GPT-kielimallien soveltaminen chat-pohjaisen asiakaspalvelun arviointiin

Turun yliopisto
Tietotekniikan laitos

Diplomityö
Tieto- ja viestintätekniikka
Maaliskuu 2025

Jukka Tuominen

Timo Maaranen (Telia Finland Oyj: ohjaaja)
Jari Björne (Turun yliopisto: ohjaaja/tarkastaja)
Antti Airola (Turun yliopisto: tarkastaja)

TURUN YLIOPISTO
Tietotekniikan laitos

Jukka Tuominen: GPT-kielimallien soveltaminen chat-pohjaisen asiakaspalvelun arviointiin

Diplomityö, 71 s.
Tieto- ja viestintätekniikka
Maaliskuu 2025

Tässä diplomityössä tarkastellaan teleoperaattorin asiakaspalvelun keskustelujen ja niihin liittyvien tapahtumien arvioinnin automatisointia GPT-kielimalleja käyttävän toteutuksen avulla. Työn lähtökohtana oli toimeksiantajan eli teleoperaattorin tarve arvioida enemmän, tarkemmin ja tehokkaammin asiakaspalvelussa tapahtuvia tilanteita sekä havaita niissä kehityskohteita asiakaspalvelun laadun ja toiminnan parantamista varten.

Työn keskeisenä tavoitteena on arvioida ja vertailla eri GPT-kielimallien ymmärrys- ja arviointikykyjä sekä käyttökustannuksia ja niiden perusteella kehittää samalla toimiva työkalu, jota voidaan käyttää asiakaspalvelun arviointiprosessin automatisointiin. Asiakaspalvelun keskustelujen tekstien ja lisätietojen perusteella tilanteista voidaan kerätä kuvaavia tietojoukkoja, joita voidaan käyttää asiakaspalvelun tilanteiden arviointiin. Teleoperaattorin asiakaspalvelun asiakaskohtaamismallin ja muiden arviointikriteerien perusteella GPT-malleille voidaan muotoilla kehotteita erilaisin oppimismenetelmin ohjaamaan tehtävää arviointityötä.

Tutkimuksessa käytetään kvantitatiivisia tilastollisia menetelmiä mallien kykyjen vertailuun. Käytettävät mallit ovat GPT-3.5 Turbo, GPT-4o Mini ja GPT-4o, joita arvioidaan kahden luokittelutehtävän perusteella. Ensimmäisessä tehtävässä analysoidaan mallien kykyjä ymmärtää keskusteluissa esiintyviä kategorioita ja toisessa puolestaan mallien vastauksina tuottamien arvioinnin arvosanojen oikeellisuutta numeroina. Mallien vertailua täydentää lisäksi niiden käyttökustannusten vertailu kuvaajien kautta.

Tutkimustulokset viittaavat siihen, että GPT-mallit pystyvät tulkitsemaan asiakaspalvelussa tapahtuvia tilanteita ja niitä voidaan käyttää luotettavaan ja objektiiviseen arviointiin. Malleista GPT-4o suoriutuu tehtävistä parhaiten ja GPT-4o Mini on puolestaan selvästi kustannustehokkain. Tuloksien perusteella voidaan tunnistaa myös useita mallien ja toteutuksen käyttöön liittyviä haasteita, jotka on hyvä ottaa huomioon jatkokehitystä varten.

Asiasanat: generatiivinen, tekoäly, automatisoitu arviointi, palaute, laadunvarmistus, chat, asiakaspalvelu, teleoperaattori

UNIVERSITY OF TURKU
Department of Computing

Jukka Tuominen: GPT-kielimallien soveltaminen chat-pohjaisen asiakaspalvelun arviointiin

Master of Science (Tech) Thesis, 71 p.
Information and Communication Technology
March 2025

This thesis examines the automation of evaluating telecom customer service conversations and related interaction events through an implementation leveraging GPT language models. The work was initiated to meet the telecom's need for more frequent, precise, and efficient evaluations of customer service situations. It also aims to identify opportunities to improve both service quality and performance through scoring metrics and written feedback.

The primary goal of the thesis is to evaluate and compare the comprehension and evaluation capabilities of different GPT language models and their usage costs. A key outcome of this research is the development of a functional tool designed to automate the customer service evaluation process by applying insights gained from the comparison results. Descriptive datasets about the customer service situations can be formed and used for evaluating interactions by analyzing conversation texts and additional metadata information related to them. Prompts for the GPT models can be engineered by using various prompting methods with the telecom's customer interaction model and other additional evaluation criteria to set model guidelines of the evaluation task.

The research applies quantitative statistical methods to compare the capabilities of the models. The models used in the study are GPT-3.5 Turbo, GPT-4o Mini and GPT-4o. The models are evaluated based on their performance in two classification tasks. In the first task, the models' abilities related to understanding categories in conversations are analyzed. In the second task, the numerical evaluation score accuracy is assessed based on how close to truth they are. The model comparison is further complemented by a cost comparison illustrated with charts.

The research results indicate that GPT models can interpret customer service situations and can be used for creating reliable and objective evaluations with written feedback. GPT-4o model performs the best in all tasks, while GPT-4o Mini is the most cost-effective solution. The results also reveal several challenges related to the usage of the models and the implementation, which should be considered for further development of the evaluation system.

Keywords: generative, artificial intelligence, automated evaluation, feedback, quality assurance, chat, customer service, telecom

Tässä diplomityössä on käytetty generatiivista tekoälyä kielioppivirheiden, sanojen oikeinkirjoituksen ja lauserakenteiden tarkastamiseen ja korjaamiseen. Työn sisällöstä ja sen oikeellisuudesta vastaa itse työn tekijä.

Sisällys

Kuvat

Taulukot

Lyhenteet

ANN	Artificial Neural Network
AUC	Area Under the Curve
BART	Bidirectional and Auto-Regressive Transformers
BERT	Bidirectional Encoder Representations from Transformers
CSAT	Customer Satisfaction Score
CoT	Chain-of-Thought
GDPR	General Data Protection Regulation
GenAI	Generative Artificial Intelligence
GPT	Generative Pre-trained Transformer
JSON	JavaScript Object Notation
LDA	Latent Dirichlet Allocation
LLM	Large Language Model
LSA	Latent Semantic Analysis
LSTM	Long Short-Term Memory
MCC	Matthews Correlation Coefficient
MLM	Masked Language Modeling
NLG	Natural Language Generation
NLP	Natural Language Processing
NLU	Natural Language Understanding
PaLM	Pathways Language Model
ReLU	Rectified Linear Unit
RLHF	Reinforcement Learning from Human Feedback
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic
T5	Text-To-Text Transfer Transformer

1 Johdanto

Monissa eri chat-asiakaspalveluissa syntyy päivittäin valtavia määriä tekstidataa. Jatkuvasti kasvava tiedon määrä aiheuttaa ongelman, missä esihenkilöiden on lähes mahdotonta arvioida, mitä asiakaspalvelun keskusteluissa arkisin tapahtuu. Kattavaa arviointia varten heidän pitäisi ehtiä lukea satoja tai jopa tuhansia keskusteluja sekä niihin liittyviä tietoja päivittäin, mikä ei ole mahdollista. Tämän vuoksi on tarpeen kehittää automaattisia ratkaisuja, joiden avulla voidaan analysoida, luokitella ja arvioida asiakaspalvelun keskusteluja tarkasti, väsymättä ja nopeasti.

Tekoälyyn perustuvat suuret kielimallit pystyvät lukemaan erittäin suuria määriä tekstiä tehokkaasti. Ne kykenevät tuottamaan johdonmukaisia vastauksia niille syötetyistä teksteistä, mikä mahdollistaa esimerkiksi tiivistelmien ja palautteiden luomisen automaattisesti. Automaattinen arviointi vapauttaa asiakaspalvelun henkilöstöä arjessa tärkeämpiin ja vaativampiin tehtäviin. Se myös varmistaa, että jokainen keskustelu tulee arvioiduksi ainakin jollakin tavalla, vaikka keskusteluissa muodostuukin valtavia määriä tekstidataa.

Tämän työn tarkoituksena on suunnitella ja kehittää toteutus GPT-kielimalleja käyttäen sekä tutkia, miten näitä malleja voidaan soveltaa yhdessä teleoperaattorin eri tietolähteiden kanssa chat-pohjaisen asiakaspalvelun arviointiin, analysointiin ja palautteen luomiseen koulutustarkoituksia varten. Arvioinnin tavoitteena on tuottaa objektiivista palautetta asiakaspalvelun laadun parantamiseksi sekä samalla luoda vertailukelpoisia mittareita laadun varmistuksen tueksi.

1.1 Tutkimuskysymykset

Tämä työ pyrkii vastaamaan seuraaviin tutkimuskysymyksiin:

1. Voiko GPT-kielimalleilla suorittaa asiakaspalvelun tilanteiden arviointia?
2. Kuinka tarkasti ja luotettavasti GPT-kielimallit pystyvät tunnistamaan ja erittelemään erilaisia asiakaspalvelun tilanteita chat-keskusteluista?
3. Millaisia kustannuksia arviointityökalun käyttämisesta muodostuu ja onko työkalun käyttäminen kannattavaa suuressa mittakaavassa?
4. Mikä GPT-malleista on soveltuvin arviointitehtävää varten?
5. Mitkä ovat keskeisimmät tekniset, eettiset ja käytännön haasteet, jotka liittyvät GPT-kielimallien hyödyntämiseen asiakaspalvelun arvioinnissa?

Tutkielman toisessa luvussa tarkastellaan luonnollisen kielen käsittelyn historiaa, generatiivisen tekoälyn käyttökohteita ja GPT-kielimallien toiminnan keskeisiä taustateorioita. Kolmannessa luvussa puolestaan tarkastellaan chat-asiakaspalvelun haasteita ja ongelmia sekä sitä, miten asiakaspalvelun säännöstöä voidaan soveltaa kielimallien käyttöön. Neljännessä luvussa syvennytään suurten kielimallien ohjaamiseen kehotteilla ja niihin liittyviin menetelmiin. Viidennessä luvussa esitellään työssä kehitetyn työkalun arkkitehtuuri ja toimintoja yksinkertaisesti sekä siihen liittyviä oikeudellisia näkökulmia. Kuudennessa luvussa käydään läpi GPT-mallien suorituskykyä erilaisin mittarein työkalun arkkitehtuurin testauksen tulosten perusteella. Seitsemännessä luvussa esitetään työn keskeiset johtopäätökset kehitetyn työkalun toimivuudesta sekä sen soveltuvuudesta asiakaspalvelun arvionnin tehostamiseksi ja pohditaan myös mahdollisia jatkokehityssuuntia, jotka voisivat parantaa työkalun käytettävyyttä ja laajentaa sen soveltamismahdollisuuksia.

2 Teoreettinen viitekehys

2.1 NLP:n ja generatiivisen tekoälyn historiaa

Luonnollisen kielen käsittely *(engl. Natural Language Processing)* eli NLP on yksi tekoälyn vanhimmista osa-alueista, jonka avulla tietokoneet pystyvät nykyään tulkitsemaan ja ymmärtämään sekä myös tuottamaan tekstiä, joka vaikuttaa ihmisen kirjoittamalta. Erilaisten algoritmien ja prosessointimenetelmien avulla tietokoneet ja niiden ohjelmistot voivat oppia kieliopillisia rakenteita, sanojen merkityksiä ja erilaisia asiayhteyksiä tiettyjen kirjainten, merkkien, sanojen ja virkkeiden välillä. NLP on kehittynyt valtavasti verrattuna sen historian alkuaikoihin ja kehitys vaikuttaa vain kiihtyvän.

NLP:n varhaisimmat vaiheet alkoivat 1950-luvulla, kun Alan Turing kehitti kuuluisan Turingin testin, jonka hän esitteli vuonna 1950. Kyseisen testin tarkoituksena oli määrittää tapa ja antaa vastaus siihen, voiko kone jäljitellä ihmisen älykkyyttä, keskustelukykyjä ja kykeneekö se ajattelemaan [1]. Myös toinen 1950-luvulla tapahtunut kehitysaskel oli Georgetownin yliopiston ja teknologiayritys IBM:n yhdessä toteuttama kokeellinen tutkimus vuosina 1953–1954. Tutkimuksessa tiettyjä valittuja venäjänkielisiä lauseita käännettiin englanniksi onnistuneesti, mikä osoitti jo silloin, että tietokoneiden avulla oli mahdollista suorittaa kirjoitettujen kielten kääntämistä [2].

Vuonna 1966 Joseph Weizenbaum puolestaan kehitti ensimmäisen varhaisen chatbotin nimeltään ELIZA, joka pystyi jäljittelemään ihmisen kanssa käytävää keskustelua, mikä oli merkittävä saavutus varhaisen tekoälyn ja NLP:n historiassa [3]. Massachusettsin teknillisen korkeakoulun *(engl. Massachusetts Institute of Technology, MIT)* tutkija Terry Winograd puolestaan kehitti 1970-luvulla *SHRDLU*-ohjelman, joka kykeni ymmärtämään luonnollista kieltä ja vastaanottamaan komentoja rajatussa ympäristössä. Tämä ympäristö oli geometrisiin kuvioihin ja kappaleisiin perustuva interaktiivinen lohkomaailma, jossa oli esimerkiksi siirreltäviä erikokoisia ja -värisiä laatikoita, kartioita, palloja ja kuutioita. [4]

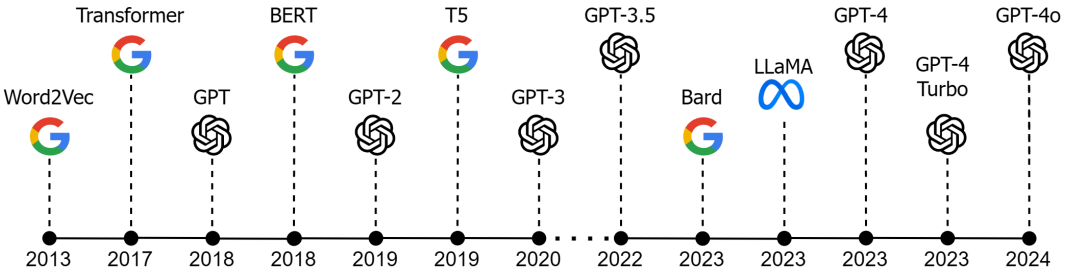
Tietokoneiden laskentatehon kasvu 1980- ja 1990-luvulla mahdollisti erilaisten tilastollisten menetelmien ja koneoppimismallien hyödyntämisen NLP:ssä. Brownin yliopisto kehitti maailman ensimmäisen suuren englanninkielisen tekstikorpuksen, joka on tarkoin määritelty ja koottu kokoelma kirjoitetun kielen tekstejä. Tämä korpus mahdollisti tilastollisten kielimallien kehittämisen. Tilastolliset kielimallit perustuvat algoritmeihin, jotka tuottavat todennäköisyysjakaumia tietyille sanajonoille. Muodostuneiden todennäköisyysjakaumien perusteella nämä mallit pystyvät esimerkiksi ennustamaan lauseen seuraavan sanan sekä täydentämään lauseita ja luokittelemaan sanoja. [5], [6]

Tekoälyn, NLP:n ja koneoppimisen alalla tapahtui merkittäviä ja mullistavia läpimurtoja 2010-luvulla. Vuonna 2013 Tomas Mikolov kehitti *Word2Vec*-tekniikan Googlen kollegoidensa kanssa, jonka avulla sanoja voitiin esittää numeerisessa vektorimuodossa. Tämän tekniikan avulla kielimallit pystyivät ymmärtämään sanojen merkityksiä niiden vektorimuotoisten esitysten kautta. *Word2Vec*-tekniikan etuna oli mahdollisuus oppia suurista tekstiaineistoista sanojen esityksiä *(engl. word embeddings)* siten, että sanat voitiin sijoittaa tietyn ulottuvuuden vektoriavaruuteen toistensa suhteen. Näiden vektorien avulla voitiin laskea sanojen välisiä merkityksiä, samankaltaisuuksia ja vastaavuuksia erilaisissa asiayhteyksissä. [7]

Vuonna 2017 Google Researchin tutkijat esittelivät Ashish Vaswanin johdolla Transformer-arkkitehtuurin, joka johti jo heti seuraavana vuonna Googlen omaan BERT-malliin *(engl. Bidirectional Encoder Representations from Transformers)* ja hieman aikaisemmin samana vuonna 2018 myös tekoälytutkimuskeskus OpenAI esitteli oman Transformer-arkkitehtuuriin perustuvan dekooderi-kielimallinsa nimeltään *Generative Pre-trained Transformer* eli GPT. Vain hieman yli puoli vuotta myöhemmin sitä seurasi jo edistyneempi GPT-2 -dekooderimalli vuoden 2019 alussa, kuten kuvan 2.1 aikajanalta voidaan havaita. 2020-luvulla OpenAI:n GPT-mallien jatkokehityksen myötä niitä seurasi vielä lisää uusia edistyksellisimpiä GPT-malleja. Googlen ja OpenAI:n lisäksi myös useat muut tahot ovat kehittäneet vastaavan tasoisia suuria Transformer-arkkitehtuuriin perustuvia kielimalleja, jotka kykenevät onnistumaan monimutkaisissa ja vaativissa NLP-tehtävissä. [8]

Transformer-arkkitehtuuriin perustuvat suuret kielimallit *(engl. Large Language Model, LLM)* ovat tällä hetkellä uraauurtava läpimurto luonnollisen kielen käsittelyssä, jossa se on syrjäyttänyt kyvykkyydellään erilaisissa tehtävissä monet muut NLP:ssä aiemmin käytetyt neuroverkkoja käyttävät malliarkkitehtuurit. Nämä uudet kielimallit ovat esimerkiksi korvanneet takaisinkytkettyjä neuroverkkoja *(engl. Recurrent Neural Network, RNN)* käyttäviä pitkäkestoiseen lyhytkestomuistiarkkitehtuuriin *(engl. Long Short Term Memory, LSTM)* perustuvia malleja erilaisissa käyttötarkoituksissa. Näistä käyttötarkoituksista esimerkkeinä tekstin generointi, chatbotit ja konekääntäminen. [8]–[10]

2.2 Suuret kielimallit



Kuva 2.1: Suurten kielimallien aikajanaa

Viimeisten muutamien vuosien aikana vuodesta 2017 eteenpäin kielimallien kohdalla ja luonnollisen kielen käsittelyssä on tapahtunut merkittäviä läpimurtoja generatiiviseen tekoälyn suhteen. Laitteistojen ja niiden kasvaneet laskentatehot, ohjelmistojen kehittyminen ja laaja-alaisen ja kattavan koulutusdatan saatavuuden lisääntyminen ovat kaikki yhdessä mahdollistaneet erilaisten arkkitehtuurien suurten kielimallien olemassaolon, toiminnan ja käyttökelpoisen suorituskyvyn käytännössä. Viimeisimmät suuret kielimallit ovat poikkeuksellisia, sillä ne voivat parhaimmillaan saavuttaa erilaisissa tehtävissä jopa saman tarkkuuden ja suorituskyvyn kuin ihminen. [8], [11]

Suuret kielimallit ovat nousseet valtavaan suosioon. Niihin liittyvät erilaiset palvelut, sovellukset, tuotteet ja ratkaisut ovat tällä hetkellä nousussa, sillä niitä yritetään tarjota ja käyttää työkaluina todella moniin erilaisiin käyttötarkoituksiin. Näitä tarjottuja ratkaisuja ja työkaluja on kaikkien saatavilla niin ilmaiseksi kuin myös maksullisina versioina. Esimerkiksi OpenAI:n GPT-kielimalleihin ja niiden eri versioihin perustuvat ChatGPT ja Microsoft Copilot keskustelevat suuret kielimallit pystyvät tavallisten keskusteluiden lisäksi suorittamaan monenlaisia tehtäviä, kuten ohjelmakoodien kirjoittamista, yhteenvetojen tai esitysten laatimista, ideointia sekä taulukoiden ja kuvaajien muodostamista. [12]

2.3 Generatiivisen tekoälyn tehtävät NLP:ssä

Generativiinen tekoäly *(engl. Generative Artificial Intelligence, GenAI)* nousi laajaan julkisuuteen vuosien 2022-2023 aikana uutena tehokkaana työkaluna erilaisten monimuotoisten ja luovuutta vaativien tehtävien ratkaisemiseksi. Nimensä mukaisesti tämä tekoälytyyppi on erittäin tehokas työkalu tuottamaan oppimansa perusteella sisältöä erilaisissa muodoissa. Generatiivisen tekoälyn erilaisia malleja voidaankin jakaa erilaisiin alaluokkiin riippuen siitä, minkälaiseen koulutusainestoon ne perustuvat ja mitä sisältöä ne voivat generoida. Generoitavan sisällön tyyppi ja laatu voivat olla hyvinkin vaihtelevia riippuen mallin syötteestä, parametrien määrästä, koulutusdatasta sekä sen käyttökohteesta. Mallien syötteinä ja koulutusaineistoina voidaan käyttää esimerkiksi tekstiä, kuvia, ääntä, tiedostoja, animaatioita, 3D-malleja sekä jopa näiden yhdistelmiä. Malleilla voidaan siten luoda nopeasti ja erittäin uskottavasti täysin uusia tekstejä, kuvia, videoita, ääniä, musiikkia ja muita tietotyyppejä. Generatiivisen tekoälyn kehittyessä kaikin puolin yhä kyvykkäämmäksi, se tulee vaikuttaamaan entistä enemmän siihen, miten kulutamme palveluita arjessa sekä miten tulkitsemme kaikkea jaettavaa sisältöä sekä informaatiota esimerkiksi sosiaalisessa mediassa tai uutismedioissa. [13], [14]

2.3.1 Tekstin generointi

Luonnollisen kielen generointi *(engl. Natural Language Generation, NLG)* on monien nykyaikaisten generatiivisten kielimallien keskeisin tehtävä. Sitä pidetään myös nykyisten tekoälymallien edistyneimpana ja samalla vanhimpana osa-alueena. NLG on ollut olemassa siitä lähtien, kun *ELIZA* kehitettiin 1960-luvun puolivälissä, vaikka kaupallisesti käyttökelpoiset NLG:ä käyttävät järjestelmät tulivat käyttöön laajemmin vasta 1990-luvulla. [3], [8]

Tämän hetken tekstiä generoivista tekoälymalleista suosituimpia ovat neuroverkkomalleihin *(engl. Artificial Neural Network, ANN)* perustuvat suuret kielimallit.

Suuria kielimalleja voidaan hyödyntää moniin erilaisiin tekstiin liittyviin tehtäviin, kuten esimerkiksi asiakirjojen tiivistelmien luomiseen, ohjelmakoodin kirjoittamiseen, kielten kääntämiseen, verkon selailussa avustimena, tekstitysten laatimisessa, geneettisten sekvenssien analysoinnissa, terveydenhuollossa potilastietojärjestelmissä potilastietojen tulkintaan, hoito-ohjeiden laatimiseen, biolääketieteessä analysoimaan ja jopa simuloimaan aineistoja. Suuret kielimallit ovat kyvyiltään poikkeuksellisen monipuolisia ja joustavia, mikä mahdollistaa niiden laajat käyttömahdollisuudet monilla eri aloilla. [12], [15]

Tavallisesti nämä generatiivisen tekoälyn suuret kielimallit ovat erinomaisia käsittelemään tekstidataa, sillä ne ovat usein koulutettu pelkästään teksipohjaisella koulutusaineistolla, mikä tekee niistä erityisen soveltuvia tekstinkäsittelyyn. Tämän tyypin suuret kielimallit omaavat siis puutteita muita tietotyyppejä kohdatessaan. Pelkästään tekstipohjaiset mallit kuten esimerkiksi GPT-3.5 ja BERT suoriutuvat erinomaisesti tekstiin liittyvistä tehtävistä, joita ovat muun muassa tekstin ja ohjelmakoodien luominen, mutta niiltä puuttuu kyky ymmärtää ja käsitellä muita syötettävien tietotyyppien muotoja.[16], [17]

Tämän kohdatun ongelman ratkaisivat uusimmat multimodaaliset suuret kielimallit *(engl. multimodal large language model)*, jotka ovat koulutettu monipuolisilla koulutusaineistoilla. Niiden koulutusaineistoihin sisältyy esimerkiksi tekstiä, kuvia, ääntä ja videoita, mikä mahdollistaa niiden kyvyn käsitellä ja yhdistellä erilaista tietoa syötteissään ja tuotoksissaan. Tämä kehitysaskel pelkkien tekstimallien rajoitusten ylittämiseksi on tuonut mahdollisuuksia erilaisten tietotyyppien käsittelyyn suurilla kielimalleilla, joista tämän hetken tunnetuin on OpenAI:n GPT-4o *(lyh. Generative Pre-trained Transformer 4 Omni)*, joka ymmärtää kuvia, tekstiä, ääntä ja erilaisia tiedostomuotoja. Multimodaalisille kielimalleille on monenlaisia käyttökohteita, ja ne pystyvät tunnistamaan kuvista esimerkiksi eläin- ja kasvilajeja, varoitusmerkkejä, logoja, tekstiä ja jopa kemiallisia yhdisteitä, mistä kielimalli voi

tehtävästään riippuen selostaa kuvassa olevaa sisältöä tai sen tarkoitusta. Kuvan perusteella kielimalli voi vastata käyttäjän laatimiin kysymyksiin ja tehtäviin, esimerkiksi tunnistamaan kasvin ja kertomaan onko se myrkyllinen, onko esimerkiksi laitteeseen kiinnitetty johto oikein kiinnitetty sekä rajoitetulla kyvykkyydellä myös tulkitsemaan erilaisia lääketieteeseen liittyviä kuvia. [18]–[20]

2.3.2 Tekstin yhteenvedon laatiminen

Yhteenvetojen laatiminen *(engl. text summarization)* malleille syötetyistä teksteistä tai dokumenteista on suurten kielimallien yksi tärkeimmistä tehtävistä NLP:ssä, ja se kuuluu luonnollisen kielen ymmärtämisen *(engl. Natural Language Understanding, NLU)* tehtävien joukkoon [8]. Tekstin ja dokumenttien yhteenvedon laatiminen voi tapahtua eri tavoin: abstraktiivisesti, ekstraktiivisesti tai niiden yhdistelmänä. Abstraktiivisessa yhteenvedossa malli yrittää tuottaa uudelleen sanoitetun, ihmismäisesti kirjoitetun tekstitiivistelmän, joka sisältää alkuperäisen tekstin tärkeimmät pääkohdat. Ekstraktiivisessa yhteenvedossa malli sen sijaan valitsee ja yhdistää suoraan alkuperäisten tekstin osia yhteenvedoksi ja tekstiä ei muotoilla uudelleen. Hybridimenetelmässä tavoitteena on yhdistää molempien menetelmien parhaat puolet, missä ekstraktiivisella menetelmällä tunnistetaan tärkeimmät pääkohdat tekstistä ja abstraktiivisen menetelmän avulla luodaan paremmin luettavissa oleva ja selkeä uudelleen sanoitettu pääkohtien mukainen tiivistelmäteksti. [8], [21], [22]

Abstraktiivinen tekstin yhteenveto on ekstraktiivista vaativampi menetelmä, mutta samalla se voi tuottaa lukukelpoisempia ja laadukkaampia tekstien tiivistelmiä. Esimerkiksi Googlen kielimalleista T5 *(lyh. Text-To-Text Transfer Transformer)* sekä BART *(lyh. Bidirectional and Auto-Regressive Transformers)* ja OpenAI:n viimeisimmät GPT-mallit pystyvät kaikki tuottamaan abstraktiivisia yhteenvetotekstejä syötteistään. Näitä malleja voidaan soveltaa laajasti erilaisien tiivistelmien luontiin, kuten esimerkiksi uutisartikkeleiden tiivistämiseen, erilaisten raporttien ja dokumen-

taatioiden laatimiseen, muistiinpanojen tai tieteellisten julkaisujen tuloksien yhteenvedon kirjaamiseen tai esimerkiksi tämän tutkielman käyttötarkoituksessa asiakaspalvelun chattien tapahtumien tiivistämiseen erilaisiksi arvioiksi. [22], [23]

2.3.3 Tekstin tunnetilojen analyysi

Tekstissä esiintyvien tunnetilojen analyysi *(engl. sentiment analysis)* on myös yksi merkittävä ja keskeinen tehtävä NLP:ssä, ja se kuuluu tekstin yhteenvetojen laatimisen kanssa myös NLU-tehtävien joukkoon [8]. Tunnetilojen analyysin tarkoituksena on tunnistaa ja tulkita tekstin asiayhteydestä, sanoista ja niiden muotoiluista ilmeneviä erilaisia tunteita, asenteita sekä merkityksiä. Kiinnostuksen, innostuneisuuden, ärtyneisyyden, tyytyväisyyden tai tyytymättömyyden tunnistaminen voi tilanteista riippuen olla hyödyllistä, sillä sen avulla esimerkiksi yritykset voivat saada arvokasta tietoa asiakaspalautteista ja tuotearvioista. Tunnetilojen analyysi tukee muun muassa päätöksentekoa ja asiakaspalvelun kehittämistä tarjoamalla lisätietoa ja ymmärrystä asiakkaiden kokemuksista, mahdollisista ongelmista, onnistumisista ja kehityksen kohteista asiakaspalvelussa.[24]

OpenAI:n GPT-mallit sekä muut suuret kielimallit ovat osoittautuneet tutkimuksissa tehokkaiksi erilaisissa tunnetilojen analysointiin liittyvissä tehtävissä. Mallit pystyvät yhdistämään syötteksteissä esiintyvistä asioista, termeistä ja sanoista erilaisia näkökulmia, jonka avulla ne voivat tunnistaa ja luokitella teksteissä käytettyjä sanavalintoja tai kirjoitustyylejä erilaisiin niitä vastaaviin tunnetiloihin. Näitä malleja voidaankin siksi hyvin soveltaa asiakaspalautteiden luokitteluun, asiakkaiden tyytyväisyyden tulkintaan, yhteydenottojen aiheiden tunnistamiseen ja myös asiakaspalvelijan ystävällisyyden sekä palvelukeskeisyyden arviointiin. [25], [26]

2.3.4 Aiheiden mallinnus tekstistä

Perinteisesti aiheiden mallintamiseen tekstistä on käytetty erilaisia menetelmiä, kuten esimerkiksi avainsanojen luokitteluun perustuvia LDA-malleja *(lyh. Latent Dirichlet Allocation)* ja avainsanojen löytämiseen perustuvia LSA-malleja *(lyh. Latent Semantic Analysis)*. Näillä menetelmillä voi kuitenkin usein olla haasteita ymmärtää koko tekstin ydinaihetta ja ne voivat sekoittaa välillä aiheita keskenään. Suuret kielimallit eivät kohtaa näitä ongelmia yhtä usein, sillä ne pystyvät yleensä tunnistamaan ja tulkitsemaan tekstin aiheen ja sävyt dynaamisesti koko syötteen asiayhteydestä, vaikka ne voivatkin joskus hallusinoida vaikeasti tulkittavien tekstien ja kielten kohdalla. Tutkimuksessa [27] havaittiin aihemallinnuksen onnistuvan erinomaisesti GPT-4-mallilla, ohittaen tuloksissa esimerkiksi tyypillisen LDA-aihemallin. Tämän lisäksi GPT-4 tunnisti aiheita, jotka olivat merkittävästi paremmin linjassa ihmisten arvioimien aihetunnisteiden kanssa verrattuna muihin aihemalleihin [27].

Toisessa tutkimuksessa [28] havaittiin myös, että suurilla kielimalleilla todettiin olevan erinomaiset kyvyt tunnistaa ja valita syötetekstistä oikeita aiheita. Tämä kuitenkin vaatii kielimalleille syötettävien ohjekehotteiden tarkan tehtävää koskevan muotoilun ja sen, että valittavat aiheet on rajattu tiettyä tarkoitusta varten.

2.4 Transformer-arkkitehtuuri

Transformer-arkkitehtuuria edeltäneet perinteiset RNN- ja LSTM-arkkitehtuurit käyttävät sisäistä hetkellistä tilaa ajallisten askelien yli eli ajallista muistia edellisistä tiloista. Kyseisen muistin avulla ne voivat käsitellä väliaikaisia riippuvaisuuksia syötesekvenssien eri vaiheiden välillä niin koulutusaineistoistaan kuin tuotoksissaan. Tämä on erityisen tärkeää esimerkiksi tekstiin liittyvien tehtävien kanssa, koska sanojen järjestys ja esiintyminen lauseissa ja virkkeissä on aina riippuvainen edellisten sanojen, lauseiden ja virkkeiden asiayhteyksistä. Takaisinkytketyt neuroverkot kou-

lutetaan useimmiten vastavirta-algoritmeja *(engl. backpropagation)* ja gradienttimenetelmiä *(engl. gradient descent)* käyttäen, missä tavoitteena mallia kouluttaessa on minimoida määritetty virhe- tai hukkafunktion tulos. Virhefunktion tulos voidaan minimoida laskemalla virhefunktion gradientin arvoja painokertoimien suhteen kerros kerrokselta neuroverkossa ja siten säätämällä painokertoimia virheen minimoimiseksi. [29]

Tämä kuitenkin aiheuttaa katoavan gradientin *(engl. vanishing gradient)* ongelman, missä vastavirta-algoritmin aikana gradientit päätyvät olemaan arvoiltaan katoavan pieniä. Erityisesti pitkien sekvenssien kohdalla voi käydä niin, että gradienttien arvot voivat kutistua lähes olemattoman pieniksi, mikä vaikeuttaa pitkäaikaisten riippuvuuksien oppimista ja käsittelyä, eli malli ei opi tai ymmärrä kontekstin kannalta merkityksellisiä asioita etäällään olevista sekvenssien osista. [30], [31]

Vuonna 2017 Googlen tutkijat (Vaswani et al., 2017) esittelivät uuden Transformer-arkkitehtuurin, joka ei hyödynnä lainkaan perinteisiä toistuvia takaisinkytkettyjä neuroverkkorakenteita. Kyseinen arkkitehtuuri luottaa kokonaan huomiomekanismeihin muodostaakseen laajoja yhteyksiä mallin syötteiden osien välillä riippumatta sekvenssien pituudesta toisin kuin edeltäneet kielimalliarkkitehtuurit. [9]

Transformer-malleihin perustuvat kielimallit ovat mullistaneet NLP-alan tarjoamalla entistä tehokkaamman tavan käsitellä laajoja tekstien välisiä riippuvuuksia huomiomekanismien *(engl. attention mechanism)* avulla. Vaikka Transformer-mallien edeltäjät, kuten RNN- ja LSTM-mallit, pystyivät myös esittämään ja ymmärtämään sanoja vektorimuodoissa, niiden kyky säilyttää sanojen laajempi konteksti heikkeni tekstin pidentyessä. Huomiomekanismien avulla Transformer-mallit voivat tehokkaasti tarkastella kaikkia tekstin osia kerralla, mahdollistaen näin paremman tekstin kontekstin ymmärryksen ja tehokkaamman mallien kouluttamisen ja käytön grafiikkasuorittimien avulla. [8]–[10]

Transformer-arkkitehtuurin kielimallit voivat olla kaksiosaisia enkooderi-dekooderi-

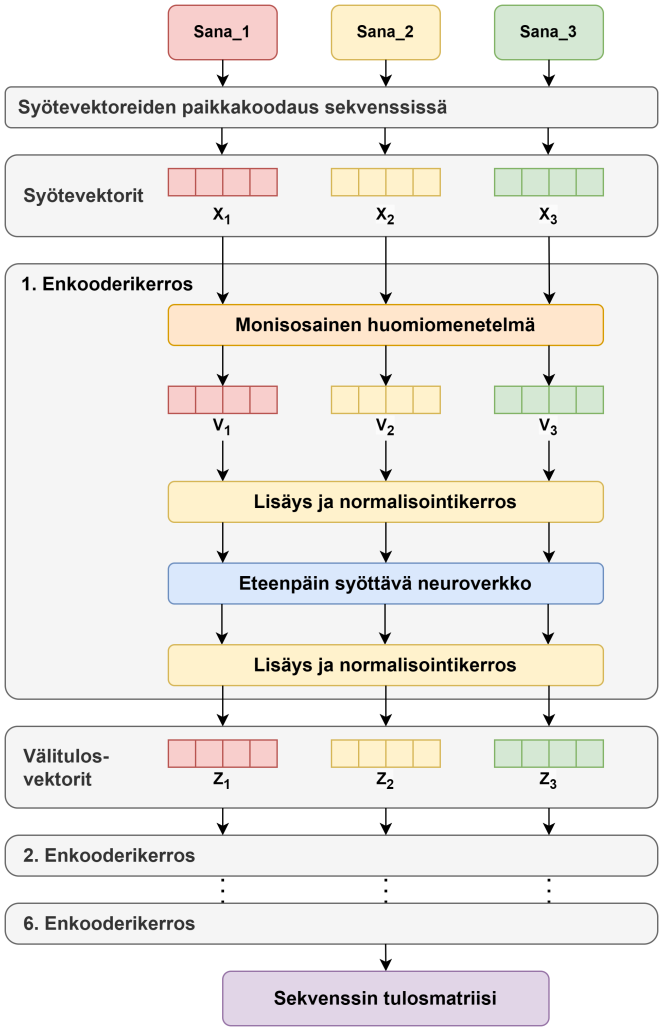
mallirakenteen avulla *(engl. encoder-decoder)*, mutta kaikki kyseiseen arkkitehtuuriin perustuvat mallit eivät käytä molempia osia. Esimerkiksi Googlen BERT-malli käyttää vain arkkitehtuurin enkooderipuoliskoa, kun taas OpenAI:n GPT-mallit hyödyntävät puolestaan vain dekooderipuoliskoa. Vuonna 2017 esitelty alkuperäinen mallin arkkitehtuuri sisälsi kuusi identtistä päällekkäistä kerrosta enkooderin ja dekooderin puolella, mutta nykyään kerrosten määrä voi vaihdella huomattavasti riippuen mallin käyttötarkoituksesta ja rakenteesta.[9]

Transformer-arkkitehtuurin rakenteessa ei siis ole takaisinkytkentää, mikä on sen yksi suurimmista eduista. Rakenne mahdollistaa tehokkaan rinnakkaislaskennan, koska sen neuroverkkokerrokset pystyvät prosessoimaan kaikki syötesekvenssin sanojen tokenit samaan aikaan matriisina, koska niiden laskeminen ei vaadi tiettyä laskujärjestystä. Tiedot syötesekvenssin järjestyksestä on jo koodattu valmiiksi erikseen matriisin vektoreihin. Tämä rinnakkaisen laskennan kyky mahdollistaa sen, että mallit pystyvät käsittelemään syötteenään tekstidataa paljon enemmän ja nopeammin kuin vanhempien arkkitehtuurien kielimallit. [9], [10]

2.4.1 Enkooderi

Enkooderin tehtävänä ja tarkoituksena on vektorisoida syötettävä sekvenssi eli tekstin sanat ja tavut erillisiksi vektoreikseen, jotta jokainen sana saa omanlaisensa kontekstiin liitetyn numeerisen arvonsa vektorin muodossa. Vektorin arvot kuvaavat jokaisen tokenin kohdalla kielellistä merkitystä sekä sen sijaintikohtaista arvoa suhteessa muihin sekvenssin tokeneihin. Enkooderin vastaus syötteeseen sisältää siis täsmälleen yhtä monta vektoria kuin syötteessä on ollut tokeneita tai sanoja, kuten esimerkkikuvassa 2.2 on havaittavissa. Itse jokaisen vektorin pituus ja ulottuvuuksien määrä riippuu puolestaan mallin arkkitehtuurista, esimerkiksi Googlen vuonna 2018 julkaistussa BERT-enkooderimallissa *(lyh. Bidirectional Encoder Representations from Transformers)* vektorin ulottuvuuksia on 768. [9], [16]

Enkooderimallia kouluttaessa malli oppii muodostamaan koulutusaineiston eli korpuksen avulla moniulotteisen vektoriavaruuden, missä kielellisesti ja merkityksellisesti samankaltaiset sanat sijoittuvat toistensa lähelle matemaattisesti sanoja vastaavien vektoreiden numeroarvojen perusteella. Kun enkooderimalli saa syötteenään lauseen, se muuntaa lauseen sanojen tokenit vektorimuotoon. Vektorimuotoisia tokeneiden esityksiä voidaan näin verrata mallin oppiman vektoriavaruuden sanojen vektoriesityksiin, jolloin malli voi analysoida syötettävien sanojen merkityksiä ja kontekseja ja siten muodostaa lopullisia sekvenssin vektoriesityksiä vastauksena.



Kuva 2.2: Enkooderin toimintaa

Enkooderin yhtenä ominaisuutena on kontekstin tiedon kaksisuuntaisuus, jossa syötettävän matriisin eli sekvenssin vektorit sisältävät tietoa sekvenssin kokonaisvaltaisesta kontekstista. Syötesekvenssin vektoreita laskettaessa vektoreiden sisältämät numeeriset arvot lasketaan itse niitä vastaavien sanojen raa'an kirjaimellisen arvon kuin myös niiden välisten etäisyyksien mukaan sekvenssistä, jolloin vektorit sisältävät sanojen välisiä sijainti-informaatioita. Jokainen sana alkuperäisessä syötelauseessa vaikuttaa kaikkiin sekvenssissä olevien vektorien muodostumiseen ja niiden numeerisiin arvoihin eli jokainen vektori saa numeeriset arvonsa ympäröiviltä sanoilta, mikä määrittää kyseisen ja ympäröivien sanojen kontekstin. Vektorimuodossaan tietty syötelauseen sana ei siis oikeastaan ole enää esitys itse sanasta, vaan itse sanan merkityksestä ja sen esiintyvyyden todennäköisyydestä ja samankaltaisuudesta ympäröiviin sanoihin ja niiden kontekstiin. [9]

Itsehuomio *(engl. self-attention, intra-attention)* on puolestaan enkooderin huomiomekanismi, joka yhdistää sekvenssin eri kohdat laskelmoiduilla vakioarvoilla vektoreissa koko sekvenssin esityksen tuloksen laskemiseksi. Tämän ominaisuuden ansiosta enkooderit ovat erittäin tehokkaita poimimaan syötteistään merkityksellistä tietoa, mitä voidaan käyttää erilaisissa tehtävissä. Näistä hyviä esimerkkejä ovat esimerkiksi tekstin luokittelu esimerkiksi arvosanoin tai termein, kysymykseen vastaaminen kontekstiin liittyen sekä naamioidun kielen mallintaminen *(engl. Masked Language Modeling, MLM)*, jossa osa sanoista korvataan merkeillä, joiden tilalle mallin tulee yrittää ennustaa alkuperäiset sanat syötteen kokonaisvaltaisen kontekstin perusteella.[9], [16]

2.4.2 Dekooderi

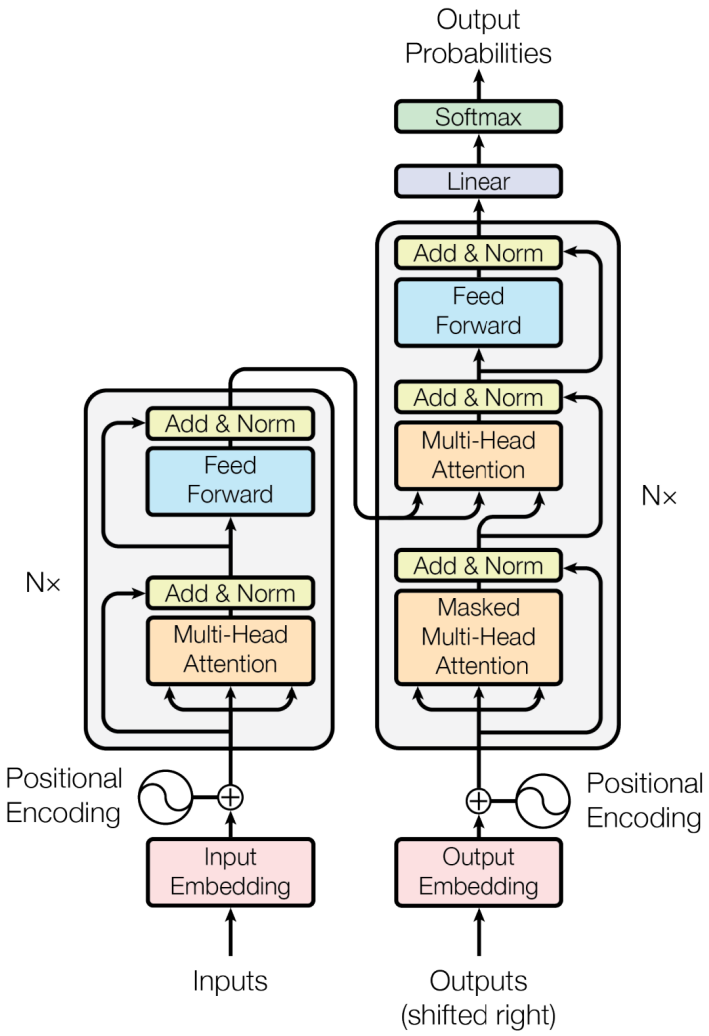
Dekooderit eroavat hieman enkoodereista sekä rakenteeltaan että toimintaperiaatteeltaan. Dekooderi-mallit vektorisoivat niille syötettävät sekvenssit muuttamalla sen tokenit erillisiksi vektoreiksi niin kuin enkooderitkin. Dekooderi-mallit eroavat

enkoodereista siinä, etteivät ne ole vektoreiden välisiltä konteksteiltaan kaksisuuntaisia vaan yksisuuntaisia. Tämä tarkoittaa, että dekooderit voivat ottaa huomioon vain niitä edeltäneet syötteen tokeneiden vektorit ennustaessaan seuraavaa uutta tokenia, eli ne ovat auto-regressiivisiä malleja ennustaessaan tulevia arvoja edeltävien arvojen perusteella. Dekoodereiden tapauksessa vain edeltäviä arvoja huomioiva ominaisuus sisältyy myös naamioidun itsehuomion *(engl. masked self-attention)* menetelmän toimintaperiaatteeseen, jonka avulla dekooderimallit ovat erinomaisia generoimaan tekstiä sana kerrallaan, muodostaen järkeviä sanoja, lauseita ja lopulta virkkeitä tiettyyn syötteen kontekstiin liittyen. [9]

Tämän ominaisuuden ansiosta dekooderimallit soveltuvat hyvin tehtäviin, missä tekstin syy-seuraussuhteen ymmärtäminen on kontekstin lisäksi tärkeää. Näitä tehtäviä ovat esimerkiksi lauseiden kirjoittaminen, yhteenvedon laatiminen tekstistä, keskusteluiden ymmärtäminen, tekstin tai ohjelmakoodin täydentäminen, vastaaminen asiakkaiden kysymyksiin tai jopa potilastietojen analysointi. Tämän hetken tunnetuimpia ja käytetyimpiää dekooderimalleja ovat todennäköisimmin OpenAI:n GPT-3.5 ja sen suorat seuraajat GPT-4 ja GPT-4o, joita käytetään muun muassa OpenAI:n omassa ChatGPT-palvelussa. [8], [10], [32]

Nämä enkooderi- ja dekooderi-mallien ominaisuudet ovat kriittisiä ja ratkaisevan tärkeitä itse Transformer-mallien suorituskyvyn ja toiminnan kannalta useissa erilaisissa tehtävissä, kuten esimerkiksi kielten kääntämisessä, yhteenvedon laatimisessa tekstistä sekä tunteiden ja tilanteiden analysoinnissa. Kaikissa näissä tehtävissä mallin on osattava oppimansa perusteella ymmärtää sanojen, lauseiden, virkkeiden ja niiden asiayhteyksiä keskenään todennäköisyyksiin perustuen niin syötteessä kuin myös vastauksessa, minkä malli itse tuottaa tekstinä takaisin.

2.4.3 Enkooderi ja dekooderi yhdistettynä



Kuva 2.3: Transformer-arkkitehtuuri yksinkertaistettuna (Vaswani et al., 2017)

Transformer-arkkitehtuurin mallit koostuvat siis tyypillisesti joko enkooderista, dekooderista tai näiden yhdistelmästä mallin käyttökohteesta riippuen [33]. Enkooderin puolisko koostuu useasta eri kerroksesta, jotka voidaan jakaa vielä alikerroksiin, jotka ovat itsehuomio- ja eteenpäin syöttävä neuroverkkokerros. Enkooderin kerrokset ovat identtisiä keskenään, mutta niiden parametrien painoarvot ovat erisuuruisia. Aluksi enkooderin puolella syötesekvenssistä muodostetaan vektoreita, missä jokaisesta sekvenssin sanojen tokeneista luodaan numeerinen esitys eli syötevekto-

rin esitys *(engl. input embedding)*. Syötevektoreihin lasketaan lisäksi niiden välinen sijainti-informaatio paikkakoodattuna vektorina *(engl. positional encoding)*, jonka avulla syötesekvenssin sanojen väliset kontekstit ja suhteelliset etäisyydet säilyvät numeerisesti vektorissa ja ovat arvoiltaan ainutlaatuisia. [9], [33]

Transformer-arkkitehtuurissa huomiomekanismit on toteutettu monipäisen huomiomekanismin *(engl. Multi-Head Attention)* kerroksilla. Huomiokerrokset ovat koko Transformer-mallien arkkitehtuurin kriittisin osa. Näiden kerroksien tehtävänä on määrittää, mitkä syötematriisin sisältämistä vektoreista ovat kontekstiltaan huomion arvoisia. Monipäisessä huomiossa syötevektori jaetaan monelle huomiopäälle, joista jokainen tarkastelee syötteen osia erilaisista näkökulmista. Näin huomiomekanismi pystyy keskittymään syötteen moneen osaan ja muodostamaan paremman käsityksen syötteen osien tärkeydestä riippuen eri näkökulmista. Lopuksi nämä käsitykset yhdistetään, jolloin malli saa muodostettua laajemman käsityksen kokonaisuudessaan. [9]

Enkooderin ja dekooderin kerroksissa monipäistä huomiomekanismia käytetään skaalatun pistetulon huomiona *(engl. Scaled Dot-Product Attention)*, jota voidaan kutsua myös itsehuomioksi. Kun syötematriisi kulkee itsehuomio-kerroksen läpi, se auttaa enkooderia tarkastelemaan sekvenssin sanojen olennaisia osia keskenään samalla kun se valitsee tärkeimpiä ydinsanoja sekvenssistä. Itsehuomiossa lasketaan kolme vektoria jokaisesta syötematriisin vektorista. Jokaiselle vektorille vuorostaan lasketaan erilliset kyselyvektori Q, avainvektori K ja arvovektori V. Nämä vektorit lasketaan kertomalla alkuperäinen vektori kolmella erilaisella enkooderin kerroksen i omilla parametrimatriiseilla eli painomatriiseilla W_i^Q, W_i^K ja W_i^V, jotka on laskettu ja muodostunut mallia kouluttaessa. [9], [33]

Kertomalla jokainen syötevektori W_i^Q-painomatriisilla saadaan kyselyvektori Q. Samalla tavalla lasketaan myös avainvektorit K ja arvovektorit V kertomalla ne vastavilla painomatriiseilla W_i^K ja W_i^V. Nämä vektorit auttavat mallia sisäistämään

syötelauseen eri sanojen välisiä suhteita. Laskussa 2.1 kyselyvektoreiden Q ja avainvektoreiden K välinen pistetulo tuottaa huomiomatriisin pistearvot. Nämä arvot määrittävät, kuinka paljon sana ja sen tokenit voivat vaikuttaa kyselyvektorien annetussa esityksessä. Laskussa pistetulon arvot jaetaan avainvektorien $(K_1, K_2, ..., K_n)$ pituutta vastaavalla arvolla d_k. *Softmax*-funktiolla lasketun todennäköisyysjakauman arvot rajoittuvat lukujen 0 ja 1 välille. Softmax-funktion kautta lasketuilla arvoilla saadaan painotettua tiettyjen tokenien painoarvoja, missä korkeampi arvo tarkoittaa suurempaa painoarvoa. [9], [33]

$$Huomiomatriisi(Q, K, V) = softmax(\frac{Q_i K_i^T}{\sqrt{d_k}})V_i \tag{2.1}$$

Kun arvovektorit $(V_1, V_2, ..., V_n)$ kerrotaan muodostuneen todennäköisyysjakauman arvoilla, tavoitteena on säilyttää ennallaan tai korostaa vain niiden tokenien arvot, joilla on selkeästi eniten merkitystä ja painoarvoa sekä samalla poistaa merkityksettömät tokenit todella pienten kertoimien arvojen avulla. Lopulta muodostuneiden huomioarvojen avulla lasketaan arvovektoreiden $(V_1, V_2, ..., V_n)$ painotettu summa, joka tuottaa huomiokerroksen lopullisen tulosvektorin. [9], [33]

Muodostunut lopputulos jatkaa kerroksessa eteenpäin syöttävälle neuroverkolle. Eteenpäin syöttävässä neuroverkossa jokainen tulosvektori käy läpi useita neuroverkon kerroksia, joihin sisältyy epälineaarisia aktivointifunktioita, kuten esimerkiksi ReLU *(engl. Rectified Linear Unit)*. Aktivointifunktioiden kautta malli oppii monimutkaisia epälineaarisia riippuvuuksia syötevektoreiden arvoista, samalla kun muodostuneen huomiomatriisin painotukset ohjaavat mallia keskittymään syötteen olennaisimpiin osiin. Nämä neuroverkot ovat jokaisessa huomiokerroksessa rakenteeltaan identtisiä ja jokainen huomiokerroksen matriisin vektori syötetään yksilöllisesti neuroverkolle joka kerroksessa. Tässä vaiheessa neuroverkkojen tarkoituksena on oppia monimutkaisia suhteita sanojen, lauseiden, muotoilun ja niiden kontekstien välillä, mitä pelkillä huomiomekanismeilla ei pystytä suoraan mallintamaan. [9], [33]

Googlen tutkijoiden julkaisun mukaan (Vaswani et al., 2017) yhden ainoan huomiomekanismin käyttöön liittyy kuitenkin rajoituksia. Yksittäinen huomiomekanismin huomiopää *(engl. attention head)* on rajoittunut keskittymään vain yhteen tietynlaiseen näkökulmaan matriisin vektoreista, eikä se välttämättä pysty havaitsemaan syötteestä monimutkaisia yhteyksiä. Transformer-arkkitehtuurin huomiomekanismin kyvykkyyttä voidaan kuitenkin tehostaa monipäisellä huomiolla *(engl. Multi-Head Attention)*. Monipäisellä huomiolla tarkoitetaan tapaa, missä monta eri huomiopäätä toimivat keskenään samanaikaisesti rinnakkain. Monipäisessä huomiossa kysely-, avain- ja arvovektorien (eli Q, K ja V) arvoista projisoidaan lineaarisesti matalampien aliulottuvuuksien projektioita d_k, d_k ja d_v yhteensä h-kertaa, esimerkiksi $h = 8$ [9]. Nämä erilaiset projektiot antavat huomiomekanismille erilaisia näkökulmia syötedataan vektoriavaruudessa, jolloin myös huomion näkökulma voi kohdistua eri osiin syötteen vektoriesitystä. Kun monipäiselle huomiomekanismille syötetään aliulottuvuuksien projektioita, jokainen niistä korostaa eri osia datasta rinnakkain, jolloin huomio kiinnittyy niin yksittäisiin kuin rinnakkaisiin esityksiin datasta. Lopputuloksena on d_v-ulottuvuuden tulosarvoja, jotka yhdistetään vielä kerran projisoimalla ne vielä uudelleen yhdeksi lopulliseksi tulokseksi, joka voidaan syöttää eteenpäin arkkitehtuurin rakenteen neuroverkolle. [33]

Transformerin enkooderin kerroksien tapaan dekooderin puoliskon kerrokset koostuvat huomiokerroksista ja eteenpäin syöttävistä neuroverkoista. Dekooderin kerrosten monipäisen huomiokerroksen ja neuroverkkokerroksen välissä on kuitenkin erityinen dekooderin ja enkooderin yhdistävä huomiokerros. Tässä erillisessä monipäisessä huomiokerroksessa dekooderi saa enkooderilta lopullisen tulosmatriisin, mikä auttaa dekooderia keskittymään tiettyihin alkuperäisen syötesekvenssin painoitettuihin sanoihin ja lauseyhteyksiin enkooderin huomion laskelmien perusteella. Näin dekooderi ei ainoastaan käytä enkooderilta saatuja piirrevektoreita, vaan yhdistää ne jo luotuihin sanoihin ja sekvensseihin käsiteltäväksi kerroksissaan. Tämä

varmistaa, että dekooderin tuottamat sekvenssit ovat johdonmukaisia ja kontekstuaalisesti tarkkoja alkuperäisten syötteiden kanssa. Tämä rakenne on erityisen tärkeä esimerkiksi tekstin kääntämisessä toiselle kielelle, missä sanojen ja lauseiden osien merkityksien säilyttäminen ennallaan on käännöstehtävän kannalta tärkeää. [9]

Transformer-arkkitehtuurin rakenteessa ei siis ole takaisinkytkentää, mikä on sen yksi suurimmista eduista. Rakenne mahdollistaa tehokkaan rinnakkaislaskennan, koska sen neuroverkkokerrokset pystyvät prosessoimaan kaikki syötesekvenssin sanat samaan aikaan matriisina, koska niiden laskeminen ei vaadi tiettyä laskujärjestystä. Tiedot syötesekvenssin järjestyksestä on jo koodattu valmiiksi erikseen matriisin vektoreihin. Tämä rinnakkaisen laskennan kyky mahdollistaa sen, että mallit pystyvät käsittelemään syötteenään tekstidataa paljon enemmän ja nopeammin kuin vanhempien arkkitehtuurien kielimallit. [9], [10]

2.4.4 GPT-mallit

OpenAI:n GPT-kielimallit ovat yleisiä perusmalleja *(engl. foundation model)*. Tällä määritelmällä tarkoitetaan todella suuren parametrimäärän omaavia kielimalleja, jotka on koulutettu valtavan laajalla ja monipuolisella koulutusaineistoilla niin, että niitä voidaan käyttää sellaisenaan moniin erilaisiin tehtäviin ilman jatkokoulutusta. GPT-mallit *(lyh. Generative Pre-trained Transformer)* ovat nimensä mukaisesti esikoulutettuja Transformer-arkkitehtuuriin perustuvia dekooderimalleja. [17], [19]

GPT-mallien koulutusprosessi koostuu kahdesta päävaiheesta: valvomattomasta esikoulutuksesta ja valvotusta hienosäädöstä. Ensimmäisen vaiheen valvomattomassa esikoulutuksessa koulutusaineistona käytetään suuria määriä tekstiä, jota ei ole tarkoitettu mitään erityistä tehtävää varten. Tämän vaiheen tarkoituksena on muodostaa kyvykäs perusmalli, joka kykenee ymmärtämään koulutusaineistossa olevia erilaisia kieliä ja aiheita. [17]

OpenAI:n GPT-perusmallit koulutetaan ennustamaan lauseiden seuraavia sanoja niiden kontekstiin liittyen. Koulutuksessa käytettään usein julkisesti saatavilla olevia sekä erikseen lisensoituja aineistoja. Tämä laaja ja monipuolinen koulutusaineisto eli korpus sisältää esimerkiksi oikeita ja vääriä ratkaisuja erilaisiin ongelmiin, eritasoista ajatuksenjuoksua ja päättelyä, ristiriitaisia ja johdonmukaisia väitteitä sekä laajan kirjon erilaisia asiatekstejä, ideologioita, näkökulmia ja historiaa erilaisiin asioihin liittyen. Tavoitteena on kattaa mahdollisimman monia erilaisia aiheita, tyylejä, näkökulmia ja luoda mallille kyky soveltaa kaikkea oppimaansa mahdollisimman monipuolisesti ja kyvykkäästi.[17], [34]

Koulutusaineiston tekstikorpus tokenisoidaan malleja vanhemmalla erillisellä menetelmällä nimeltään *Byte Pair Encoding (lyh. BPE)*, joka on kehitetty alun perin 1990-luvulla [35]. BPE mahdollistaa tehokkaan tekstin käsittelyn ja opitun sanastoavaruuden luomisen. Menetelmä auttaa vähentämään sanaston kokoa yhdistämällä yleisimpiä tokeneita vähitellen pidemmiksi jonoiksi tokeneita, jolloin harvinaisia tai tuntemattomia sanoissa olevia merkkijonoja voidaan käsitellä tehokkaammin jakamalla ne pieniin osiin ja vastaamaan yleisimmin esiintyviä tokeneita jos se vain on mahdollista. Yleisimmät tokenit saavat siis omat numeeriset esityksensä vektorimuodossa, jotka lopulta muodostavat mallin sanastoavaruuden. [17], [36]

Koulutuksien aikana GPT-mallit käyvät läpi lukuisia eri iteraatioita, joiden kautta niiden parametreja ja neuroverkkojen painokertoimia lasketaan uudelleen. Mallien parametreja säädetään ja optimoidaan erilaisin optimointimenetelmin, kuten esimerkiksi stokastisten gradienttien laskemisella parametreille. Ideana on käyttää syötteenä pieniä satunnaisia koulutusaineiston eriä kerrallaan ja näin ohjata parametreja oikeaan suuntaan pieni päivitysaskel kerrallaan sen sijaan, että malleille yritettäisiin kouluttaa koko aineisto kerralla, mikä ei ole kannattavaa. Tämä vaihe lopetetaan kun mallien hukkafunktio on saavuttanut konvergenssin, eli kun ennustevirhettä mittaava hukkafunktio ei enää laske huomattavasti ja mallien suorituskyky

ei parane vaikka koulutusta yritettäisiin jatkaa koulutusaineistolla. [17], [19]

Mallien koulutuksen toisessa vaiheessa eli valvotussa hienosäädössä esikoulutettuja malleja hienosäädetään uudella koulutusaineistolla käyttämällä vahvistusoppimista ihmisen palautteen avulla *(engl. Reinforcement Learning from Human Feedback, RLHF)*. Menetelmässä mallien vastauksia annetaan erikseen koulutetulle palkintomallille, joka edustaa ihmismäisiä mieltymyksiä oikeista ja vääristä vastauksista. Koulutettavaa mallia palkitaan pistein, minkä kautta malli oppii parantamaan käytöstään ja ennusteitaan iteratiivisesti. Prosessin jälkeen malli onnistuu tuottamaan parempia ja inhimillisempiä vastauksia, jotka ovat yhdenmukaisempia ihmisten odotusten kanssa. [17], [34], [37]

GPT-mallien generoimat tokenit ja niistä muodostuvat sanat perustuvat todennäköisyyksien sarjaan. Mallien oppimat asiat sekä tuotokset voivat sisältää vääristymiä ja puolueellisuuksia, joita voi esiintyä toistuvasti niiden valtavissa koulutusaineistoissa. Lisäksi niiden vastaukset voivat sisältää väitteitä, jotka eivät ole lainkaan tosiasioihin perustuvia. Näitä virheitä kutsutaan ilmiöinä hallusinaatioiksi, jotka syntyvät, kun malli ennustaa uusien generoitavien tokenien todennäköisyydet väärin. Tämä puolestaan johtaa siihen, että myös kaikilla niitä seuraavilla tokeneilla on silloin myös suurempi riski olla virheellisiä ja hallusinointi ei lopu. [38]

Hallusinointi voi lopulta johtaa outoihin, epätarkkoihin tai virheellisiin mallin generoimiin vastauksiin, joissa sanat tai aihe eivät ole enään järkeviä. Tämä ilmiö johtuu suoraan mallien tokeneiden todennäköisyysjakaumiin perustuvasta ennustamistavasta. Pahimmassa tapauksessa esimerkiksi vaikeiden kielien ja numeroiden kohdalla selkeä johdonmukaisuus voi kokonaan puuttua mallien tuotoksista. Vaikeiden aiheiden kohdalla malleilla on myös taipumus alkaa toistaa itseään sitä enemmän mitä pidemmäksi vastauksen pituus kasvaa. [39]

Näiden heikkouksien vuoksi esimerkiksi monet mediassa esitetyistä vaikuttavista tehtäväkohtaisisa tuloksista voivat usein olla tarkan valinnan ja toistuvan kokeilun

tuloksia, missä valitaan vain parhaimmat tulokset. Tietty valittu tulos saadaan suorittamalla kielimallia samalla syötteellä tai sen uudelleenmuotoilulla versiolla muutaman kerran, minkä jälkeen voidaan valita mielekkäin mallin aikaansaama tulos ja sivuuttaa kaikki edelliset liian yksinkertaiset, puutteelliset, vääristyneet tai vialliset vastaukset. Tämä ei kuitenkaan tarkoita, että GPT-mallit tai muutkin suuret kielimallit olisivat hyödyttömiä ja jatkuvasti epätarkkoja. Niiden kyvyt ja onnistumiset ovat erittäin tehtäväkohtaisia ja riippuvaisia niille syötetyistä kehotteista ja niissä olevasta informaatiosta. Tietyissä konteksteissa mallien käyttäminen voi myös olla ongelmallista ja riskialtista. Esimerkiksi ihmisten terveyteen liittyvät käyttötarkoitukset voivat pahimmillaan aiheuttaa vakavia riskejä tai onnettomuuksia, sillä näissä tarkoituksissa malleilla ei ole varaa tehdä merkittäviä virheitä. [38], [39]

3 Asiakaspalvelun arvionti

3.1 Chat-pohjaisen asiakaspalvelun ongelmat

Esihenkilöiden on arvioitava manuaalisesti asiakaspalvelussa tapahtuneita tilanteita lukemalla niihin liittyyviä tekstejä varmistaakseen, että asiakaspalvelun agentit pyrkivät aktiivisesti tarjoamaan ystävällisyyden kautta asiakkaille mahdollisimman hyviä positiivisia kokemuksia sekä myymään tuotteita ja palveluita. Käytännön arjessa on kuitenkin mahdotonta, että esihenkilöt pystyisivät tarkastelemaan ja arvostelemaan kaikkia alaistensa tilanteita ja niistä suoriutumista päivittäin. Tilanteisiin liittyviä viestiketjuja ehditään usein lukea korkeintaan kourallinen satunnaisella otannalla asiakaspalvelun agenttia kohden ja tämä voi pahimmillaan kuluttaa useita tunteja työpäivästä. Tämä tarkoittaa, että todella pieni murto-osa tilanteista päätyy lopulta arvioiduiksi. GPT-kielimallien avulla tätä prosessia on mahdollista tehostaa huomattavasti vähemmän aikaa vieväksi, minkä lisäksi jokainen tilanne tulee arvioitua tasa-vertaisesti ja yhtä huolellisesti.

GPT-kielimallit ylittävät moninkertaisesti ihmisen luku- sekä kirjoitusnopeuden ja tarkkaavaisuuskyvyn tehtävien toistojen kasvaessa. Ne ovat uupumattomia lukemaan tekstiä ja pystyvät luomaan vastauksia oikeaan asiayhteyteen liittyen, vaikka alkuperäisessä tekstissä olisi kirjoitusvirheitä. Lisäksi ne ovat erinomaisia tekemään käskyinä huomioimaan asioita, tunnetilojen analyysiä, ymmärtämään kokonaisuuksia, järkeilemään loogisia asioita ja ratkomaan ongelmia. Nämä kaikki ominaisuudet

ovat kriittisiä tarkkojen arvioiden luomista varten ja ovat avain esihenkilöiden arviointityön tehostamiseksi.[8], [9], [40]

Jokaisen asiakaspalvelijan tulisi työskennellessään noudattaa tiettyä kaavamaista yrityksessä käytettävää toimintamallia sekä muita yleisiä toiminnan ohjeita. Näitä ovat esimerkiksi yrityksen yleinen äänensävy, joka on yrityksen koko kommunikaation perusta niin asiakkaiden kanssa kommunikoidessa kuin yleisessä viestinnässäkin. Tämä korostuu erityisesti chat-pohjaisessa asiakaspalvelussa, missä pelkästään tekstin kirjoitusasu ja kysymysten asettelu voivat merkittävästi vaikuttaa asiakaskokemukseen.

Taitavat ja kokeneet asiakaspalvelijat pystyvät kiinnittämään asiakkaan huomion tekstin välityksellä eri tavoilla kuin esimerkiksi vasta aloittaneet asiakaspalvelijat. Kommunikaation sulavuus, ystävällisyys, ymmärtäväisyys ja asiakaslähtöisyys ovat tärkeitä osia asiakkaan tarpeiden kartoitusta ja ratkaisun löytämistä, jotta asiakas tuntee olonsa ymmärretyksi, olevansa oikeassa paikassa ja että hänen asiansa tulevat järjestymään kuntoon. [41]

Tyytyväinen asiakas käyttää ja kuluttaa jatkossakin lisää yrityksen tarjoamia palveluja ja tuotteita sekä suosittelee niitä myös eteenpäin läheisilleen ja sosiaalisessa mediassa. Kun asiakkaat ovat luottavaisia ja tyytyväisiä palveluihinsa, heille muodostuu luottavainen ja positiivinen suhde yrityksen brändiä kohtaan, mikä on keskeinen osa asiakassuhdemarkkinointia. [42]

Asiakaspalvelijan tekemä palvelu voi kuitenkin asiakkaan kontaktin aiheesta ja tilanteesta riippuen olla erittäin vaihtelevaa ja siten myös esihenkilöiden voi välillä olla vaikeaa tai jopa mahdotonta arvioida asiakkaan kohtaamiseen liittyvää suoritusta tarkasti tai puolueettomasti. Asiakaspalvelijalla voi esimerkiksi olla hankalia asiakkaita, teknisiä ongelmia tai asiakkaalla voi olla jokin kysymys mihin asiakaspalvelija ei yksinkertaisesti pysty vastaamaan tai vaihtoehtoisesti jokin ongelma mitä asiakaspalvelija ei voi ratkaista. Asiakaspalvelijan kokeneisuus ja tietotaito tietyissä

tilanteissa voi olla myös erittäin ratkaisevaa lopputuloksen kannalta. [43]

Henkilökohtaiset syyt • Esihenkilön omat käytännöt, asenteet tai mieltymykset voivat heikentää arvioinnin laatua. • Stressi ja työpaineet voivat vaikuttaa arviointityön määrään, laatuun ja objektiivisuuteen. • Kiire ja väsymys voivat heikentää arvioinnin tarkkuutta, kun ajatukset ja huomio ovat toisaalla. • Suhteet työntekijän ja esihenkilön välillä voivat aiheuttaa puolueellisuutta.
Työsuhteen sisäiset syyt • Rakentavan ja säännöllisen palautteen puute heikentää työntekijän kehitysmahdollisuuksia. • Kommunikaation puute voi johtaa väärinkäsityksiin työtehtävien suorittamisessa. • Yrityksen toimintamallit ja ohjeistukset voivat vaikuttaa arviointikriteereihin. • Epäselvät tai epäsopivat arviointikriteerit voivat johtaa epäluotettaviin tuloksiin.
Satunnaiset syyt • Käytettyjen palvelukielten erot voivat vaikuttaa arviointiin. • Kulttuuritaustojen eroavaisuudet voivat vaikuttaa arvioinnin tulkintaan. • Asiakaspalvelutilanteiden ja asiakkaiden vaihtelu voi hankaloittaa yhdenmukaista arviointia.

Taulukko 3.1: Vääristyneiden arviointien mahdollisia syitä

Asiakaspalvelijan suoritusta arvioidessa esihenkilön asemassa olevalla työntekijällä voi puolestaan olla taustalla monia erilaisia henkilökohtaisia syitä, jotka voivat vaikuttaa arvion laatimiseen ja sen objektiivisuuteen. Näitä voivat olla esimerkiksi kiire, väsymys, stressi tai aamukahvien juomatta jääminen. Henkilökohtaiset syyt voivat näin vaikuttaa voimakkaastikin eri tilanteista tehtyihin arviointeihin, johtopäätöksiin sekä niihin liittyviin muihin huomioihin. [44]

Lisäksi tilanteessa esihenkilön ja asiakaspalvelijan väliset suhteet voivat vääristää tehtyjä arvioita erilaisiin suuntiin. Arviointitilanteessa esihenkilö voi antaa esimerkiksi hyvälle ystävälleen tai pitkän uran tehneelle työntekijälle alitajuisesti parempia arvioita. Samalla esimerkiksi nuorempi, kokemattomampi tai epämieluisaksi koettu työntekijä voi saada matalampia arvosanoja tehdystä työstään, vaikka hänen suoriutumisensa olisi aivan yhtä laadukasta kuin jonkun toisen työntekijän. Toisaalta se, että esihenkilöt antavat työntekijöilleen hieman parempia arvioita kuin he ansaitsisivat ei ole välttämättä yhtä haitallista kuin ankarampi arviointi ja heikot arvosanat. Marchegiani et al. (2016) tutkimuksen mukaan lempeämpien arviointien kautta työntekijät pysyvät motivoituneempina ja tehokkaina, koska itsensä aliarvioiduiksi kokevilla työntekijöillä on tapana olla myös vähemmän motivoituneita. [44]

Arvioiden laatimiseen voivat vaikuttaa myös erilaiset kommunikaatioon liittyvät asiat, joita voivat olla esimerkiksi esihenkilön riittämätön palaute tehdystä työstä tai tiettyjen tavoitteiden kertomatta jättäminen. Se, millä kielellä asiakaspalvelua on tehty, voi myös vaikuttaa arviointiin. Suomi on kaksikielinen maa, joten asiakaspalvelua on aina oltava saatavilla suomeksi ja ruotsiksi, joiden lisäksi usein vaihtoehtona on myös englannin kieli. Tämä kuitenkin voi tuottaa ongelmia arviointeja tehdessä, sillä etenkin tekstipohjaisesta asiakaspalvelussa voi olla erittäin vaikeaa havaita onko asiakaspalvelija ollut tietyllä palvelukielellä ystävällinen asiakasta kohtaan vai ei. Tämä voi tehdä objektiivisen arvioinnin haastavaksi, jos arviointia tekevä esihenkilö ei hallitse asiakaspalvelun tilanteen palvelukieltä tarpeeksi hyvin.

Tämän ongelman ratkaisemisessa GPT-mallit ovat erittäin eteviä ja ne omaavat erinomaisia ominaisuuksia asiakaspalvelun tilanteiden tulkitsemista ja arviointia varten. Niitä voidaan ohjeistaa tulkitsemaan teksteistä löytyviä tilanteiden tärkeimpiä pääkohtia, arvioimaan asiakkaiden tunnetiloja ja tyytyväisyyttä, asiakaspalvelijoiden ystävällisyyttä ja koko tilanteen lopputuloksia niin suomen, ruotsin kuin englannin kielellä.

3.2 Asiakaskohtaamismalli

Asiakaskohtaamismalli kuvaa yleisiä tavoitteita ja toimintatapoja asiakaspalvelun arjen tilanteiden toteutukseen, jotta asiakkaiden kokemukset asiakaspalvelun laadusta olisivat mahdollisimman positiivisia. Hyvien asiakaskohtaamisten kautta organisaatioiden erilaiset palvelut ja tuotteet tuottavat asiakkaille arvoa. Asiakaskohtaamismallin useiden vaiheiden kautta on tarkoituksena luoda yhtenäinen, ystävällinen ja laadukas asiakaskokemus, joka tukee organisaation tai yrityksen arvoja, brändiä ja yrityksen viestinnän yleistä äänensävyä. Asiakaskohtaamismallin eri vaiheiden toteutumista ja noudattamista voidaan käyttää asiakaspalvelun agentin tapausten arvioinnin perustana. Se kuvastaa kokonaisuudessaan kuinka hyvin agentti on onnistunut noudattamaan yrityksen yleistä äänensävyä, yleisiä ohjeistuksia ja erilaisia malleja. Asiakaskohtaamismalli koostuu neljästä päävaiheesta:

1. vaihe: kohtaaminen

Ensimmäisessä vaiheessa asiakaspalvelun agentin tehtävänä on luoda ystävällinen ja luottamuksellinen olo asiakkaalle. Tämä voidaan tehdä monin eri tavoin riippuen henkilökohtaisuuksista sekä ajankohtaisista asioista, mutta kaikkein tärkeintä on, että asiakas ymmärtää olevansa oikeassa paikassa ja asiat saadaan järjestymään kuntoon yhdessä agentin kanssa.

2. vaihe: kartoitus

Toisessa vaiheessa agentin tulisi kartoittaa asiakkaan tarpeita niin nykyhetken kuin tulevaisuuden näkökulmasta. Agentin tulee esittää avoimia ja tarkentavia kysymyksiä ja huomioida asiakkaan koko talous eli esimerkiksi palveluiden ja laitteiden määrä koko taloudessa. Lisäksi asiakkaan syy kontaktiin on painava tekijä esimerkiksi erilaisissa ongelmatilanteissa.

3. vaihe: ratkaisu

Kolmannessa vaiheessa tarkoituksena on ratkaista asiakkaan sen hetkiset tarpeet tai ongelmat. Agentin tehtävänä on tarjota kokonaisvaltainen ratkaisu käyttäen edellisissä vaiheissa ilmi tulleita asioita ja tarpeita. Riippumatta siitä onko asiakas on päättänyt hyväksyä tarjouksen tai ratkaisun, agentin tulee aina mainita erilaisia hyötyjä, etuja ja tarjouksia asiakkaan ymmärryksen, ilmenneiden tarpeiden ja käytössä olevien tuotteiden ja palveluiden nojalla.

4. vaihe: ymmärryksen vahvistaminen ja lopetus

Neljännessä eli viimeisessä vaiheessa asiakkaan ymmärrystä vahvistetaan nykyisten sekä uusien palveluiden tai tuotteiden hyödyistä ja eduista. Tässä asiakaspalvelijan kuuluu käydä läpi yhteenveto kaikista sovituista asioista ja mitä seuraavaksi tapahtuu. Asiakkaan ymmärrys vahvistetaan kysymällä ja kertaamalla. Asiakaspalvelun agentin täytyy säilyttää tilanteen tunnelma positiivisena loppuun saakka. Agentin täytyy myös viimeiseksi tehdä asiakkaan kohtaamisesta kirjaus järjestelmään ja kirjoittaa tapauksesta lyhyt raportti siitä mistä on ollut kysymys ja mitä asiakkaan kanssa on sovittu.

Tapahtuman kirjaukset arvioinnin tukena

Arvioinnissa voidaan asiakaskohtaamismallin noudattamisen ja toteutumisen lisäksi hyödyntää järjestelmään kirjattuja raportteja tapahtumista. Näistä raporteista ilmenee usein esimerkiksi, onko asiakaspalvelun agentti onnistunut tuotteiden ja palveluiden myynnissä, miten asiakaspalvelija on suhtautunut asiakkaan tilanteeseen sekä onko asiakkaan yhteydenottoon liittyen kirjattu muita tapahtumia tai toimenpiteitä jo aiemmin tai saman päivän aikana.

Muita lisänäkökulmia arviointia varten

Manuaalisesti kirjattujen tapahtumien raporttien lisäksi agenttien tekstejä voidaan arvioida monista eri näkökulmista. Esimerkiksi ystävällisyys on erittäin tärkeä mittari, joka välittyy chat-asiakaspalvelussa selkeästi agentin tekstien sanavalinnoista sekä kokonaisvaltaisesti esimerkiksi kysymyksiä ja tarjouksia tehtäessä tai asiakkaiden ongelmatilanteita ratkaistaessa. Vaikka agentti noudattaisi asiakaskohtaamismallia ja yrityksen yleistä äänensävyä, nämä eivät kuitenkaan itsestään takaa ystävällisyyden välittymistä asiakkaalle.

Asiakaskohtaamismallista kehotteiksi

Asiakaskohtaamismalli ja siihen liittyvät eri arviointinäkökulmat on koostettava johdonmukaisiksi GPT-malleille annettaviksi ohjekehotteiksi, joita mallit noudattavat jokaisen arvioitavan tapahtuman tekstien kohdalla. Tavoitteena on luoda kehotteiden avulla tasa-arvoisia ja -laatuisia arvioita, joissa otetaan huomioon niin asiakaskohtaamismallin toteutuminen, tilanteen ratkaisu, myynnin näkökulma sekä kirjauksista saatavien lisätietojen perusteella tapahtuman lopputulos. Mallien ohjekehotteiden muotoiluun liittyviin menetelmiin perehdytään tarkemmin seuraavassa luvussa.

4 GPT-mallien ohjaaminen

4.1 Kielimallien kehotteiden muotoileminen

Suurien kielimallien uudelleenkouluttaminen tiettyä tehtävää ja kontekstiä varten vaatii valtavasti resursseja ja aikaa. Sen sijaan voi olla kannattavampaa muotoilla mallille erillisiä ohjekehotteita, jotka määrittävät miten mallin tulisi toimia. Laadukkaat mallille syötetyt ohjeet saavat mallin generoimaan haluttuja tuloksia ja vastauksia ilman, että mallin toimintaa pitää vastauksen jälkeen ohjata lisää. Tämä voi kuitenkin tehtävästä ja valitusta mallista riippuen vaatia erittäin tarkasti muotoiltuja kehotteita.

Kielimalleille syötettävät kehotteet voidaan jakaa kahteen ryhmään järjestelmän ja käyttäjän kehotteiksi. Järjestelmän kehote määrittää mallin käytöksen yleiset toimintatavat, roolin ja säännöt, ja se syötetään mallille aina ensimmäisenä ennen käyttäjän kehotteita. Järjestelmäkehotteissa mallille on kannattavaa antaa tehtävän asiayhteys sekä määrittää mallille siihen liittyvä tarkka rooli, mitkä voivat ohjata mallia tuottamaan parempia vastauksia tai käyttäytymään tietyllä tavalla, esimerkiksi noudattamaan tiettyjä sääntöjä tai suhtautumaan syötetekstiin tietyin tavoin.

Mallin roolin ja kontekstin määrittämisen jälkeen järjestelmän kehotteeseen voidaan koostaa yksi tai useampi tehtävä, mitä mallin tulisi tehdä ja antaa vastaus. Tehtävien kuvauksia ja vaiheita lisätessä voi olla hyödyllistä lisätä ohjeina lyhyitä kontekstiin liittyviä esimerkkejä tai ehtoja, jotka voivat auttaa mallia suoriutumaan

tehtävistä toivotulla tavalla. Tehtävälistauksen jälkeen järjestelmän kehotteeseen voidaan määritellä tietty lopullinen vastausformaatti, esimerkiksi listana, taulukkona tai useana rivinä tekstiä tai erilaisina tiedostomuotoina. Työssä GPT-malleja ohjeistettiin aina vastaamaan tehtävään *JSON*-tiedostomuodossa, jotta arvioinnin arvosanoja ja tekstejä voitiin tallentaa avain–arvo-pareina.

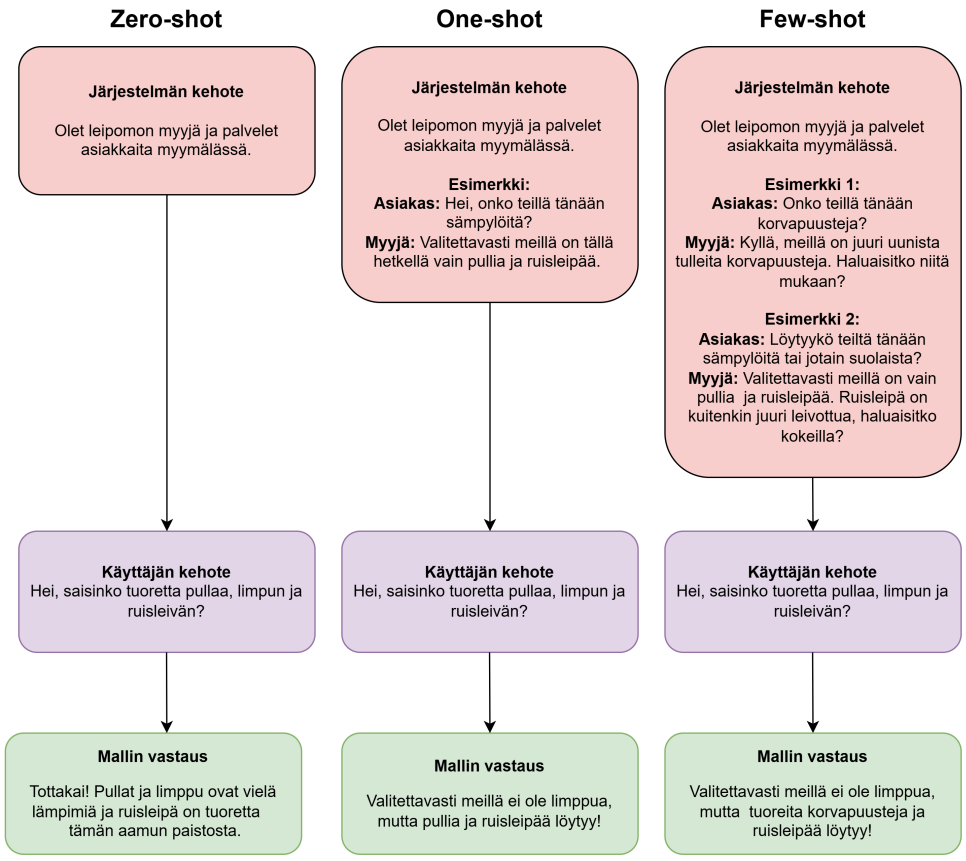
Mallien käytöksen ohjaaminen järjestelmän kehotteiden kautta voi parhaimmillaan olla erittäin tehokasta sekä nopeaa, mutta käytetyistä malleista ja niiden tehtävistä riippuen se voi olla myös erittäin hankalaa. Esimerkiksi kehote joka toimii erinomaisesti jollain mallilla, ei välttämättä toimi lainkaan samalla tavalla toisen mallin tai jopa saman mallin uudemman tai vanhemman version kanssa. [8], [40]

4.2 Zero, One-Shot ja Few-Shot -menetelmät

Generatiivisen tekoälyn suurimmat kielimallit, kuten esimerkiksi OpenAI:n GPT-4-mallit, ovat koulutettu valtavan laajoilla koulutusaineistoilla ja niiden neuroverkoissa on jopa satoja miljardeja parametreja. Mallien suuren koon takia kyseiset mallit pystyvät parhaimmillaan suorittamaan tehtäviä ilman, että niitä on erikseen koulutettu juuri kyseiseen tehtävään liittyvällä koulutusaineistolla. Mallien kehittyessä ja niiden parametrimäärän kasvaessa niiden kyky käsittää syötteiden ja kehotteiden kontekstia joustavammin ja paremmin kasvaa, mikä puolestaan johtaa laadukkaampiin mallien vastauksiin. Esimerkiksi OpenAI:n tutkimuksessa [40] vuonna 2020 havaittiin, että suurten kielimallien kyvyt paranivat monilla eri osa-alueilla tasaisesti koko ajan yli sadan miljardin mallin parametrin jälkeenkin. Näiden mallien huipputulokset ja onnistumiset erilaisissa NLP -tehtävissä liitetään usein erityisesti Zero- ja Few-shot -oppimistekniikoihin, joiden avulla mallit voivat ratkaista tehtäviä joko suoraan pelkkien annettuiden ohjeiden tai myös niiden lisäksi tehtävää vastaavien esimerkkien avulla. [45]

Zero-, one- ja few-shot -oppimismenetelmät voivat merkittävästi lyhentää sitä ai-

kaa ja resurssien tarvetta, mitä vaaditaan mallin riittävän suorituskyvyn aikaansaamiseksi sen sijaan, että mallia varten pitäisi kerätä kattava määrä lisää tarkentavaa koulutusaineistoa tiettyä tehtävää varten. Näillä oppimismenetelmillä on kuitenkin rajoituksensa. Monet mallit voivat olla todella herkkiä syötteiden ja kehotteiden muotoilulle, joista riippuen mallit voivat antaa yllättävänkin vaihtelevia vastauksia. Tarkan ja suunnitelmallisen kehotteiden muotoilun tavoitteena on erilaisten menetelmien avulla saada aikaan järkevästi käyttäytyvä ja vastaava malli. Mallille annettavista tehtävistä riippuen kehotteiden muotoilua voidaan tehdä erilaisten esimerkkien, toimintamallien, ohjelistojen ja jopa tarinankerronnan avulla.[40], [46]



Kuva 4.1: Esimerkkejä kehotepohjaisista oppimismenetelmistä

Kuvan 4.1 kaltaisesti zero-shot -oppimismenetelmässä mallille annetaan vain tehtävää kuvaava ohje, mitä mallin halutaan tekevän ilman esimerkkejä. Syöte voi olla kysymys, käännöksen tekeminen tekstistä toiselle kielelle tai jokin muu tehtävä. Mallit voivat generoida vastauksia vain koulutusaineistoista opittujen asiayhteyksien kautta. Annetun tehtävän aiheesta ja luonteesta riippuen voi kuitenkin olla, että malli kohtaa tehtävän, aiheen tai kielen joka ei ole ollut osana sen koulutusaineistoa, jolloin mallin vastaukset voivat olla väärin tai ohi aiheen. Toisaalta mallien massiivisissa koulutusaineistoissa voi kuitenkin olla useita erilaisia tietojoukkoja ja tehtävien aihepiirien lähellä olevia esimerkkejä, joiden avulla mallit pystyvät tiettyyn pisteeseen asti korvaamaan puutteellisia tietojaan ja mukautumaan annettuihin tehtäviin riippuen mallien parametrien määristä ja koulutusaineistojen tehtäviin osuvista konteksteista. [8], [40], [46]

GPT-mallien ratkaisukyky ja tehtävissä onnistumisen todennäköisyys voi kasvaa moninkertaisesti niiden parametrimäärän kasvaessa. Esimerkiksi GPT-3-mallin kohdalla zero-shot-menetelmän kanssa onnistumisen todennäköisyys useimmissa tehtävissä kasvoi noin 30 prosentista yli 70 prosenttiin mallin parametrien määrän kasvaessa 100 miljoonasta noin 13 miljardiin parametriin [40]. One- ja few-shot -menetelmät olivat tehtäväkohtaisesti vielä parempia suoritumaan samoista tehtävistä, saavuttaen esimerkiksi jopa yli 85 prosentin onnistumisen todennäköisyyden LAMBADA-testiaineiston kanssa sekä monissa muissa Winograd-tyylisissä testitehtävissä, kun mallin parametrien määrää kasvatettiin 175 miljardiin parametriin saakka [40]. Tutkimuksissa (Wei et al., 2022) ja (Brown et al., 2020) myös osoitettiin, että yli 100 miljardin parametrin mallit tukevat poikkeuksellisen hyvin few-shot-oppimismenetelmää. [40], [47]

One-shot -menetelmässä mallille syötetään kehotteessaan vain yksi esimerkkitapaus tehtävästä, mistä malli voi ottaa mallia järkeilyyn tuloksen päättelemiseksi. Tämä usein parantaa mallin kykyä päätellä yksinkertaisia ongelmia ja ratkaisuja,

mutta yksi esimerkki ei tehtävästä riippuen välttämättä ole riittävä onnistumisen takaamiseksi, minkä takia one-shot -menetelmää tulisikin käyttää vain tehtävissä, jotka ovat todella yksinkertaisia [40]. Tämän ongelman ratkaisemiseksi voidaan kuitenkin käyttää few-shot -oppimismenetelmää, missä mallille annetaan useita tehtävään liittyviä esimerkkejä ennen varsinaisen tehtävän ohjeen esittämistä kehotteessa. Toisin kuin zero- ja one-shot -menetelmissä, tämä menetelmä tarjoaa mallille useita erilaisia kuvauksia tehtävästä, jotka toimivat mallille kaavana tehtävän rakenteen ja sen kontekstin muodostuksessa. Tämän tavoitteena on mahdollistaa, että malli kykenisi ratkaisemaan tehtävän onnistuneesti, vaikka se olisi monimutkainen ja -muotoinen tai esitykseltään vaihtelevaa laatua. [8], [40], [48]

Zero-, one- ja few-shot -menetelmien tarkkuus erilaisissa tehtävissä voi vaihdella paljonkin riippuen tehtävästä ja mallien parametrien määrästä. Lähtökohtaisesti pienikokoisilla malleilla kehotteiden tarkentava muotoilu esimerkkien kanssa voi parantaa mallin ratkaisukyvykkyyttä noin 10-20 prosentin verran esimerkiksi GPT-3-mallin kohdalla. Mallin parametrien määrän kasvaessa useisiin kymmeniin ja satoihin miljardeihin oppimismenetelmien väliset erot eri tehtävissä suoriutumisen välillä pienenevät. Joidenkin aihealueiden kohdalla zero-shot -menetelmän avulla malli voi suoriutua tehtävistä muutamalla prosentilla paremmin kuin one-shot -menetelmää käyttäessä. Tämä ilmiö johtuu siitä, että zero-shot -menetelmässä käytettävät kehotteet generalisoituvat usein one-shot -menetelmää paremmin tehtävissä, joissa on jonkinlaista vaihtelua. One-shot -kehotteet voivat siis rajata mallille annettavaa tehtävää väärään suuntaan tai liian tiukaksi, jolloin mallin käyttäytyminen erilaisissa tilanteissa voi olla rajoittunutta ja puutteellista. [40], [45], [49]

Few-shot -menetelmässä tätä ongelmaa ei ole samalla tavalla havaittavissa, koska kehotteessa mallille annetaan monen erilaisen tilanteen esimerkillinen kuvaus tehtävää varten ja malli osaa näiden saamiensa esimerkkien perusteella soveltaa laajemmin ja valita tehtävään parhaiten sopivia päätelmiä ja ratkaisuja. Parannus näkyy

erityisesti tehtävissä, jotka vaativat monimutkaisempaa päättelyä tai alakohtaista tietämystä, kuten esimerkiksi luonnollisen kielen ymmärtämisessä, aritmeettisessa päättelyssä, tavallisessa maalaisjärkeilyssä tai arvioinnin tekemisessä. [45]

Parhaimmillaan mallien ratkaisukykyä voidaan parantaa erilaisissa testeissä ja tehtävissä useilla kymmenillä prosenteilla few-shot -menetelmän kehotteiden kautta. Parhaiten kuitenkin few-shot -menetelmän kehotteista hyötyvät mallit, joiden parametrimäärät ovat suuria, kooltaan useita satoja miljardeja, jolloin on havaittavissa poikkeuksellisen suurta kasvua mallin kyvyissä pohtia tehtävänantoa ja suoriutua tehtävistä onnistuneesti. Mallin parametrien määrän kasvaessa tietyn pisteen yli malleissa on havaittu poikkeuksellisia uusia kykyjä ja käyttäytymistapoja, joita kutsutaan esiin nouseviksi kyvyiksi *(engl. emergent abilities)*. Näitä kykyjä havaitaan erityisesti few-shot -menetelmää käytettäessä mallien kyvykkyyksien kasvuna erilaisten tehtävien kohdalla, vaikka kykyjen esiintymistä ei ole voitu ennustaa ja järkeillä pelkästään mallien koulutusaineistojen perusteella. [40], [47]

4.3 Ajatusketjut

Edellisiä oppimismenetelmiä voidaan soveltaa yhdessä mallien esille nousevien kykyjen kanssa vielä pidemmälle, kun kielimallin parametrien määrä eli koko ylittää noin 100 miljardia. Ajatusketjumainen päättely voidaan helposti saada käynnistettyä näissä riittävän suurissa kielimalleissa yksinkertaisesti sisällyttämällä erilaisia ajatusketjuja käynnistäviä muotoiluja kielimallin kehotteeseen few-shot -menetelmän esimerkkien muodossa.[47], [50]

Ajatusketjut *(engl. Chain-of-Thought, CoT)* mahdollistavat tehtävissä olevien monivaiheisten ongelmien jakamisen pienempiin välivaiheisiin, joita mallien on helpompi ymmärtää, ratkaista ja muodostaa niistä parempia vastauksia. Ajatusketjut auttavat myös tulkitsemaan kielimallien käyttäytymistä paremmin. Ketjutuksen avulla voidaan havaita tarkemmin, kuinka mallit ovat saattaneet päätyä tiettyihin

vastauksiin. Vastauksien ja niiden välivaiheiden kautta on mahdollista huomata sekä korjata kehotteiden virheitä ja sanoituksia tehokkaasti, jos mallien päättelyn ajatusketjut eivät ole tuottaneet toivottuja tai järkeviä lopputuloksia vastauksina. Ajatusketjumaista päättelyä voidaan käyttää erilaisissa tehtävissä, kuten esimerkiksi sanallisissa ongelmissa, luovassa ideoinnissa tai aivan arkisessa maalaisjärkeilyssä. Ajatusketjuja voidaan soveltaa erityisesti sellaisiin tehtäviin, jotka voidaan ratkaista tavallisella kirjoitetulla kielellä ja vaiheittaisella järkeilyllä. [45], [50]

Ajatusketjutuksen käyttämisellä kielimallin kehotteissa on useita hyödyllisiä ominaisuuksia lähestymistapana helpottaa ja tehostaa kielimallin päättelykykyä ja tuotoksia ilman erillistä koulutusaineistoa tai uudelleenkoulutusta mallin hienosäätöä varten. Tutkimuksessa (Wei, Jason, et al., 2022) erilaisissa tehtävissä erityisesti GPT-3 175B ja PaLM 540B -mallit saavuttivat ratkaisukykyyn liittyviä parannuksia ajatusketjujen avulla monilla eri osa-alueilla, parhaimmillaan saavuttaen sen aikaisia huipputuloksia niin kirjoitetuissa matemaattisissa ongelmissa (mm. MAWPS ja GSM8K-aineistoissa) kuin myös sanallisissa pulmatehtävissä. [50]

Myös toisessa vuonna 2022 tehdyssä Tokion yliopiston ja Googlen yhteisessä tutkimuksessa (Kojima et al., 2022) havaittiin, että yksinkertaisetkin ajatusketjut voivat parantaa mallin ratkaisukykyä moninkertaisesti. Tätä yksinkertaisempaa few-shot -oppimismenetelmästä poikkeavaa ajatusketjumenetelmää kutsutaan zero-shot -ajatusketjuiksi *(engl. Zero-shot-Chain-of-Thought)* missä kehotteeseen lisätään yksinkertaisesti viimeiselle riville käsky *"ajattele yksi askel kerrallaan"*, tai englanniksi *"let's think step by step"*, ilman että malli tarvitsee kehotteisiin yhtäkään erillistä esimerkkiä tavallisen zero-shot -menetelmän tehtävänannon lisäksi. [45]

Tämä on samanlaista kuin miten myös ihmiset kehittävät ja tajuavat ratkaisuja tietyssä hetkessä kohdattuun ongelmaan. Tiettyyn kysymykseen tai ongelmaan tarvitaan ratkaisu, joten yleensä on järkevää pohtia tai kokeilla erilaisia mahdollisia vaihtoehtoja ja työvaiheita parhaimpien tulosten saamiseksi. Tämän jälkeen voidaan

päättää, mikä menettely olisi ratkaisun kannalta paras vaihtoehto ja mitkä puolestaan eivät sovellu ratkaisuiksi. Tällä tavoin kielimallit voivat saavuttaa huipputuloksia monissa tehtävissä ilman, että niiden kehotteisiin vaaditaan manuaalisesti ihmisen laatimia kaikkia tehtävään liittyviä esimerkkejä. [45], [51]

Kokonaisuudessaan oikeanlaisten, huolellisten ja tarkkojen kehotteiden rakenteiden suunnittelu tarjoaa mahdollisuuden parantaa suurten kielimallien suorituskykyä merkittävästi erilaisissa tehtävissä. Tästä on merkittävästi hyötyä etenkin silloin, kun tehtävään liittyvää koulutusaineistoa tai resursseja ei ole riittävästi tarjolla, tai jos niitä on mahdotonta saada mallin uudelleenkoulutusta varten. Erilaiset kehotteiden muotoilumenetelmät eivät tee minkäänlaista muutosta itse suurten kielimallien sisäisiin toimintapoihin, mutta ne voivat parhaimmillaan ohjata malleja toimimaan tarkasti halutuilla tavoilla. Kehotteiden muotoilumenetemien etuna koko mallin uudelleenkoulutuksen sijaan on myös se, että se ei vaadi käyttäjiltään erityistä ymmärrystä mallin oikeasta toimintavasta tai sen kouluttamisesta, mikä mahdollistaa nopean, tehokkaan ja joustavan tavan ohjata GPT-malleja, vaikka tiettyjä tehtäviä tai aiheita varten koulutettujen mallien taso onkin korkeampi.

Kielimallin tehtäväkohtainen ajatusketjutus on erityisen tärkeää esimerkiksi asiakaspalvelun tilanteita arvioidessa, sillä jokainen tilanne on uniikki ja käytettävissä oleville kielimalleille ei yksinkertaisesti voida koostaa tarvittavaa määrää uutta, tarkkaa ja varmistettua koulutusaineistoa, koska se voisi mahdollisesti sisältää arkaluonteista tietoa niin asiakkaasta kuin yrityksestä itsestään, mikä voisi huono-onnisesti vuotaa mallin vastauksina eteenpäin tahattomasti. Tämän työn ja toteutuksen kaltaista ongelmaa varten onkin siis paljon järkevämpää suunnitella kehotteita, jotka käyttävät yhdistelmänä kielimallien suorituskykyä korottavia kehotteiden muotoilumenetelmiä, minkä lisäksi GPT-mallille syötettäviä tietoja pystytään tarkasti järjestelemään ja sisällyttämään järjestelmän ja käyttäjien kehotteisiin siten, että malli ymmärtää tehtävän ja syötteensä kontekstin mahdollisimman tarkasti.

5 Menetelmät ja arviointityökalu

5.1 Aineisto ja työkalun arkkitehtuuri

Tämän työn tarkoituksena on tutkia ja selvittää kehitettävää työkalua varten, onko asiakaspalvelun tekemästä työstä mahdollista luoda GPT-kielimallien avulla luotettavia, tasalaatuisia ja objektiivisia arvioita tapahtumiin liittyvien tietojen perusteella. Samalla työssä pohditaan myös, miten GPT-kielimalleja voitaisiin mahdollisesti soveltaa myös muihin erilaisiin arviointitehtäviin tai niitä vastaaviin käyttötarkoituksiin, missä arvioinnin puolueettomuus, kaavamaisuus ja toistettavuus on tärkeää toistojen kasvaessa.

Tutkimusta aloittaessa luvussa 3 esitelty asiakaspalveluun liittyvä asiakaskohtaamismalli toimi arvoinnin tärkeimpänä perustana, minkä perusteella kehotteiden kohtia voitiin aloittaa muotoilemaan ja järjestelemään. Asiakaskohtaamismallin kriteerit oli tärkeää ymmärtää selkeästi ja kattavasti, jotta niistä voitiin koota tarpeeksi tarkkoja kehotteita. Arvioinnin eri vaiheiden kriteerien perusteella kehotteita muotoillessa myös tapahtumiin liittyvien tietojoukkojen ominaisuudet ja vaaditut tiedot tarkentuivat selkeämmiksi.

Kehotteita ja tietojoukkoa suunnitellessa työkalun arkkitehtuuri alkoi myös muodostua. Ensimmäisenä kaikki arvioinnin vaiheet ja kriteerit koottiin yhteen järjestelmän kehotteeseen. Järjestelmän kehote muodostui pituudeltaan pitkäksi, kun siihen lisättiin arvioinnin eri vaiheita ja näkökulmia. Heti ensimmäisiä kehotteita testates-

sa oli mahdollista havaita, että yhtä järjestelmän kehotetta käytettäessä malleilla oli selvästi ongelmia arvioinnin kanssa. Mallien tuottamat vastaukset eivät olleet rakenteiltaan tai laaduiltaan tasaisia ja jotain osia arvioinnista jäi puuttumaan tai sekottui muiden osien kanssa. Yhden kehotteen sijasta oli siis järkevämpää luoda monta pienempää ja tarkempaa kehotetta arvioimaan tiettyä näkökulmaa tapahtumien arviointia varten. Kehotteiden jakaminen osiin mahdollisti myös mallille syötettävien tietojen pilkkomisen näkökulmakohtaisesti.

5.2 Yleinen tietosuoja-asetus ja Suomen lait

Asiakaspalvelun asiakkaiden tietojen käsittelyn, työntekijöiden työnteon valvonnan ja arvioinnin prosessien tulee noudattaa Suomen lakeja ja Euroopan unionin määrittämiä asetuksia. Euroopan unionin yleinen tietosuoja-asetus *(engl. General Data Protection Regulation, GDPR)* asettaa tiukat vaatimukset henkilötietojen käsittelylle. Henkilötiedoiksi voidaan luokitella nimet, osoitteet, henkilökortin, ajokortin ja passin numerot, tulot, kulttuurinen profiili, IP-osoitteet sekä sairaalan tai lääkärin hallussa olevat tiedot, jotka yksilöivät henkilön terveydenhuollon piirissä [52].

Henkilötietoja käsitellessä on aina varmistettava prosessin avoimuus, lainmukaisuus ja ennen kaikkea tietoturvallisuus. Generatiivista tekoälyä käytettäessä ja asiakkaiden tietoja käsitellessä vaaditaan erityistä huolellisuutta näiden periaatteiden noudattamisessa. Erityisesti tietoturvallisuuteen liittyen tietojen minimoinnin periaatetta on noudatettava tarkasti, jonka mukaan vain tarpeellisia tietoja saa tarvittaessa käsitellä. Tämän lisäksi GDPR:n artikla 22 rajoittaa automatisoitua päätöksentekoa asiakaspalvelun agenttien tekemän työn automatisoituun arviointiprosessiin liittyen, minkä mukaan jokaisella henkilöllä on oikeus olla joutumatta sellaisen päätöksen kohteeksi, joka perustuu pelkästään automaattiseen käsittelyyn, kuten esimerkiksi profilointiin, ja jolla on häntä koskevia oikeusvaikutuksia tai joka vaikuttaa häneen vastaavalla tavalla merkittävästi [52].

Työnantajilla on kuitenkin työsopimuslain 1 luvun 1 § (55/2001) nojalla aina direktio- eli työnjohto- ja valvonta-oikeus työsuhteessa oleviin työntekijöihinsä, eli työnantajalla on oikeus valvoa ja johtaa työtä. Työsuhteessa työntekijä sitoutuu tekemään työtä työnantajan johdon ja valvonnan alaisena noudattaen niitä määräyksiä, joita työnantaja antaa toimivaltansa mukaisesti työn suorittamisesta [53]. Työnantaja voi työn johto- ja valvontaoikeutensa nojalla määrätä miten, missä ja milloin työ tulee suorittaa. Työnantajalla on siten myös oikeus valvoa työntekoa ja työn lopputuloksen laatua [53].

Direktio-oikeuden laajuutta kuitenkin puolestaan tarkentavat ja määrittelevät työ- ja työehtosopimukset sekä työlainsäädäntö. Suomen työsopimuslaissa painotetaan, että kaikki valvonta- ja arviointiprosessit on toteutettava läpinäkyvästi. Lain mukaan työntekijöille on tiedotettava, miten heidän suorituskykyään arvioidaan. Lisäksi työsopimuslain 2 luvun 2 § määritellään, että työnantajan ei saa syrjiä työntekijöitään ja työntekijöitä on kohdeltava tasapuolisesti, jollei siitä poikkeaminen ole työntekijöiden tehtävät ja asemat huomioon ottaen perusteltua, mikä voi myös vaikuttaa arviointiprosessiin. [54]

Laki yksityisyyden suojasta työelämässä 7 luvun 21 § (30.12.2021/1337) puolestaan velvoittaa, että työnantajan on myös käytävä yhteistoimintalain mukainen vuoropuhelu työntekijöiden tai heidän edustajiensa kanssa ennen kuin tekninen valvontamenetelmä voidaan ottaa työpaikalla käyttöön [55]. Lakien lisäksi yleisesti eettisiin työkäytäntöihin perustuen tekoälyn käyttö työsuoritusten arvioinnissa on aina oltava oikeudenmukaista, tasalaatuista, puolueetonta ja sen tuottamat arviot eivät saa johtaa potkuihin tai työntekijän vääränlaiseen kohteluun. Esihenkilöt ovat itse vastuussa päätöksistään ja velvoitettuja tekemään lopulliset työsuoritukseen liittyvät arviot ja päätökset ilman tekoälyn tukea.

5.3 Euroopan unionin tekoälysäädös

Euroopan unioni on toteuttanut vuonna 2024 tekoälysäädöksen, jossa määritellään oikeudelliset kehykset tekoälyn yleiselle käytölle unionin alueella. Säädös määrittää ja jakaa tekoälyjärjestelmät neljään eri riskiryhmään niiden käyttötarkoituksien ja niihin liittyvien riskien perusteella, joita ovat minimaalisen, rajallisen, suuren ja ei-hyväksyttävän riskin ryhmät. Tekoälysäädös asettaa kieltoja ja vaatimuksia tekoälyjärjestelmille, jotka liittyvät esimerkiksi ihmisten arviointiin, tunteiden tulkintaan, sosiaaliseen pisteytykseen sekä päätöksentekoon ja jotka voidaan mahdollisesti luokitella suurten riskien tekoälyjärjestelmiksi. [56]

Tekoälysäädöksen artikla 5 asettaa kieltoja tekoälyn käytölle tilanteissa, joissa sen toiminta voi aiheuttaa vakavaa haittaa yksilöille tai yhteiskunnalle. Artikla kieltää tekoälyjärjestelmät, jotka hyödyntävät manipulatiivisia tai harhaanjohtavia tekniikoita, arvioivat yksilöitä sosiaalisen pisteytyksen perusteella, käyttävät tunteiden analysointia työpaikoilla tai luokittelevat henkilöitä biometristen, poliittisten, uskonnollisten tai muiden arkaluonteisten ominaisuuksien perusteella [56].

Työssä kehitetty tekoälyjärjestelmä ei sisällä kieltoihin liittyviä riskitekijöitä, sillä sen toiminta perustuu ennalta määriteltyihin asiakaspalvelun onnistumisen kriteereihin. Nämä kriteerit ovat samat kaikille agenteille tasavertaisesti ja myös esihenkilöt käyttävät niitä itse arvioidessaan asiakaspalvelun tilanteita manuaalisesti. Toteutetun tekoälyjärjestelmän tarkoituksena ei ole missään vaiheessa suoraan analysoida tunteita keskustelujen teksteistä, vaan olivatko asiakaspalvelijoiden toimintatavat annettujen toimintaohjeiden mukaisia sekä olivatko tilanteet onnistuneita niin teleoperaattorin kuin myös asiakkaiden näkökulmista.

Artikla 5 kieltää työntekijöiden sosiaalisen pisteytyksen, luokittelun tai rankaisemisen heidän sosiaalisen käyttäytymisensä tai tunnettujen, pääteltyjen tai ennakoitujen henkilökohtaisten ominaisuuksiensa tai luonteenpiirteidensä perusteella [56]. Työkalun tuottamien numeroarvosanojen tarkoituksena ei ole pisteyttää työnteki-

jöitä sosiaalisesti, vaan arvioida erilaisten tilanteiden onnistumista asiakaspalvelijan henkilöllisyydestä riippumatta. Arvosanojen jakamisen tarkoituksena on asiakaspalvelun tasapuolinen laadun kehittäminen, parannettavissa olevien tilanteiden havaitseminen ja koulutuksellisen palautteen jakaminen suoraan työntekijöille eri asiakaspalvelun tilanteissa kehittymistä varten. Arvosanat perustuvat yksittäisten asiakaspalvelutilanteiden arvosteluun ja palautteen muotoiluun anonyymisti eikä niitä tule käyttää työntekijöiden luokitteluun. Näin ollen työkalun käyttö ei riko artiklan 5 kieltoja, mutta sitä käytettäessä sekä jatkokehittäessä pitää huomioida erityisen tarkasti kuinka lähellä tekoälysäädöksen määrittämiä rajoja järjestelmä on sen hetkisessä tilassaan, mihin suuntaan sitä ei voida laajentaa ja pitääkö joitakin järjestelmän toiminnallisuuksia muuttaa, jotta se noudattaa Euroopan unionin asettamia säädöksiä myös tulevaisuudessa.

Työssä kehitettyä tekoälyjärjestelmää ei tekoälysäädöksen artiklan 6 kohdan 3 mukaan voida pitää suuririskisenä, koska se ei aiheuta merkittävää vahingon riskiä luonnollisten henkilöiden terveydelle, turvallisuudelle tai perusoikeuksille, mukaan lukien se, että se ei vaikuta olennaisesti päätöksenteon tulokseen [56]. Artiklan 6 kohtien (3a-d) mukaan tekoälyjärjestelmän toiminta on hyväksyttävää, jos sen tarkoituksena on suorittaa suppea menettelyllinen tehtävä, parantaa aiemmin suoritetun ihmisen toiminnan tulosta, havaita päätöksentekotapoja tai poikkeamia aiemmista päätöksentekotavoista, eikä sen tarkoituksena ole korvata aiemmin tehtyä ihmisen tekemää arviota tai vaikuttaa siihen ilman asianmukaista ihmisen suorittamaa arviota [56]. Nämä seikat toteutuvat tässä työssä kehitetyn tekoälyjärjestelmän käyttötarkoituksessa, jossa tavoitteena on tehostaa ihmisten tekemää arviointityötä.

5.4 Käytettävät tiedot ja niiden käsittely

Tutkimusta varten tekstiaineistoa noudettiin teleoperaattorin kahdesta eri tietokannasta, jotka yhdistettiin yhdeksi isoksi kokonaisuudeksi taulukkomuodossa. Aineistosta poistettiin tyhjät rivit, rivien kaksoiskappaleet sekä rivit, joissa olevat tekstikeskustelut eivät sisältäneet tarpeeksi tekstiä tapahtuman arviointia varten. Näitä olivat esimerkiksi rivit, joissa keskustelu ei ollut edennyt tervehdystä pidemmälle asiakkaan puolelta tai jos keskustelu oli katkennut liian aikaisin. Kootuista ja suodatetuista aineistotaulukon kymmenistä tuhansista riveistä valittiin ositetulla otannalla vuoden 2024 toukokuun ajalta 151 riviä tapahtumien keston mukaan.

Ensimmäisestä tietokannasta saatavilla olleet tekstit oli anonymisoitu henkilötietojen osalta etukäteen. Teksteissä olleet asiakkaiden nimet, syntymäajat, sähköpostiosoitteet, kotiosoitteet ja puhelinnumerot olivat kaikki korvattu erilaisilla merkinnöillä, jotka osoittivat mahdolliset arkaluontoiset kohdat tekstissä niiden tietotyypin mukaan. Tämä anonymisointi oli toteutettu käytettävien tietojen minimoinnin periaatetta noudattaen, jolloin käsiteltävät tiedot rajoittuivat vain niihin, jotka olivat olennaisia arviointityön kannalta.

Ensimmäisestä tietokannasta kerättiin keskusteluihin liittyviä kategorioita ja keskusteluiden tekstit järjestyksessä niin asiakkaiden kuin asiakaspalvelijoiden puolelta, jotka ovat näkyvissä taulukossa 5.1. Tämän lisäksi GPT-mallien arviointiprosessien tueksi tietokannasta kerättiin keskusteluista löytyviä kategorioita ja metatietoja, joita voitiin käyttää asiakastietoihin liittyvien asiakaspalvelun tekemien kirjauksien hakemiseen toisesta tietokannasta. Kaikista tärkein kerätyistä tiedoista oli asiakkaiden ja agenttien väliset chat-keskustelut, sillä niiden avulla kielimalleille voidaan antaa tapahtumien tarkat vuoropuhelut ja yhdessä lisätietojen kanssa ne muodostavat erittäin tehokkaasti kuvauksia siitä, miksi asiakkaat ovat olleet yhteydessä, mistä on keskusteltu ja millaisiin lopputuloksiin asiakaspalvelutilanteissa lopulta päädyttiin.

Sarake	Tyyppi	Esimerkki
Päivämäärä	`datetime`	`YYYY-MM-DD HH:MM:SS`
Tapahtuman tunniste	`string`	`"esimerkki1234"`
Agentin tunniste	`string`	`"esim12345"`
Asiakasnumero	`integer`	`123456789`
Uudelleenkontakti	`boolean`	`0 tai 1`
Käsittelyaika (s)	`integer`	`300`
Keskustelun aiheet	`list`	`aihe1, aihe2, aihe3`
Agentin aiheet	`list`	`aihe4, aihe5, aihe6`
Asiakkaan ja agentin keskustelu	`string`	`"tässä keskustelua"`
Agentin kategoriat	`list`	`kategoria1, kategoria2`
Asiakkaan antama arvosana	`integer`	`tyhjä tai 1-5`
Asiakkaan antama tekstipalaute	`string`	`"palautetta"`

Taulukko 5.1: Tietokannasta kerättyjen rivien sarakkeet ja tietotyypit

Tietokannasta kerättyyn aineistoon liittyviä huomioita

Tietokannasta kerättyjen kaikkien keskusteluiden laatu oli erittäin vaihtelevaa. Keskusteluiden kesto saattoi ajallisesti vaihdella alle parista minuutista jopa useiden kymmenien minuuttien mittaisiin keskusteluihin. Keskusteluiden ajallinen kesto ei lisäksi aina ollut suoraan verrannollinen keskustelun tekstin määrään, koska asiakaspalvelun asiakkaat eivät välttämättä vastaa heti takaisin agentin viesteihin, minkä lisäksi asiakkaat voivat odottamatta sulkea keskustelun milloin tahansa.

Näiden syiden nojalla keskusteluista olikin tutkimuksen kannalta järkevää käyttää ja arvioida vain niitä keskusteluja, joissa on riittävästi keskustelua asiakkaan ja agentin välillä ja käyttää keskusteluiden ajallista kestoa vain lisätietona. Kootusta keskusteluiden aineistosta löytyi lisähuomiona, että keskustelut sisälsivät usein jonkin kategorioitavan kontaktin syyn ja täysimittaisen keskustelun asiakaspalvelijan kanssa, jos asiakaan puolen syötetekstit olivat pituudeltaan vähintään noin 137 merkkiä tai sitä enemmän.

Asiakastietojärjestelmästä noudetut kirjaukset

Toisesta tietokannasta puolestaan noudettiin ja yhdistettiin kirjauksia kyseiseen asiakaspalvelun tapahtumaan liittyen rivikohtaisesti sen päivämäärän lähettyviltä, milloin asiakaspalvelun keskustelu oli tapahtunut. Nämä lisätiedot auttavat kielimallia tulkitsemaan ja ymmärtämään paremmin onko asiakaspalvelun tapahtumaa esimerkiksi onnistunut myynnin kannalta, vai onko asiakas esimerkiksi halunut irtisanoa jonkin palvelunsa, onko asiakkaalla jokin ongelma mihin on tarvittu ratkaisu ja mitä asiakkaan asialle on tehty tapahtumahetkellä.

Sarake	Tyyppi	Esimerkki
Kirjattu huomio asiakkaasta	`string`	`"Huomio asiakkaasta"`
Kirjattu tapahtuma 1	`string`	`"pvm: kirjaus tapahtumasta"`
Kirjattu tapahtuma 2	`string`	`"pvm: kirjaus tapahtumasta"`
Kirjattu tapahtuma 3	`string`	`"pvm: kirjaus tapahtumasta"`

Taulukko 5.2: Asiakastietojärjestelmästä noudettujen tietojen rakenne

Asiakaskirjauksiin liittyvät ongelmat

Taulukon 5.2 kirjausten tekstit eivät olleet alunperin niitä kerätessä millään tavalla anonymisoituja, joten kaikki mahdolliset yhteystiedot, henkilötunnukset ja nimet piti erikseen anonymisoida jokaista kirjausta kohden ennen niiden syöttämistä GPT-kielimalleille. Anonymisointi tapahtui samalla tavalla kuin ensimmäisen tietokannan kanssa, eli korvaamalla teksteissä olevat kohdat erilaisin tyyppimerkinnöin kertomaan malleille minkälaista tietoa tietystä kohtaa oli peitetty piiloon.

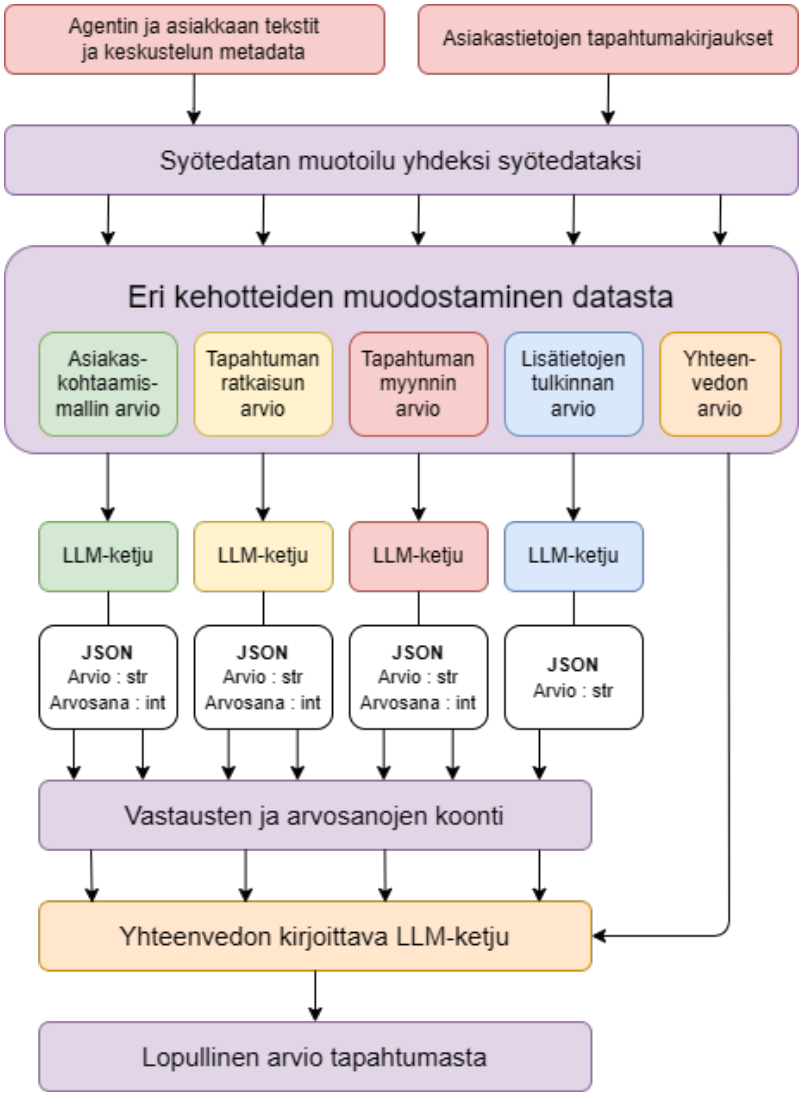
Jokaiselle tapahtumariville kerättiin korkeintaan kolme kirjausta tapahtuman päivämäärältä tai sitä edeltäviltä päiviltä, jos niitä oli. Kaikkia asiakkaisiin liittyviä kirjauksia ei ollut kannattavaa hakea, koska liian vanhat kirjaukset sekoittivat kielimallien tulkintaa arvioitavista tapahtumista. Samasta syystä myös mahdollisia arvioitavien tapahtumien päivämäärien jälkeisiä kirjauksia ei ollut järkevää käyttää,

koska tapahtumiin liittyvät kirjaukset tehdään aina tapahtumahetkellä tai välittömästi sen jälkeen.

5.5 Arkkitehtuuri

Työn aikana toteutettua ohjelmaa kehitettiin ja suoritettiin Windows 10 ja 11 käyttöjärjestelmien Linux-ympäristössä *(engl. Windows Subsystem for Linux, WSL)* erillisessä Docker-kontissa. Linuxin versiona käytössä oli Ubuntu 22.04 LTS. Ohjelmointikielenä käytössä oli Python versio 3.11.9, ja kehitysympäristön hallintaan hyödynnettiin Anacondan versiota 24.1.2. Pythonin kirjastoista erityisesti LangChain-kirjasto oli kielimallien ketjutuksensa ansiosta kriittinen koko työkalun arkkitehtuurin rakenteen toimivuuden kannalta. Kyseinen kirjasto mahdollistaa tehokkaasti eri kielimallien, niiden kehotteiden, syötteiden ja vastausten käsittelyn rinnakkain kirjaston oman LCEL-ilmaisukielen *(engl. LangChain Expression Language)* avulla. Azure OpenAI:n kielimalleista käytössä olivat GPT-3.5 Turbo 16K, GPT-4o Mini 128K ja GPT-4o 128K. Muiden OpenAI:n GPT-4 ja GPT-4 Turbo -mallien käyttäminen tutkimuksen tulosten ja toteutuksen tarkoitukseen ei ollut kannattavaa, sillä niiden hinnoittelu ja vastausajat eivät olleet uudempien GPT-4o-mallien tasolla Azure OpenAI-palvelussa.

Kuvassa 5.1 olevan toteutuksen arkkitehtuuri koostui viidestä identtisiä hyperparametrejä käyttävästä GPT-kielimallista, joista neljä suoritettiin keskenään rinnakkain suoritusajan tehostamiseksi. GPT-malleilla on erilaisia säädettäviä hyperparametreja, jotka vaikuttavat mallien toimintaan ja miten tekstiä generoidaan. Näitä hyperparametreja ovat lämpötila, token-määrä, Top-P-arvo, sanojen esiintymistiheyden ja niiden läsnäolon sakotus. Hyperparametreista ainoastaan lämpötilaa muokattiin tavallisesta oletusarvosta 0.5 alaspäin arvoon 0.0, mikä vähensi mallien vastausten satunnaisuutta ja ohjasi malleja antamaan suorempia ja lyhyempiä vastauksia.

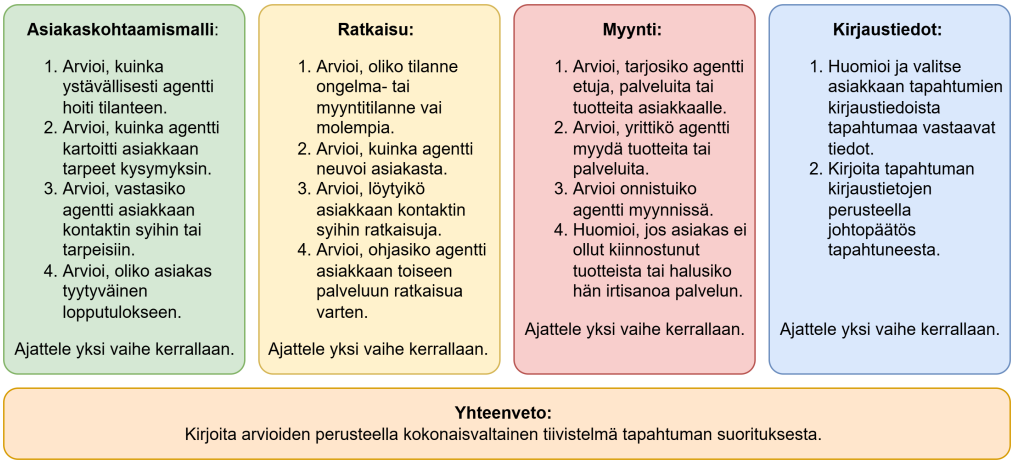


Kuva 5.1: Vuokaavio arkkitehtuurista

Kun kuvan 5.1 neljä mallia ovat vastanneet syötteksteihinsä, niiden tuottamat vastaukset jäsennellään ja eritellään, jonka jälkeen ne syötetään vielä kerran eteenpäin kielimalliputken viimeiselle mallille, joka luo kaikista vastauksista tapahtumaan liittyvän yhteenvedon agentin suorituksesta ja arvosanoista tiivistelmän muodossa. Tuotetut lopulliset arviot saadaan yhteenvetona tekstimuodossa tietystä tapahtumasta ja siihen liittyvät arvosanat erilaisista kategorioista tai vaihtoehtoisesti voidaan käyttää yhtä arvosanaa. Toteutuksessa käytettiin yhtä lopullista arvosanaa väliltä 1-5, jonka avulla voidaan arvioida agentin onnistumista tapahtumassa.

Käytetyt kehotteet

Luvussa 3 esitetyt asiakaspalvelun asiakaskohtaamismallin vaiheet ja muut arvioinnin näkökulmat muotoiltiin mallien kehotteiksi luvussa 4 esitetyn Zero-shot Chain-of-Thought -oppimismenetelmän mukaan. Kuvan 5.1 toteutuksen arkkitehtuurin mukaisesti kehotteita oli muotoiltu viisi erilaista jokaista arvioinnin näkökulmaa varten, joiden yksinkertaistetut versiot ovat kuvassa 5.2.



Kuva 5.2: Muodostetut kehotteet yksinkertaistettuna

Zero-shot Chain-of-Thought -menetelmän käyttämisen etuna on sen ominaisuus auttaa mallia mukautumaan paremmin vaihteleviin tilanteisiin, jotka vaativat mallilta joustavuutta, kuten esimerkiksi sisällöiltään erilaisten asiakaspalvelun tilanteiden arvioinnissa. Menetelmää käytettäessä malleille ei tarvitse syöttää kehotteiden tehtävänannon lisäksi erillisiä esimerkkejä, mikä vähentää mallien käyttämien kehotteiden tokenien käyttömääriä ja siten myös kustannuksia. Menetelmän ajatusketjutus mahdollistaa sen, että GPT-mallit voivat pohtia syötetekstien tärkeimpiä kohtia ja vaiheita yksi kerrallaan vastatakseen kehotteiden arviointitehtävien eri vaiheisiin mahdollisimman kattavasti. Ajatusketjujen avulla mallit pystyvät myös tuottamaan perusteluita ja palautetta tuottamiinsa arviointiteksteihin asiakaspalvelun laadun parantamista varten.

6 Tulokset

6.1 Mittausmenetelmät

Toteutetun työkalun tuottamat rivikohtaiset arviot ovat uniikkeja ja tapahtumakohtaisia palautetekstejä. Mitattavia tuloksia varten tekstimuotoiset arvioinnit muutettiin erilaisten kategorioiden totuusarvoja sekä numeroarvoja vastaaviksi vektoreiksi, joita voitiin mitata tuloksien kautta. Jokainen vastausvektori sisälsi 17 arvoa, joista kolme oli numeroarvosanoja ja loput 14 olivat binäärimuotoisia totuusarvoja. Totuusarvot vastasivat chat-asiakaspalvelun tapahtumasta löydettäviä kategorioita.

GPT-mallien tehtävänä oli tehdä näiden erilaisten kategorioiden luokittelua sen perusteella, mitä niistä ne pystyivät havaitsemaan ja ymmärtämään tapahtumiin liittyvistä syötteksteistään. Totuusarvoisten kategorioiden kohdalla käytettiin binääriluokitukseen soveltuvia mittareita mallien suorituskyvyn arvioimiseksi, joita olivat tarkkuus, osuvuus, herkkyys, F1-pistemäärä, Cohenin Kappa, Matthewsin korrelaatiokerroin ja ROC-käyrän käyränalainen pinta-ala.

Mallien tuottamat numeroarvosanat olivat kokonaislukuja välillä yhdestä viiteen samalla tavalla kuin asiakastyytyväisyyden CSAT-mittarilla *(lyh. Customer Satisfaction Score)*. Numeroarvosanojen tarkkuuden ja hajonnan jakautumista eri mallien tuottamien arvosanojen välillä tarkasteltiin puolestaan lämpökarttojen avulla.

Tarkkuus

Tarkkuus *(engl. accuracy)* on mittareista kaikista yksinkertaisin ja tavallisin. Tarkkuus kuvaa kaikkien oikein luokiteltujen tapausten osuutta kaikkien luokiteltujen tapausten kokonaismäärästä. Tarkkuus voi kuitenkin olla harhaanjohtava mittari, jos aineistossa on epätasapainoa kategorioiden luokkien välillä.

$$\text{Tarkkuus} = \frac{OP + ON}{OP + ON + VP + VN} \tag{6.1}$$

Missä:

OP = oikeat positiiviset, ON = oikeat negatiiviset

VP = väärät positiiviset, VN = väärät negatiiviset

Osuvuus

Osuvuus *(engl. precision)* mittaa ja kuvaa, kuinka suuri osa kaikista positiivisista ennusteista on oikeita positiivisia. Korkeat osuvuuden arvot viittaavat siihen, että malli pystyy tarkasti ennustamaan oikeita positiivisia tapauksia ja vain vähän vääriä positiivisia.

$$\text{Osuvuus} = \frac{OP}{OP + VP} \tag{6.2}$$

Herkkyys

Herkkyyden *(engl. recall)* arvot puolestaan kuvaavat, kuinka suuren osan aidoista positiivisista tapauksista malli onnistuu luokittelemaan oikeiksi positiivisiksi.

$$\text{Herkkyys} = \frac{OP}{OP + VN} \tag{6.3}$$

F1-pistemäärä

F1-pistemäärä on osuvuuden ja herkkyyden harmoninen keskiarvo. Arvo kertoo, kuinka hyvin malli onnistuu luokittelemaan kaikkia positiivisia ja oikeita positiivisia tapauksia tasapainoisesti. F1-pistemäärän arvo voi olla korkea vain silloin, jos osuvuuden ja herkkyyden arvot ovat myös korkeita keskenään.

$$\text{F1-pistemäärä} = 2 \times \frac{\text{Osuvuus} \times \text{Herkkyys}}{\text{Osuvuus} + \text{Herkkyys}} \tag{6.4}$$

Cohenin Kappa

Cohenin Kappa -arvo mittaa mallin ennusteiden ja todellisten arvojen välistä yhteneväisyyttä, ottaen samalla huomioon myös sattumalta tapahtuvan yhtenevien arvojen todennäköisyyden. Tiettyjen kategorioiden kohdalla voi käydä niin, että suurin osa ennusteista ja todellisista arvoista osuvat vain yhteen ja samaan luokkaan. Tämän seurauksena mallin ennusteiden tavallinen tarkkuus saattaa näyttää todella korkealta, vaikka ennusteet perustuisivat vain pelkän yhden saman arvon ennustamiseen binääriluokittelussa. Cohenin Kappan arvot vähentävät tätä vääristymää, mikä tekee siitä hyvän mittarin tilanteissa, joissa luokat ovat epätasapainoisia. [57]

$$\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e} \tag{6.5}$$

Missä:

p_o = havaittu yhteneväisyys, eli luokittelijan ja todellisten luokkien välinen samaa mieltä oleminen.

p_e = sattumanvaraisesti odotettu yhteneväisyys.

Cohenin Kappa on erityisen hyödyllinen epätasapainoisten luokkien kohdalla, sillä se ottaa huomioon tarkkuuden vääristymät tilanteissa, joissa binääriluokittelija voi saavuttaa korkean tarkkuuden yksinkertaisesti ennustamalla enemmistöluokkaa.

Tekstistä tulkittavien kategorioiden luokittelussa Cohenin Kappa tarjoaa luotettavamman arvion mallien luokittelukyvystä kuin tarkkuuden, osuvuuden ja herkkyyden arvot, koska eri kategorioiden luokkien jakaumat voivat olla todella epätasapainoiset, jolloin muut mittarit saattavat antaa liian optimistisia tulosarvoja. Kappa-arvo voi vaihdella välillä [-1,1] missä arvo -1.0 tarkoittaa täydellistä erimielisyyttä kahden eri luokittelijan välillä, kun taas puolestaan arvo 1.0 kuvaa täydellistä yhteisymmärrystä luokittelua tekevän mallin ja manuaalisesti arvioitujen luokittelujen välillä.

Matthewsin korrelaatiokerroin

$$\text{MCC} = \frac{(OP \times ON) - (VP \times VN)}{\sqrt{(OP + VP)(OP + VN)(ON + VP)(ON + VN)}} \tag{6.6}$$

Matthewsin korrelaatiokerroin (MCC) on yksi yleisimmistä mittareista binääriluokittelijan arvioinnissa. Se on binääriluokittelijan suorituskyvyn arvioinnin kannalta luotettavampi mittari kuin esimerkiksi tavallisemmat tarkkuus ja F1-pistemäärä, koska sen arvoissa huomioidaan molempien oikeiden sekä väärien positiivisten ja negatiivisten ennusteiden määrät ja niiden suhteet toisiinsa. MCC voi saada arvoja välillä [-1, 1], missä 1 tarkoittaa täydellistä ennustetta, 0 satunnaisen arvauksen tasoa ja arvo -1 täysin virheellistä ennustetta. [58]

ROC-käyrän käyränalainen pinta-ala

Receiver Operating Characteristic -käyrä eli ROC-käyrä kuvaa mallin luokittelukykyä todellisten positiivisten osuuksien (TPR) ja väärien positiivisten osuuksien (FPR) arvojen kautta. ROC-käyrä kuvaa kuinka hyvin malli tunnistaa todelliset positiiviset tapaukset suhteessa siihen, kuinka usein malli ennustaa negatiiviset tapaukset väärin positiivisiksi. Käyrä piirtää nämä kaksi arvoa toisiaan vasten ja käyrän muoto kertoo mallin kyvykkyydestä erotella kahden luokan välillä. Muodostuneen käyrän alle jäävä pinta-ala (AUC) on yksittäinen arvo, joka tiivistää mallin

kyvyn erottaa kahden luokan välillä. Pinta-alan arvot voivat saada arvoja väliltä [0,1]. Mitä lähempänä AUC-arvo on arvoa 1, sitä paremmin malli kykenee erottamaan kaikki positiiviset ja negatiiviset tapaukset toisitaan, kun taas arvo 0.5 on satunnainen arvauksen tasoa. Jos saatu AUC-arvo jää 0.5 alapuolelle, malli ennustaa järjestelmällisesti väärin.

Mittareiden rooli mallien havaintokykyjen arvioinnissa

Chat-keskusteluissa tapahtuneiden kategorioiden tulkitseminen totuusarvojen kautta on binäärinen luokittelutehtävä, jossa tavoitteena on luokitella tapaukset kahteen luokkaan jokaisen kategorian kohdalla. Edellä esitetyt mittarit soveltuvat hyvin mallien suorituskyvyn arviointiin eri kategorioiden tunnistamisessa.

Tarkkuus antaa yleisen käsityksen mallien suorituskyvystä havaita mitä keskusteluissa on tapahtunut, mutta se voi olla harhaanjohtava epätasapainoisten kategorioiden kanssa. Osuvuuden ja herkkyyden arvot havainnollistavat mallien kykyjä välttää ennustamasta vääriä positiivisia tapauksia ja kykyjä tunnistaa oikeita positiivisia tapauksia mahdollisimman paljon oikein. F1-pistemäärä yhdistää osuvuuden ja herkkyyden arvot harmoniseksi keskiarvoksi ja ottaa siten huomioon sekä väärien positiivisten että väärien negatiivisten ennusteiden vaikutuksen arvoissaan. Cohenin Kappa- ja MCC-arvot ottavat huomioon epätasapainoiset binääristen luokkien jakaumat ja varmistavat, että mallit eivät ole painottuneita ennustamaan sokeasti pelkkää enemmistöluokan arvoja. Viimeisenä ROC AUC-arvojen avulla saadaan näkemys siitä, kuinka hyvin mallit erottelevat erilaisia kategorioita luokkiinsa.

6.2 Mallien tulokset

6.2.1 Kategorioiden tunnistaminen

Kategoriat	Tarkkuus	Osuvuus	Herkkyys	F1	CK	MCC	AUC
Agentti hoiti tilanteen ystävällisesti	0.31	1.00	0.27	0.42	0.04	0.14	0.63
Agentti kartoitti asiakkaan tarpeet	0.87	0.92	0.94	0.93	0.15	0.15	0.57
Agentti vastasi asiakkaan tarpeisiin	0.90	0.92	0.98	0.95	0.50	0.52	0.70
Asiakas oli tyytyväinen lopputulokseen	0.85	0.84	0.97	0.90	0.52	0.55	0.72
Tilanne oli ongelmatilanne	0.62	0.81	0.70	0.75	-0.06	-0.07	0.46
Tilanne oli myyntitilanne	0.77	0.50	0.08	0.14	0.08	0.13	0.53
Agentti neuvoi asiakasta	0.79	0.95	0.80	0.87	0.36	0.40	0.76
Agentti löysi ratkaisun asiakkaan kanssa	0.79	0.92	0.82	0.87	0.35	0.37	0.72
Agentti ohjasi asiakkaan toiseen palveluun	0.92	1.00	0.20	0.33	0.31	0.43	0.60
Agentti mainosti etuja ja palveluita	0.73	0.61	0.61	0.61	0.41	0.41	0.70
Agentti yritti myydä tuotteita tai palveluita	0.79	0.71	0.67	0.69	0.53	0.53	0.76
Agentti onnistui myynnissä	0.90	0.60	0.50	0.55	0.49	0.49	0.73
Asiakas ei ollut kiinnostunut	0.37	0.89	0.20	0.33	0.06	0.13	0.56
Asiakas halusi irtisanoa palvelun	0.83	0.50	0.11	0.18	0.13	0.17	0.54

Taulukko 6.1: GPT-3.5 Turbon tuloksia

Taulukon 6.1 tulosarvojen perusteella GPT-3.5 Turbo suoriutui eri kategorioiden tunnistamisessa erittäin vaihtelevasti. Joidenkin kategorioiden tulokset ovat arvoiltaan hyviä, mutta samalla tiettyjen kategorioiden kanssa tulokset olivat heikkoja. Malli onnistuu muutamissa yksinkertaisissa kategorioissa, kuten *"Agentti vastasi asiakkaan tarpeisiin"* ja *"Asiakas oli tyytyväinen lopputulokseen"*, mutta vaikeampien ja monitulkintaisten kategorioiden kohdalla mallin suorituskyky on heikkoa. Malli suoriutui huonoiten ongelmatilanteiden havaitsemisessa, minkä kategorian kohdalla tulosarvoista käy ilmi, että GPT-3.5 Turbo ei pystynyt tulkitsemaan kyseisen kategorian tilanteita ollenkaan, ja tulokset jäävät satunnaisen arvauksen tasolle. Myös myyntitilanteiden tunnistamisessa ja luokittelussa mallilla oli selvästi vaikeuksia.

Suurimmassa osassa kategorioista CK- ja MCC-arvot olivat arvoiltaan noin 0.30 ja 0.50 välillä muutamaa poikkeusta lukuunottamatta. Molempien mittareiden arvojen kohdalla kyseinen väli merkitsee kohtuullista luokittelijan suorituskykyä. Samalla arvot antavat viitteita siitä, että GPT-3.5 Turbo ei ollut kovin hyvä erottamaan tapauksia ja niiden kategorioita. Mallin suoritukset olivat silti kaikkien paitsi kahden kategorian kohdalla parempia kuin täysin satunnainen arvaus.

Kategoriat	**Tarkkuus**	**Osuvuus**	**Herkkyys**	**F1**	**CK**	**MCC**	**AUC**
Agentti hoiti tilanteen ystävällisesti	0.96	1.00	0.96	0.98	0.73	0.76	0.98
Agentti kartoitti asiakkaan tarpeet	0.85	0.93	0.89	0.91	0.25	0.25	0.65
Agentti vastasi asiakkaan tarpeisiin	0.87	0.95	0.89	0.92	0.51	0.52	0.80
Asiakas oli tyytyväinen lopputulokseen	0.79	1.00	0.72	0.84	0.56	0.62	0.86
Tilanne oli ongelmatilanne	0.83	0.83	1.00	0.91	0.00	0.00	0.50
Tilanne oli myyntitilanne	0.85	0.83	0.42	0.56	0.47	0.52	0.70
Agentti neuvoi asiakasta	0.81	0.95	0.82	0.88	0.39	0.42	0.77
Agentti löysi ratkaisun asiakkaan kanssa	0.69	0.89	0.73	0.80	0.16	0.18	0.61
Agentti ohjasi asiakkaan toiseen palveluun	0.90	0.00	0.00	0.00	0.00	0.00	0.50
Agentti mainosti etuja ja palveluita	0.87	0.72	1.00	0.84	0.73	0.76	0.90
Agentti yritti myydä tuotteita tai palveluita	0.83	0.67	1.00	0.80	0.66	0.70	0.87
Agentti onnistui myynnissä	0.94	0.71	0.83	0.77	0.74	0.74	0.89
Asiakas ei ollut kiinnostunut	0.40	0.91	0.25	0.39	0.09	0.17	0.58
Asiakas halusi irtisanoa palvelun	0.85	1.00	0.11	0.20	0.17	0.31	0.56

Taulukko 6.2: GPT-4o Minin tuloksia

Malleista uusin GPT-4o Mini suoriutui myös vaihtelevasti erilaisten kategorioiden välillä, kuten taulukon 6.2 kategorioiden tulosarvoista voidaan havaita. GPT-4o Mini-malli suoriutui kuitenkin GPT-3.5 Turbo-mallia selvästi paremmin erityisesti tunne- ja tilannetajua vaativien kategorioiden luokittelussa. Esimerkiksi kategorioiden *"Agentti hoiti tilanteen ystävällisesti"*, *"Agentti mainosti etuja ja palveluita"* ja *"Agentti onnistui myynnissä"* tulosarvot ovat erittäin korkeita kaikkien mittareiden osalta. Cohenin Kappan ja MCC korkeat arvot viittaavat hyvään tarkkuuteen ja yhteensopivuuteen sekä vahvaan korrelaatioon mallin ennusteiden ja käsin arvioitujen

luokkien välillä, mikä kertoo hyvästä kategorioiden luokittelukyvystä ja siten myös keskusteluiden tapahtumien ymmärtämisestä.

GPT-4o Mini-mallin suorituskyky kategorioiden tunnistamisessa ja luokittelussa ei kuitenkaan ollut aivan täydellistä. Tuloksien taulukon 6.2 kategoriassa *"Agentti ohjasi asiakkaan toiseen palveluun"* moni rivin arvoista näyttää nollaa. Tämä tarkoittaa sitä, että malli ei pystynyt lainkaan tunnistamaan tai luokittelemaan kategoriaa oikein, ja tulkitsi kaikki tapaukset aiheen ohi luokittelemalla ne vain yhteen luokkaan. Tässä tapauksessä malli luokitteli niin, ettei yksikään tapaus liittynyt ongelmatilanteisiin tai asiakkaan ohjaamiseen toiseen palveluun.

Kategoriat	**Tarkkuus**	**Osuvuus**	**Herkkyys**	**F1**	**CK**	**MCC**	**AUC**
Agentti hoiti tilanteen ystävällisesti	0.92	0.96	0.96	0.96	0.29	0.29	0.65
Agentti kartoitti asiakkaan tarpeet	0.92	0.94	0.98	0.96	0.46	0.48	0.69
Agentti vastasi asiakkaan tarpeisiin	0.90	0.92	0.98	0.95	0.50	0.52	0.70
Asiakas oli tyytyväinen lopputulokseen	0.83	0.94	0.82	0.88	0.59	0.61	0.83
Tilanne oli ongelmatilanne	0.81	0.84	0.95	0.89	0.09	0.10	0.53
Tilanne oli myyntitilanne	0.83	1.00	0.25	0.40	0.34	0.45	0.62
Agentti neuvoi asiakasta	0.75	0.97	0.73	0.84	0.35	0.42	0.80
Agentti löysi ratkaisun asiakkaan kanssa	0.81	0.89	0.89	0.89	0.26	0.26	0.63
Agentti ohjasi asiakkaan toiseen palveluun	0.94	0.75	0.60	0.67	0.64	0.64	0.79
Agentti mainosti etuja ja palveluita	0.88	0.80	0.89	0.84	0.75	0.75	0.89
Agentti yritti myydä tuotteita tai palveluita	0.90	0.81	0.94	0.87	0.80	0.80	0.91
Agentti onnistui myynnissä	0.98	1.00	0.83	0.91	0.90	0.90	0.92
Asiakas ei ollut kiinnostunut	0.69	0.90	0.68	0.77	0.33	0.36	0.71
Asiakas halusi irtisanoa palvelun	0.87	0.57	0.89	0.70	0.61	0.64	0.87

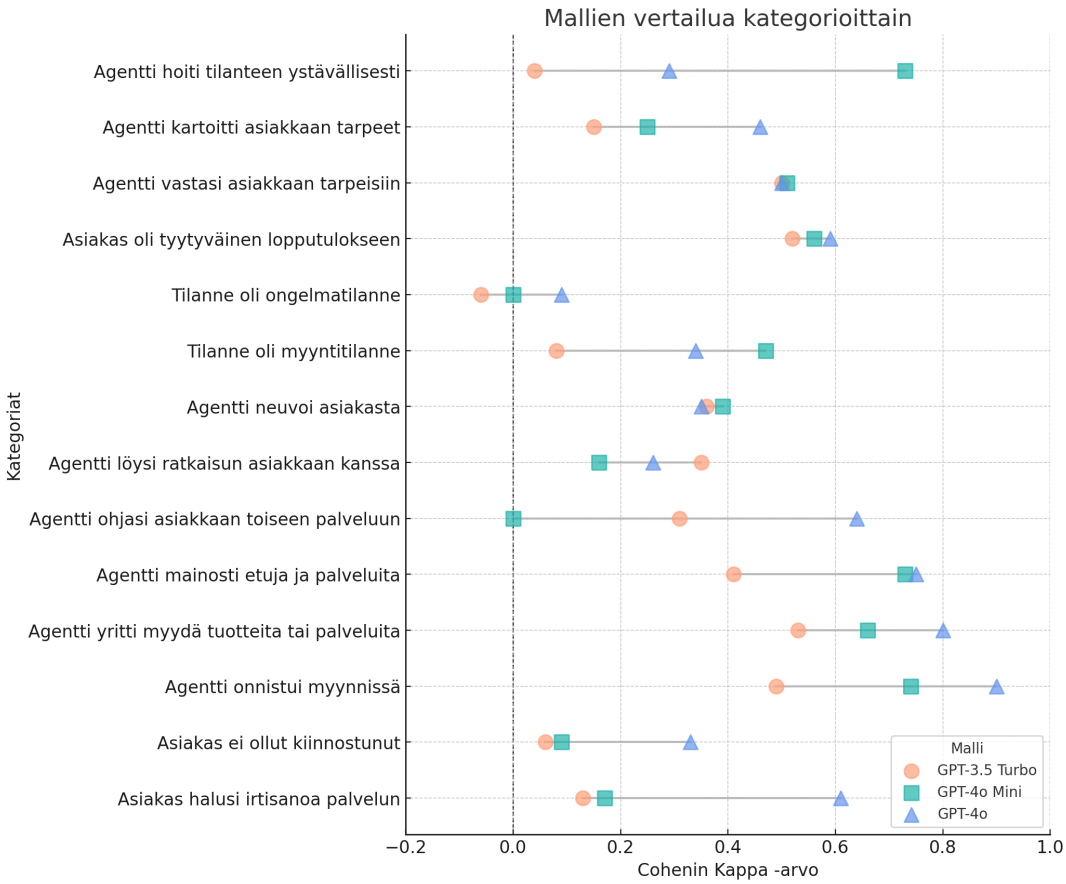
Taulukko 6.3: GPT-4o mallin tuloksia

Myyntiin liittyvissä kategorioissa GPT-4o Mini-malli oli todella lähellä suuremman GPT-4o-mallin tulosten tasoa, mutta vaikeammissa kategorioissa sen kyvykkyys ei aivan yltänyt samalle tasolle. Molemmat mallit suoriutuivat silti hyvin ja tulosarvot olivat molempien mallien välillä todella lähellä toisiaan. GPT-4o-malli suoritui kaikissa kategorioiden luokittelussa ja tunnistamisessa silti selvästi parhaiten.

Mallin F1-pistemäärän, CK- ja MCC-arvot taulukossa 6.3 antavat hyvin kuvaa siitä, että malli pystyi tunnistamaan, ymmärtämään ja luokittelemaan tapahtumien kategorioita vähintään kohtuullisen hyvin ja joidenkin kategorioiden kohdalla jopa erinomaisesti.

Taulukon 6.3 eri tulosarvoista voidaan havaita, että GPT-4o-mallin suorituskyky vaihteli jonkin verran kategoriasta riippuen. Malli osasi selkeästi tunnistaa syötteteksteistä monia erilaisia kategorioita, mutta samaan aikaan mallilla oli selkeästi jonkin verran vaikeuksia tunnistaa kategorioita, jotka voivat olla tilannekohtaisesti erittäin monitulkintaisia pelkän tekstin perusteella. Lisäksi binäärinen luokittelu ei välttämättä toimi parhaalla tavalla kuvaamaan monitulkintaisia ja uniikkeja chat-keskustelussa olevia tilanteita ja kategorioita tulkittaessa karkeasti luokkina. GPT-4o-mallin parhaat tulokset olivat jopa hieman yllättävästi myyntiin liittyvissä kategorioissa, joissa malli onnistui luokittelemaan oikein suurimman osan kaikista oikeista tapauksista. Myös tunnetajua vaativat kategoriat asiakkaan tyytyväisyyteen ja tarpeisiin vastaamiseen liittyen olivat tulosarvoiltaan todella hyviä.

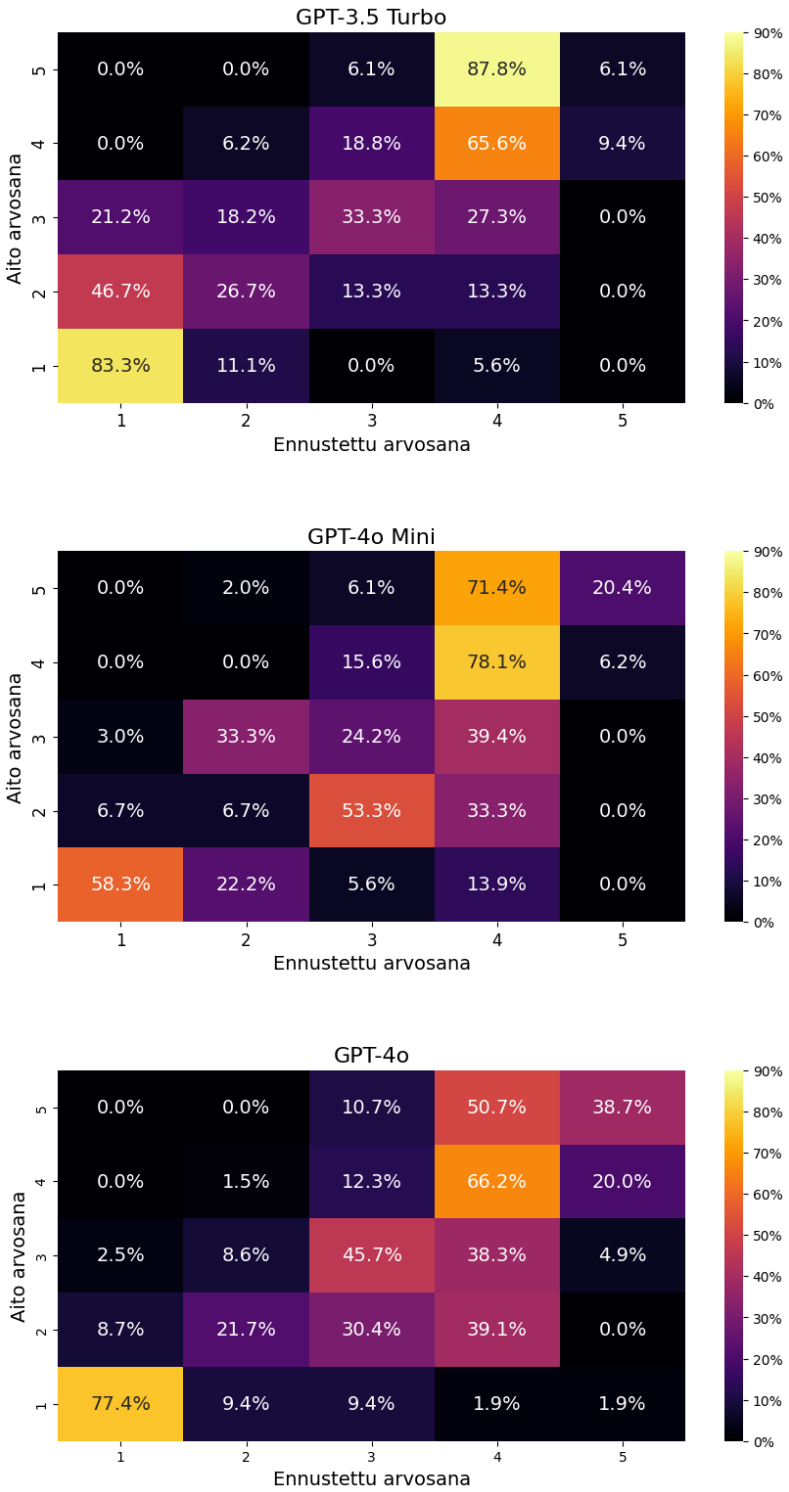
Muista edellisistä malleista poiketen GPT-4o-mallilla puolestaan ei ole minkään kategorian kohdalla vastaavanlaisia kategorioiden havaitsemiseen tai tulkintaan liittyviä ongelmia, vaikka senkään tulosarvot olleet täydellisiä, kuten taulukossa 6.3 olevista arvoista on mahdollista havaita. Kaikkien mallien heikoin kategoria oli kokonaisvaltaisesti *"Tilanne oli ongelmatilanne"* taulukoiden 6.1, 6.2 ja 6.3 sarakkeiden CK-, MCC- ja AUC-arvojen perusteella. Tämä voi johtua siitä, että ongelmatilanteita oli aineistossa vähän. Toisaalta ohjaavien kehotteiden perusteella mallit eivät välttämättä täysin käsitä, mitä kaikkea ongelmatilanteet voivat olla. Malleille vaikuttaa olevan vaikeaa havaita tiettyjä tilanteita, jotka voivat olla samaan aikaan sekä ongelma- että mahdollisia myyntitilanteita. Tämä on nähtävissä taulukoissa 6.2 ja 6.3 sekä kuvassa 6.1, koska molemmat GPT-4o-mallit pystyvät kuitenkin ymmärtämään ja luokittelemaan oikein myyntiin liittyviä kategorioita ja tilanteita.



Kuva 6.1: Mallien Cohenin Kappa -arvojen vertailua kategorioittain

Mallien tuloksien Cohenin Kappa -arvojen avulla voidaan arvioida, kuinka yhteneviä luokitteluja mallit tekevät ihmisen tekemien luokittelujen kanssa. Mallien tulosarvot kuvassa 6.1 kuvaavat selvästi joidenkin kategorioiden olevan vaikeita malleille, joista erityisesti GPT-3.5 Turbo-mallilla on vaikeuksia matalien arvojen kanssa monen kategorian kohdalla. GPT-4o Mini-mallin tulosarvot olivat monissa tilanteissa yhtä suuria tai selvästi GPT-3.5 Turbo-mallia korkeampia. Malleista GPT-4o oli useimpien kategorioiden kohdalla selvästi kyvykkäin, sillä suurin osa sen tulosarvoista oli vähintään kohtuullisen hyviä ja osa jopa erittäin hyviä. Erityisesti mainostamiseen ja myyntiin liittyvien kategorioiden kohdalla GPT-4o-mallin luokittelun tulokset olivat useimmin hyviä ja yhdenmukaisia ihmisen tekemän luokittelun kanssa.

6.2.2 Mallien arvosanojen tuloksia



Kuva 6.2: Mallien antamien arvosanojen jakaumia

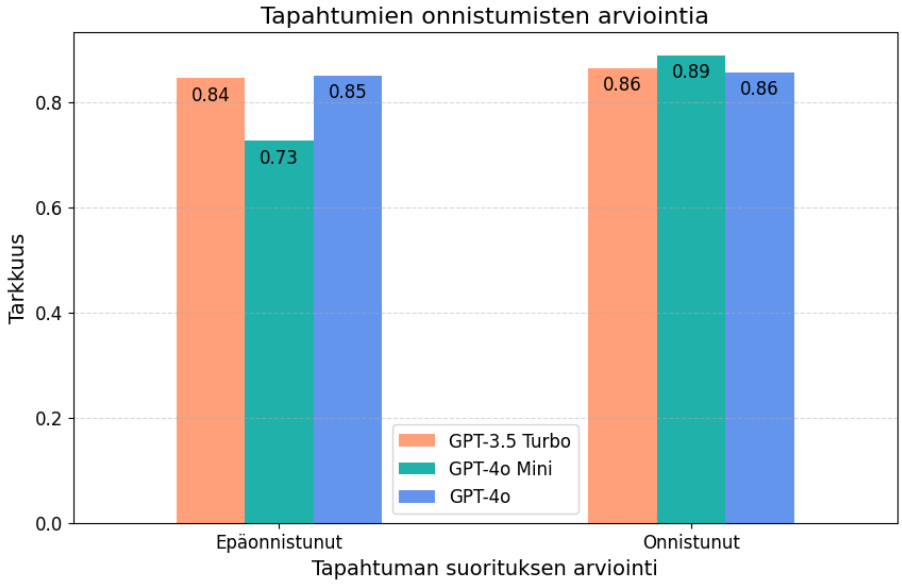
Kuvassa 6.2 olevat lämpökartat kuvaavat kaikkia mallien tuottamien ja aitojen arvosanojen välistä jakaumaa sekä niiden tarkkuuksia. Lämpökarttojen perusteella voidaan selvästi todeta mallien antavan arvosanoja riippuen siitä kuinka onnistuneita tapahtumat ovat. Eniten lämpökartalla nousevat esiin niihin tapahtumiin liittyvät arvosanat, kun tapahtuma on ollut selvästi onnistunut tai epäonnistunut. Tämä ilmiö on näkyvissä suurina prosenttimäärinä ja kirkkaina väreinä lämpökartoilla, joissa erityisesti arvosanoilla 1 ja 4 on eniten osumia.

GPT-3.5 Turbo-mallin tuottamat arvosanat olivat selvästi jakautuneet matalien ja korkeiden arvosanojen välille. Mallilla oli selvästi taipumusta arvioida tilanteita arvosanoilla 1 tai 4, minkä lisäksi korkein arvosana 5 saa todella vähän osumia. Mallit GPT-4o ja GPT-4o Mini olivat puolestaan huomattavasti vähemmän ankaria arvioinneissaan. Molemmat mallit tuottivat arvosanoja täsmällisemmin ja niillä oli samalla selvästi taipumus antaa hieman korkeampia arvosanoja.

Numeroarvosanojen 1 ja 5 välille jäävien lukujen arvioinnissa kaikilla malleilla oli selvästi vaikeuksia osua täysin oikeaan arvosanaan. Etenkin arvosanojen 2 ja 3 kohdalla mallit antoivat vaihdellen noin 9-53 % tapauksista yhden numeron verran liian korkean arvosanan ja puolestaan yhden numeron verran liian matalan arvosanan noin 8-18 % tapauksista, jotka voidaan nähdä kuvan 6.2 lämpökartoilla tumman violetin eri sävyinä.

Arvosanojen 4 ja 5 kohdalla mallit suoriutuivat hieman paremmin. Ensisilmäyksellä GPT-4o Mini vaikuttaa olevan arvosanan 4 kanssa huomattavasti isompaa GPT-4o-mallia tarkempi, mutta kuvan 6.2 lämpökarttojen tarkemmalla tarkastelulla voidaan todeta, että GPT-4o Mini ja GPT-3.5 Turbo -mallit antoivat todella harvoin oikein tapauksille arvosanaksi 5. Tämä näkyy selvästi myös kuvan 6.2 lämpökartoilla, missä molempien mallien kohdalla arvosana 4 vaikuttaisi olevan korkein arvosana mitä mallit antavat arvioiksi arvosanan 5 sijasta. GPT-4o-mallilla oli myös vaikeuksia erotella ja arvioida tapahtumien suorituksia arvosanojen 4 ja 5 välillä.

Noin 20 % aidon arvosanan 4 arvoisista tapauksista sai arvokseen 5 ja puolestaan noin puolet aidon 5 arvoisista tapauksista sai arvokseen arvosanan 4. GPT-4o-malli onnistui arvioimaan arvosanojen 5 arvoisista tapauksista oikein noin 39 % tapauksista. GPT-4o Mini-mallilla tämä osuus jäi kuitenkin vain noin 20 % tapauksista ja GPT-3.5 Turbo onnistui arvioimaan arvosanalla 5 oikein vain noin 6 % tapauksista.



Kuva 6.3: Tapahtuman onnistumisen arviointi kahteen luokkaan

Mallien antamat arvosanat ja niiden lämpökartat kuvassa 6.2 osoittavat, että GPT-malleja voidaan soveltaa käytettäväksi erilaisiin arviointitehtäviin. Vaikka arvosanat eivät ole osumatarkkuuksiltaan erittäin hyviä lämpökartoilla, voidaan kuitenkin todeta, että mallit eivät anna arvosanoja sattumanvaraisesti, vaan ne pystyvät antamaan järkeviä ja suuntaa antavia arvosanoja hyväksyttävällä tarkkuudella käytetystä mallista riippuen. Tuloksien perusteella voidaan myös todeta, että arvosanojen välille 1-5 perustuva arvionti ei välttämättä ole kielimalleille parhaiten soveltuva tapa antaa arvosanoja asiakaspalvelussa tapahtuvista tilanteista.

Tehtävä arviointityö voi olla tarkempaa ja malleille mahdollisesti helpompaa, jos tilanteita luokitellaan numeroiden sijaan esimerkiksi joko epäonnistuneiksi, vain osittain onnistuneiksi tai täysin onnistuneiksi. Kuvassa 6.3 mallien tuottamat numeroarvosanat on muotoiltu ja jaettu uudelleen jakamalla matalat (arvosanat alle 3) ja korkeat arvosanat (arvosanat yli 3) kahteen eri luokkaan. Kuvan 6.3 arvojen perusteella voidaan todeta, että hieman yksinkertaisempi mahdollisten arvosanojen jakauma voisi parantaa mallien tuottamien arviointien tulosten tarkkuutta entisestään ja tukea generoitujen tekstiarvioiden muodostumista numeroiden sijasta inhimillisemmiksi ja järkevämmiksi lopullista työelämän arjen käyttöä varten.

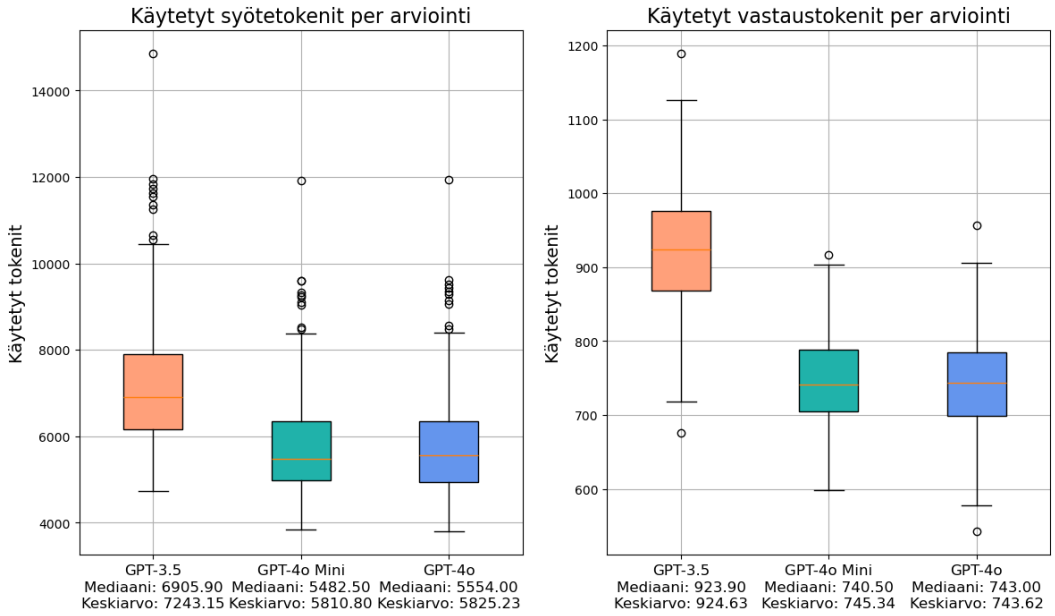
Mallin valinnan vaikutukset kustannuksiin

Käytettävien GPT-kielimallien eri versioiden valinnalla on vaikutus kustannuksiin Microsoftin Azure OpenAI palvelussa. Mallit on hinnoiteltu eri tavoin tilaustyypistä, kielimallien rajoituksista ja käyttöasteesta riippuen. Alla olevassa taulukossa 6.4 on listattu esimerkkinä erilaisten GPT-mallien käyttökustannuksia tuhatta tokenia kohden. Kustannukset muodostuvat käytetyistä tokeneista, kun käytössä olevalle mallille syötetään tokeneita kehotteiden muodossa ja se tuottaa vastauksen takaisin myös tokeneina. [59]

Mallit	**Konteksti**	**Syötteen hinta** (€/1K)	**Vastauksen hinta** (€/1K)
GPT-4o	128K	0.0045	0.0135
GPT-4o Mini	128K	0.000149	0.0006
GPT-4 Turbo	128K	0.009	0.027
GPT-4	32K	0.054	0.108
GPT-3.5 Turbo	16K	0.0005	0.0014

Taulukko 6.4: Azure OpenAI-palvelun mallien hinnasto syyskuussa 2024 [59]

Mallia käytettäessä tokeneiden määrällä ei ole niin paljon merkitystä, jos mallia on silti edullista käyttää suurellakin määrällä tokeneita. Azure OpenAI-palvelussa saa-



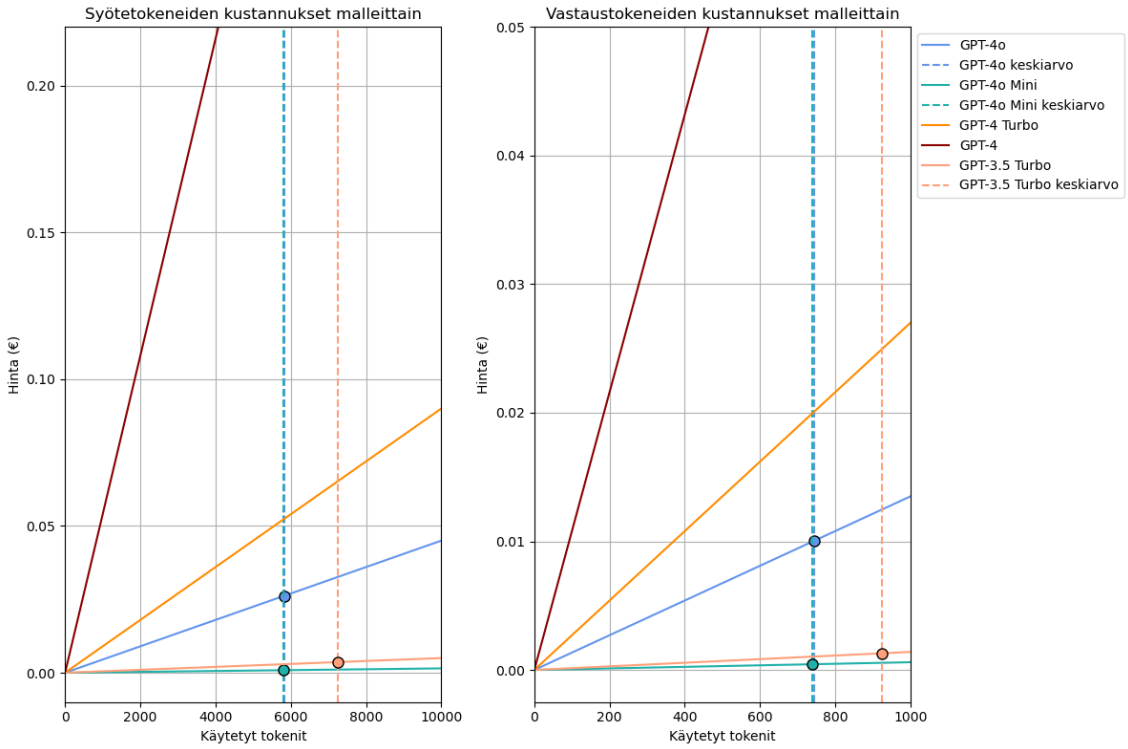
Kuva 6.4: Mallien tokeneiden käyttömääriä

tavilla olevat mallit voivat erota kustannuksiltaan paljonkin jos ja kun käytettyjen tokeneiden määrä kasvaa nopeasti. Mallien keskimääräisiä tokeneiden käyttömääriä kuvataan laatikko-janakuvioiden avulla kuvassa 6.4. Molemmat GPT-4o-mallit käyttivät niin syötteissään kuin vastauksissaan lähes täysin saman verran tokeneita, kun taas vanhempi GPT-3.5 Turbo-malli käytti noin 24 % enemmän tokeneita.

Työn toteutuksessa yhden arviointitapahtuman kustannus oli usein yhteensä yli kuusi tuhatta tokenia, kuten kuvasta 6.4 voidaan havaita. Käytettyjen tokeneiden määrä nousee korkeaksi toteutuksen laatiessa päivittäin satoja tai tuhansia arviointeja. Tämän takia on tärkeää valita malli, joka on tokeneiden käytön kannalta mahdollisimman edullinen, mutta tuottaa samalla riittävän tarkkoja arviointeja.

Malli	**Syöte (€)**	**Vastaus (€)**	**Kokonaiskustannus (€)**
GPT-3.5 Turbo	0.003622	0.001294	0.004915
GPT-4o Mini	0.000858	0.000444	0.001302
GPT-4o	0.026212	0.010044	0.036256

Taulukko 6.5: Mallikohtaiset keskiarvokustannukset per tapahtuman arvio



Kuva 6.5: Mallien kustannukset käyttömäärän mukaan

Kuvan 6.5 kuvaajien avulla voidaan havaita taulukkoa paremmin, kuinka suuria kustannuseroja eri Azure OpenAI-mallien välillä voi syntyä käytettyjen tokeneiden määrän kasvaessa. Vanhemmat mallit kuten GPT-4 ja GPT-4 Turbo ovat selvästi kustannuksiltaan joukon kalleimmat, vaikka vielä vanhempi GPT-3.5 Turbo-malli on kustannuksiltaan puolestaan todella edullinen. Työssä käytetyt GPT-4o ja GPT-4o Mini ovat kustannuksiltaan myös edullisia. GPT-4o on jatkokehitetty versio GPT-4 Turbosta ja sen on tarkoitus olla nopeampi ja 50 % edullisempaa käyttää kuin GPT-4 Turbo, mutta samalla myös ymmärtää ja tuottaa paremmin erilaisia kieliä, tekstin muotoilua sekä kuvaa ja ääntä [60]. Tämä pitää paikkansa, sillä kuvaajien pisteitä tarkastellessa GPT-4o on kustannuksiltaan puolet GPT-4 Turbon kustannuksista. Taulukon 6.5 keskiarvokustannukset ja kuvaajan 6.5 perusteella voidaan todeta GPT-4o Mini-mallin olevan malleista kaikkein edullisin merkittävällä erolla. GPT-3.5 Turbon yhden tuotetun arvion kustannukset olivat noin 3,7-kertaa suuremmat ja GPT-4o-mallin kustannukset olivat puolestaan noin 27,8-kertaa suuremmat kuin GPT-4o Mini-mallilla.

7 Johtopäätökset ja pohdinta

Tämän työn tulokset osoittavat, että GPT-kielimallit ovat potentiaalisia ja käyttökelpoisia asiakaspalvelun tilanteiden arvioiden ja palautteiden tuottamisen automatisoinnissa. Azure OpenAI-palvelussa saatavilla olevista malleista GPT-3.5 Turbo, GPT-4o Mini ja GPT-4o-mallit olivat tutkimuksen kohteina kehitettävää työkalua varten. Työkalun arkkitehtuurin jokaisen osan GPT-mallilla oli erilainen näkökulma arviointitehtävässä. Näkökulmat ja arviointitehtävät muotoiltiin jokaiselle osalle erikseen kehotteita muotoilemalla. Mallit generoivat vastauksia osissa, minkä jälkeen mallien tuottamat vastaukset voitiin koostaa lopullisiksi tiivistetyiksi tapahtumien arvioiksi.

Asiakaspalvelun tilanteiden ja kategorioiden tunnistamista käytettiin mallien suorituskyvyn mittarina tapahtumissa esiintyvien kategorioiden binääriluokittelun avulla. Luokittelun tulosten arviointiin käytettiin erilaisia mittareita, joiden arvojen perusteella arvioitiin mallien kykyä havaita ja ymmärtää keskusteluissa esiintyviä kategorioita ja niihin liittyviä tilanteita. Tämän lisäksi mallien arviointikykyjä arvioitiin niiden tuottamien numeroarvosanojen tarkkuuden ja jakautumisen perusteella arvonsanojen lämpökarttojen avulla.

Hinnoittelua vertailtiin käytettyjen tokeneiden keskiarvojen perusteella, jotta mallien käytöstä aiheutuvia kustannuksia ja kannattavuutta voitiin arvioida. Suuremmilla käyttömäärillä mallien kustannustehokkuus korostuu erityisesti silloin, kun käytettyjen tokenien kustannukset ja mallin suorituskyky ovat tasapainossa. Työn havaintojen ja tuloksien perusteella löydettiin GPT-mallien toimintaan liittyviä rajoituksia sekä erilaisia jatkokehityksen kohteita lopullista tuotannollista työkalua varten.

7.1 Vastaukset tutkimuskysymyksiin

Voiko GPT-kielimalleilla suorittaa asiakaspalvelun tilanteiden arviointia?

GPT-kielimallit pystyvät suorittamaan objektiivista arviointia syöteteksteistään tarkan kehotteiden muotoilun avulla. Generoitujen arvioiden tasalaatuisuuden ja objektiivisuuden kannalta on tärkeää, että arvioinnin eri tehtävät ja näkökulmat on jaettu erillisiksi malleille syötettäviksi kehotteiksi. Tilanteista saatavan tekstinformaation ja kyseistä informaatiota pilkkovien ja jalostavien kehotteiden yhdistelmän avulla mallit pystyvät tuottamaan tasavertaisia ja joustavia arviointeja sekä antamaan kirjallista palautetta tilannekohtaisesti erilaisista näkökulmista.

Kuinka tarkasti ja luotettavasti GPT-kielimallit pystyvät tunnistamaan ja erittelemään erilaisia asiakaspalvelun tilanteita keskusteluista?

GPT-mallit pystyvät tulkitsemaan oikein useimpia suomalaisessa asiakaspalvelussa tapahtuvia tilanteita käytetystä palvelukielestä riippumatta. Joidenkin keskusteluissa esiintyyvien vaikeiden ja monitulkintaisten kategorioiden kohdalla tulokset olivat käytetystä mallista riippuen vaihtelevia. Suurin osa kategorioista oli onnistuneesti tunnistettu kohtuullisen hyvällä tai erinomaisella tarkkuudella, mutta pieni osa kategorioista jäi satunnaisen arvauksen tasolle etenkin GPT-3.5 Turbo-mallia käytettäessä. Molemmat GPT-4o-mallit suoriutuivat tilanteiden tulkinnassa huomattavasti paremmin.

Kategoriakohtaisesti kaikki GPT-mallit suoriutuivat parhaiten myyntitilanteisiin yhdistettävien kategorioiden tulkitsemisen kanssa. Keskusteluiden numeroarvosanoja tuottaessa malleilla oli selvästi enemmän vaikeuksia. Arvosanoissa oli hajontaa etenkin niissä asiakaspalvelun keskusteluissa, joissa asiakaspalvelu oli onnistunut vain osittain. Selvästi onnistuneiden ja epäonnistuneiden asiakaspalvelun tilanteiden arvosanoja mallit arvioivat kuitenkin suuntaa-antavammin ja tarkemmin.

Millaisia kustannuksia arviointityökalun käyttämisesta muodostuu ja onko työkalun käyttäminen kannattavaa suuressa mittakaavassa?

Työkalun käytön kasvaessa arvioitavien rivien määrä päivässä voi helposti kasvaa moneen tuhanteen, minkä takia on tärkeä valita kustannustehokkaita malleja arviointitehtävän suorittamista varten. Valittujen mallien mahdollisimman korkea kustannustehokkuus ei kuitenkaan ole tuotettujen arviointien laadun kannalta välttämättä paras vaihtoehto, vaan tärkeintä on löytää tasapaino kustannusten ja arviointien laadun välillä.

Azure OpenAI-palvelussa saatavilla olevista malleista GPT-4o oli arvioinnissa kaikista laadukkain ja kyvykkäin, mutta samalla GPT-4o Mini oli kustannuksiltaan edullisin malli merkittävällä erolla kaikkiin muihin malleihin. Vaikka GPT-4o Minin arviointikyvykkyys ei ollut GPT-4o-mallin tuloksien tasolla, sen edullisuus mahdollistaa arviointitehtävän jakamisen tarvittaessa entistä pienempiin ja tarkempiin osatehtäviin, jotka voidaan muotoilla malleille uusiksi kehotteiksi ilman, että kustannukset nousevat merkittävästi korkeammaksi työkalun alkuperäiseen arkkitehtuuriin verrattuna. Tämä mahdollistaa myös joustavuutta, koska työkalun arkkitehtuurin mallien tekemää arviointiprosessia voidaan hallita ja kohdistaa vain tiettyihin osatehtäviin arvioinnin tarpeiden mukaan.

Tämän lisäksi eri mallien käyttäminen arvioinnin eri osatehtäviin on myös vaihtoehto arvioinnin laadun ja kustannustehokkuuden optimointiin, jos esimerkiksi pelkillä GPT-4o Mini-malleilla tuotetut arviot eivät ole enää jossain vaiheessa laadultaan riittäviä. Suuremman mittakaavan käyttöä ja aikaväliä arvioidessa esimerkiksi kuukauden ja tuhansien päiväkohtaisten arvioiden kohdalla mallien kustannukset vaihtelevat edullisimmillaan kymmenistä euroista moniin satoihin euroihin kuukaudessa.

Mikä GPT-malleista on soveltuvin arviointitehtävää varten?

Saatavilla olevista malleista GPT-4o Mini-mallin arviointikyky, generoitujen arviointien tekstin laatu ja edulliset kustannukset ovat erinomaisen hyvin tasapainossa, minkä takia se soveltuu tutkituista GPT-malleista selkeästi parhaiten suuren mittakaavan arviointien tuottamiseen. GPT-4o voi kuitenkin olla parempi vaihtoehto tulevaisuudessa, jos jossain vaiheessa sen käyttökustannukset muuttuvat huomattavasti nykyisiä edullisemmiksi.

Mitkä ovat keskeisimmät eettiset, tekniset ja käytännön haasteet, jotka liittyvät GPT-kielimallien hyödyntämiseen asiakaspalvelun arvioinnissa?

Asiakaspalvelun keskusteluiden teksteissä esiintyy usein arkaluonteisia tietoja, jotka kuuluvat Euroopan unionin tietosuoja-asetusten piiriin. GPT-malleja käytettäessä onkin siis erityisen tärkeää varmistaa, että keskusteluissa esiintyvät henkilötiedot anonymisoidaan ennen kuin ne syötetään malleille, koska niillä ei ole merkitystä mallien tuottamiin arviointeihin. Lisäksi on tärkeää huomioida Euroopan unionin tekoälyasetuksen asettamat rajoitukset ja kiellot tekoälyjärjestelmän toimintaan liittyen.

Jatkuvalla syötöllä tapahtuva automatisoitu arviointiprosessi voi asiakaspalvelun työntekijöiden mielestä tuntua tilanteelta missä heillä ei olisi yksityisyyttä tai varaa tehdä virheitä valvonnan alla. Valvonta voi lisätä suorituspaineita ja stressiä, mitkä puolestaan voivat heikentää työn sujuvuutta etenkin asiakaspalvelutilanteissa ja muutenkin työssä jaksamista. Siksi onkin erityisen tärkeää, että asiakaspalvelun esihenkilöt ja agentit ovat kaikki tietoisia ja yhteisymmärryksessä siitä, että automatisoidun arvioinnin tarkoituksena on tukea heidän kehittymistään.

Arvioinnin kannalta on tärkeää huomioida, että kielimallit eivät ole puolueettomia, ja ne sisältävät niiden koulutusaineistoista opittuja vääristymiä. Poikkeuksellisen asiakaspalvelun tilanteen kohdatessaan mallit voivat tuottaa virheellisiä tulkintoja, jotka voivat suoraan heijastua arviointiin. Asiakaspalvelijoiden kirjoitustyyli, keskustelun aiheet sekä palvelukieli voivat joissain tapauksissa vaikuttaa tilanteen arvioon ja arvosanaan, mitkä tietenkin heikentävät arvioiden oikeudenmukaisuutta ja objektiivisuutta.

Tarkkojen ohjekehotteiden, erilaisten muotoilumenetelmien ja niiden sanoituksien kautta mallien käyttäytymistä voidaan kuitenkin parannella toivottuun suuntaan. Näin mallit voidaan saada ymmärtämään tiettyjä tilanteita paremmin ja tuottamaan parempia, tasalaatuisia ja luotettavampia arvioita. Malleille syötettävää kehotetta ei kuitenkaan voida parannella tai kasvattaa loputtomiin kasvavien kustannusten sekä mallien omien käyttörajojen takia, minkä seurauksena jotkin asiakaspalvelun tilanteiden kategoriat voivat olla malleille erittäin vaikeita tulkita ja arvioida oikein.

7.2 Jatkotyöt

Nykyisessä muodossaan työkalun jatkokehityksen kannalta voi olla kannattavampaa vaihtaa GPT-mallien tekemän numeroarvioinnin arvosanat sanalliseen muotoon esimerkiksi onnistuneiden, parannusta vaativien ja epäonnistuneiden suoritusten väliseksi luokittelutehtäväksi, koska GPT-mallit ja myös muutkin kielimallit pystyvät toimintatapansa takia arvioimaan ja kuvailemaan asiakaspalvelun tilanteita paremmin sanallisin arvosanoin.

Arvioinnin kannalta toisena vaihtoehtona jatkotöille on selkeämpi pisteytysjärjestelmä arvosanojen tuottamista varten. Käyttökohteesta riippuen mallien tarkoituksena olisi kehotteissä määritettyjen tehtävien mukaan tunnistaa tiettyjä tapahtumia tai vaiheita teksteistä ja kirjata niitä ylös esimerkiksi totuusarvojen avulla. Koko keskustelun aikana kertyneet pisteet voitaisiin tämän jälkeen suhteuttaa suurimpaan mahdolliseen pistemäärään. Tämä arviointitapa selventäisi arvosanojen muodostumista, mutta se vaatii selkeästi ja monipuolisesti määritetyt kriteerit ja pisteytyksen kehotteita varten, mitä ei välttämättä ole käytettävissä tai ne eivät ole sopivia sovellettavaksi kehotteissa.

Työkalun arkkitehtuurin rakenteen muokattavien osien takia työkalu on sovellettavissa muihin tekstin analysointiin liittyviin käyttötarkoituksiin joustavasti kielimalleille syötettäviä tietoja, kielimalliketjuja ja kehotteita muokkaamalla. Tulevien ja kehittyneempien suurten kielimallien avulla voidaan puolestaan jatkossa parantaa generoitujen arviointien tarkkuutta ja laatua vielä entisestään. Toisaalta nykyisten eri mallien yhdistelmien käyttäminen eri tehtävissä, kuten edullisien mallien käyttäminen yksinkertaisissa ja kalliimpien monimutkaisemmissa osatehtävissä, voi myös optimoida kustannuksia ja parantaa työkalun tuottamia arvioita jatkossa. Jatkokehittäessä työkalua on syytä huomioida ajantasaiset Euroopan unionin tekoälyasetuksen määrittämät oikeudelliset kehykset, riskit ja ehdottomat kiellot tekoälyjärjestelmiin liittyen. Ne voivat vaikuttaa siihen, mitä tekoälyjärjestelmän toimintoja voidaan kehittää laajemmiksi, mitä toimintoja pitää muuttaa säädöksen mukaisiksi tai mitä toimintoja pitää lopettaa.

Lähdeluettelo

[1] A. M. Turing, *Computing machinery and intelligence*. Springer, 2009, s. 23–24.

[2] W. J. Hutchins, "The Georgetown-IBM experiment demonstrated in January 1954", teoksessa *Conference of the Association for Machine Translation in the Americas*, Springer, 2004, s. 102–114.

[3] J. Weizenbaum, "ELIZA—a computer program for the study of natural language communication between man and machine", *Communications of the ACM*, vol. 9, nro 1, s. 36–45, 1966.

[4] T. Winograd, "What does it mean to understand language?", *Cognitive science*, vol. 4, nro 3, s. 209–241, 1980.

[5] W. N. Francis ja H. Kucera, "Brown corpus manual", *Letters to the Editor*, vol. 5, nro 2, s. 7, 1979.

[6] N. Ide ja C. Macleod, "The american national corpus: A standardized resource of american english", teoksessa *Proceedings of corpus linguistics*, Lancaster University Centre for Computer Corpus Research on Language ..., vol. 3, 2001, s. 1–7.

[7] T. Mikolov, K. Chen, G. Corrado ja J. Dean, "Efficient estimation of word representations in vector space", *arXiv preprint arXiv:1301.3781*, 2013.

[8] H. Naveed, A. U. Khan, S. Qiu et al., "A comprehensive overview of large language models", *arXiv preprint arXiv:2307.06435*, 2023.

[9] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention is all you need", *Advances in neural information processing systems*, vol. 30, 2017.

[10] K. S. Kalyan, "A survey of GPT-3 family large language models including ChatGPT and GPT-4", *Natural Language Processing Journal*, s. 100 048, 2023.

[11] T. Wolf, "Transformers: State-of-the-Art Natural Language Processing", *arXiv preprint arXiv:1910.03771*, 2020.

[12] S. Feuerriegel, J. Hartmann, C. Janiesch ja P. Zschech, "Generative ai", *Business & Information Systems Engineering*, vol. 66, nro 1, s. 111–126, 2024.

[13] H. S. Sætra, "Generative AI: Here to stay, but for good?", *Technology in Society*, vol. 75, s. 102 372, 2023.

[14] NVIDIA. "Generative AI". (2023), url: `https://web.archive.org/web/20241226201135/https://www.nvidia.com/en-us/glossary/generative-ai/` (viitattu 21. 11. 2023).

[15] A. Lee, "What Are Large Language Models Used For?", *NVIDIA Blog*, tammikuu 2023. url: `https://web.archive.org/web/20241222222910/https://blogs.nvidia.com/blog/what-are-large-language-models-used-for/` (viitattu 21. 11. 2023).

[16] J. Devlin, M.-W. Chang, K. Lee ja K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding", *arXiv preprint arXiv:1810.04805*, 2018.

[17] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever et al., "Improving language understanding by generative pre-training", 2018.

[18] J. Wu, W. Gan, Z. Chen, S. Wan ja P. S. Yu, "Multimodal large language models: A survey", *arXiv preprint arXiv:2311.13165*, 2023.

[19] J. Achiam, S. Adler, S. Agarwal et al., ”GPT-4 Technical Report”, *arXiv preprint arXiv:2303.08774*, 2023.

[20] R. C. King, V. Bharani, K. Shah, Y. H. Yeo ja J. S. Samaan, ”GPT-4V passes the BLS and ACLS examinations: An analysis of GPT-4V’s image recognition capabilities”, *Resuscitation*, vol. 195, 2024.

[21] W. S. El-Kassas, C. R. Salama, A. A. Rafea ja H. K. Mohamed, ”Automatic text summarization: A comprehensive survey”, *Expert systems with applications*, vol. 165, s. 113 679, 2021.

[22] M. Lewis, Y. Liu, N. Goyal et al., ”BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension”, *arXiv preprint arXiv:1910.13461*, 2020.

[23] C. Raffel, N. Shazeer, A. Roberts et al., ”Exploring the limits of transfer learning with a unified text-to-text transformer”, *Journal of Machine Learning Research*, vol. 21, nro 140, s. 1–67, 2020.

[24] M. Wankhade, A. C. S. Rao ja C. Kulkarni, ”A survey on sentiment analysis methods, applications, and challenges”, *Artificial Intelligence Review*, vol. 55, nro 7, s. 5731–5780, 2022.

[25] W. Zhang, Y. Deng, B. Liu, S. J. Pan ja L. Bing, ”Sentiment analysis in the era of large language models: A reality check”, *arXiv preprint arXiv:2305.15005*, 2023.

[26] P. F. Simmering ja P. Huoviala, ”Large language models for aspect-based sentiment analysis”, *arXiv preprint arXiv:2310.18025*, 2023.

[27] C. M. Pham, A. Hoyle, S. Sun, P. Resnik ja M. Iyyer, ”Topicgpt: A prompt-based topic modeling framework”, *arXiv preprint arXiv:2311.01449*, 2023.

[28] Y. Mu, C. Dong, K. Bontcheva ja X. Song, "Large Language Models Offer an Alternative to the Traditional Approach of Topic Modelling", *arXiv preprint arXiv:2403.16248*, 2024.

[29] Y. Bengio, N. Boulanger-Lewandowski ja R. Pascanu, "Advances in optimizing recurrent networks", teoksessa *2013 IEEE international conference on acoustics, speech and signal processing*, IEEE, 2013, s. 8624–8628.

[30] A. H. Ribeiro, K. Tiels, L. A. Aguirre ja T. Schön, "Beyond exploding and vanishing gradients: analysing RNN training using attractors and smoothness", teoksessa *International conference on artificial intelligence and statistics*, PMLR, 2020, s. 2370–2380.

[31] S. Hochreiter, Y. Bengio, P. Frasconi, J. Schmidhuber et al., *Gradient flow in recurrent nets: the difficulty of learning long-term dependencies*, 2001.

[32] Y. Liu, T. Han, S. Ma et al., "Summary of ChatGPT-related research and perspective towards the future of large language models", *Meta-Radiology*, s. 100 017, 2023.

[33] T. Lin, Y. Wang, X. Liu ja X. Qiu, "A survey of transformers", *AI open*, vol. 3, s. 111–132, 2022. DOI: `https://doi.org/10.1016/j.aiopen.2022.10.001`.

[34] OpenAI, *GPT-4 System Card*, 2023. url: `https://cdn.openai.com/papers/gpt-4-system-card.pdf` (viitattu 23. 11. 2024).

[35] P. Gage, "A new algorithm for data compression", *The C Users Journal*, vol. 12, nro 2, s. 23–38, 1994.

[36] R. Sennrich, B. Haddow ja A. Birch, "Neural machine translation of rare words with subword units", *arXiv preprint arXiv:1508.07909*, 2015.

[37] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg ja D. Amodei, "Deep reinforcement learning from human preferences", *Advances in neural information processing systems*, vol. 30, 2017.

[38] R. Dale, "GPT-3: What's it good for?", *Natural Language Engineering*, vol. 27, nro 1, s. 113–118, 2021.

[39] A. Tamkin, M. Brundage, J. Clark ja D. Ganguli, "Understanding the capabilities, limitations, and societal impact of large language models", *arXiv preprint arXiv:2102.02503*, 2021.

[40] T. Brown, B. Mann, N. Ryder et al., "Language models are few-shot learners", *Advances in neural information processing systems*, vol. 33, s. 1877–1901, 2020.

[41] I. Santouridis ja A. Veraki, "Customer relationship management and customer satisfaction: the mediating role of relationship quality", vol. 28, s. 1122–1133, 2017. DOI: 10.1080/14783363.2017.1303889.

[42] C. Gronroos, "From marketing mix to relationship marketing: Towards a paradigm shift in marketing", *Asia-Australia Marketing Journal*, vol. 2, nro 1, s. 9–29, 1994.

[43] B. Cheng, Y. Dong, X. Zhou, G. Guo ja Y. Peng, "Does customer incivility undermine employees' service performance?", *International Journal of Hospitality Management*, vol. 89, s. 102 544, 2020.

[44] L. Marchegiani, T. Reggiani ja M. Rizzolli, "Loss averse agents and lenient supervisors in performance appraisal", *Journal of Economic Behavior & Organization*, vol. 131, s. 183–197, 2016.

[45] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo ja Y. Iwasawa, "Large language models are zero-shot reasoners", *Advances in neural information processing systems*, vol. 35, s. 22 199–22 213, 2022.

[46] J. Wei, M. Bosma, V. Y. Zhao et al., "Finetuned language models are zero-shot learners", *arXiv preprint arXiv:2109.01652*, 2021.

[47] J. Wei, Y. Tay, R. Bommasani et al., "Emergent abilities of large language models", *arXiv preprint arXiv:2206.07682*, 2022.

[48] H. W. Chung, L. Hou, S. Longpre et al., ”Scaling instruction-finetuned language models”, *Journal of Machine Learning Research*, vol. 25, nro 70, s. 1–53, 2024.

[49] H. Touvron, T. Lavril, G. Izacard et al., ”Llama: Open and efficient foundation language models”, *arXiv preprint arXiv:2302.13971*, 2023.

[50] J. Wei, X. Wang, D. Schuurmans et al., ”Chain-of-thought prompting elicits reasoning in large language models”, *Advances in neural information processing systems*, vol. 35, s. 24 824–24 837, 2022.

[51] J. Huang, S. S. Gu, L. Hou et al., ”Large language models can self-improve”, *arXiv preprint arXiv:2210.11610*, 2022.

[52] European Parliament ja Council of the European Union, *Regulation (EU) 2016/679 of the European Parliament and of the Council*, of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), Article 4 & Article 22, 4. toukokuuta 2016. url: `https://eur-lex.europa.eu/legal-content/FI/TXT/?uri=CELEX:32016R0679` (viitattu 11. 08. 2024).

[53] *Työsopimuslaki (55/2001)*, 1 luku 1 § Soveltamisala. url: `https://www.finlex.fi/fi/laki/ajantasa/2001/20010055` (viitattu 11. 08. 2024).

[54] *Työsopimuslaki (55/2001)*, 2 luku 2 § Tasapuolinen kohtelu ja syrjintäkielto (30.12.2014/1331). url: `https://www.finlex.fi/fi/laki/ajantasa/2001/20010055` (viitattu 11. 08. 2024).

[55] *Laki yksityisyyden suojasta työelämässä (759/2004)*, 7 luku 21 § Yhteistoiminta teknisin menetelmin toteutetun valvonnan ja tietoverkon käytön järjestämisessä. url: `https://www.finlex.fi/fi/laki/ajantasa/2004/20040759` (viitattu 24. 10. 2024).

[56] *Euroopan parlamentin ja neuvoston asetus (EU) 2024/1689*, https://eur-lex.europa.eu/legal-content/FI/TXT/?uri=CELEX:32024R1689, 2024. (viitattu 09. 02. 2025).

[57] M. L. McHugh, "Interrater reliability: the kappa statistic", *Biochemia medica*, vol. 22, nro 3, s. 276–282, 2012.

[58] D. Chicco ja G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation", *BMC genomics*, vol. 21, s. 1–13, 2020.

[59] "Azure OpenAI Service Pricing". (2024), url: https://web.archive.org/web/20241001162753/https://azure.microsoft.com/en-us/pricing/details/cognitive-services/openai-service/ (viitattu 24. 09. 2024).

[60] OpenAI, *GPT-4o System Card*, 2024. url: https://web.archive.org/web/20241004003638/https://openai.com/index/gpt-4o-system-card/ (viitattu 01. 10. 2024).