislukuarvoja väliltä [3, 5]. Kysymyksessä neljä (Q4, BPH) hyväksytyt vastausvaihtoehdot olivat “kyllä/yes”, “ei/no/na” ja “na” eli yhdestä kolmeen tilanteesta riippuen. Syy miksi “na” hyväksyttiin “ei/no” tilanteessa oli se, että kuten aiemmin mainittiin, jos näytteestä löydetään Gleason-luokitus, ei BPH ole enää keskeinen tieto ja näin sitä ei luultavimmin mainita myöskään potilaslausunnossa.

Kuvasta (6a) nähdään, että GPT-4 -mallin oikeellisuus on paras sekä kysymyksessä kaksi että kysymyksessä kolme. GPT-4 -mallilla oikeiden vastausten määrä on korkein ja väärin vastausten määrä on matalin. Myös GPT-3.5 - ja Gemini1.0-pro-001 -mallien oikeellisuus on lähellä GPT-4 -mallia. Jokaisessa kohdassa Gemini1.0-

	GPT-3.5	GPT-4	Gemini1.0-pro-001	Llama3-70b
Q3 (lkm)	536	588	543	327
Q3 (%)	74.4	81.7	75.4	45.4
Q4 (lkm)	607	654	618	351
Q4 (%)	84.3	90.8	85.8	48.8

(a) Oikeat vastaukset lukumäärinä ja prosenttiosuuksina malleittain. Tässä tarkastelussa mitä suurempi arvo sitä parempi.

	GPT-3.5	GPT-4	Gemini1.0-pro-001	Llama3-70b
Q3 (lkm)	184	129	144	153
Q3 (%)	25.6	17.9	20.0	21.3
Q4 (lkm)	113	63	69	129
Q4 (%)	15.7	8.8	9.6	17.9

(b) Väärät vastaukset lukumäärinä ja prosenttiosuuksina malleittain. Tässä tarkastelussa mitä pienempi arvo sitä parempi.

	GPT-3.5	GPT-4	Gemini1.0-pro-001	Llama3-70b
lkm	0	3	33	240
%	0.0	0.4	4.6	33.3

(c) JSON-syntaksivirheitä sisältävät vastaukset lukumäärinä ja prosenttiosuuksina malleittain. Tässä tarkastelussa mitä pienempi arvo sitä parempi.

Kuva 6: Histopatologisten potilaslausuntojen rakenteellistamistehtävän oikeellisuustarkastelu. Vastauksia oli yhteensä 720 kappaletta per malli.

pro-001 on suoriutunut jonkin verran paremmin verrattuna GPT-3.5 -malliin. Kaikkien mallien oikeellisuus oli parempi kysymyksessä neljä (Q4) kysymykseen kolme (Q3) verrattuna. Llama3-70b-mallin vastaukset eroavat kolmen muun mallin vastauksista huomattavasti. Tämä johtuu siitä, että Llama3-70b -mallin vastauksista kolmasosassa oli JSON-syntaksivirheitä, mikä vähentää mahdollisten oikeiden vastausten määrää. Vääriä vastauksia tarkastellessa Llama3-70b -mallin suoritus ei eroa yhtä paljon muista malleista.

Kysymyksessä kolme (Q3, Gleason-luokitus) yleisimpiä virheitä kaikkien mallien vastauksissa olivat ylimääräiset merkit, väärät muotoilut ja vain osavastauksen antaminen kokonaisuena vastauksena. Ylimääräiset merkit olivat esimerkiksi laajempia selityksiä vastauksesta tai pienempiä eroavaisuuksia, kuten sulkeita vastauksen ympärillä. Molemmat tilanteet ovat haastavia, kun vastauksia pyritään käsittelemään automaattisesti. Jos vastaukset sisältävät ennalta-arvaamattomia ylimääräisiä merkkejä, ei niitä voida käsitellä automaattisesti. Tällöin, vaikka ihmisen käsittelemänä vastaus olisikin sisällöltään oikeellinen, ei sitä voida merkintätavasta johtuen merkitä oikeaksi. Tämän lisäksi mallit antoivat joissakin tilanteissa vastauksena myös vain major- tai minor-luvun esittäen sen koko summana. Tämän takia

sum-merkinnän hyväksyminen on haastavaa, kun vastauksia käsitellään automaattisesti. Tämän tutkielman manuaalisessa tarkastelussa oli kuitenkin mahdollista erottaa oikeelliset summat ja osasummat toisistaan. Oikeellisuusprosenttia voisi yrittää parantaa antamalla malleille kehoitteissa tarkemmat ohjeet toivottujen vastausten merkintätavasta.

Kysymystä neljä (Q4, BPH) tarkastellessa yleisimpiä virheitä olivat täysin väärät vastaukset. Merkintätapojen vaihtelevuus ei ollut niin suurta kuin Gleasonluokitukseen liittyvissä vastauksissa. Tämä oli odotettavissa aiemmin mainitun kysymysten erilaisten vastausavaruuksien ja yleisesti kysymysten erilaisen luonteen takia. Kuitenkin väärissä vastauksissa oli myös jonkin verran vääriä merkintätapoja, kuten ylimääräisiä sulkeita tai lainausmerkkejä. Täysin väärät vastaukset ovat erittäin haitallisia, sillä rakenteellistettujen lausuntojen jatkokäsittelyssä vastauksiin tulisi voida luottaa. Tämän takia täysin väärät vastaukset olisi saatava eliminoitua ennen mahdollista jatkokäyttöä. Muuten vastaukset ovat harhaanjohtavia ja ne aiheuttavat haittaa myöhemmin.

Seuraavaksi käydään läpi kehoitteiden välisiä eroja mallien vastauksissa oikeellisuustarkastelussa. Kuvassa (7a) on esitetty väärrien vastausten määrä malli- ja kehotekohtaisesti kysymyksessä kolme. Kehotteiden yleistä vaikutusta tarkastellessa on otettava huomioon, että Llama3-70b -mallilla kaikki kehoitteiden P1 ja P4 vastaukset olivat JSON-syntaksivirheitä sisältäviä. Tästä huolimatta voidaan kuitenkin nähdä yleisiä trendejä kehoitteiden vaikutukselle mallien suorituksissa.

Kuvasta (7a) nähdään, että kysymyksessä kolme kehoitteilla on jonkin verran vaikutusta mallien suorituksiin. Lähes kaikissa tapauksissa kehoitteella saadut tulokset ovat samankaltaisia mallista riippumatta. Tämän lisäksi kehotteiden välillä on nähtävissä yhteneväisyyksiä tuloksissa. Kehotteet P1 ja P5 suoriutuivat parhaiten tehtävästä. Nämä kaksi ovat saman kehotteen variaatioita. Kehotteet P0 ja P4 taas ovat suoriutuneet huomattavasti heikommin. Nämä kaksi ovat toisen kehotteen variaatioita. Kehoteparien P2 - P6 ja P3 - P7 tapauksissa taas nähdään huomattavasti heikompia tuloksia alan asiantuntija -tekniikalla tehtyjen kehoitteiden (P6 ja P7) käytössä alkuperäisiin kehoitteisiin verrattuna. Kuvasta (7b) nähdään, että kysymyksessä neljä ei ole nähtävissä niin selkeitä trendejä kehoitteiden vaikutuksesta mallien suorituksiin kysymykseen kolme (7a) verrattuna. Yleisesti kysymyksessä neljä väärät vastaukset olivat jakautuneet tasaisemmin ja epäjohdonmukaisemmin kehoitteiden välillä kysymykseen kaksi verrattuna. Tämäkin selittyy ainakin osaksi aiemmin mainitulla kysymysten erilaisella vastausavaruudella.

	GPT-3.5	GPT-4	Gemini1.0-pro-001	Llama3-70b	yht.
P0	35	27	30	27	119
P1	9	6	3		18
P2	6	14	6	42	68
P3	9	15	12	9	45
P4	28	24	27		79
P5	7	6	3	12	28
P6	38	21	39	30	128
P7	52	16	24	33	125
yht.	184	129	144	153	

(a) Väärien vastausten määrä kysymyksessä kolme kehote- ja mallikohtaisesti.

	GPT-3.5	GPT-4	Gemini1.0-pro-001	Llama3-70b	yht.
P0	13	1	6	9	29
P1	14	8	15		37
P2	12	8	9	42	71
P3	8	10	3	15	36
P4	26	8	9		43
P5	15	9	12	15	51
P6	12	8	9	39	68
P7	13	10	6	9	38
yht.	113	62	69	129	

(b) Väärien vastausten määrä kysymyksessä neljä kehote- ja mallikohtaisesti.

	GPT-3.5	GPT-4	Gemini1.0-pro-001	Llama3-70b	yht.
P0	0	0	6	12	18
P1	0	0	0	90	90
P2	0	3	0	30	33
P3	0	0	0	0	0
P4	0	0	21	90	111
P5	0	0	0	0	0
P6	0	0	0	12	12
P7	0	0	6	6	12
yht.	0	3	33	240	

(c) JSON-syntaksivirheitä sisältävien vastausten määrä kehote- ja mallikohtaisesti. Kysymykset kolme ja neljä on käsitelty yhdessä.

Kuva 7: Väärien vastausten ja JSON-syntaksivirheitä sisältävien vastausten määrä malli- ja kehotekohtaisesti. Käsiteltäviä vastauksia oli yhteensä jokaisella mallilla 720 kappaletta ja jokaisella kehotteella 360 kappaletta.

Kuvasta (7c) nähdään, miten JSON-syntaksivirheitä sisältävät vastaukset ovat jakautuneet kehote- ja mallikohtaisesti. Parhaiten kaikilla malleilla ovat suoriutuneet kehotteet P3 ja P5, joilla JSON-syntaksivirheitä sisältäviä vastauksia ei ole yhtäkään. Huonoiten taas ovat suoriutuneet kehotteet P4 ja P1. Tarkastelussa on kuitenkin otettava huomioon, että koska GPT-3.5 - ja GPT-4 -malleilla JSON-syntaksivirheitä sisältäviä vastauksia on hyvin vähän tai ei lainkaan, kuvaa tämä tarkastelu lähinnä Gemini1.0-pro-001 - ja Llama3-70b -mallien toimintaa. Tämän lisäksi, suurin osa JSON-syntaksivirheitä sisältävistä vastauksista on Llama3-70b

-mallin vastauksia. Kuitenkin, kehote P4 on suoriutunut huonoiten sekä mallilla Gemini1.0-pro-001 että mallilla Llama3-70b.

Kuten aiemmin mainittiin, Llama3-70b - mallilla JSON-syntaksivirheitä sisältävien vastausten määrä oli huomattavasti suurempi kuin kolmella muulla mallilla. Llama3-70b -mallilla kaikki kehotteen P4 vastaukset olivat JSON-syntaksivirheitä sisältäviä. Käytännössä kaikissa tämän tutkielman kokeiluosuudessa käsiteltävissä Llama3-70b-mallin antamissa vastauksissa oli ylimääräistä sisältöä JSON-vastauksen ulkopuolella. Kuten aiemmin mainittiin, tämä kuitenkin hyväksyttiin kohtuullisuuden rajoissa tarkastelussa. Kehotteen P4 avulla saadut Llama3-70b -vastaukset eivät noudattaneet JSON-formaattia, eivätkä tutkielmassa ennalta määriteltä, JSON-formaatista eroavaa, mutta hyväksyttyä, formaattia. Tässä formaatissa merkkien “ ` ` ` ” ja “ ` ` ` ” välissä olevaa tekstiä voitiin käsitellä, jos se oli JSON-formaatissa. Kehotteella P4 saadut vastaukset eivät kuitenkaan antaneet merkkejä vastauksen molemmilla puolilla, jolloin ylimääräiset merkkijonot, useimmiten selitykset vastauksista, rikkoivat JSON-formaattia. Jos tavoitteena olisi vastausten manuaalinen tarkastelu, tällainenkin tiedosto voitaisiin hyväksyä, mutta tutkielmassa saatuja tuloksia tarkastellessa pidettiin mielessä, että rakenteellistamistehtävän lopullinen tavoite on vastausten automaattinen käsittely, Tällöin edellä mainitunlaisia virheitä ei voida hyväksyä.

Samoin kaikki kehotteen P1 vastaukset olivat JSON-syntaksivirheitä sisältäviä Llama3-70b-mallin vastauksissa. Kehotteeseen P1 annetut vastaukset oli annettu erillisinä JSON-vastauksia muistuttavina vastauksina, joissa muotoilu kuitenkin erosi JSON-formaatista. Osaa vastauksista ei oltu annettu loppuun asti, vaan malli oli lopettanut osan vastauksista jo esimerkiksi ensimmäisen tai toisen kysymyksen jälkeen. Kehotteessa P1 pyydettiin mallia käymään kysymykset läpi askel askeleelta. Useassa vastauksessa Llama3-70b -malli pyysi antamaan lisäohjeita, kun vastausta voi jatkaa. Mielenkiintoista tässä tarkastelussa on se, että nämä kaksi kehotetta, P1 ja P4, eivät ole variaatioita samasta kehotteesta. Kehotteen P1 variaatio on kehote P5, joka on Llama3-70b -mallilla suoriutunut erittäin hyvin. Kehotteen P4 variaatio taas on P0, joka on myös suoriutunut hyvin Llama3-70b -mallilla. Mahdollisessa jatkotarkastelussa olisikin mielenkiintoista selvittää, miksi näin samanlaiset kehotteet ovat antaneet samalla mallilla näin erilaisia tuloksia.

Kehote P1 menestyi hyvin GPT-3.5 -, GPT-4 ja Gemini1.0-pro-001 -malleilla, kun taas Llama3-70b-malli tuotti vastauksina pelkkiä JSON-syntaksivirheitä sisältäviä vastauksia. Kehotteessa P1 malleja kehotetaan pohtimaan vastauksia askel askeleelta. Olisikin mielenkiintoista selvittää, toimisiko tämänlainen kehote myös Llama3-70b-mallille yhtä hyvin, jos mallin JSON-syntaksivirheet pystyttäisiin mi-

nimoimaan. Kehote P5 on kehotteen P1 variaatio. Myös kehotteessa P5 malleja pyydetään miettimään vastauksia askel askeleelta sillä erolla, että malleille annetaan ohjeet toimia histopatologisena erikoislääkärinä. Tämän tutkielman tulosten perusteella askel askeleelta -ajattelu eli ajatusketjutekniikka olisi siis yleisesti ainakin hieman toimivampi muihin tekniikoihin ja kehotetyyleihin verrattuna.

Muuten kehotteiden välillä ei ole suuria eroja oikeellisuustarkastelussa. Erot ovat niin pieniä ja vaihtelevia eri tarkastelujen välillä, että kaikkien kehotteiden voidaan ajatella olevan lähestulkoon yhtä hyviä. Tämän tutkielman perusteella voidaan kuitenkin todeta, että erilaiset kehotesuunnittelutekniikat, kuten tässä tutkielmassa käytetyt ajatusketju- ja alan asiantuntijatekniikat, vaikuttavat mallien suorituksiin. Kuitenkin esimerkiksi Llama3-70b-mallille jotkut kehotesuunnittelutekniikat tai -rakenteet ovat mahdollisesti fataaleita. Tälle ei kuitenkaan tämän työn puitteissa löytynyt vielä mitään selvää logiikkaa.

Jotta kehotteiden välillä oltaisiin saatu selkeämpiä ja suurempia vaikutuksia mallien antamiin vastauksiin, oltaisiin voitu käyttää hyvin erilaisia ja selvästi huonojakin kehotteita. Kaikki tässä tutkielmassa käytetyt kehotteet oli kokeiltu ainakin jollakin tapaa toimiviksi kaikille malleille ennen niiden valitsemista tutkielmaan. Tämän lisäksi on haastavaa vertailla kehotteiden vaikutusta vastauksiin neljän eri mallin toiminnassa. Tässä tutkielmassa tavoitteena oli saada yleiskäsitys kehotteiden vaikutuksesta mallien toimintaan. Jos jotakin tiettyä mallia haluttaisiin käyttää toisessa käytännön sovelluksessa, hyödyllisintä olisi vertailla kehotteiden vaikutusta tietyn mallin toiminnassa. Kuten tämänkin tutkielman tuloksista nähdään, eri kehotteet vaikuttavat mallien toimintaan eri tavalla.

5.3 Mallien vertailu

Tässä aliluvussa vertaillaan tutkielmassa käytettyjen neljän suuren kielimallin toimintaa annetussa rakenteellistamistehtävässä. Keskitytään erityisesti kunkin kielimallin vahvuuksiin, heikkouksiin ja keskeisiin asioihin, jotka tulee ottaa huomioon näitä suuria kielimalleja käytettäessä tiedon rakenteellistamisessa.

Tämän tutkielman implementointiosuudessa GPT-3.5 -malli suoritus oli vertailukelpoinen muiden mallien kanssa, vaikka se ei tutkielman kirjoittamisen aikaan ollut uusin versio malliperheestään. GPT-3.5 -mallin vahvuus oli erityisesti vastausten oikeellisessa muotoilussa. Yksikään mallin vastauksista ei sisältänyt JSON-syntaksivirheitä, vaan kaikki vastaukset oli esitetty JSON-formaatissa. GPT-3.5 -mallin heikkous oli väärin vastausten suuri määrä. Jotta tämä voitaisiin välttää, voitaisiin GPT-3.5 -mallille esimerkiksi kokeilla kehotteita, joissa mallilta vaaditaan

erityistä tarkkuutta syötteitä käsitellessä. Myös “na” vaihtoehdon antamista vastauksena epäselvissä tilanteissa voisi painottaa.

Koska suurten kielimallien kehittäminen on kovan kilpailun alaisena, kehitetään jatkuvasti sekä kokonaan uusia suuria kielimalleja että uusia versioita jo olemassaolevista malleista. Usein uudet versiot tuovat mukanaan muun muassa rakenteellistamistehtävissä hyödyllisiä ominaisuuksia. Tämän takia tulisikin mahdollisuuksien mukaan harkita uudemman version käyttämistä erilaisissa sovelluskohteissa. Uusin versio ei aina kuitenkaan suoraan tarkoita parasta versiota, sillä kuten aiemmin tässä tutkielmassa todettiin, suuria kielimalleja kehitetään erilaisiin käyttötarkoituksiin. Tällöin tulee ottaa huomioon mallin uutuuden lisäksi myös ratkaisu tehtävä. GPT-3.5 -mallia on käytetty laajasti eri sovelluskohteissa. Se onkin yksi ensimmäisistä suurista kielimalleista, jotka mahdollistivat laajan käytön sekä arkielämän sovelluksissa että erinäisissä tutkimuskohteissa. GPT-3.5 -mallin vahvuus on siitä saatavilla oleva tieto julkaisuissa esimerkiksi kehotesuunnittelussa.

GPT-4 -mallin vahvuuksiksi tämän tutkielman implementointiosuudessa osoitautuivat vastausten oikeellisuus ja oikeellinen muotoilu. Myös GPT-4 -mallilla lähes kaikki vastaukset oli esitetty pyydettyssä JSON-formaatissa. GPT-4 -malli suoriutui edeltäjänsä paremmin kaikissa tarkasteluissa JSON-syntaksivirheitä sisältäviä vastauksia lukuunottamatta. Toistettavuustarkastelussa ero oli melko pieni, mutta oikeellisuustarkasteluissa ero oli huomattava. GPT-4 -mallin heikkous oli toistettavuus. Keskenään eroavien replikaattien määrä oli suurempi verrattuna Gemini1.0-pro-001 ja Llama3-70b -malleihin. Haastavimmissa tapauksissa replikaatit antavat samalla ajolla täysin eri vastaukset, esimerkiksi “kyllä” ja “ei”. Näissä tapauksissa olisi tärkeää yhtenäistää replikaattien antamia vastauksia esimerkiksi sopivalla kehotesuunnittelulla, jolloin vastaukset olisivat deterministisempiä. Kehotesuunnittelulla voitaisiin esimerkiksi antaa entistä tarkempia ohjeita mallille vastausten odotetusta sisällöstä ja muodosta. Tämän tutkielman kokeiluosuuden perusteella ei kuitenkaan voida sanoa, millaisia replikaattien väliset erot olivat. Kuitenkin, edes toistettavuustarkastelussa parhaiten menestyneen Gemini1.0-pro-001 -mallin vastaukset eivät oleet absoluuttisen toistettavia. Tutkielman perusteella voidaan näin vetää johtopäätös, että stabiileimmiltakaan malleilta ei voida odottaa täydellistä tieteellistä toistettavuutta.

Llama3-70b -mallin heikkous oli, että monet sen vastauksista sisälsivät JSON-syntaksivirheitä. JSON-syntaksivirheitä sisältävien vastausten suuri määrä vaikutti mallin antamien oikeiden vastausten määrään niin, että se oli huomattavasti matalampi kolmeen muuhun malliin verrattuna. JSON-syntaksivirheitä sisältävien vastausten määrää voitaisiin mahdollisesti vähentää kehotesuunnittelulla, joka painot-

taa vastausten oikeellista muotoilua. Tämän lisäksi Llama-malleista voitaisiin käyttää sellaista versiota, jolle on saatavilla JSON-formaattia painottavia työkaluja.

Llama3-70b -malli oli muita malleja sensitiivisempi sille annetuille kehoitteille ja tällöin kehoitesuunnittelu onkin erityisen tärkeää. Oikeanlaisilla kehoitteilla voidaan ohjata mallia antamaan toivottuja vastauksia. Llama3-70b -mallin käytössä tuleekin olla erityisen tarkka käytetyistä sanoista ja muotoiluista. Myös erilaisten kehoitteiden kokeilulla saadaan enemmän tietoa siitä, millaiset asiat vaikuttavat mallin toimintaan positiivisesti tai negatiivisesti. Tässä tutkielmassa käytetyt kehoitteet vaihtelivat suhteellisen vähän eikä tarkoituksena ollut systemaattisesti vertailla merkittävästi eroavia kehoitteita vaan lähinnä tarkastella erilaisten mallien suoritusta rakenteellistamistehtävässä ja selvittää, onko kehoitteiden pienillä vaihteluilla merkittävää eroa mallien suorituksiin.

Llama3-70b -mallin vahvuus on se, että se on avoin suuri kielimalli, jonka parametriarvot ja neuroverkkoarkkitehtuuri ovat avoimesti ladattavissa. Tämä tarkoittaa, että malli on käyttäjälleen joustavampi ja sitä voi muuttaa oman käyttökohteen mukaisesti sopivalla tavalla. Tällainen käyttö vaatii kuitenkin usein hienosäätöä, mikä on erilainen lähestymistapa tämän tutkielman implementointiosuuteen verrattuna. Avoimien suurten kielimallien etuna on myös aiemmin tutkielmassa mainittu käyttömahdollisuus käyttäjän omassa sisäverkossa, jolloin arkaluonteista tai muuten tärkeää tietoa ei tarvitse jakaa ulkopuolisille toimijoille.

Gemini1.0-pro-001 -mallin vahvuuksia oli muita malleja selkeästi parempi toistettavuus. Gemini1.0-pro-001-mallilla on myös nähtävissä sensitiivisyyttä käytettyjä kehoitteita kohtaan, vaikkei yhtä vahvasti kuin Llama3-70b -mallilla. Mallin suoriutuminen eri kehoitteilla on kaikista malleista johdonmukaisinta. Tämä tukee johtopäätöstä mallin vastausten deterministisyydestä. Gemini1.0-pro-001 -mallin suoriutuminen kehoitevariaatioiden välillä oli yleisesti samankaltaista.

Myös Gemini-malleista on saatavilla versioita, joilla on käytettävissä työkaluja, jotka painottavat JSON-formaatin käyttöä mallin antamissa vastauksissa. Vaikka Gemini1.0-pro-001 -mallin antamien JSON-syntaksivirheitä sisältävien vastausten määrä oli huomattavasti pienempi kuin Llama3-70b -mallilla, oli niitä silti jonkin verran. Mallin heikkous oli kuitenkin toistuvat muotoiluvirheet kaikissa mallin vastauksissa. Tämän tutkielman kokeiluosuudessa kaikista vastauksista poistettiin jo toistuvat muotoiluvirheet. Rakenteellistamistehtävässä olisi kuitenkin hyödyllisintä, jos vastaukset olisi muotoiltu täysin pyydetyn formaatin mukaisesti. Parhaan tuloksen saamiseksi tällaisten työkalujen käyttäminen voisikin olla järkevää.

Vaikka kaikista tutkielmassa käytetyistä malliperheistä julkaistaan uusia versioita nopealla tahdilla ja uusien versioiden ominaisuudet ja mahdollisuudet saattavat

olla houkuttelevia erilaisissa sovelluksissa, on myös vanhemmilla versioilla omat hyvät puolensa. Uudempien versioiden oikeellisuus on usein vanhoja versioita parempi, mutta niiden käyttökustannukset saattavat nousta korkeiksi varsinkin suurilla aineistoilla. Tällöin vanhempien versioiden käyttäminen saattaa olla kustannustehokkaampaa sekä ajallisesti että rahallisesti, ja myös niillä saadut tulokset ovat usein riittävän hyviä.

6 Yhteenveto

Suurten kielimallien kehitys on nopeaa ja niiden käyttötarkoitukset ovat laajoja arkisista pienistä tehtävistä suuriin monimutkaisiin tehtäviin, joita ei aiemmin kehitettyillä tekniikoilla tai tavoilla ole voitu ratkaista riittävän hyvin. Tässä tutkielmassa esitellyssä tehtävässä tarkasteltiin neljän suuren kielimallin suorituksia lääketieteellisten lausuntojen rakenteellistamisessa.

Rakenteellistamistehtäviä on monenlaisia. Tässä tutkielmassa käsiteltiin lääketieteellisiä lausuntoja, jotka tuovat mukanaan alalle ominaisia haasteita. Eettisyys ja yksityisyys ovat tärkeässä asemassa lääketieteessä ja ne tuovat oman haasteensa myös silloin, kun tehtäviä pyritään ratkaisemaan suurilla kielimalleilla. Tämän tutkielman implementointiosuudessa käytettiin tästä syystä edustavia synteettisiä eturauhassyövästä kirjoitettuja histopatologisia potilaslausuntoja. Jos tehtävää halutaan toistaa alkuperäiselle aineistolle, on käytettävien mallien valinta kiinni myös siitä, millaisia malleja on saatavilla niin, että potilastietoja voidaan käsitellä eettisesti ja potilaiden yksityisyyden säilyttäen.

Ideaalitilanteessa rakenteellistamistehtävässä kaikki vastaukset olisivat oikein ja täysin toistettavissa. Todellisuudessa tämä harvoin toteutuu. Kehotesuunnittelulla voidaan kuitenkin ohjata mallien antamia vastauksia toivottuun suuntaan. Kehotesuunnittelu on usein iteratiivinen prosessi, jossa mallien vastausten perusteella kehoitteita kehitetään edelleen. Se vaatii paljon resursseja ja tietotaitoa, mutta esimerkiksi hienosäätämiseen verrattuna se on tietoteknisesti halvempi prosessi. Tämän lisäksi se on monille helpommin lähestyttävämpi mallien hienosäätämiseen verrattuna, sillä se ei vaadi yhtä paljoa tietoteknistä osaamista käyttäjältään.

Suurten kielimallien tuomat mahdollisuudet leviävät laajemmin uusiin sovelluskohteisiin, kun joka vuosi niiden kehitys menee eteenpäin harppauksin. Suurten kielimallien kehitys on tällä hetkellä niin nopeaa, että tutkielmaa kirjoitettaessa monista malleista ilmestyi uusia versioita ja myös täysin uusia malleja julkaistiin. Suurten kielimallien rakenteessa keskeisessä osassa oleva transformer-arkkitehtuuri julkaistiin vuonna 2017. Alle kymmenessä vuodessa suuret kielimallit ja tekoäly ovat mullistaneet useita aloja. Suuria kielimalleja halutaan hyödyntää eri yritysten ja julkisen sektorin toiminnassa niiden tuoman tehokkuuden ja laajojen mahdollisuuksien takia.

Tutkielman kirjoittamisen aikaan yksi laajempi suurten kielimallien kehitystrendi oli multimodaalisen datan käsittely eli tekstin lisäksi esimerkiksi kuvan, äänen ja videon yhdistämistä samalla mallilla. Tästä yhtenä ensimmäisenä esimerkkinä julkiseen käyttöön tuli GPT-4o [19], joka voi käyttää syötteenä ja antaa ulostulona minkä tahansa yhdistelmän tekstiä, ääntä, kuvaa tai videota.

Uusien suurten kielimallien kehittäminen on usein myös nopeuskilpailua. Alalle on vakiintunut joitakin päätekijöitä, mutta myös uusilla tekijöillä on mahdollisuus nousta esille. Tämä nähtiin muun muassa vuoden 2025 alussa julkaistun kiinalaisen DeepSeekin [14] tapauksessa. Uusien ominaisuuksien kehittäminen ja markkinoille tuonti on niin kovan kilpailun alaisena, että joskus suuret kielimallit herättävät kysymyksiä niiden eettisyydestä. Suuret kielimallit kehittyvät niin nopeasti ja kansainvälisesti, että lainsäädännön on haastavaa pysyä kehityksessä ajan tasalla. Esimerkiksi opetusaineistot, suurten kielimallien käyttötarkoitukset ja niille annettujen tietojen käyttötarkoitukset ovat herättäneet huolta tasaisin väliajoin.

Tässä tutkielmassa käytetyt suuret kielimallit, GPT-3.5, GPT-4, Gemini1.0-pro-001 ja Llama3-70b kuuluvat tämän hetken seuratuimpien ja suosituimpien suurien kielimallien perheisiin, ja ovat ainakin osittain edustavia niiden uudemmissa versioista, kuten GPT 4o -, Gemini 2.5 - ja Llama3.3 -sarjoista. Nyt esitettyjä tuloksia ei kuitenkaan voi suoraan ekstrapoloida näihin seuraavan sukupolven malleihin. Vaikka esimerkiksi Llama3-70b -mallin tapauksessa kehoteohjeistuksessa oli selkeästi vielä varaa lisätutkimukselle, kaikki neljä mallia suoriutuivat hyvin suomenkielisten lääketieteellisten lausuntojen rakenteellistamistehtävästä ja edelleen kehitetyllä kehotesuunnittelulla niiden suorituksilla on potentiaalia olla erinomaisia kaikissa tarkasteluissa.

Viitteet

- [1] M. Alkhalaf, P. Yu, M. Yin, and C. Deng. Applying generative ai with retrieval augmented generation to summarize and extract key clinical information from electronic health records. *Journal of Biomedical Informatics*, page 104662, 2024.
- [2] E. Alpaydin. *Introduction to machine learning*. MIT press, 2020.
- [3] R. Anil, A. M. Dai, O. Firat, M. Johnson, D. Lepikhin, A. Passos, S. Shakeri, E. Taropa, P. Bailey, Z. Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.
- [4] R. Arumugam and R. Shanmugamani. *Hands-On Natural Language Processing with Python: A practical guide to applying deep learning architectures to your NLP applications*. Packt Publishing Ltd, 2018.
- [5] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.
- [6] Bray. Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. 2024.
- [7] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [8] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 2023.
- [9] H. Dalianis. *Clinical text mining: Secondary use of electronic patient records*. Springer Nature, 2018.
- [10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [11] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- [12] W. Ertel. *Introduction to artificial intelligence*. Springer, 2018.

- [13] M. A. Fink, A. Bischoff, C. A. Fink, M. Moll, J. Kroschke, L. Dulz, C. P. Heußel, H.-U. Kauczor, and T. F. Weber. Potential of chatgpt and gpt-4 for data mining of free-text ct reports on lung cancer. *Radiology*, 308(3):e231362, 2023.
- [14] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- [15] T. Han, S. Nebelung, F. Khader, T. Wang, G. Müller-Franzes, C. Kuhl, S. Försch, J. Kleesiek, C. Haarbuerger, K. K. Bressemer, et al. Medical large language models are susceptible to targeted misinformation attacks. *NPJ digital medicine*, 7(1):288, 2024.
- [16] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [17] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. d. L. Casas, L. A. Hendricks, J. Welbl, A. Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.
- [18] Y. Hu, Q. Chen, J. Du, X. Peng, V. K. Keloth, X. Zuo, Y. Zhou, Z. Li, X. Jiang, Z. Lu, et al. Improving large language models for clinical named entity recognition via prompt engineering. *Journal of the American Medical Informatics Association*, page ocad259, 2024.
- [19] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- [20] D. Jurafsky and J. H. Martin. *Speech and language processing : an introduction to natural language processing, computational linguistics and speech recognition*. Prentice Hall series in artificial intelligence. Prentice Hall, Upper Saddle River, N.J, 2nd ed. edition, 2009.
- [21] I. Kaleva, T. Mirtti, A. Rannikko, and T. Laajala. P6 practical comparison of gpt 3.5, gemini and llama 2 llms for structuring biomedical data from synthetic finnish prostate cancer pathology statements. *European Urology Open Science*, 63:S6, 2024.

- [22] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.
- [23] D. Khurana, A. Koli, K. Khatter, and S. Singh. Natural language processing: State of the art, current trends and challenges. *Multimedia tools and applications*, 82(3):3713–3744, 2023.
- [24] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.
- [25] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.
- [26] J. Lei Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. *ArXiv e-prints*, pages arXiv–1607, 2016.
- [27] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.
- [28] S. Locke, A. Bashall, S. Al-Adely, J. Moore, A. Wilson, and G. B. Kitchen. Natural language processing in medicine: a review. *Trends in Anaesthesia and Critical Care*, 38:4–9, 2021.
- [29] R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon, and T.-Y. Liu. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6):bbac409, 2022.
- [30] R. Luukkonen, J. Burdge, E. Zosa, V. Komulainen, P. Sarlin, and S. Pyysalo. Viking: A family of nordic llms, 2025.
- [31] R. Luukkonen, J. Burdge, E. Zosa, A. Talman, V. Komulainen, V. Hatanpää, P. Sarlin, and S. Pyysalo. Poro 34b and the blessing of multilinguality. *arXiv preprint arXiv:2404.01856*, 2024.
- [32] R. Luukkonen, V. Komulainen, J. Luoma, A. Eskelinen, J. Kanerva, H.-M. Kupari, F. Ginter, V. Laippala, N. Muennighoff, A. Piktus, et al. Fingpt: Large generative models for a small language. *arXiv preprint arXiv:2311.05640*, 2023.

- [33] S. Marsland. *Machine learning: an algorithmic perspective*. Chapman and Hall/CRC, 2011.
- [34] B. Meskó. Prompt engineering as an important emerging skill for medical professionals: tutorial. *Journal of Medical Internet Research*, 25:e50638, 2023.
- [35] M. A. Nielsen. *Neural networks and deep learning*, volume 25. Determination press San Francisco, CA, USA, 2015.
- [36] OpenAI. Searchgpt prototype, 2024.
- [37] Y. Peng, A. Rios, R. Kavuluru, and Z. Lu. Extracting chemical–protein relations with ensembles of svm and deep learning models. *Database: The Journal of Biological Databases and Curation*, 2018.
- [38] J. Piskorski and R. Yangarber. Information extraction: Past, present and future. *Multi-source, multilingual information extraction and summarization*, pages 23–49, 2013.
- [39] J. Pitkäniemi, N. Malila, S. Heikkinen, and K. Seppä. Syöpä 2022: Tilastoraportti Suomen syöpätilanteesta. *Suomen Syöpäyhdistys*, 2024.
- [40] S. Pyysalo, J. Kanerva, A. Virtanen, and F. Ginter. Wikibert models: deep transfer learning for many languages. *arXiv preprint arXiv:2006.01538*, 2020.
- [41] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang. Pre-trained models for natural language processing: A survey. *Science China technological sciences*, 63(10):1872–1897, 2020.
- [42] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al. Improving language understanding by generative pre-training. 2018.
- [43] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- [44] A. Rahali and M. A. Akhloufi. End-to-end transformer-based models in textual-based nlp. *AI*, 4(1):54–110, 2023.
- [45] Future of Life Institute. Pause giant ai experiments: An open letter, 2023.
- [46] M. Renze and E. Guven. The effect of sampling temperature on problem solving in large language models. *arXiv preprint arXiv:2402.05201*, 2024.

- [47] K. Sandeman, A. Rannikko, A. Kenttämies, and T. Mirtti. Eturauhassyövän diagnostiikka murroksessa. *Lääketieteellinen Aikakauskirja Duodecim*, 2016.
- [48] O. Sanseviero, L. Tunstall, P. Schmid, S. Mangrulkar, Y. Belkada, and P. Cuenca. Mixture of experts explained, 2023.
- [49] T. Schick and H. Schütze. It’s not just size that matters: Small language models are also few-shot learners. *arXiv preprint arXiv:2009.07118*, 2020.
- [50] R. Sennrich, B. Haddow, and A. Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.
- [51] J. Song, S. H. Min, S. Chae, K. H. Bowles, M. V. McDonald, M. Hobensack, Y. Barrón, S. Sridharan, A. Davoudi, S. Oh, et al. Uncovering hidden trends: identifying time trajectories in risk factors documented in clinical notes and predicting hospitalizations and emergency department visits during home health care. *Journal of the American Medical Informatics Association*, 30(11):1801–1810, 2023.
- [52] H. Suominen. Towards an international electronic repository and virtual laboratory of open data and open-source software for telehealth research: comparison of international, australian and finnish privacy policies. In *Global Telehealth 2012*, pages 153–160. IOS Press, 2012.
- [53] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- [54] G. Team, P. Georgiev, V. I. Lei, R. Burnell, L. Bai, A. Gulati, G. Tanzer, D. Vincent, Z. Pan, S. Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.
- [55] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting. Large language models in medicine. *Nature medicine*, 29(8):1930–1940, 2023.
- [56] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

- [57] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- [58] R. J. Turner, F. Coenen, F. Roelofs, K. Hagoort, A. Härmä, P. D. Grünwald, F. P. Velders, and F. E. Scheepers. Information extraction from free text for aiding transdiagnostic psychiatry: Constructing nlp pipelines tailored to clinicians’ needs. *BMC psychiatry*, 22(1):407, 2022.
- [59] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [60] J. Wang, E. Shi, S. Yu, Z. Wu, C. Ma, H. Dai, Q. Yang, Y. Kang, J. Wu, H. Hu, et al. Prompt engineering for healthcare: Methodologies and applications. *arXiv preprint arXiv:2304.14670*, 2023.
- [61] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.
- [62] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, and D. C. Schmidt. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*, 2023.
- [63] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, et al. Google’s neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.
- [64] H. Xiong, J. Bian, Y. Li, X. Li, M. Du, S. Wang, D. Yin, and S. Helal. When search engine services meet large language models: visions and challenges. *IEEE Transactions on Services Computing*, 2024.
- [65] Y. Yang, Z. Wu, Y. Yang, S. Lian, F. Guo, and Z. Wang. A survey of information extraction based on deep learning. *Applied Sciences*, 12(19):9691, 2022.
- [66] X. Zhang, N. Talukdar, S. Vemulapalli, S. Ahn, J. Wang, H. Meng, S. M. B. Murtaza, D. Leshchiner, A. A. Dave, D. F. Joseph, et al. Comparison of prompt engineering and fine-tuning strategies in large language models in the classification of clinical notes. *medRxiv*, pages 2024–02, 2024.

- [67] Y. Zhang, Y. Li, L. Cui, D. Cai, L. Liu, T. Fu, X. Huang, E. Zhao, Y. Zhang, Y. Chen, et al. Siren’s song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*, 2023.
- [68] Y. Zhang, H. Lin, Z. Yang, J. Wang, S. Zhang, Y. Sun, and L. Yang. A hybrid model based on neural networks for biomedical relation extraction. *Journal of biomedical informatics*, 81:83–92, 2018.

Liitteet

A Tutkielman implementointiosuudessa käytetyt kehotteet.

P0

Merkkien ''' välissä on annettu histopatologin kirjoittama lausunto eturauhassyöpänäytteestä. Vastaa esitettyihin kysymyksiin. Anna vastaukset JSON-formaatissa. Anna vastauksissa avaimena merkkien () välissä oleva sisältö kysymyskohtaisesti.

'' ''

1. Analysoi annettua lääkärin kirjoittamaa lausuntoa, ja vastaa sen perusteella kysymyksiin a.-c. Anna kaikkien lausunnosta löydettyjen arvojen kanssa niihin liittyvä määre, jos sellainen on. Anna vastauksena vain numeerisia arvoja. Jos lausunnosta ei löydy selkeää vastausta, anna vastaus 'NA'.

\\

a. Kuinka monta näytettä potilaalta on otettu? (Näytteiden määrä)\\

b. Minkä kokoisia otetut näytteet ovat? (Näytteiden koko) \\

c. Mikä on näytteiden Gleason-luokitus? (Gleason-luokitus) \\

2. Analysoi annettua lääkärin kirjoittamaa lausuntoa, ja vastaa sen perusteella 'kyllä' tai 'ei' kysymyksiin d.-f., jos lausunnosta löytyy selkeä vastaus.

Jos lausunnosta ei löydy selkeää vastausta, anna vastaus 'NA'. \\

d. Onko näytteessä havaittu hyvälaatuista eturauhasen liikakasvua? (Eturauhasen hyvälaatuinen liikakasvu)\\

e. Onko näytteessä havaittu perineuraalista kasvua? (Perineuraalinen kasvu) \\

f. Onko näytteessä havaittu atypiaa? (Atypia)

P1

Merkkien "" välissä on annettu histopatologin kirjoittama lausunto eturauhassyöpänäytteestä. Tehtäväsi on toimia lääketieteellisenä assistenttina ja etsiä lausunnosta alla listatut tiedot. Käy kysymykset läpi askel kerrallaan. Anna vastaukset JSON-formaatissa.\\

'' '' \\

1. Näytteiden määrä: \\

-Etsi lausunnosta maininta siitä, kuinka monta näytettä potilaalta on otettu.

\\

- Anna vastaukseksi otettujen näytteiden määrä. \\

-Jos lausunnosta ei löydy selvää mainintaa näytteiden määrästä, anna vastaukseksi 'NA' \\

2.Näytteiden koko: \\

-Etsi lausunnosta maininta siitä, minkä kokoisia potilaalta otetut näytteet ovat. \\

-Anna vastaukseksi näytteiden koko numeerisena arvona. Anna numeerisen arvon lisäksi lausunnossa mainittu määre, jos sellainen on. \\

-Jos lausunnosta ei löydy selvää mainintaa näytteiden koosta, anna vastaukseksi 'NA'. \\

3. Gleason-luokitus: \\

-Etsi lausunnosta maininta näytteiden Gleason-luokituksesta. \\

-Anna vastaukseksi näytteiden Gleason-luokitus muodossa major+minor=summa tai pelkkä summa. \\

-Jos lausunnosta ei löydy selvää mainintaa näytteiden Gleason-luokituksesta, anna vastaukseksi 'NA'. \\

4.Hyväläatuinen eturauhasen liikakasvu: \\

-Etsi maininta siitä, onko näytteessä havaittu hyväläatuista eturauhasen liikakasvua vai ei. \\

-Jos lausunnossa on selvästi kerrottu, että näytteessä on hyväläatuista eturauhasen liikakasvua, anna vastaukseksi 'kyllä'. Jos lausunnossa on selvästi kerrottu, että näytteessä ei ole hyväläatuista eturauhasen liikakasvua, anna vastaukseksi 'ei'. \\

-Jos lausunnosta ei löydy selvää mainintaa, anna vastaukseksi 'NA'. \\

5.Perineuraalinen kasvu: \\

-Etsi maininta siitä, onko näytteessä havaittu perineuraalista kasvua vai ei. \\

-Jos lausunnossa on selvästi kerrottu, että näytteessä on perineuraalista kasvua, anna vastaukseksi 'kyllä'. Jos lausunnossa on selvästi kerrottu, että näytteessä ei ole perineuraalista kasvua, anna vastaukseksi 'ei'.

\\

-Jos lausunnosta ei löydy selvää mainintaa, anna vastaukseksi 'NA' \\

6.Atypia: \\

- Etsi maininta siitä, onko näytteessä havaittu atypiaa vai ei. \\

-Jos lausunnossa on selvästi kerrottu, että näytteessä on atypiaa, anna vastaukseksi 'kyllä'. Jos lausunnossa on selvästi kerrottu, että näytteessä ei ole atypiaa, anna vastaukseksi 'ei'. \\

-Jos lausunnosta ei löydy selvää mainintaa, anna vastaukseksi 'NA'", \\

P2

Merkkien ''' välissä on annettu histopatologin kirjoittama lausunto eturauhassyöpänäytteestä. Tehtäväsi on etsiä tietoa annetusta lausunnosta. Etsi vastaukset alla oleviin kysymyksiin. Anna vastaukset () merkkien välissä olevien ohjeiden mukaan. Anna vastaukset JSON-formaatissa. \\

''' \\

1. Kuinka monta näytettä potilaalta on otettu? (näytteiden määrä)\\
2. Mikä on otettujen näytteiden koko? (näytteiden koko)\\
3. Mikä on näytteiden Gleason-luokitus? (major+minor=summa, summa tai 'NA')\\
4. Onko näytteessä havaittu hyvälaatuista eturauhasen liikakasvua? ('kyllä', 'ei' tai 'NA')\\
5. Onko näytteessä havaittu perineuraalista kasvua? ('kyllä', 'ei' tai 'NA') \\
6. Onko näytteessä havaittu atypiaa? ('kyllä', 'ei' tai 'NA')", \\

P3

You will be provided with a histopathological prostate cancer sample statement delimited by quotes (symbols“”). The text is a medical note written by a pathologist and it is in Finnish. Your job is to extract information from the given text. Find the answers to the questions below. Give the answers based on the instructions given between the parenthesis. Give the answers in JSON-format. \\

“ ” \\

1. How many samples have been taken? (number of samples)\\
2. What is the size of the samples taken? (size of samples)\\
3. What is the Gleason score of the samples? (major+minor=sum, sum or 'NA')\\
4. Can BPH be detected in the samples taken? ('yes', 'no' or 'NA')\\
5. Can perineural growth be detected in the samples taken? ('yes', 'no' or 'NA')\\
6. Can atypia be detected in the samples taken? ('yes', 'no' or 'NA') \\

P4

Merkkien “ ” välissä on annettu histopatologin kirjoittama lausunto eturauhassyöpänäytteestä. Olet erittäin kokenut histopatologiaan erikoistunut erikoislääkäri ja tehtäväsi on vastata esitettyihin kysymyksiin annetun lausunnon perusteella. Anna vastaukset JSON-formaatissa. Anna vastauksissa avaimena merkkien () välissä oleva sisältö kysymyskohtaisesti.\\

“ ”\\

Analysoi annettua histopatologin kirjoittamaa lausuntoa, ja vastaa sen perusteella kysymyksiin 1.-3. Anna kaikkien lausunnosta löydettyjen arvojen kanssa niihin liittyvä määre, jos sellainen on. Anna vastauksena vain numeerisia arvoja. Jos patologisesta lausunnosta ei löydy selkeää vastausta, anna vastaukseksi 'NA'. \\

1. Kuinka monta näytettä potilaalta on otettu? (Näytteiden määrä) \\
2. Minkä kokoisia otetut näytteet ovat? (Näytteiden koko) \\
3. Mikä on näytteiden Gleason-luokitus? (Gleason-luokitus) \\

Analysoi annettua histopatologin kirjoittamaa lausuntoa, ja vastaa sen perusteella kysymyksiin 4.-6., Anna vastaukseksi 'kyllä' tai 'ei', jos lausunnosta löytyy selkeä vastaus. Jos potilaslausunnosta ei löydy selkeää vastausta, anna vastaukseksi 'NA'. \\

4. Onko näytteessä havaittu hyvälaatuista eturauhasen liikakasvua? (Eturauhasen hyvälaatuinen liikakasvu) \\
5. Onko näytteessä havaittu perineuraalista kasvua? (Perineuraalinen kasvu) \\
6. Onko näytteessä havaittu atypiaa? (Atypia)\\

P5

Merkkien “ ” välissä on annettu histopatologin kirjoittama lausunto eturauhassyöpänäytteestä. Olet erittäin kokenut histopatologiaan erikoistunut erikoislääkäri, jonka tehtävä on etsiä lausunnosta alla listatut tiedot. Käy kohdat läpi askel kerrallaan. Anna vastaukset yhtenä objektina JSON-formaatissa. \\

“ ” \\

- 1.Näytteiden määrä: \\

- Etsi lausunnosta maininta siitä, kuinka monta näytettä potilaalta on otettu. \\
 - Anna vastaukseksi otettujen näytteiden määrä \\
 - Jos lausunnosta ei löydy selvää mainintaa näytteiden määrästä, anna vastaukseksi 'NA' \\
- 2.Näytteiden koko: \\
- Etsi lausunnosta maininta siitä, minkä kokoisia potilaalta otetut näytteet ovat \\
 - Anna vastaukseksi näytteiden koko numeerisena arvona. Anna numeerisen arvon lisäksi lausunnossa mainittu määre, jos sellainen on \\
 - Jos lausunnosta ei löydy selvää mainintaa näytteiden koosta, anna vastaukseksi 'NA' \\
3. Gleason-luokitus: \\
- Etsi lausunnosta maininta näytteiden Gleason-luokituksesta \\
 - Anna vastaukseksi näytteiden Gleason-luokitus muodossa major+minor=summa tai pelkkä summa \\
 - Jos lausunnosta ei löydy selvää mainintaa näytteiden Gleason-luokituksesta, anna vastaukseksi 'NA' \\
- 4.Hyväläatuinen eturauhasen liikakasvu: \\
- Etsi maininta siitä, onko näytteessä havaittu hyväläatuista eturauhasen liikakasvua vai ei \\
 - Jos lausunnossa on selvästi kerrottu, että näytteessä on hyväläatuista eturauhasen liikakasvua, anna vastaukseksi 'kyllä'. Jos lausunnossa on selvästi kerrottu, että näytteessä ei ole hyväläatuista eturauhasen liikakasvua, anna vastaukseksi 'ei'. \\
 - Jos lausunnosta ei löydy selvää mainintaa, anna vastaukseksi 'NA' \\
- 5.Perineuraalinen kasvu: \\
- Etsi maininta siitä, onko näytteessä havaittu perineuraalista kasvua vai ei \\
 - Jos lausunnossa on selvästi kerrottu, että näytteessä on perineuraalista kasvua, anna vastaukseksi 'kyllä'. Jos lausunnossa on selvästi kerrottu, että näytteessä ei ole perineuraalista kasvua, anna vastaukseksi 'ei'. \\
 - Jos lausunnosta ei löydy selvää mainintaa, anna vastaukseksi 'NA' \\
- 6.Atypia: \\
- Etsi maininta siitä, onko näytteessä havaittu atypiaa vai ei \\
 - Jos lausunnossa on selvästi kerrottu, että näytteessä on atypiaa, anna

vastaukseksi 'kyllä'. Jos lausunnossa on selvästi kerrottu, että näytteessä ei ole atypiaa, anna vastaukseksi 'ei'.\\

- Jos lausunnosta ei löydy selvää mainintaa, anna vastaukseksi 'NA' \\

P6

Merkkien “ ” välissä on annettu histopatologin kirjoittama lausunto eturauhassyöpänäytteestä. Olet erittäin kokenut histopatologiaan erikoistunut erikoislääkäri, jonka tehtävä on etsiä tietoa annetusta lausunnosta. Etsi vastaukset alla oleviin kysymyksiin. Anna vastaukset () merkkien välissä olevien ohjeiden mukaan. Anna vastaukset JSON-formaatissa.\\

“ ” \\

1. Kuinka monta näytettä potilaalta on otettu? (näytteiden määrä) \\
2. Minkä kokoisia otetutu näytteet ovat? (näytteiden koko) \\
3. Mikä on näytteiden Gleason-luokitus? (major+minor=summa/summa/NA) \\
4. Onko näytteessä havaittu hyvälaatuista eturauhasen liikakasvua? (kyllä/ei/NA) \\
5. Onko näytteessä havaittu perineuraalista kasvua? (kyllä/ei/tai/NA) \\
6. Onko näytteessä havaittu atypiaa? (kyllä/ei/NA)" \\

P7

You will be provided with text delimited by quotes. The text is a medical note written by a histopathologist and it is in Finnish. Your job is to act as a very experienced doctor that specializes in pathology and extract information from the given text. Find the answers to the questions below. Give the answers based on the instructions given between the parenthesis. Give the answers in JSON-format. \\

“ ” \\

1. How many samples have been taken? (number of samples) \\
2. What is the size of the samples taken? (size of samples) \\
3. What is the Gleason score of the samples? (major+minor=sum/sum/NA) \\
4. Can BPH be detected in the samples taken? (kyllä/ei/NA) \\

5. Can perineural growth be detected in the samples taken? (yes/no/NA)
\\
6. Can atypia be detected in the samples taken? (kyllä/ei/NA)\\

B Esimerkkejä saaduista vastauksista.

Esimerkkilausunto

"Tutkittavana systeemibiopsianeulanäytteitä pituudeltaan yhteensä 4+6+3 vasemmalta ja 5+7+8mm oikealta. Näytteiden indeksoinnissa kuitenkin käynyt fiba, ja vasen on oikea ja vice versa. Korjauksen jälkeen oikean puolen ensimmäisessä ja kolmannessa näytteessä näkyy maligniin viittaavaa kasvua ja marginaalit positiivit. Vahva adenokarsinooma-suspisio siis. Näytteiden laatu kuitenkin kyseenalainen ja neulat eivät todnäk täysin edustuskelpoisia, joten suositellaan biopsiointia uudelleen ohjatusti fuusiona oikealle suspisio-alueelle."

Esimerkki oikeasta vastauksesta (GPT-4-mallin vastaus)

```
{
  "1": 6,
  "2": "4+6+3 mm vasemmalta, 5+7+8 mm oikealta",
  "3": "NA",
  "4": "NA",
  "5": "NA",
  "6": "NA"
}
```

Esimerkki väärästä vastauksesta (GPT-3.5-mallin vastaus)

```
{
  "number of samples": 6,
  "size of samples": {
    "left": [4, 6, 3],
    "right": [5, 7, 8]
  },
}
```

```
"Gleason score": "4+3=7/NA",
"BPH detected": "ei",
"perineural growth detected": "kyllä",
"atypia detected": "NA"
}
```

Esimerkki JSON-syntaksivirheestä (Llama3-70b-mallin vastaus)

Here are the answers in JSON format:

```
““
{
  "1. Number of samples": 6,
  "2. Size of samples": "4+6+3 mm and 5+7+8 mm",
  "3. Gleason score": "NA",
  "4. BPH detection": "NA",
  "5. Perineural growth detection": "NA",
  "6. Atypia detection": "kyllä" (note: this is likely referring to the
  "maligniin viittaavaa kasvua" mentioned in the text, which suggests malignant
  growth, but not specifically atypia)
}
““
```

Note: The Gleason score, BPH detection, and perineural growth detection are not mentioned in the provided text, so I've marked them as "NA" (not available).