Puhelinasiakaspalvelun laadun arviointi tekoälyn avulla

Christel Baker

Haaga-Helia ammattikorkeakoulu
Tradenomi, Tietojenkäsittelyn koulutusohjelma
Opinnäytetyö
2024

Tiivistelmä

Tekijä(t)
Christel Baker

Tutkinto
Tradenomi

Raportin/Opinnäytetyön nimi
Puhelinasiakaspalvelun laadun arviointi tekoälyn avulla.

Sivu- ja liitesivumäärä
32 + 0

Opinnäytetyö toteutettiin syksyllä 2024 toimeksiantona 020202 Palvelut Oy:lle, joka on keskisuuri kotimainen yritys ja tunnetaan parhaiten sen kuluttajille tarjoamista nimi- ja numeropalveluista. Työn tavoitteena oli kehittää ratkaisu, jonka avulla toimeksiantaja voi tehostaa puhelinasiakaspalvelun laadun arviointia tekoälyn avulla, mikä mahdollistaisi palveluarvioinnin tekemisen useammin, sekä suuremmalle määrälle puheluita. Opinnäytetyön kohteeksi valittiin taksipalvelut-yksikön puhelinasiakaspalvelu. Työssä keskityttiin nykytilanteen kartoitukseen, puheluiden litterointiin, aineiston anonymisointiin, tekoälykehotteen suunnitteluun sekä tekoälymallien vertailuun. Opinnäytetyössä ei käsitelty opinnäytetyön ohessa toteutettavan sovelluksen kehittämistä. Työssä käytettiin OpenAI:n Whisper- ja ChatGPT-tekoälymalleja, koska haluttiin hyödyntää yrityksen olemassa olevia resursseja.

Työn keskeisenä tietoperustana käytettiin OpenAI:n omaa dokumentaatiota, sekä alan artikkeleita ja aikaisempia tutkimuksia aiheesta. Tärkeää tietoa palveluarviointiprosessista, sekä työssä tarvittavaa aineistoa saatiin toimeksiantajalta.

Opinnäytetyöraportti toteutettiin vetoketjumallilla, sillä työ koostuu useasta eri osioista. Jokaisessa osiossa on oma tietoperusta, toteutus sekä pohdintaosio. Opinnäytetyön viimeisessä luvussa pohditaan opinnäytetyön onnistumista kokonaisuutena

Työssä litteroitiin toimeksiantajalta saatuja asiakaspalvelupuheluita Whisper-mallin avulla. Litteroiduista puheluista laskettiin sanavirheet, ja tuloksia verrattiin aikaisempiin tutkimuksiin. Tekoälykehote luotiin palveluarviointilomakkeen ja asiakaspalvelun ohjeistusten pohjalta. ChatGPT:tä pyydettiin luomaan palveluarvio, ja sen luomaa arviota verrattiin manuaalisesti tuotettuun palveluarvioon.

Opinnäytetyön kaikkia suunniteltuja vaiheita ei ehditty toteuttaa. Opinnäytetyöstä saatiin kuitenkin paljon uutta tietoa, ja sitä voidaan hyödyntää ratkaisun jatkokehityksessä. Puhelutallenteiden litterointi onnistui hyvin, ja tulokset olivat verrattavissa aikaisempaan tutkimukseen. Tekoälykehotteen suunnittelussa aika riitti vain yhden iteraation toteuttamiseen. Tietoperustan ja empirian avulla saatiin hyvä käsitys siitä, miten tekoälykehotetta ja arviointikriteerejä tulisi kehittää, jotta tekoälyn tuottama palveluarvio vastaisi manuaalisesti tuotettuja. Aineiston anonymisointia ja kielimallien vertailua ei ehditty toteuttamaan. Työn tulokset esiteltiin toimeksiantajalle lokakuussa 2024. Työn tärkeimpänä antina saatiin vahvistus siitä, että tekoäly pystyy tuottamaan palveluarviointeja, mikä avaa uusia mahdollisuuksia palvelun laadun arvioinnin tehostamiselle.

Asiasanat
Tekoäly, puheentunnistus, ChatGPT, Whisper

Sisällys

1 Johdanto

Forbes Advisorin mukaan asiakaskokemuksen laiminlyönti vaikuttaa negatiivisesti yhtiön tuottavuuteen ja maineeseen. Sen sijaan erinomainen asiakaspalvelu vahvistaa asiakkaiden uskollisuutta brändille, saa heidät palaamaan sekä lisää todennäköisyyttä, että he suosittelevat tuotetta tai palvelua läheisilleen. Hyvä asiakaspalvelu auttaa myös yritystä erottumaan kilpailijoistaan. (Iwuozor 2024.) Toisessa Forbesin artikkelissa kerrotaan, että yritykset ympäri maailman saattavat menettää jopa 3,7 biljoonaa dollaria vuodessa huonon asiakaspalvelun seurauksena. Tieto perustuu Qualtricsin vuoden 2023 kolmannella neljänneksellä tuottamaan tutkimuksen, johon osallistui noin 28400 kuluttajaa 26 eri maasta. Kyselyssä käsiteltiin asiakkaiden negatiivisia kokemuksia 20 eri toimialalla. Asiakaspalvelun laatu nousee kilpailuvaltiksi erityisesti silloin, kun vastaavaa tuotetta tai palvelua tarjoaa useampi eri yritys. (Hyken 2024.) Erinomainen asiakaspalvelu voi siis olla kilpailuvaltti, mutta palvelun laatua on seurattava ja arvioitava jatkuvasti, jotta voidaan tunnistaa kehityskohteet ja pysyä kilpailijoiden edellä. Tämän opinnäytetyön tavoitteena on selvittää, miten puhelinasiakaspalvelun laadunarvioimista voitaisiin tehostaa tekoälyn avulla.

1.1 Toimeksiantajan esittely

Opinnäytetyön toimeksiantaja, 020202 Palvelut Oy, on keskisuuri kotimainen yritys, joka tunnetaan parhaiten sen kuluttajille tarjoamista nimi- ja numeropalveluista. Yritys työllistää noin 80 työntekijää, joista suurin osa työskentelee asiakaspalvelussa. 020202 Palvelut Oy oli aiemmin osa Fonecta Oy:tä, mutta se on erkaantunut omaksi yrityksekseen vuonna 2018. Muita sisaryhtiöitä ovat 020202 Ratkaisut ja 02 Taksi. Yhtiön omistaa pääomasijoitusyhtiö Sponsor Capital Oy. Sisaryhtiöt tekevät keskenään paljon yhteistyötä, esimerkiksi 020202 Palvelut Oy hoitaa 02 Taksin puhelinpalvelun. (020202 Palvelut s.a.)

02 Taksi on valtakunnallinen taksintilauspalvelu, jossa kuluttajat voivat vertailla ja valita, minkä yhteistyökumppanin auton haluavat tilata. Palvelu on käytettävissä puhelimitse, tekstiviestillä tai mobiilisovelluksella. Mikäli tarkka lähtöosoite ja määränpää on tiedossa, voidaan palvelussa vertailla saatavilla olevia yhteistyökumppaneita saapumisajan, hinnan tai asiakasarvioiden perusteella. (02 Taksi s.a.) 02 Taksin tärkeimmät kilpailuvaltit ovat valtakunnallisuus, kyytivaihtoehtojen vertailumahdollisuus sekä laadukas asiakaspalvelu (020202 Palvelut 2021).

1.2 Projektin kuvaus ja tavoitteet

Opinnäytetyön tavoitteena on kehittää ratkaisu, jonka avulla toimeksiantaja voi tehostaa asiakaspalvelun laadun arviointia tekoälyn avulla. Nykytilassa esihenkilöt tekevät palveluarviointia manuaalisesti kuuntelemalla alaistensa puheluita ja pisteyttämällä ne lomakkeen avulla.

Palveluarviointiprosessia tehostamalla mahdollistetaan palveluarvioinnin tekeminen useammin, sekä suuremmalle määrälle puheluita, tai vaihtoehtoisesti voidaan vapauttaa resursseja muihin tehtäviin. Ratkaisuun on tarkoitus käyttää OpenAI:n ChatGPT- ja Whisper-malleja. Whisper-mallin avulla litteroidaan puhelu, jotta se voidaan syöttää ChatGPT:lle. ChatGPT:tä ohjeistetaan luomaan palveluarvio litteroidusta puhelusta. Opinnäytetyö toteutetaan vetoketjumallilla, sillä työ koostuu useasta eri osiosta. Jokaisessa osiossa on oma tietoperusta, toteutus sekä pohdintaosio. Opinnäytetyön viimeisessä luvussa pohditaan opinnäytetyön onnistumista kokonaisuutena.

Työn alussa on tärkeätä kartoittaa nykytilanne ja tutustua kohteena olevaan prosessiin, sekä selvittää, miten palveluarviot tällä hetkellä tehdään ja mitä mittareita niissä käytetään. Palveluarviointiprosessiin tutustuminen tapahtuu perehtymällä asiakaspalvelijoiden ohjeistukseen yrityksen intranetissä, sekä kuuntelemalla puheluita ja analysoimalla niistä tehtyjä arviointeja. Tämän työvaiheen aikana kerätyn tiedon avulla määritellään, kuinka tarkasti tekoälyn on suoriuduttava tehtävästä, jotta sen käyttö on hyödyllistä. Nykytilan kartoittaminen käsitellään luvussa 2.

Jotta aineisto voidaan prosessoida ChatGPT-mallin avulla, tullaan se ensin litteroimaan Whisper-mallia hyödyntäen. Litteroimisella tarkoitetaan, että nauhoitettu puhe muutetaan tekstimuotoon. Tämän työvaiheen tavoite on selvittää, miten litterointi Whisper-mallilla käytännössä toimii, sekä arvioida, kuinka hyvin Whisper-malli selviytyy suomenkielisten puheluiden litteroimisesta. Lisäksi selvitetään, minkälaisia haasteita luonnollisen puheen tunnistamisessa yleensä kohdataan. Näistä tiedoista on hyötyä seuraavien työvaiheiden suunnittelussa. Litterointi käsitellään luvussa 3.

Tietoturvan kannalta on mielenkiintoista selvittää, miten litteroidut puhelut voidaan anonymisoida. Anonymisointi tarkoittaa, että aineistossa esiintyvät henkilötiedot poistetaan tai muutetaan niin, että henkilöä ei voi enää tunnistaa aineistosta. Anonymisoinnin tarkoitus on varmistaa, että henkilötietoja ei päädy kolmansien osapuolien haltuun. Toimeksiantajalla on käytössä ChatGPT:n maksullinen versio, ja se on tehnyt OpenAI:n kanssa tietojenkäsittelysopimuksen, joten anonymisointi ei ole opinnäytetyön kannalta välttämätöntä. OpenAI:n (5.4.2023) mukaan ChatGPT:n ilmainen versio voi käyttää käyttäjien keskusteluja koulutusaineistona, joten ilmaisversiota käytettäessä anonymisointi olisi erittäin tärkeää. Anonymisointi käsitellään luvussa 4.

Työn aikana tutustutaan tekoälykehotteen suunnitteluun, ja pyritään luomaan tekoälylle kehote, jolla se tuottaa mahdollisimman korkealaatuisen palveluarvion annettujen kriteerien perusteella litteroidusta puhelusta. Lopuksi vertaillaan eri ChatGPT-mallien (3.5, 3.5 Turbo, 4.0) välisiä eroja, sekä arvioidaan, kuinka hyvin ne suoriutuvat palveluarvioiden tuottamisessa verrattuna manuaalisesti tehtyihin arvioihin. Opinnäytetyön yhteydessä kehitetään yrityksen sisäiseen käyttöön sovellus, jonka avulla esihenkilöt voivat syöttää puhelutallenteet ja saada palveluarviot automaattisesti

tekoälyn tuottamina. Tekoälykehotteen suunnittelu käsitellään luvussa 5 ja eri kielimallien vertailu luvussa 6.

Opinnäytetyön tavoitteita tarkastellaan myös vastuullisuuden näkökulmasta. Koska työssä käsitellään luottamuksellisia puhelutallenteita, tietoturva on keskeisessä asemassa. Tarvittaessa kaikki asiakastiedot anonymisoidaan, ja huolehditaan siitä, että henkilötietoja ei säilytetä tarpeettomasti tai välitetä kolmansille osapuolille. Palveluarvioiden avulla arvioidaan työntekijöiden työsuorituksia, joten on tärkeää varmistaa, että tekoälyn tuottamat palveluarviot ovat luotettavia.

1.3 Rajaukset

Opinnäytetyöraportissa keskitytään nykytilanteen kartoitukseen, puheluiden litterointiin, aineiston anonymisointiin, tekoälykehotteen suunnitteluun sekä tekoälymallien vertailuun. Raportissa ei käsitellä opinnäytetyön ohessa toteutettavan sovelluksen kehittämistä. Työn kohteeksi on valittu taksipalvelut-yksikön puhelinasiakaspalvelu. Puhelut valitaan esimerkiksi keston perusteella. Tarkemmat valintakriteerit sovitaan toimeksiantajan kanssa myöhemmin. Kielimallien vertailu rajoittuu yrityksessä jo käytössä oleviin ChatGPT -malleihin (3.5, 3.5 Turbo ja 4.0). Muita tekoälymalleja ei oteta mukaan, koska tavoitteena on löytää paras vaihtoehto yrityksen nykyisistä resursseista. Puheluiden litteroinnissa vertaillaan Whisper-mallin *small* ja *medium* versioita, jotta voidaan arvioida, vaikuttaako laajemman Whisper-mallin käyttäminen oleellisesti lopullisen palveluarvion laatuun.

Käytettäviä teknologioita valittaessa suositaan yrityksessä ennestään käytössä olevia teknologioita. Tämä varmistaa, että ratkaisun käyttöönotto on sujuvaa ja yhteensopivaa yrityksen IT-infrastruktuurin kanssa. Ohjelmointikielenä käytetään pythonia, puhelutallenteiden litterointiin käytetään OpenAI:n Whisper-mallia ja palveluarviot generoidaan OpenAI:n ChatGPT-mallin avulla.

1.4 Keskeiset käsitteet

Automaattinen puheentunnistus, automatic speech recognition system, ASR (Open AI 21.9.2022.)

ChatGPT, OpenAI:n kehittämä keskustelubotti, joka perustuu GPT-kielimalliin (OpenAI 30.11.2022).

Kehotesuunnittelu, prompt engineering, menetelmä, jonka avulla pyritään optimoimaan tekoälylle syötettävä kehote, jotta saataisiin mahdollisimman hyvälaatuinen vastaus.

Palveluarvio, toimeksiantajan sisäinen arviointimenetelmä, jolla tarkastellaan asiakaspalvelun laatua määriteltyjen kriteerien perusteella.

Reinforcement Learning from Human Feedback, RLHF, OpenAI:n kehittämä, vahvistusoppimiseen perustuva menetelmä, jolla koulutetaan tekoälyä (OpenAI 30.11.2022).

Sanavirhearvo, Word Error Rating, WER. Yleisin tapa mitata puheentunnistuksen laatua. Sanavirhearvossa huomioidaan väärin tunnistettujen sanojen lisäksi puuttuvat ja ylimääräiset sanat (Sallinen 2017, 29).

Suhteellinen sanavirheiden vähennys, $WERR_{A \rightarrow B}$, Relative Word Error Rate Reduction. Arvo, jolla kuvataan kahden eri mallin välistä suhteellista sanavirheiden vähennystä (Heikinheimo 2023). Voidaan laskea kaavalla $WERR_{A \rightarrow B} = (WER_A - WER_B) / WER_A$.

Tietoaineisto, dataset. ”Yksilöitävissä oleva kokoelma tietoja” (Tieteen termipankki 2024).

Whisper, Open AI:n kehittämä automaattinen puheentunnistusjärjestelmä (Open AI 21.9.2022).

2 Nykytilanteen kartoittaminen

Opinnäytetyön alussa on tärkeätä kartoittaa, miten palvelunarviointiprosessi yrityksessä toimii nykytilassa, jotta voidaan suunnitella seuraavia työvaiheita. Tämän vaiheen tavoitteena on selvittää, miten palvelua arvioidaan, ja mitä lopullinen palveluarvio sisältää.

2.1 Palveluarviointi 020202 Palvelut Oy:ssä

020202 Palvelut Oy:ssä asiakaspalvelun esihenkilöt tekevät palveluarvioita kuuntelemalla alaistensa puheluita. Palveluarviot tehdään manuaalisesti pisteyttämällä puhelut palveluarviointilomakkeen avulla. Palveluarviot ovat yritykselle erittäin tärkeitä, sillä niiden avulla seurataan palvelun laatua ja kehityskohteita.

Asiakaspalvelun esihenkilöt hyötyvät merkittävästi palveluarviointiprosessin automatisoinnista. Tekoälyn avulla on tarkoitus helpottaa heidän työtään nopeuttamalla ja automatisoimalla asiakaspalvelun laadun arviointiprosessia. Sen avulla voidaan seurata ja kehittää palvelun laatua entistä tehokkaammin, tunnistaa kehityskohteita ja helpottaa palautteen antamista työntekijöille. Sovelluksesta tulee olemaan konkreettista hyötyä yrityksen toiminnan tehostamisessa.

Nykytilan kartoituksessa tavoitteena on tutustua palveluarviointiprosessiin ja selvittää, miten palveluarviot tällä hetkellä tehdään ja mitä mittareita niissä käytetään. Lisäksi määritellään, kuinka tarkasti tekoälyn on suoriuduttava tehtävästä, jotta sen käyttö on hyödyllistä.

2.2 Nykytilanteen kartoituksen toteutus

Nykytilanteen kartoitus suoritettiin tutustumalla yrityksen sisäiseen dokumentointiin ja ohjeistukseen, sekä tutustumalla manuaalisesti tuotettuihin palveluarvioihin. Tarkennuksia ja lisätietoja saatiin tarvittaessa ICT-johtajalta ja palvelupäälliköltä. Työn kohteeksi on valittu taksipalvelut-yksikön puhelinasiakaspalvelu. Puhelut valitaan esimerkiksi keston perusteella. Tarkemmat valintakriteerit sovitaan toimeksiantajan kanssa myöhemmin.

2.2.1 Tiedonhankinta

Sovimme Teams-kokouksen ICT-johtajan ja palvelupäällikön kanssa pe 27.9.2024. Palaverissa esittelin työn palvelupäällikölle ja kävimme läpi työn aikataulua. Kyselin palvelupäälliköltä tarkentavia kysymyksiä palveluarviointiprosessista. Sain tietää, että yhdellä esihenkilöllä on alaisuudessaan keskimäärin 15 työntekijää. Palveluarviossa esihenkilö kuuntelee ja pisteyttää arviointilomakkeen avulla 10 asiakaspalvelupuhelua jokaista työntekijää kohti. Yhden henkilön palveluarvion (, 10 puhelua,) tekeminen kestää noin 45–60 minuuttia. Asiakaspalvelija voidaan palkita hyvistä arvioinneista.

Keskustelimme myös siitä, minkälaista aineistoa tarvitsen työtä varten. Palvelupäällikkö lupasi toimittaa minulle tyhjän palveluarviointilomakkeen, sekä puhelunauhoitteita, joista on tehty palveluarvio. Sovimme että hän toimittaa noin 10 kpl puhelunauhoitteitta, joista on annettu vaihtelevia pistemääriä. Päätimme luoda Teams-kanavan, jotta kaikki opinnäytetyöhön liittyvät materiaalit ja kysymykset löytyvät yhdestä paikasta.

Tutustuttuani arviointilomakkeeseen ja nauhoitteisiin, minulle heräsi lisää kysymyksiä, jotka lähetin palvelupäällikölle Teams-viestinä. Kaipasin tarkennuksia kysymyksiin, kuinka usein palveluarvioita tehdään? Miten puhelut valitaan? Mistä selviää, kuinka monta yhteistyökumppania tilaushetkellä oli tarjolla? Sekä saako työntekijä lisäksi sanallisen palautteen? Palvelupäällikön vastauksesta selvisi, että esihenkilö tekee yhdessä asiakaspalvelijan kanssa virallisen pisteytettävän palveluarvion kerran vuodessa. Palveluarvioon valitaan sattumanvaraisesti puheluita viimeisimmän kuukauden ajalta. Puhelinjärjestelmästä näkee puheluiden ajankohdan, pituuden, sekä kuinka monta kumppania on ollut valittavissa tilaushetkellä. Arviointiin otetaan mukaan puheluita, joissa oli saatavilla eri määrä kumppaneita, koska toimintaohjeet vaihtelevat sen mukaan, kuinka monta kumppania on saatavilla tilausta tehdessä. Lisäksi palveluarvioon pyritään sisällyttämään eri pituisia puheluita, sekä puheluita päivä- ja yövuorosta. Esihenkilö ja työntekijät keskustelevat puheluista samalla, kun ne pisteytetään ja lopuksi kiteyttävät yhdessä mitkä ovat asiakaspalvelijan vahvuudet palvelussa sekä miettivät tärkeimpiä kehityskohteita.

2.2.2 Palveluarviointilomake

Palveluarviointilomakkeella on 6 kohtaa, jotka pisteytetään 0–2 pisteellä. Paras mahdollinen tulos on 10/10 pistettä. Lomakkeen tarkoitus on selvittää, tuliko kaikki tilauksen ja palvelun kannalta oleelliset asiat käytyä läpi puhelun aikana. Tilauksen onnistumisen kannalta tärkeitä seikkoja ovat esimerkiksi asiakkaan nimen, lähtöosoitteen, paikkakunnan ja määränpään selvittäminen. Palvelun laadun kannalta on tärkeää, että asiakkaalle esitetään selkeästi eri vaihtoehdot ja kerrotaan seurantalinkistä. On myös tärkeätä, että puhelussa on ystävällinen tunnelma ja asiakasta muistetaan kiittää puhelun lopussa.

2.2.3 Aineisto

Palvelupäällikkö valitsi aineistoksi yhteensä 13 puhelua kahdelta eri työntekijältä ja toimitti myös manuaalisesti tuotetun palveluarvion näistä puheluista. Henkilön A puheluita oli 10 kpl, niistä oli tehty pisteytys arviointilomakkeen avulla. Henkilön B puheluita oli yhteensä 3 kpl, ja niistä kaikista oli annettu täydet pisteet, joten erillistä pisteytyslomaketta ei toimitettu. Puheluiden kesto vaihtelee välillä 59–118 sekuntia. Aineistoksi valittiin tarkoituksella puheluita, josta on annettu erilaisia pisteitä, jotta voitaisiin varmistua siitä, että ChatGPT antaa eri tilanteissa oikean määrän pisteitä.

2.3 Yhteenveto palveluarviointiprosessin nykytilasta

Nykytilan kartoituksen tarkoituksena oli saavuttaa parempi ymmärrys palveluarviointiprosessista sen nykytilassa, jotta voidaan suunnitella tulevia työvaiheita tarkemmin.

Palveluarvioinnista opittiin seuraavaa:

- Esihenkilö tekee palveluarvion yhdessä asiakaspalvelijan kanssa
- Puheluista keskustellaan pisteytyksen yhteydessä
- Lopuksi kiteytetään yhdessä asiakaspalvelijan vahvuudet ja kehityskohteet palvelussa
- Palveluarvioon valitaan 10 puhelua samalta asiakaspalvelijalta
    - jotka ovat tulleet viimeisen kuukauden aikana
    - joissa oli tilaushetkellä saatavilla eri määrä kumppaneita
    - jotka ovat eri pituisia puheluita
    - mahdollisuuksien mukaan puheluita sekä päivä- että yövuorosta

Arviointilomake sisältää yhteensä 6 kohtaa, joista suurin osa on melko yksinkertaisia, tilauksen kannalta oleellisia kysymyksiä, kuten ”kysyttiinkö asiakkaan nimi” tai ”selvisikö lähtöosoite ja paikkakunta”. Uskon tekoälyn suoriutuvan näistä kysymyksistä hyvin. Palvelun laadun varmistamisen kannalta on oleellista selvittää, onko asiakaspalvelija noudattanut ohjeistusta. Asiakaspalvelijoiden odotetaan kertovan asiakkaalle saatavilla olevista vaihtoehdoista. Ohjeistus vaihtelee sen mukaan, kuinka monta vaihtoehtoa on saatavilla. Tämä tieto ei välttämättä selviä puhelun sisällöstä, joten on tärkeä sisällyttää se tekoälylle lähetettävään litteroituun puheluun, jotta ChatGPT:n on mahdollista arvioida, seurattiinko ohjeistusta. Tilaushetkellä saatavilla olleiden kumppanien määrä näkyy puhelinjärjestelmästä. Lomakkeen viimeisessä kohdassa pisteytetään puhelun tunnelma ja asiakkaan kiittäminen. On mielenkiintoista nähdä, saako ChatGPT kiinni puhelun tunnelmasta.

ChatGPT:tä on ohjeistettava riittävän tarkasti, että se tulee useimmissa tapauksissa samaan lopputulokseen kuin manuaalisissa palveluarvioissa. Palveluarvio lasketaan 10 puhelun keskiarvosta, joten yksittäinen puuttuva tai ylimääräinen piste ei muuta arviointia merkittävästi. Palveluarvio on riittävän hyvä, mikäli 10 puhelun keskiarvo on sama kuin manuaalisesti tuotetussa palveluarviossa.

Aiemmin esihenkilö ja asiakaspalvelija ovat tehneet palveluarvion yhdessä, jolloin asiakaspalvelijalla on mahdollisuus perustella valintojaan. On syytä pohtia, miten tämä voitaisiin ottaa huomioon jatkossa, jos palveluarvion tekeekin tekoäly, joka ei tunne palvelua yhtä hyvin kuin esihenkilö ja asiakaspalvelija. Lisäksi tulee ottaa huomioon mahdollisuus, että tekoäly tuottaa virheellisiä, tai perusteettomia vastauksia. Asiakaspalvelijoita voidaan palkita hyvistä palveluarvioista, joten on tärkeää varmistaa, että asiakaspalvelijoilla on mahdollisuus pyytää, että tekoälyn tekemä palveluarvio tarkistetaan.

3 Puhelutallenteiden litterointi Whisper-mallin avulla

Tässä työvaiheessa tavoitteena on litteroida toimeksiantajalta saadut puhelutallenteet Whisper-mallilla ja analysoida litteroinnin laatua. Opinnäytetyön kannalta on tärkeää selvittää, kuinka hyvin Whisper-mallilla pystytään litteroimaan suomenkielistä puhetta. Lisäksi verrataan tuloksia aiempiin tutkimuksiin, joissa on arvioitu Whisper-mallin suoriutumista eri kielillä ja eri tietoaineistoilla. Seuraavia työvaiheita varten on tärkeää tunnistaa ja analysoida, minkälaisia virheitä kielimalli tekee, jotta virheet voidaan huomioida seuraavien työvaiheiden suunnittelussa. Whisper-malli on avoimen lähdekoodin ratkaisu, joka voidaan asentaa tietokoneelle paikallisesti, jolloin sen käyttäminen ei aiheuta ylimääräisiä kustannuksia eikä tietoja päädy kolmansien osapuolten haltuun.

3.1 Puheentunnistuksen perusteet ja litterointi Whisper-mallilla

Tässä osiossa tutustutaan puheentunnistukseen yleisellä tasolla, jotta saadaan käsitys siitä, minkälaisia ongelmia puheentunnistuksessa yleensä kohdataan. Lisäksi tullaan tutustumaan Whisper-malliin, sekä aikaisempiin tutkimuksiin, joissa on verrattu, kuinka hyvin Whisper-malli litteroi suomenkielistä puhetta verrattuna muihin kieliin.

3.1.1 Puheentunnistus

Mikko Kurimon artikkelissa kerrotaan, että nykyaikaisessa puheentunnistuksessa analysoidaan ääninäytteitä ja verrataan niitä kielimallin avulla aikaisempiin näytteisiin. Kielimallin avulla lasketaan mikä on ääninäytteen todennäköisin puhuttu sisältö. Kurimon mukaan puheentunnistus on haastavaa, koska ”puhesignaali on luonteeltaan jatkuvaa”. Haasteita tuottaa esimerkiksi yksittäisten sanojen, lauseiden tai puheenvuorojen erottelu toisistaan. Luonnollisessa puheessa haasteita lisää myös se, että ihmiset ääntävät sanoja eri tavoin, puhuvat eri nopeuksilla sekä päästelevät puhuessaan ääniä, kuten yskimistä ja epäröintiä, jotka eivät kuulu varsinaiseen puheeseen. Myös taustamelu, kohina ja tilan akustiikka saattaa hankaloittaa puheentunnistusta. (Kurimo 2008, 73–74).

Sanavirhearvo (Word Error Rate, WER) on yleisin tapa arvioida puheentunnistusjärjestelmiä. Sanavirhearvossa huomioidaan väärin tunnistettujen sanojen lisäksi puuttuvat ja ylimääräiset sanat. (Sallinen 2017, 29.) Kurimon (2008, 78) mukaan parhaat englanninkieliset puheentunnistimet saavuttavat keskimäärin alle 10 % sanavirheitä, kun kohteena on radio- tai tv-uutislähetys. Vertailuksi Kurimo (2008, 79) kertoo, että TKK:lla kehitetty ’jatkuvan puheen tunnistimen tarkkuus on noin 20 % sanavirheitä.

3.1.2 Whisper-malli

Whisper on OpenAI:n kehittämä automaattinen puheentunnistusjärjestelmä, joka julkaistiin syyskuussa 2022. Sen avulla voidaan litteroida useita eri kieliä, sekä kääntää eri kielistä puhetta englanniksi. OpenAI:n mukaan Whisper-malli tunnistaa englanninkielistä puhetta lähes yhtä hyvin kuin ihminen. Whisper-mallin koulutukseen on käytetty 680 000 tuntia monikielistä aineistoa, josta noin kolmasosa on muuta kuin englanninkielistä. (Open AI 21.9.2022.) Brockmanin ja muiden (2022, 27) mukaan monikielistä aineistoa on 117 113 tuntia, josta esimerkiksi suomenkielistä aineistoa on 1 066 tuntia, ruotsinkielistä 2 119 tuntia, saksankielistä 13 344 tuntia ja arabian kielistä 739 tuntia.

Whisper-mallin dokumentaatiosta selviää, että siitä on tehty useita versioita, nimeltään *tiny*, *base*, *small*, *medium*, *large* ja *turbo*. Pienemmät mallit ovat nopeampia ja vaativat vähemmän näyttömuistia (VRAM). Kaikista malleista, large- ja turbo-malleja lukuun ottamatta, on olemassa sekä monikielinen että pelkästään englanninkielinen versio. Large-mallista on lisäksi julkaistu parannellut versiot *large-v2* ja *large-v3*. *Turbo* on large-v3-mallista optimoitu kevyempi ja nopeampi versio. (OpenAI s. a.) Eri versioiden välisiä eroja havainnollistetaan taulukossa 1.

Taulukko 1. Whisper-mallin eri versiot (mukaillen OpenAI s.a. a)

Versio	**Parametrejä**	**Vaadittu VRAM**	**Suhteellinen nopeus**
tiny	39 M	~1 Gt	~10x
base	74 M	~1 Gt	~7x
small	244 M	~2 Gt	~4x
medium	769 M	~5 Gt	~2x
large	1550 M	~10 Gt	1x
turbo	809 M	~6 Gt	~8x

Whisper-malli voidaan asentaa paikallisesti, jolloin ylimääräisiä käyttökustannuksia ei synny. Whisper-mallia voidaan myös käyttää API-rajapinnan kautta, jolloin käytöstä veloitetaan opinnäytetyön tekohetkellä 0,006 $ per minuutti (OpenAI s.a. b).

3.1.3 Aikaisempi tutkimus suomenkielisen puheen litteroinnista Whisper-mallilla

Heikinheimo vertaili Whisper-mallin suoriutumista YouTube-videoiden litteroinnissa. Vertailussa oli mukana 19 kieltä. Aineistona käytettiin yhteensä viisi tuntia manuaalisesti litteroituja YouTube-videoita. Vertailun mittarina käytettiin sanavirhearvoa (Word Error Rating, WER). Vertailussa tutkittiin myös, miten Whisper-mallin eri versiot selviytyivät tehtävästä. Vertailun mukaan Whisper-malli selvisi suomenkielisen aineiston litteroinnista keskivertoa huonommin. Vertailussa selvitettiin myös small- ja medium-mallien välistä suhteellista sanavirheiden vähennystä (Relative Word Error Rate Reduction, WERR). Suurin osa kielistä ei saa merkittävää hyötyä käytettäessä raskaampaa

medium-mallia, mutta suomen kielen kohdalla ero on Heikinheimon mukaan merkityksellinen. (Heikinheimo 2023.) Taulukossa 2 esitetään poimintoja Heikinheimon (2023) vertailun tuloksista. Sarakkeessa ’WERR: S → M’ kuvataan small- ja medium-mallien välistä suhteellista sanavirheiden vähennystä.

Taulukko 2. Whisper-mallin eri versioiden sanavirhearvo eri kielten litteroinnista

Kieli / malli WER	**Large**	**Medium**	**Small**	**Base**	**Tiny**	**WERR: S → M**
Englanti	0,15	0,17	0,17	0,20	0,23	0,00
Saksa	0,18	0,18	0,21	0,27	0,37	0,14
Ruotsi	0,29	0,31	0,38	0,51	0,64	0,19
Suomi	0,41	0,43	0,53	0,70	0,85	0,19
Arabia	0,52	0,53	0,61	0,75	0,88	0,14

Brockman ja kumppanit (2022, 23–26) vertasivat eri Whisper-mallien suoriutumista eri kielten litteroinnista, kun käytettiin eri tietoaineistoja (dataset). Taulukoihin 3 ja 4 on poimittu small- ja medium-mallien sanavirhearvo muutamasta eri kielestä. Tuloksista käy ilmi, että käytetyllä tietoaineistolla on suuri vaikutus lopputulokseen. Lisäksi tässäkin tutkimuksessa Whisper-mallit näyttävät suoriutuvan suomen kielen litteroinnista keskivertoa heikommin.

Taulukko 3. Whisper small-mallin sanavirhearvo eri tietoaineistoilla

Kieli / tietoaineisto (WER %)	**Common Voice 9**	**VoxPopuli**	**FLEURS**
Englanti	14,5	8,2	6,1
Saksa	13,0	14,8	10,2
Ruotsi	22,1	-	20,8
Suomi	30,5	24,9	24,0
Arabia	66,4	-	30,6

Taulukko 4. Whisper medium-mallin sanavirhearvo eri tietoaineistoilla

Kieli / tietoaineisto (WER %)	**Common Voice 9**	**VoxPopuli**	**FLEURS**
Englanti	11,2	7,6	4,4
Saksa	8,5	12,4	6,5
Ruotsi	13,7	-	11,2
Suomi	18,8	16,6	13,9
Arabia	60,3	-	20,4

3.2 Litteroinnin toteutus

Tässä osassa valmistellaan Whisper-mallin käyttöönotto tutustumalla sen asennukseen ja järjestelmävaatimuksiin. Kun Whisper-malli on asennettu, voidaan litteroida toimeksiantajalta saatuja puhelutallenteita, ja analysoida litteroinnin onnistumista. Tässä työvaiheessa vertaillaan myös litteroinnin kestoa Whisper small- ja medium-malleilla.

3.2.1 Valmistelut

Opinnäytetyön suorittamista varten minulla oli käytössä toimeksiantajalta saatu kannettava tietokone. Koneessa on 16 Gt järjestelmämuistia (RAM) sekä integroitu grafiikkapiiri, joten erillistä näyttömuistia (VRAM) ei ole, vaan näytönohjain käyttää järjestelmämuistia.

Aluksi varmistin, että koneelle oli esiasennettu Python. Whisper-mallin (OpenAI s.a. a) dokumentaation mukaan Pythonin versiot 3.8–3.11 ovat yhteensopivia Whisper-mallin kanssa. Koneelle oli valmiiksi asennettu python 3.12. Asensin myös komentorivityökalu ffmpeg:in. Ffmpeg on avoimen lähdekoodin komentorivityökalu, joka mahdollistaa multimedia tiedostojen muuntamisen eri formaatteihin (FFmpeg s.a.). Asensin koneelle Whisper-mallin OpenAI:n (s.a. a) dokumentaation ohjeiden mukaan.

3.2.2 Litterointi

Whisperiä voidaan ajaa suoraan komentoriviltä. Alla olevassa esimerkissä komennossa on litteroitavan audiotiedoston lisäksi määritelty litterointiin käytettävä Whisper-malli, sekä määritelty näytteen kieli. Whisper-malli tunnistaa kielen automaattisesti, jos sitä ei ole määritelty.

Esimerkki komennosta

```
whisper nauhoite.mp3 --model small --language Finnish
```

Small-mallilla litteroitaessa litterointi kävi nopeasti joka kerta. Tekstiä oli kuitenkin hankala lukea, sillä se sisälsi huomattavan määrän virheitä. Suurin osa virheistä oli pieniä, esimerkiksi sanasta oli yksi kirjain väärin tai sanassa oli käytetty väärää taivutusmuotoa. Medium-mallilla litterointi kesti huomattavasti kauemmin, mutta lopputulos oli vastaavasti huomattavasti helpompi lukea, sillä virheitä oli paljon vähemmän. Medium-mallia käytettäessä kuitenkin joistain äänitteestä jäi pois pitempiä pätkiä tai sitten pitemmän hiljaisuuden aikana litteraatissa toistui sama lause useita kertoja peräkkäin. Kokeilin litteroida saman puhelun heti uudestaan, mutta tulos ei muuttunut merkittävästi. Näissä tapauksissa kokeilin litteroida myös turbo-mallilla. Osa äänitteistä onnistui paremmin turbo-mallilla. Osa äänitteistä onnistui myöhemmin myös medium-mallilla. Alla olevista näytteistä (A, B) näkyy ote eri ajankohtina litteroidusta puhelusta. Tulosteen aikaleimoista huomataan, että näytteestä A puuttuu paikoitellen sekunnin mittaisia pätkiä, sekä neljän sekunnin pätkä välillä 00:32–

0:36, jolloin nauhoitteelta kuuluu vain näppäimistön ääniä. Näytteessä B taas välillä 00:28–00:36 toistuu useaan otteeseen ”Kuopion saana”, ja asiakkaan arvailut katuosoitteesta on jäänyt pois.

Näyte A

```
[00:09.000 --> 00:13.000]  Mitä tuota Saanalle saataisko taakse neljä hen-
kilöä?
[00:14.000 --> 00:15.000]  Lähesitään.
[00:15.000 --> 00:16.000]  Joo, missä olitte?
[00:17.000 --> 00:18.000]  Saanalla.
[00:18.000 --> 00:19.000]  Onks se tosiaan on tekijöllä?
[00:20.000 --> 00:21.000]  Anteeks.
[00:22.000 --> 00:25.000]  Joo, mikä oli se tarkempi osote?
[00:26.000 --> 00:27.000]  Mikä se oli tää tarkempi osote?
[00:27.000 --> 00:30.000]  Ois ootte Pellaranta vai mikä, Siikaranta vai
mitä?
[00:31.000 --> 00:32.000]  Saana, Kuopio.
[00:36.000 --> 00:37.000]  Onks tää Kuopion Saana?
[00:38.000 --> 00:39.000]  Joo.
[00:39.000 --> 00:42.000]  Joo, yes. Ja oliks te sit tuota noin, minne
matka sieltä?
[00:43.000 --> 00:47.000]  No sieltä ihan viisi kilometriä sen keskustan
ollaan menossa.
```

Näyte B

```
[00:08.000 --> 00:14.000]  Mitä tuota Saanalle saataisko taksi neljä henki-
löä?
[00:14.000 --> 00:16.000]  Joo, missä olitte?
[00:16.000 --> 00:18.000]  Saanalla.
[00:18.000 --> 00:20.000]  Onks se tosiaan on tekijöllä?
[00:20.000 --> 00:22.000]  Anteeks.
[00:22.000 --> 00:26.000]  Joo, mikä oli se tarkempi osote?
[00:26.000 --> 00:28.000]  Mikä se oli tää tarkempi osote?
[00:28.000 --> 00:30.000]  Kuopion saana.
[00:30.000 --> 00:32.000]  Kuopion saana.
[00:32.000 --> 00:34.000]  Kuopion.
[00:34.000 --> 00:36.000]  Kuopion saana.
[00:36.000 --> 00:38.000]  Onks tää Kuopion saana?
[00:38.000 --> 00:40.000]  Joo.
[00:40.000 --> 00:42.000]  Ja oliks tei sit tuota noin, minne matka sieltä?
[00:42.000 --> 00:48.000]  No sieltä ihan 5 kilometriä sen keskustaan ol-
laan menossa.
```

Taulukossa 5 on mitattu litteroinnin kestoa sekunneissa eri puhelutallenteilla. Vertailua tehtiin Whisper small- ja medium-malleilla. Ajoin jokaisen puhelun kolme kertaa molemmilla malleilla, sekä laskin kolmen puhelun keskiarvon. Laskin myös small- ja medium-mallien keskiarvojen välisen eron prosentteina.

Taulukko 5. Litteroinnin kesto (s) eri Whisper-malleilla

		Litteroinnin kesto Whisper small-mallilla				**Litteroinnin kesto Whisper medium-mallilla**				
Tal-lenne	**Kesto (s)**	**I**	**II**	**III**	**ka.**	**I**	**II**	**III**	**ka.**	**Ero (%)**
1	78	42,1	59,3	45,7	49,0	104,3	102,0	156,4	120,9	147
3	72	22,2	22,0	21,2	21,8	56,9	59,5	63,4	59,9	175
5	78	25,5	26,7	24,6	25,6	84,1	84,9	86,1	85,0	232
7	59	26,5	28,1	28,0	27,5	79,3	85,6	77,8	80,9	194
9	91	46,8	52,3	50,7	49,9	99,5	103,3	102,8	101,9	104
10	95	23,1	23,8	29,5	25,5	69,4	69,8	72,1	70,4	177
A	60	19,4	17,8	19,0	18,7	61,0	65,4	62,3	62,9	236
B	118	39,1	38,3	34,9	37,4	304,4	294,1	276,8	291,8	680
ka.	84	30,6	33,5	31,7	31,9	107,4	108,1	112,2	109,2	243

3.2.3 Whisper small- ja medium-mallien vertailu

Vertaillakseni small- ja medium-mallien välisiä eroja, litteroin 8 kpl toimeksiantajalta saatuja puhelutallenteita kummallakin mallilla. Tallenteet 2 ja 3 litteroin myös turbo-mallilla, sillä ne eivät aluksi onnistuneet medium-mallilla. Laskin ylimääräiset, puuttuvat ja virheelliset sanat sekä sanavirheprosentin. Laskin myös suhteellisen sanavirheiden vähennyksen small- ja medium-mallien välillä. Alla olevassa taulukossa (6) on kuvattu litteroinnin tulokset eri Whisper-malleilla.

Taulukko 6. Puhelutallenteiden litteroinnin tulokset

				sanavirhearvo			
Tallenne	**kesto (s)**	**tiedoston koko (Mt)**	**sanoja**	**small**	**medium**	**turbo**	**$WERR_{(S\rightarrow M)}$**
1	78	1,20	118	0,31	0,22		0,29
2	72	1,12	79	0,46	0,31	0,23	0,33
3	78	1,20	127	0,64	-	0,40	
4	59	0,94	73	0,37	0,18	-	0,51
5	91	1,39	121	0,37	0,21	-	0,43
7	95	1,46	121	0,39	0,17		0,56
A	60	0,96	98	0,15	0,07		0,53
B	118	1,82	227	0,27	0,16		0,41
ka.	84	1,30	121	0,37	0,19		0,44

3.2.4 Litteroinnin analysointi

Analysoidakseni litteraatin virheitä tarkemmin, laskin litteroitujen puheluiden pienet virheet, suuremmat virheet, puuttuvat sanat sekä ylimääräiset sanat. Pieniksi virheiksi laskin sanat, joissa on esimerkiksi yksi kirjain väärin, tai väärä taivutusmuoto. Taulukoissa 7 ja 8 on laskettu litteroitujen puheluiden sanavirheiden jakauma Whisper small- ja medium-malleilla.

Taulukko 7. Sanavirheiden jakauma Whisper small-mallilla

Tallenne	Sanoja	Pienet virheet	Isot virheet	Puuttuvat sanat	Ylim. sanat	Virheet yht.	Virheet (%)
1	118	15	10	8	3	36	31
2	79	11	9	19	5	44	56
3	127	30	12	33	6	81	64
4	73	12	5	9	2	28	38
5	121	26	2	12	5	45	37
7	121	27	5	15	3	50	41
A	98	8	3	4	0	15	15
B	227	20	20	16	5	61	27
ka.	121	19	8	15	4	45	39

Taulukko 8. Sanavirheiden jakauma Whisper medium-mallilla

Tallenne	Sanoja	Pienet virheet	Isot virheet	Puuttuvat sanat	Ylim. sanat	Virheet yht.	Virheet (%)
1	118	7	9	10	0	26	22
2	79	10	2	14	0	26	33
3	127	21	12	6	10	49	39
4	73	10	2	1	0	13	18
5	121	16	4	6	0	26	21
7	121	13	2	5	3	23	19
A	98	2	3	2	1	8	8
B	227	14	4	16	3	37	16
ka.	121	12	5	8	2	26	22

Vertailun vuoksi laskin suhteellisen sanavirheiden vähennyksen myös Brockmanin ja kumppaneiden (2022, 23–26) aineistolle seuraavalla kaavalla: $WERR_{S \rightarrow M} = (WER_S - WER_M) / WER_S$. Tulokset on kuvattu alla olevassa taulukossa (9).

Taulukko 9. Whisper small- ja medium-mallien välinen suhteellinen sanavirheiden vähennys ($WERR_{S \rightarrow M}$) eri tietoaineistoilla

Kieli	Common Voice 9	VoxPopuli	FLEURS
Englanti	22,8 %	7,3 %	27,9 %
Saksa	34,6 %	16,2 %	36,3 %
Ruotsi	38,0 %	-	46,2 %
Suomi	38,4 %	33,3 %	42,1 %
Arabia	9,2 %	-	33,3 %

3.3 Litteroinnin tulokset ja päätelmät

Tässä osiossa summataan, minkälaisia ongelmia puheentunnistuksessa yleensä kohdataan, sekä pohditaan, kohdattiinko opinnäytetyön tekemisen yhteydessä samoja ongelmia. Lisäksi käydään läpi litteroinnin tulokset ja verrataan niitä aikaisempiin tutkimuksiin.

3.3.1 Litteroinnin keskeiset haasteet

Kurimon (2008) artikkelista selvisi, että puheentunnistuksen yleisiä haasteita on esimerkiksi yksittäisten sanojen, lauseiden tai puheenvuorojen erottelu toisistaan. Tunnistin omassa työssäni samat ongelmat, virheitä esiintyi paljon silloin, kun asiakas ja asiakaspalvelija puhuivat päällekkäin. Kurimon (2008) mukaan haasteita aiheuttaa myös se, että ihmiset ääntävät sanoja eri tavoin, puhuvat eri nopeuksilla sekä päästelevät puhuessaan ääniä, kuten yskimistä ja epäröintiä, jotka eivät kuulu varsinaiseen puheeseen. Omassa työssäni huomasin selvästi, että puhenopeuden vaihtelu aiheutti Whisper-mallille ongelmia. Nopea puhetapa saattoi johtaa siihen, että sanat yhdistyivät uudeksi tunnistamattomaksi sanaksi. Hidas puhetapa tai epäröinti kesken sanan taas johti usein siihen, että yhdyssana ilmestyi litteraattiin kahtena erillisenä sanana.

Työssä minulla oli käytettävissä kannettava tietokone, jossa ei ollut erillistä näytönohjainta. Whisper small-mallilla litterointi onnistui nopeasti. Oletin, että medium-mallilla litterointikin onnistuisi, sillä OpenAI:n (s.a. a) dokumentoinnin mukaan suositeltu määrä näyttömuistia (VRAM) on 5 Gt. Litterointi medium-mallilla oli kuitenkin hidasta ja takkuilevaa, erityisesti silloin kun koneella oli käynnissä muitakin ohjelmia, kuten tekstinkäsittelyohjelma, verkkoselain ja viestintäohjelma. Huomasin suuria eroja, kun mittasin kuinka kauan litterointi kestää eri malleilla. Medium- ja large-mallien litteroinnissa litteraatista saattoi jäädä pois jopa 30 sekunnin pituinen pätkä, tai siinä saattoi toistua yksi lause tai sana useaan otteeseen kohdassa, jossa oli pitempi hiljaisuus. OpenAI:n (s.a. a) dokumentoinnin mukaan medium-malli on 2 kertaa nopeampi kuin large-malli ja vastaavasti small-malli on 4 kertaa nopeampi kuin large-malli. Tämä vaikuttaisi pitävän paikkansa omien mittausteni perusteella. Sovelluksen arkkitehtuuria suunniteltaessa on tärkeää ottaa huomioon

Whisper medium-mallin vaatima korkea määrä näyttömuistia, mikäli Whisper-malli halutaan asentaa esimerkiksi paikalliselle palvelimelle.

3.3.2 Whisper-mallin suoriutuminen suomenkielisen puheen litteroinnissa

Saadakseni käsityksen siitä, kuinka haastavaa suomenkielisen aineiston litterointi on, tutustuin aiempiin tutkimuksiin, joissa oli vertailtu Whisper-mallin suoriutumista eri kielisten tietoaineistojen litteroimisesta. Valitsin vertailun kieliksi saksan ja englannin, koska ne ovat maailmalla laajasti puhuttuja, joten oletin että kyseisistä kielistä on saatu paljon koulutusaineistoa. Valitsin suomen kielen, koska se on opinnäytetyön kannalta kriittisin kieli. Lisäksi valitsin ruotsin, koska arvioin, että ruotsin kieli olisi suomen kielen kanssa suunnilleen yhtä tavallinen. Viimeiseksi valitsin arabian kielen, koska arvioin, että arabiankielistä koulutusaineistoa olisi vähemmän, ja arabiankielisen aineiston litteroiminen voisi olla Whisper-mallille haastavampaa kuin suomenkielisen. Vertailuista kävi nopeasti ilmi, että käytetty aineisto vaikuttaa merkittävästi litteraatin sanavirhearvoon. Whisper-mallia valittaessa on täten erityisen tärkeää tehdä koeajoja oikealla aineistolla.

Heikinheimon (2023) vertailussa small- ja medium-malleilla litteroitaessa englanti (WER 0,17 molemmissa) ja saksa (WER 0,21 ja 0,18) onnistuvat huomattavasti paremmin kuin suomi (WER 0,53 ja 0,43). Ruotsi (WER 0,38 ja 0,31) onnistui hieman paremmin kuin suomi ja arabia (WER 0,61 ja 0,53) hieman huonommin. Brockman ja kumppanit (2022, 23–26) vertailivat eri kieliä eri tietoaineistoilla. Poimin vertailuuni tietoaineistot ”Common Voice 9”, ”VoxPopuli” ja ”FLEURS”, koska niissä oli mukana suomen kieli. VoxPopuli-aineistossa ei ollut mukana ruotsin ja arabian kieliä. Näissäkin vertailuissa englanti, saksa ja ruotsi pärjäsivät paremmin kuin suomi, kun taas arabia oli kaikilla aineistoilla suomea haastavampi. Brockmanin ja muiden vertailussa suomi sai small-mallilla litteroitaessa eri tietoaineistoilla sanavirhearvon 30,5, 24,9 ja 24,0 ja medium-mallilla 18,8, 16,6 ja 13,9. Omassa työssäni small-mallilla litteroitaessa sanavirhearvon keskiarvo oli 0,37 ja medium-mallilla litteroitaessa 0,19. Tulokset ovat jopa parempia kuin Heikinheimon vertailussa, mutta hieman huonompia kuin Brockmanin ja kumppaneiden vertailussa. Parhaat tulokset sain puheluista, joissa puhe oli rauhallista ja selkeää. Haastavimmat puhelut olivat sellaisia, joissa puhe oli epäselvää tai asiakaspalvelija ja asiakas puhuivat paljon toistensa päälle.

Heikinheimon vertailussa oli myös laskettu small- ja medium-mallien litteraattien välinen suhteellinen sanavirheiden vähennys (WERR) kuvaamaan, kuinka paljon eri kielet hyötyvät Whisper medium-mallin käyttämisestä. Heikinheimon vertailussa kaikki vertailuun poimimani kielet saivat WERR lukemaksi 14–19 %, paitsi englanti, jonka tulos oli 0 %. (Heikinheimo 2023.) Brockmanin ja kumppaneiden (2022, 23–26) vertailussa sanavirheiden suhteellinen vähennys oli paljon vaikuttavampi kaikilla kielillä ja tietoaineistoilla (7,3–46,2 %). Tässäkin vertailussa suomen kieli hyötyi muita enemmän medium-mallista (muutos 33,3–42,1 %). Omassa vertailussani keskiarvo oli 44 %,

mikä on vielä hieman parempi kuin Brockmanin ja kumppaneiden vertailussa ja huomattavasti parempi kuin Heikinheimon vertailussa. Tämä viittaa siihen, että suomenkielisten puhelutallenteiden litteroinnissa olisi merkittävää hyötyä Whisper-medium mallin käyttämisestä.

3.3.3 Litteroinnin havainnot ja kehitysmahdollisuudet

OpenAI:n (s.a. a) mukaan Whisper small-malli on kaksi kertaa niin nopea kuin medium-malli. Tämä vaikuttaisi pitävän paikkansa. Tekemieni vertailujen (taulukko 9) perusteella puheluiden pituuden keskiarvo oli 84 sekuntia. Litterointi small-mallilla kesti keskimäärin 31,9 sekuntia ja medium-mallilla 109,2 sekuntia. Small- ja medium-mallien keskiarvojen välinen ero oli yli 200 %, tämä selittynee sillä, että mukana oli puheluita, joiden kohdalla litterointi kangerteli, ja oli huomattavasti keskiarvoa hitaampaa.

Suurin osa litteroitujen puhelutallenteiden sanavirheistä oli pieniä virheitä, jotka todennäköisesti voitaisiin korjata tekoälyn avulla. En kuitenkaan usko, että tekoälyllä pystyttäisiin luotettavasti korvaamaan puuttuvia sanoja litteraateissa. Kun luodaan palveluarvioita tekoälyn avulla luvussa 5, voitaisiin myös selvittää, parantaisiko korjatun litteraatin käyttö palveluarvioiden laatua.

Puhelut ovat melko kaavamaisia, ja niissä toistuu tietyt palvelulle ominaiset käsitteet. Olisi mielenkiintoista selvittää, miten Whisper-mallia voitiaisiin kouluttaa omalla aineistolla, jolloin se tunnistaisi helpommin palvelulle tyypillisiä termejä ja fraaseja.

Haasteita esiintyi, kun nauhoitteessa oli pitkiä hiljaisuuksia, hälinää taustalla tai kun asiakas ja asiakaspalvelija puhuivat päällekkäin. Pitkän hiljaisuuden aikana litteraatti saattoi toistaa yhtä lausetta useaan otteeseen, tai jättää pitkän pätkän kokonaan pois litteraatista. Koska valtakunnallisessa palvelussa kohdataan paljon erilaisia murteita ja puhetapoja, voidaan olettaa, että palveluarvioihin käytettävässä aineistossa tulee olemaan paljon laadunvaihtelua. Litteraatin aikaleimojen perusteella voidaan ohjelmallisesti tarkistaa, puuttuuko litteraatista pitempiä osioita. Mahdollisesti voitaisiin myös pyytää ChatGPT:tä arvioimaan, onko litteraatti epäonnistunut, jolloin sitä ei käytettäisi palveluarviossa. Haasteita esiintyi, kun litteroitiin paikallisesti asennetulla Whisper medium-mallilla, mikä viittaa siihen, ettei työhön käytetyssä kannettavassa tietokoneessa ollut riittävästi järjestelmämuistia. En usko, että vastaavia ongelmia esiintyy, mikäli Whisper-mallia käytetään API-rajapinnan kautta.

4 Litteroidun aineiston anonymisointi

Opinnäytetyön tavoitteita tarkastellaan myös vastuullisuuden näkökulmasta. Koska työssä käsitellään luottamuksellisia puhelutallenteita, tietoturva on keskeisessä asemassa. Anonymisointi tarkoittaa, että aineistossa esiintyvät henkilötiedot poistetaan tai muutetaan niin, että henkilöä ei voi enää tunnistaa aineistosta. Anonymisoinnilla varmistetaan, että yksittäistä asiakasta ei voi tunnistaa aineistosta. OpenAI:n (5.4.2023) mukaan ChatGPT:n ilmainen versio voi käyttää käyttäjien keskusteluja koulutusaineistona, joten ilmaisversiota käytettäessä anonymisointi olisi erittäin tärkeää tehdä ennen kuin aineisto lähetetään API-rajapinnan yli ChatGPT:lle prosessoitavaksi. Näin varmistetaan, ettei arkaluontoista tietoa päädy kolmansien osapuolten haltuun. Toimeksiantajalla on käytössä ChatGPT:n maksullinen versio, ja se on tehnyt OpenAI:n kanssa tietojenkäsittelysopimuksen, joten anonymisointi ei ole opinnäytetyön kannalta välttämätöntä. Tässä työvaiheessa on tarkoitus selvittää erilaisia keinoja toteuttaa anonymisointi, mikäli siihen jää aikaa. Opinnäytetyön aineistoa on vähän, ja se voidaan tarvittaessa anonymisoida manuaalisesti muuttamalla tai poistamalla aineistosta henkilötiedot. Anonymisointia ei ehditty toteuttaa opinnäytetyön aikana.

5 Litteraatin muuttaminen palveluarvioksi ChatGPT:n avulla

Tässä osiossa on tarkoitus suunnitella kehote, jonka avulla voidaan luoda palveluarvioita ChatGPT-tekoälymallia hyödyntäen. Lisäksi tullaan arvioimaan tekoälyn suoriutumista vertaamalla sen tekemiä palveluarvioita manuaalisesti tuotettuihin. Tietoperustassa tutustutaan lyhyesti ChatGPT-kielimalliin, sekä syvennytään tekoälykehotteen suunnitteluun. Tekoälykehotteen suunnittelun tavoite on luoda kehote tekoälylle, jolla se tuottaa mahdollisimman laadukkaan palveluarvion annettujen kriteerien perusteella litteroidusta puhelusta. ChatGPT:n ohjeistuksessa käytetään toimeksiantajalta saatua palveluarviointilomaketta sekä yrityksen intranetistä löytyviä ohjeita.

5.1 ChatGPT ja tekoälykehotteen suunnittelu

Tietoperustassa tutustutaan ChatGPT-malliin, sekä syvennytään tekoälykehotteen suunnitteluun. Opinnäytetyön tekohetkellä toimeksiantajalla oli käytössä ChatGPT 4o-malli, ilmaiskäyttäjillä oli käytössä ChatGPT 4o-mini-malli, API-rajapinnan kautta pystyttiin myös käyttämään vanhempia malleja, kuten GPT-3.5 Turbo ja GPT-4. Eri GPT-malleihin tutustutaan tarkemmin luvussa 6.

5.1.1 ChatGPT

ChatGPT on OpenAI:n kehittämä tekoälymalli, jonka kanssa käyttäjä voi käydä vuorovaikutteista keskustelua. Dialogipohjaisen vuorovaikutuksen avulla ChatGPT pystyy kysymään tarkentavia kysymyksiä, myöntämään virheensä, haastamaan virheelliset lähtötiedot, sekä hylkäämään sopimattomia pyyntöjä. ChatGPT perustuu OpenAI:n GPT-3 kielimalliin ja se koulutettiin Reinforcement Learning from Human Feedback -menetelmän avulla. (OpenAI 30.11.2022.)

ChatGPT:n perusversio on ilmainen ja sen käyttöönotto vaatii vain rekisteröitymisen. OpenAI:n (30.11.2022) mukaan ChatGPT halutaan pitää saavutettavana, jotta voidaan kerätä käyttäjäpalautetta mallin ongelmista ja virheistä. Helmikuussa 2023 OpenAI (1.2.2023) julkaisi ChatGPT Plus nimisen, GTP-4 malliin perustuvan, maksullisen version. Elokuussa 2023 OpenAI (28.8.2023) ilmoitti julkaisevansa ChatGPT Enterprise version, joka perustuu GPT-4-kielimalliin, ja lupaa maksua vastaan yrityksille parempaa tietosuojaa, tehokkaampaa suorituskykyä, rajattomasti pyyntöjä ym. uusia ominaisuuksia. Marraskuussa 2024 Reuters ja Tivi uutisoivat, että OpenAI suunnittelee muuttavansa yhtiörakenteensa voittoa tavoittelemattomasta organisaatiosta voittoa tavoittelevaksi (Nyman 2024; Reuters 2024).

5.1.2 Tekoälykehotteen suunnittelu

Kehotesuunnittelun (prompt engineering) avulla pyritään luomaan tekoälylle kehote siten, että saadaan tekoälyltä mahdollisimman hyvä vastaus. Hyvässä kehotteessa tulisi olla selkeä tehtävä,

sekä konteksti ja esimerkki, jotka auttavat tarkentamaan pyyntöä. Lisäksi kehotteessa voidaan määritellä toivottu vastaustyyli. kuten persoona, formaatti tai sävy. (Gupta 19.3.2024; Su 1.8.2023.) Myös Open AI:n dokumentaation mukaan tekoäly tarvitsee selkeät ohjeet, jotta se pystyy antamaan relevantin vastauksen. Ohjeissa tulee selkeästi ilmaista konteksti ja tärkeät yksityiskohdat, muuten tekoälymalli joutuu arvaamaan mitä kehotteella haetaan. Lisäksi dokumentaatiossa esitellään useita strategioita ja taktiikoita, joiden avulla kehotetta voidaan tehostaa. OpenAI:n mukaan monimutkaiset kehotteet kannattaa pilkkoa yksinkertaisemmiksi vaiheiksi, joita tekoälymallin on helpompi seurata. Kehotteessa voidaan käyttää erottimia rajaamaan tekstistä osioita, joita halutaan käsitellä eri tavalla. Tekstikappale voidaan erottaa ohjeistuksesta kolmella lainausmerkillä (esimerkki 1). Erottimina voidaan myös käyttää XML-elementtejä (esimerkki 2). Erottimien käyttäminen auttaa tekoälyä hahmottamaan tehtävän, siitä on erityisen paljon hyötyä monimutkaisissa ja pitkissä kehotteissa.

```
Esimerkki 1

            """tekstikappale"""

Esimerkki 2

            <otsikko>Otsikkoteksti</otsikko>
```

Tekoälyn voi myös ohjeistaa järkeilemään vastaus ennen kuin se antaa vastauksensa, esimerkiksi matemaattisen pulman tarkistamisessa tekoäly ei välttämättä huomaa matkan varrella sattunutta ajatusvirhettä, jos sitä pyydetään vain tarkistamaan, onko vastaus oikein. Vastauksen pituus voidaan määritellä esimerkiksi sanoina, lauseina tai kappaleina. Tekoälymallille voidaan syöttää esimerkkejä toivotusta tulosteesta, joista se voi oppia. Tästä menetelmästä käytetään englanninkielistä termiä ”Few-Shot Prompting”. (OpenAI. s.a. c.)

5.2 Palveluarvion toteutus

Tässä osiossa luodaan ensin mahdollisimman hyvä kehote tekoälylle, ja sen jälkeen luodaan palveluarvio, jota voidaan verrata manuaalisesti tuotettuun palveluarvioon. Osiossa 2 todettiin, että tekoälyn tuottama palveluarvio on riittävän hyvä, jos 10 puhelun keskiarvo on sama kuin manuaalisesti tuotetussa palveluarviossa.

5.2.1 Tekoälykehotteen luominen

Tässä työvaiheessa käytin selaimessa ChatGPT 4o-versiota. Syötin ChatGPT:lle ensin arviointilomakkeen kysymykset, sen jälkeen syötin vielä taksiasiakaspalvelun ohjeistuksen, jonka sain 020202 Palvelut Oy:n (2021) intranetistä. Pyysin ChatGPT:tä luomaan arviointikriteerit lomakkeen kysymysten pohjalta, ja sisällyttämään niihin tarkennuksia ohjeista. Tämän jälkeen pyysin

ChatGPT:tä tekemään palveluarvion puheluista, jotka oli litteroitu small- ja medium-malleilla. Halusin kokeilla molemmilla malleilla, koska epäilin, että Whisper small-mallilla litteroitujen puheluiden lukuisat virheet voivat johtaa virheisiin palveluarvioissa.

Koska asiakaspalvelijoiden toimintaohjeet vaihtelevat saatavilla olevien yhteistyökumppaneiden perusteella, tuli ChatGPT:n kehotteeseen lisätä tieto siitä, kuinka monta yhteistyökumppania tilaushetkellä oli saatavilla. Pyysin ChatGPT:tä palauttamaan pisteet taulukkomuodossa, sekä antamaan sanallisen palautteen asiakaspalvelijan vahvuuksista ja kehityskohteista. Numeroin puhelut, jotta niihin olisi mahdollista viitata palautteessa.

Loin ensimmäiset palveluarviot käyttäen sekä Whisper small- että medium-mallilla luotuja litteraatteja. Manuaalisen palveluarvion kokonaisarvosana oli 5,5. Small-mallin litteraateilla sain kokonaisarvosanan 8 ja medium-mallin litteraateilla 7. Molemmissa versioissa ChatGPT toi esiin samat vahvuudet ja kehityskohteet. Tuloksia vertaillessani, huomasin, että ChatGPT suoriutui parhaiten kohdissa, joissa sai vain yhden pisteen. Näissä kohdissa vastaus oli yksinkertainen kyllä tai ei. Kohdissa, joissa voi saada kaksi pistettä ChatGPT antoi useimmiten täydet pisteet, vaikka manuaalisessa arvioinnissa oli annettu yksi piste. Poikkeuksena kohta 2, jossa medium-mallin litteraatteja käytettäessä ChatGPT antoi kolmessa puhelussa vähemmän pisteitä kuin manuaalisessa palveluarviossa. Whisper small-mallin litteraatteja käytettäessä, ChatGPT oli antanut kahdessa puhelussa perusteettomia pisteitä kohdassa 5. Taulukossa 10 esitetään palveluarviointilomakkeen eri kohtien pisteiden keskiarvot, kun arvioinnit tehtiin Whisper small- ja medium-malleilla litteroitujen puheluiden perusteella. Vertailun vuoksi taulukossa on mukana myös täydet pisteet ja manuaalisesti tuotetun palveluarvion pisteet.

Taulukko 10. Palveluarvioiden vertailu, keskiarvo per kohta

	Whisper small	**Whisper medium**	**Manuaalinen**	**Täydet pisteet**
Kohta 1	1,0	0,9	1,0	1
Kohta 2	1,8	1,6	1,9	2
Kohta 3	1,7	1,6	0,6	2
Kohta 4	1,0	1,0	1,0	1
Kohta 5	0,6	0,2	0,1	2
Kohta 6	1,9	1,7	0,9	2
Arvosana	8	7	5,5	10

Päätin kokeilla, antaako ChatGPT tarkemman palveluarvion, jos kahden pisteen kysymykset pilkotaan kahdeksi yhden pisteen kysymykseksi. Tällä kertaa sekä Whisper small- että medium-mallilla litteroiduista puheluista tehdyn palveluarvion arvosana oli 7,8. Molemmat mallit antoivat samat

pisteet pilkottujen kysymysten a ja b kohdissa. Tällä kertaa Whisper small-mallin litteraatteja käytettäessä, ChatGPT ei antanut perusteettomia pisteitä kohdassa 5. Taulukossa 11 esitetään palveluarviointilomakkeen eri kohtien pisteiden keskiarvot, kun kahden pisteen kysymykset on pilkottu erillisiksi kysymyksiksi. Arvioinnit tehtiin Whisper small- ja medium-malleilla litteroitujen puheluiden perusteella. Vertailun vuoksi taulukossa on mukana myös täydet pisteet ja manuaalisesti tuotetun palveluarvion pisteet.

Taulukko 11. Palveluarvioiden vertailu, kahden pisteen kysymykset pilkottu, keskiarvo per kohta

	Whisper small	**Whisper medium**	**Manuaalinen**	**Täydet pisteet**
1	1,0	1,0	1,0	1
2a	1,0	1,0	1,9	1
2b	1,0	1,0	-	1
3a	0,8	0,8	0,6	1
3b	0,8	0,8	-	1
4	1,0	1,0	1,0	1
5a	0,1	0,1	0,1	1
5b	0,1	0,1	-	1
6a	1,0	1,0	0,9	1
6b	1,0	1,0	-	1
Arvosana	7,8	7,8	5,5	10

5.3 ChatGPT:n luoman palveluarvion onnistumiset ja kehityskohteet

ChatGPT:tä kehotettiin luomaan palveluarvio kymmenestä puhelusta annettujen kriteerien pohjalta. Ohjeistuksessa kerrottiin, että puhelut on numeroitu ja numeron jälkeen suluissa on ilmoitettu, kuinka monta yhteistyökumppania tilaushetkellä oli saatavilla. Kehotteessa pyydettiin palauttamaan pisteet taulukkomuodossa, sekä antamaan sanallinen palaute asiakaspalvelijan vahvuuksista ja kehityskohteista.

ChatGPT onnistui parhaiten lomakkeen niissä kohdissa, joista annettiin vain yksi piste. Näissä kohdissa vastaus oli yksinkertainen kyllä tai ei. Kohdissa, joissa voi saada kaksi pistettä ChatGPT antoi useimmiten täydet pisteet, vaikka manuaalisessa arvioinnissa oli annettu vain yksi piste. Esimerkiksi kohdassa, jossa arvioitiin puhelun tunnelmaa ja asiakkaan kiittämistä ChatGPT antoi kaikissa puheluissa täydet kaksi pistettä, kun manuaalisissa palveluarvioinneissa oli usein annettu vain yksi piste. Vaikka kahden pisteen kysymykset pilkottiin erillisiksi kysymyksiksi ChatGPT antoi molemmista kysymyksistä pisteen, eikä lopputulos muuttunut merkittävästi. Kahden pisteen

kysymyksissä ChatGPT tarvitsee tarkempaa ohjeistusta siitä, milloin annetaan täydet pisteet ja milloin vain yksi piste.

Palveluarvion onnistumisen kannalta litteraatin ei tarvitse olla täysin virheetön, esimerkiksi osoitteen ei tarvitse olla oikein, kunhan litteraatista käy ilmi, että sitä on puhelun aikana kysytty. Olin olettanut, että Whisper small-mallilla litteroiduissa puheluissa olisi liian paljon virheitä, jotta palveluarvio onnistuisi hyvin, mutta käytännössä ne vaikuttavat olevan riittävän hyviä. Jatkokehityksen kannalta olisi hyödyllistä pohtia, millä tavalla kaikista huonoimmat litteraatit voitaisiin sulkea pois palveluarvioinnista. Litteraatissa näkyy segmenttien alun ja lopun aikaleimat, joten tätä tietoa hyväksikäyttämällä voidaan ohjelmallisesti varmistaa, että litteraatti on eheä. On kuitenkin haastavampaa tarkistaa, ettei joku sana ole muuttunut niin, että se olennaisesti muuttaa puhutun sisällön merkitystä.

Kehotetta voitaisiin parantaa OpenAI:n (s.a. c) dokumentaatiossa mainituilla keinoilla, esimerkiksi puhelut voitaisiin erotella toisistaan XML-elementeillä. Few-Shot prompting-menetelmän mukaisesti ChatGPT:lle voitaisiin syöttää manuaalisesti tehtyjä palveluarvioita malliksi, jolloin se saattaisi saada paremman käsityksen kriteereiden tulkintatavasta.

Uskon, että selkeyttämällä pisteytyskriteerejä ja kehittämällä kehotetta edelleen ChatGPT pystyy tuottamaan riittävän tarkkoja palveluarvioita.

6 Eri GPT-mallien laadullinen vertailu

Kielimallien vertailun tavoitteena on selvittää eri GPT-mallien (3.5, 3.5 Turbo, 4.0) välisiä eroja, sekä arvioida, kuinka hyvin ne suoriutuvat palveluarvioiden tuottamisessa verrattuna manuaalisesti tehtyihin arvioihin. Vertailu rajoittuu yrityksessä jo käytössä oleviin GPT-malleihin (3.5, 3.5 Turbo ja 4.0). Muita tekoälymalleja ei oteta mukaan, koska tavoitteena on löytää paras vaihtoehto yrityksen nykyisistä resursseista. Vertailun avulla liiketoiminta voi päättää mikä kielimalli on sopivin palveluarvioiden tuottamiseen.

6.1 GPT-mallien eri versiot ja hinnoittelu

Tässä osiossa tutustutaan OpenAI:n julkaisemiin eri GPT-versioihin, sekä tutustutaan OpenAI:n API-rajapinnan hinnoitteluun. Tieto eri mallien hintaeroista on oleellista, kun päätetään mitä GPT-mallia sovelluksessa tullaan käyttämään.

6.1.1 GPT-mallien hinnoittelu

OpenAI:n hinnastosta selviää, että API-rajapinnan yli käytettäessä GPT-mallien hinnoittelu perustuu rahakkeisiin (token). 1000 rahaketta vastaa noin 750 sanaa. Tilanteissa, joissa rajapintaan lähetettyihin kyselyihin ei tarvita vastausta välittömästi, voidaan säästää kustannuksissa lähettämällä pyynnöt eräajona (batch) käyttäen OpenAI:n Batch API-rajapintaa. Batch API-rajapinta palauttaa vastaukset 24 tunnin kuluessa, ja hinnasta myönnetään 50 % alennus. (OpenAI s.a. b). Taulukkoon 12 on poimittu hintatietoja OpenAI:n (s.a. b) hinnastosta. Hinnat on kerrottu dollareina miljoonaa rahaketta kohden ja olivat voimassa marraskuussa 2024.

Taulukko 12. Eri kielimallien hinta dollareissa per miljoona rahaketta

Malli	**Syöte ($)**	**Tuloste ($)**
GPT-4o	2,50	10,00
GPT-4o-mini	0,15	0,60
o1-preview	15,00	60,00
o1-mini	3,00	12,00
GPT-4	30,00	60,00
GPT-4 Turbo	10,00	30,00
GPT-3.5 Turbo	0,50	1,50

6.1.2 Uusimmat GPT-mallit: GPT-4o ja 4o-mini

Opinnäytetyön tekohetkellä OpenAI:n lippulaivatuote oli toukokuussa 2024 julkaistu GPT-4o. Tässä versiossa sekä syöte, että palaute voi olla multimodaalista, eli siinä voidaan yhdistää tekstiä,

ääntä, kuvaa ja videota. Tavoite on ollut kehittää ihmisen ja koneen välisestä vuorovaikutuksesta luonnollisempaa. OpenAI:n mukaan GPT-4o-mallin vasteaika äänisyötteisiin vastaa ihmisten välisen keskustelun vasteaikoja. Aikaisempiin malleihin verrattuna GPT-4o-malli tulkitsee kuvaa ja ääntä huomattavasti paremmin. Englanninkielisen tekstin ja koodin käsittelyssä GPT-4o-mallin suorituskyky on samalla tasolla kuin GPT-4 Turbo-mallissa, mutta sen kyky käsitellä ei-englanninkielistä tekstiä on merkittävästi parempi. (OpenAI 13.5.2024.)

Heinäkuussa 2024 julkaistu GPT-4o-mini on kevyempi ja edullisempi versio GPT-4o-mallista. Se on suunniteltu nopeisiin ja kevyisiin tehtäviin. OpenAI:n mukaan se soveltuu hyvin sovelluksiin, jotka tekevät lukuisia kutsuja API-rajapintaan tai syöttävät mallille paljon taustatietoa. Mallin koulutusdata ulottuu lokakuuhun 2023 saakka. (OpenAI 18.7.2024.)

6.1.3 Vanhemmat GPT-mallit: GPT-4, GPT-4 Turbo ja GPT-3.5 Turbo

GPT-4-malli julkaistiin maaliskuussa 2023. Se on multimodaalinen malli, joka vastaanottaa syötteen tekstinä tai kuvana sekä palauttaa tulosteen tekstinä. OpenAI:n mukaan se suoriutuu ammatillisista ja akateemisista tehtävistä yhtä hyvin kuin ihminen, esimerkiksi läpäisemällä yhdysvaltalaisen asianajokokeen (Bar Exam) ja sijoittumalla parhaan 10 % joukkoon. Vertailun vuoksi GPT-3.5-malli sijoittui huonoimman 10 % joukkoon. (OpenAI 14.3.2023.) Mallin koulutusdata ulottui alun perin syyskuuhun 2021 asti, mutta päivitysten myötä joulukuuhun 2023 asti (OpenAI s.a. d). GPT-4 Turbo-malli julkaistiin marraskuussa 2023 ja sen koulutusdata ulottuu huhtikuuhun 2023 asti. Se on suorituskyvyltään optimoitu versio GPT-4-mallista. (OpenAI 6.11.2023a.) GPT 3.5-malli ei ole enää saatavilla API-rajapinnan kautta, GPT 3.5 Turbo malli on edelleen saatavilla, mutta heinäkuusta 2024 lähtien suositellaan 4o-mini mallin käyttöä (OpenAI s.a. d).

6.2 Vertailun tulokset ja analyysi

Kielimallien laadullista vertailua ei ehditty toteuttamaan. Tutustuttuani OpenAI:n eri GPT versioihin, suosittelen, että vertailu tehdään 4o- ja 4o-mini-malleilla, sillä ne ovat kehittyneemmät kuin suunnitelmassa mainitut GPT-3.5- , GPT-3.5 Turbo- ja GPT-4-mallit. Hinnoittelulla selvästi ohjataan käyttäjiä siirtymään uudempiin malleihin. GPT-3.5 Turbo-malli voitaisiin ottaa mukaan verrokiksi, mutta GPT-3.5-mallia ei löydy enää hinnastosta, ja GPT-4-malli on huomattavasti kalliimpi kuin 4o-malli.

7 Pohdinta

Opinnäytetyön tavoitteena oli kehittää ratkaisu, jonka avulla toimeksiantaja voi tehostaa asiakaspalvelun laadun arviointia tekoälyn avulla. Palveluarviointiprosessia tehostamalla mahdollistetaan palveluarvioinnin tekeminen useammin, sekä suuremmalle määrälle puheluita, tai vaihtoehtoisesti voidaan vapauttaa resursseja muihin tehtäviin. Tässä luvussa pohditaan opinnäytetyön onnistumista kokonaisuutena.

7.1 Yhteenveto

Opinnäytetyön alussa kartoitettiin, miten palveluarviointiprosessi yrityksessä toimii nykytilassa. Kartoitus onnistui hyvin, ja sen avulla saatiin muodostettua käsitys siitä, miten palvelua arvioidaan, ja mitä lopullinen palveluarvio sisältää.

Toimeksiantajalta saadut puhelutallenteet litteroitiin Whisper small- ja medium-malleilla, jotta voitiin vertailla ja analysoida litteraattien laatua. Litteraateille laskettiin sanavirhearvo (WER), jonka avulla tuloksia voitiin vertailla aikaisempiin tutkimuksiin. Perehtymällä puheentunnistuksen haasteisiin yleisellä tasolla, saatiin hyvä käsitys siitä, minkälaisia haasteita työssä voidaan kohdata. Puheluiden litterointi ja litteraattien analysointi onnistui hyvin, ja auttoi muodostamaan käsityksen siitä, kuinka hyvin suomenkielisen puheen litteroiminen onnistuu Whisper-mallilla.

Litteraattien analysointiin kului suunniteltua enemmän aikaa, joten päätin jättää anonymisoinnin toteuttamatta, sillä se ei ollut opinnäytetyön kannalta kriittinen työvaihe, koska toimeksiantaja on allekirjoittanut tietojenkäsittelysopimuksen OpenAI:n kanssa. Jos käytössä olisi ollut ChatGPT:n ilmainen versio, olisi aineistosta pitänyt poistaa kaikki henkilötiedot, kuten nimet ja osoitteet.

Tekoälykehotteen suunnittelussa oli tarkoitus suunnitella kehote, jonka avulla voidaan luoda palveluarvioita Chat-GPT-tekoälymallia hyödyntäen. ChatGPT:n tekemää palveluarviota verrattiin manuaalisesti tehtyyn palveluarvioon. Kehotesuunnittelussa ehdittiin toteuttaa vain yksi iteraatio. Tutustumalla kehotesuunnitteluun saatiin ideoita, joiden avulla kehotetta voidaan parantaa. Kehotteen suunnittelu jäi keskeneräiseksi, sillä toteutetulla kehotteella ChatGPT antoi palveluarvioinnin arvosanaksi 7 tai 8, kun manuaalisessa palveluarviossa arvosana oli 5,5.

Kielimallien vertailun tavoitteena oli luoda palveluarvioita eri kielimalleilla ja analysoida niiden eroavaisuuksia. Kielimallien vertailua ei ehditty toteuttaa. Tutustumalla OpenAI:n (s.a. b) API-rajapinnan hinnastoon, voitiin todeta, että kielimallien vertailu kannatta tehdä GPT-4o- ja GPT-4o-mini-malleilla, sillä ne ovat kehittyneemmät kuin suunnitelmassa mainitut GPT-3.5- , GPT-3.5 Turbo- ja GPT-4-mallit. GPT-3.5 Turbo-malli voitaisiin ottaa vertailuun mukaan verrokiksi, mutta GPT-3.5-malli ei ole enää saatavissa, ja GPT-4-malli on huomattavasti kalliimpi kuin GPT-4o-malli.

7.2 Johtopäätökset ja jatkokehitysideat

Opinnäytetyö tarjosi arvokasta tietoa tekoälyn soveltamisesta suomenkielisten puheluiden litterointiin ja palveluarviointiin. Työ nosti esiin sekä onnistumisia että kehityskohtia, joiden avulla tekoälyratkaisuja voidaan jatkossa parantaa ja laajentaa. Työn tärkeimpänä antina saatiin vahvistus siitä, että tekoäly pystyy tuottamaan palveluarviointeja, mikä avaa uusia mahdollisuuksia palvelun laadun arvioinnin tehostamiselle. Tässä luvussa esitetään keskeiset johtopäätökset ja jatkokehitysideat.

Suomenkielisten puheluiden litterointi onnistuu paremmin Whisper medium-mallilla suhteessa small-malliin. Litteraattien laadussa on paljon vaihtelua, joten kaikista huonoimmat litteraatit olisi hyvä jättää pois palveluarvioinnista. Palveluarvion onnistumisen kannalta litteraatin ei tarvitse olla täysin virheetön, esimerkiksi osoitteen ei tarvitse olla oikein, kunhan litteraatista käy ilmi, että osoitetta on puhelun aikana kysytty. Litteraateista voitaisiin saada vielä parempia kouluttamalla Whisper-mallia Fine-tuning-menetelmän avulla tunnistamaan paremmin työssä usein käytettyjä termejä ja fraaseja, kuten "arvioitu saapumisaika" tai yhteistyökumppaneiden nimet (Pawar 31.1.2024).

Opinnäytetyössä tuli esiin useita kehitysehdotuksia, joiden avulla tekoälykehotteesta saataisiin vielä parempi. Näitä ei ehditty kokeilemaan. Palveluarvioinnin kriteerejä tulisi tarkentaa ja selkeyttää, jotta tekoäly onnistuisi luomaan palveluarvion, joka vastaa manuaalisesti tuotettua palveluarviota. ChatGPT pystyy nyt jo vastaamaan hyvin yksinkertaisiin kysymyksiin, joihin riittää vastaukseksi ”kyllä” tai ”ei”. Ilman ohjelmointia voidaan luoda kustomoitu versio ChatGPT:stä (OpenAI 6.11.2023b). Tämä voisi olla mielenkiintoinen ja kustannustehokas vaihtoehto yrityksen sisäiseen käyttöön, sillä se ei vaadi erillisen sovelluksen kehittämistä.

Aiemmin esihenkilö ja asiakaspalvelija ovat tehneet palveluarvion yhdessä, jolloin asiakaspalvelijalla on mahdollisuus perustella valintojaan. On syytä pohtia, miten tämä voitaisiin ottaa huomioon jatkossa. Tekoäly saattaa tuottaa virheellisiä, tai perusteettomia vastauksia, joten asiakaspalvelijoilla tulee olla mahdollisuus pyytää tekoälyn tuottaman palveluarvion tarkistusta. Tämä on erityisen tärkeää tilanteissa, joissa palveluarvio vaikuttaa työntekijöiden palkitsemiseen.

7.3 Luotettavuus ja hyödynnettävyys

Puheluita litteroitiin kahdeksan kappaletta, ja litteroinnin tulokset olivat verrattavissa aikaisempiin tutkimuksiin. Litteroinnin osalta arvioisin, että tulokset ovat luotettavia. Palveluarvioita tehtiin vain yksi kappale, joten tulokset ovat suuntaa antavia. Saatuja tuloksia ja oppeja voidaan käyttää tekoälykehotteen jatkokehityksessä. Työn tulokset on esitelty toimeksiantajalle Teams-palaverissa 31.10.2024. Lisäksi toimeksiantajalle toimitetaan raportti, johon sisällytetään kaikki tutkimustulokset ja käytetty aineisto.

Opinnäytetyön tuloksia voidaan hyödyntää laajemmin muissa sovelluksissa, joissa tekoälyä käytetään suomenkielisen puheen litterointiin ja sen jatkoprosessointiin. Tämä laajentaa työn merkitystä myös muille aloille, joissa luonnollista puhetta halutaan litteroida ja jatkojalostaa tekoälyn avulla.

Tekoälyn luoma palveluarvio voi olla objektiivisempi verrattuna manuaalisesti tuotettuun, sillä tekoälyllä ei ole ennakkokäsityksiä, joita esimerkiksi esihenkilöillä saattaa tiedostamattaan olla alaisistaan. Tämä voi edistää arviointiprosessien tasapuolisuutta ja luotettavuutta. Tilanteissa, joissa palveluarviot vaikuttavat palkitsemiseen tai palautteenantoihin, on kuitenkin tärkeää varmistaa, että tekoälyn tekemä arvio voidaan tarvittaessa tarkistaa. Tämä lisää arviointijärjestelmän läpinäkyvyyttä ja luottamusta sen oikeudenmukaisuuteen.

7.4 Tietoturva, aineiston hallinta ja eettiset näkökohdat

Opinnäytetyössä hyödynnettiin toimeksiantajalta saatua aineistoa, kuten manuaalisesti tuotettuja palveluarvioita ja asiakaspalvelupuheluiden äänitteitä. Aineiston kerääminen tapahtui yhteistyössä toimeksiantajan kanssa. Aineiston käsittelyssä huomioitiin tietosuoja- ja eettiset periaatteet. Aineistoa analysoitiin ja käsiteltiin OpenAI:n Whisper- ja ChatGPT-tekoälymallien avulla. Toimeksiantajalla on tietojenkäsittelysopimus OpenAI:n kanssa.

Aineisto säilytettiin toimeksiantajan tietovarastossa koko opinnäytetyön ajan. Opinnäytetyön päätyttyä kaikki tallennettu data hävitetään toimeksiantajan tietosuojaohjeistuksen mukaisesti. Opinnäytetyön raportti julkaistaan Theseuksessa, mutta siihen ei sisällytetä mitään luottamuksellista tietoa.

7.5 Oman opinnäytetyöprojektin ja oppimisen arviointi

Työn alusta asti tiedostin, että opinnäytetyö edellytti paljon työtä suhteellisen lyhyessä ajassa. Aikataulun venyttäminen ei ollut mahdollista, sillä tavoitteeni oli valmistua joulukuussa. Työvaiheet oli suunniteltu hyvin, ja priorisoitu jo työn suunnitteluvaiheessa, joten oli helppoa siirtyä seuraavaksi tärkeimpään työvaiheeseen. Esimerkiksi anonymisointi olisi ollut mielenkiintoinen toteuttaa, jotta tutkimuksen tulokset olisivat laajemmin hyödynnettävissä, mutta se ei ollut toimeksiantajan näkökulmasta välttämätöntä. Erityisesti tulosten analysointiin kului arvioitua enemmän aikaa. En ollut aikaisemmin tutustunut Whisper-malliin, joten en osannut varautua siihen, että tekisin vertailua sekä small- että medium-mallilla, mikä kaksinkertaisti työmäärän litteraattien analysoinnin osalta.

Halusin tehdä nimenomaan toiminnallisen työn, koska koen oppivani parhaiten tekemällä. Vetoketjumallin käyttäminen opinnäytetyöraportin rakenteena oli toimiva ratkaisu. Sain vuorotella teoreettisten ja toiminnallisten osuuksien välillä, mikä auttoi pitämään mielenkiintoa yllä.

Työssä käytettiin tekoälyä monipuolisesti raportin kirjoittamisen tukena. Esimerkiksi Kurimon (2008) artikkeli sisälsi runsaasti matemaattisia kaavoja, mikä teki sen vaikeaselkoiseksi. Pyysin ChatGPT:tä selittämään osia artikkelista yksinkertaisemmalla tavalla, mikä auttoi minua sisäistämään artikkelin paremmin. Suurin osa käytetyistä lähteistä oli englanninkielisiä, ja sopivien suomenkielisten käännösten ja ilmaisujen keksiminen tuotti paikoitellen haasteita. Näissä tapauksissa pyysin ChatGPT:ltä apua, esimerkiksi pyytämällä sitä kääntämään kohtia tekstistä, tai muotoilemaan kirjoittamani tekstin sujuvammaksi. En käyttänyt ChatGPT:n generoimia vastauksia sellaisinaan, vaan poimin vastauksista ideoita, joiden avulla sain oman tekstini sujuvammaksi ja luonnollisemmaksi.

Siltä osin, kun työ tuli valmiiksi olen tyytyväinen lopputulokseen. Puhelutallenteiden litterointiin ja litteraattien analysointiin olen erityisen tyytyväinen. Erityisen harmissani olen siitä, että en ehtinyt kehittää tekoälykehotetta pitemmälle. Uskon kuitenkin, että projektin aikana hankkimani ymmärrys tekoälystä ja sen soveltamisesta luo vahvan pohjan jatkokehitykselle tulevaisuudessa. Opinnäytetyö tarjosi minulle arvokasta oppia tekoälysovelluksista ja projektinhallinnasta. Työn suunnittelu, priorisointi ja aikatauluttaminen antoivat käytännön kokemusta, jota voin hyödyntää tulevissa työtehtävissä.

Lähteet

020202 Palvelut. s.a. Mitä meillä tehdään? 020202 Palvelut Oy. Luettavissa: https://www.020202.fi/meista/020202-palvelut-oy/. Luettu: 27.6.2024.

020202 Palvelut 2021. 020202 Palvelut intranet. 02 Taksi ohjeet. Perustietoa. Perustietoa 02 Taksista. Luettu: 15.10.2024.

02 Taksi. s.a. Tutustu meihin. Luettavissa: https://02taksi.fi/tietoa-palvelusta/. Luettu 30.9.2024.

Brockman, G., Kim, J. W., McLeavey, C., Radford, A., Sutskever, I., Xu, T. 21.9.2022. Robust Speech Recognition via Large-Scale Weak Supervision. OpenAI. Dublin/San Francisco. Luettavissa: https://cdn.openai.com/papers/whisper.pdf. Luettu: 23.9.2024.

Ffmpeg. s.a. About FFmpeg. Luettavissa: https://ffmpeg.org/about.html. Luettu: 30.9.2024.

Gupta, D. 19.3.2024. Prompt Engineering. Medium. Luettavissa: https://medium.com/@researchgraph/prompt-engineering-21112dbfc789. Luettu: 15.10.2024.

Heikinheimo, H. 4.4.2023. Analyzing OpenAI's Whisper ASR Accuracy: Word Error Rates Across Languages and Model sizes. Speechly. Luettavissa: https://www.speechly.com/blog/analyzing-open-ais-whisper-asr-models-word-error-rates-across-languages. Luettu: 19.9.2024.

Hyken, S. 17.3.2024. Bad Customer Service Could Cost More Than $3.7 Trillion. Forbes. Luettavissa: https://www.forbes.com/sites/shephyken/2024/03/17/bad-customer-service-could-cost-more-than-37-trillion/. Luettu: 6.9.2024.

Iwuozor, J. 13.6.2024. What Is Customer Service? Definition & Best Practices. Forbes advisor. Luettavissa: https://www.forbes.com/advisor/business/what-is-customer-service-definition-best-practices/. Luettu: 6.9.2024.

Kurimo, M. 2008. Puheentunnistus. Puhe ja kieli, 2 s. 78–83.

Nyman, M. 5.11.2024. ChatGPT:n kehittänyt tekoäly-yhtiö mullistaa täysin periaatteensa. Tivi. Luettavissa: https://www.tivi.fi/uutiset/tv/d3f38467-fca2-4778-8249-523a289c05ac. Luettu: 7.11.2024.

OpenAI. s.a. a. Whisper. Luettavissa: https://github.com/openai/whisper/blob/main/README.md. Luettu 26.9.2024.

OpenAI. s.a. b. Pricing. Luettavissa: https://openai.com/api/pricing/. Luettu: 1.11.2024.

OpenAI. s.a. c. Prompt engineering. Luettavissa: https://platform.openai.com/docs/guides/prompt-engineering. Luettu: 15.10.2024.

OpenAI. s.a. d. Models. Luettavissa: https://platform.openai.com/docs/models. Luettu: 7.11.2024.

OpenAI. 21.9.2022. Introducing Whisper. Luettavissa: https://openai.com/index/whisper/. Luettu: 22.8.2024.

OpenAI. 30.11.2022. Introducing ChatGPT. Luettavissa: https://openai.com/blog/chatgpt/. Luettu: 11.9.2023.

OpenAI. 1.2.2023. Introducing ChatGPT Plus. Luettavissa: https://openai.com/blog/chatgpt-plus. Luettu: 16.10.2023.

OpenAI. 14.3.2023. GPT-4. Luettavissa: https://openai.com/index/gpt-4-research/. Luettu: 8.11.2024.

OpenAI. 5.4.2023. Our approach to AI safety. Luettavissa: https://openai.com/index/our-approach-to-ai-safety/. Luettu: 16.10.2024.

OpenAI. 28.8.2023. Introducing ChatGPT Enterprise. Luettavissa: https://openai.com/blog/introducing-chatgpt-enterprise. Luettu: 4.10.2023.

OpenAI. 6.11.2023a. New models and developer products announced at DevDay. Luettavissa: https://openai.com/index/new-models-and-developer-products-announced-at-devday/. Luettu 12.11.2024.

OpenAI. 6.11.2023b. Introducing GPTs. Luettavissa: https://openai.com/index/introducing-gpts/. Luettu: 8.11.2024.

OpenAI. 13.5.2024. Hello GPT-4o. Luettavissa: https://openai.com/index/hello-gpt-4o/. Luettu: 8.11.2024.

OpenAI. 18.7.2024. GPT-4o mini: advancing cost-efficient intelligence Luettavissa: https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/. Luettu 12.11.2024.

Pawar, S. 31.1.2024. A comprehensive guide for Custom Data Fine-Tuning with the Whisper Model. Medium. Luettavissa: https://medium.com/@shridharpawar77/a-comprehensive-guide-for-custom-data-fine-tuning-with-the-whisper-model-60e4cbce736d. Luettu: 12.11.2024.

Reuters. 5.11.2024. OpenAI in talks with California to become for-profit company, Bloomberg News reports. Luettavissa: https://www.reuters.com/technology/openai-talks-with-california-become-for-profit-company-bloomberg-news-reports-2024-11-04/. Luettu: 7.11.2024.

Sallinen, N. 2017. Development of the Finnish Spoken Dialog System for an Educational Robot. Master's Thesis. Aalto University, Department of Signal Processing and Acoustics. Luettavissa: https://urn.fi/URN:NBN:fi:aalto-201704133563. Luettu: 30.9.2024.

Su, J. 1.8.2023. Master the Perfect ChatGPT Prompt Formula (in just 8 minutes)! Video. Katsottavissa: https://www.youtube.com/watch?v=jC4v5AS4RIM. Katsottu 12.10.2024.

Tieteen termipankki. 26.9.2023: Avoin tiede:tietoaineisto. Luettavissa: https://tieteentermipankki.fi/wiki/Avoin tiede:tietoaineisto. Luettu 23.10.2024.