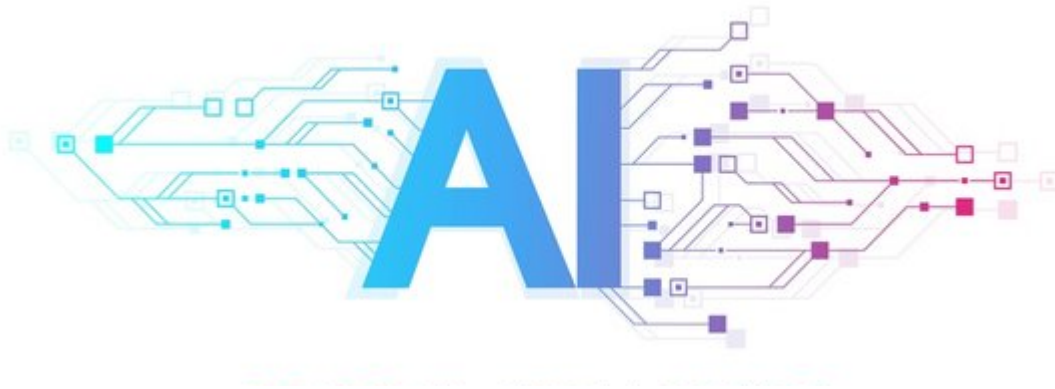


Moni osaava koneoppimisjärjestelmä työprosessien automointiin



Ryhmä Neuro

Otto Oksanen | rocksanen@gmail.com | +358505700792

Emil Ålgars | emil.algars@gmail.com | +358400672676

Mohammed Al-Jewari | mjassim18@gmail.com | +358406636764

pvm 17.06.2024

1 Johdanto	4
2 Alaan perehtyminen	4
2.1 Moodle kurssi.....	4
2.2 Yritysvierailu.....	4
3 Aiheen valinta	5
3.1 Yrityksen tarpeiden tunnistaminen.....	5
3.2 Visio.....	5
4 Projektin eteneminen	6
4.1 Ryhmän työskentely.....	7
4.2 Ideaan liittyvän materiaalin opiskelu.....	7
5 Tulokset	7
5.1 Mitä saatiin selvitettyä.....	8
5.2 Arkkitehtuuri.....	8
5.2.1 Single LLM.....	9
5.2.2 Mixture of Experts (MOE).....	11
5.2.3 Multi-Head Mixture-of-Experts (MH-MoE).....	12
5.3 RAG (Retrieval-Augmented Generation).....	14
5.4 MOE (Mixture of Experts).....	16
5.5 MH-MoE (Multi-Head Mixture-of-Experts).....	18
5.6 RLHF (Reinforcement Learning from Human Feedback).....	19
5.7 Vektori tietokannat.....	20
5.7.1 Qdrant.....	21
5.7.2 Chroma.....	21
5.7.3 Milvus.....	22
5.8 Adobe rajapinta.....	22
5.8.1 Rajapinnan arkkitehtuuri.....	23
5.8.2 Inter-Process Communication (IPC).....	23
5.9 Jatko tutkittavaa.....	24
5.9.1 Multimodal learning.....	24
5.9.2 Koulutus areenan kehittäminen.....	24
5.9.3 Hienosäätämisen ja vahvistetun oppimisen yhdistäminen.....	25
6 Toteutus	25
6.1 Tarvittava osaaminen ja henkilöstö.....	25
6.2 Aikataulu.....	26
Vaihe 1: Suunnitelma.....	26
Vaihe 2: Prototyyppi.....	26
Vaihe 3: Testaus.....	26
Vaihe 4: Käyttöönotto.....	27
Vaihe 5: Lisätoiminnallisuus.....	27
6.3 Suunnitelma.....	27
1. Projektin Alustava Määrittely.....	27
2. Arkkitehtuurisuunnittelu.....	27
3. Tekninen Tiekartta.....	28
6.4 Prototyyppi.....	28

6.5 Testaus.....	29
Testausvaiheen Tavoitteet.....	29
Testausvaiheen Vaiheet.....	30
Testausvaiheen Resurssit.....	30
6.6 Käyttöönotto.....	31
6.7 Lisätoiminnallisuus.....	32
7 Yhteenveto.....	32
7.1 Projektin Tausta.....	32
7.2 Suunnittelu ja Toteutus.....	32
7.3 Käyttöönotto ja Lisätoiminnallisuus.....	32
7.4 Johtopäätös.....	32

1 Johdanto

LSB on pitkään toiminut paino- ja mediayritys, joka on havainnut teknologian nopean kehityksen merkittävät vaikutukset toimialallaan. Perinteisesti painoala on ollut työvoimavaltaista ja manuaalista, mutta digitalisaatio on avannut uusia mahdollisuuksia tehostaa toimintaa ja parantaa kilpailukykyä.

Työnkulkujen automatisointi voi merkittävästi vähentää manuaalisen työn määrää ja vähentää virheiden riskiä. Tässä projektissa tavoitteemme oli selvittää, kuinka tekoälyä voisi hyödyntää yrityksen päivittäisissä tehtävissä ja mitä teknologioita tarvitaan, jotta yritys saisi lisäarvoa tuottavan lopputuloksen.

Projektimme aikana suunnittelimme myös ylätasoin arkkitehtuurin kompleksisen järjestelmän toteuttamiseksi. Lopuksi arvioimme projektin toteutuksen vaatiman työmäärän, henkilöstön tarpeen ja kustannukset. Näiden arvioiden perusteella voimme suunnitella tehokkaan ja kustannustehokkaan tavan hyödyntää tekoälyä LSB:n liiketoiminnassa.

2 Alaan perehtyminen

Prepress ja digipainaminen ovat kiinnostavia aloja, mutta tietoa näistä ei löydy runsaasti verkosta. Tämän vuoksi parhaat resurssit oppimiseen ovat kirjat ja kurssit. Onneksenne Metropolia tarjosi aiheeseen liittyvän verkkokurssin, jota pystyimme hyödyntämään yritysprojektin aikana.

2.1 Moodle kurssi

Projektimme aikana jokainen ryhmämme jäsen suoritti Metropolian Moodle-alustalla verkkokurssin nimeltä Digipainaminen ja prepress. Kurssin tarkoituksena oli toimia meille koulutusmateriaalina ja tutustuttaa meidät alaan. Kurssi käsitteli kaikkea työkaluista värien hallintaan ja auttoi selvittämään työnkulkua sekä joitain teknisiä asioita alaan liittyen. Joitakin tärkeitä prosesseja, kuten taittamisen työnkulku ja yleisimmät käsitellyt asiat, ei kuitenkaan käsitelty riittävästi.

2.2 Yritysvierailu

Kävimme ryhmämme kanssa LSB:n konttorilla vierailulla kahdesti projektin aikana. Ensimmäinen vierailu oli yleisluonteinen, ja sen tarkoituksena oli perehdyttää meidät yrityksen prosesseihin ja työnkulkuun. Vierailusta saimme joitain ideoita, mutta

huomasimme, että prepress-vaiheeseen on jo olemassa erittäin hyviä ohjelmistoja, ja prosessi on melko suoraviivainen. Näin ollen totesimme, että todellinen hyöty tulisi innovatiivisesta ja luovasta ideasta. Tämän vierailun perusteella syntyi ajatus kokonaisvaltaisesta AI-avustajasta.

Toisella vierailullamme esitimme ideamme yrityksen toimitusjohtaja Riku Suomalaiselle. Ideamme otettiin hyvin vastaan ja saimme siitä vahvistusta lähteä tutkimaan syvemmin ideamme kokonaisvaltaisesta ai-avustajasta.

3 Aiheen valinta

3.1 Yrityksen tarpeiden tunnistaminen

Keskusteluistamme LSB-yrityksen toimitusjohtajan kanssa, sekä tutustuessamme yrityksen työskentelyproseduureihin, ilmeni, että yrityksellä ei ollut tarkkaa käsitystä siitä, miten heidän toimintojaan voitaisiin automatisoida tai hyödyntää tekoälyä. Sen sijaan kokemuksemme mukaan heillä oli tarve saada konsultointia mahdollisuuksista, joita tekoälyn hyödyntäminen toimintojen kehittämisessä voisi tarjota ja näin ollen parantaa heidän kykyään tunnistaa alueita joissa tekoäly tuo oikeaa arvoa yritykselle.

Näiden keskustelujen perusteella keskityimme kartoittamaan tekoälykehityksen nykypäivän standardeja ja toteutusmahdollisuuksia. Tavoitteemme oli antaa yritykselle selkeämpi kuva siitä, mitä kaikkea on mahdollista saavuttaa tekoälyratkaisujen avulla, miten ne voisivat konkreettisesti tehostaa yrityksen prosesseja ja parantaa kilpailukykyä, sekä luoda ylätasoinen toimintasuunnitelma siitä kuinka tällainen järjestelmä mahdollisesti rakennetaan.

3.2 Visio

Tavoitteenamme oli suunnitella ja luoda edellytykset kehittää kokonaisvaltainen AI-avustaja (chatbot), jonka avulla yrityksen työntekijät voivat tehostaa ja yksinkertaistaa päivittäisiä työtehtäviään monipuolisesti ja tehokkaasti. Tämä AI-avustaja voisi mahdollisesti tuoda merkittävää arvoa yritykselle seuraavilla tavoilla:

Yrityksen prosessien ymmärtäminen ja hallinta:

- AI-avustajan avulla työntekijät voivat keskustella chatbotin kanssa saadakseen reaaliaikaista tietoa yrityksen prosessien eri vaiheista. Tämä parantaa prosessien ymmärrystä ja auttaa tekemään tietoon perustuvia päätöksiä, mikä tehostaa työnkulkua ja vähentää virheiden määrää.

Tietojen haku internetistä:

- AI-avustaja pystyy hakemaan ja esittämään ajankohtaista tietoa internetistä, joka voi olla hyödyllistä päätöksenteossa ja työn suorittamisessa. Tämä vähentää merkittävästi aikaa, joka kuluu tiedon etsimiseen ja kokoamiseen, ja mahdollistaa

nopeamman reagoinnin muuttuviin tilanteisiin.

Sovellusten automaation ja hallinnan:

- AI-avustaja voi automatisoida ja hallita muiden sovellusten, kuten Adobe Photoshopin ja InDesignin, toimintoja. Tämä mahdollistaa esimerkiksi graafisten elementtien muokkaamisen ja asiakirjojen luomisen suoraan chatbotin kautta, mikä säästää aikaa ja vähentää manuaalisen työn tarvetta.

Uusien työntekijöiden ja harjoittelijoiden tukeminen:

- AI-avustaja voi opastaa uusia työntekijöitä ja harjoittelijoita työtehtävissä tarjoamalla vaiheittaisia ohjeita ja vastaamalla kysymyksiin. Tämä vähentää kokeneiden työntekijöiden tarvetta käyttää aikaansa perehdyttämiseen ja koulutukseen, ja nopeuttaa uusien työntekijöiden sopeutumista.

Aineistohallinta:

- AI-avustaja voi auttaa aineistohallinnassa, kuten asiakirjojen ja tiedostojen järjestämisessä, tallentamisessa ja jakamisessa. Tämä parantaa aineiston saavutettavuutta ja organisointia, mikä helpottaa tiedonhallintaa ja vähentää tiedon hukkaamisen riskiä.

Aikataulujen hallinta:

- AI-avustaja voi avustaa aikataulujen luomisessa ja hallinnassa, muistuttaen tärkeistä määräajoista ja tapaamisista. Tämä auttaa työntekijöitä pysymään aikataulussa ja vähentää unohdettujen tehtävien määrää, mikä parantaa kokonaistuottavuutta.

Sähköpostien käsittely:

- Chatbot voi auttaa sähköpostien hallinnassa, esimerkiksi suodattamalla tärkeitä viestejä, luomalla luonnoksia ja vastaamalla yleisiin kyselyihin. Tämä parantaa viestinnän tehokkuutta ja säästää aikaa, jolloin työntekijät voivat keskittyä ydintehtäviinsä.

Näiden lisäksi AI-avustaja voi tuoda arvoa yritykselle myös muilla innovatiivisilla tavoilla, kuten analysoimalla suuria tietomääriä liiketoiminnan parantamiseksi, ennakoimalla markkinatrendejä ja tukemalla asiakaspalvelua. Kokonaisvaltainen AI-avustaja voisi näin ollen olla merkittävä kilpailuetu LSB:lle, auttaen yritystä pysymään teknologisen kehityksen kärjessä ja parantamaan operatiivista tehokkuuttaan.

4 Projektin eteneminen

Projektimme eteni johdonmukaisesti ja suunnitelmallisesti alusta alkaen. Ensimmäinen vaihe oli selvittää, mihin meidän tulisi keskittyä ja mitä tavoitteita haluamme saavuttaa. Päätimme keskittyä tarjoamaan LSB chatbotin/AI-avustajan, joka helpottaisi heidän päivittäisiä työtehtäviään ja parantaisi tehokkuutta.

4.1 Ryhmän työskentely

Aloimme tutkia eri tapoja toteuttaa AI-avustajaa. Kävimme läpi olemassa olevia teknologioita ja alustoja, joita voisimme hyödyntää projektissamme. Tarkastelimme eri chatbot-ratkaisuja ja kielimalleja, arvioiden niiden soveltuvuutta yrityksen tarpeisiin. Ryhmämme varasi joka päivä tietyn ajan, jonka omistauduimme projektityölle. Pidimme myös säännöllisiä palavereita, joissa kävimme läpi edistymistämme ja käsittelimme mahdollisia ongelmia. Näissä palavereissa päivitimme toisiamme tehtävien etenemisestä, jaoimme ideoita ja ratkoimme haasteita yhdessä. Palavereissa esittelimme myös artikkelien lukemisen tuloksia ja sovelsimme niistä saatuja oppeja projektiin.

4.2 Ideaan liittyvän materiaalin opiskelu

Projektin aikana paneuduimme syvällisesti ideaan liittyvään materiaaliin. Tutkimuksemme keskittyivät erityisesti seuraaviin aiheisiin:

- Phi-3 mallit ja Microsoft Azure AI bot Service:
Luimme näiden teknologioiden dokumentaatiota ja tarkistimme monia keskusteluketjuja, jotka käsittelivät niitä. Tämä auttoi meitä ymmärtämään paremmin niiden toimintaa ja mahdollisuuksia projektissamme.
- Kielimallien tekniset raportit:
Tutustuimme useisiin teknisiin raporteihin ja tutkimuksiin, jotka käsittelivät suurten kielimallien (LLM) suorituskykyä ja niiden soveltuvuutta erilaisiin käyttötarkoituksiin. Näiden raporttien avulla saimme arvokasta tietoa kielimallien vahvuuksista ja heikkouksista.
- RAG (Retrieval Augmented Generation):
Perehdyimme RAG-tekniikkaan, joka yhdistää tiedonhaku- ja generointimallit parantaakseen vastausten tarkkuutta ja relevanssia. Tämä tekniikka osoittautui lupaavaksi mahdollisuudeksi parantaa AI-avustajamme suorituskykyä.
- MoE (Mixture of Experts):
Tutkimme myös Mixture of Experts -mallia, joka jakaa laskennalliset tehtävät useiden asiantuntijamallien kesken parantaen tehokkuutta ja tarkkuutta. Tämä koneoppimistekniikka vaikuttaa erityisen lupaavalta, kun pyritään optimoimaan AI-avustajan suorituskykyä monimutkaisissa tehtävissä.
- Muut artikkelit jotka liittyvät koneoppimiseen ja tekoälyyn

5 Tulokset

Projektimme laajuuden vuoksi käytimme paljon aikaa modernien teknologioiden tutkimiseen, jotta voisimme ymmärtää parhaiten, kuinka kehittää vaatimusten mukainen järjestelmä ja missä määrin se vaatisi resursseja.

5.1 Mitä saatiin selvitettyä

Aloitimme tutkimuksemme pienistä kielimalleista sillä uskoimme, että pilvipalveluissa suoritettavien kielimallien hyödyntäminen paikallisesti pyörivien ohjelmistojen kanssa olisi haastavaa. Emme onneksi investoineet liikaa aikaa niiden tutkimiseen, sillä selvitimme, että suurempia kielimalleja voi hyödyntää paikallisten sovellusten suorittamiseen, vaikka ne sijaitsevat pilvipalveluissa. Tästä sitten suuntasimme tutkimuksemme suurten kielimallien hyödyntämiseen, sillä ne ovat uusinta uutta ja ovat osoittautuneet monille yrityksille hyödyllisiksi.

Tutkimme myös, kuinka kasvattaa kielimallien tietopohjaa ja perehdyimme useisiin menetelmiin, joista tunnetuin on RAG (Retrieval-Augmented Generation). Totesimme, että tästä olisi useita hyötyjä kohdeyritykselle, kuten:

- Kieli mallin tietämyksen kasvattaminen
- Kieli mallin perehdyttäminen työprosesseihin
- Asiakkaiden vaatimusten kommunikointi kielimallille

Perehdyimme myös moni-eksperttijärjestelmiin (MoE, Mixture of Experts) ja niiden mahdolliseen hyödyntämiseen tuotteessa. Keskityimme etuihin sekä haittoihin, ja tämä teknologia vaikuttaa lupaavalta tavoitteidemme saavuttamisessa.

5.2 Arkkitehtuuri

Projektin arkkitehtuuri koostuu useista eri tasoista ja komponenteista, jotka yhdessä mahdollistavat kokonaisvaltaisen AI-avustajan toiminnan. Arkkitehtuuri suunniteltiin joustavaksi ja modulaariseksi, jotta se voi vastata LSB-yrityksen muuttuviin tarpeisiin ja laajentua tulevaisuudessa.

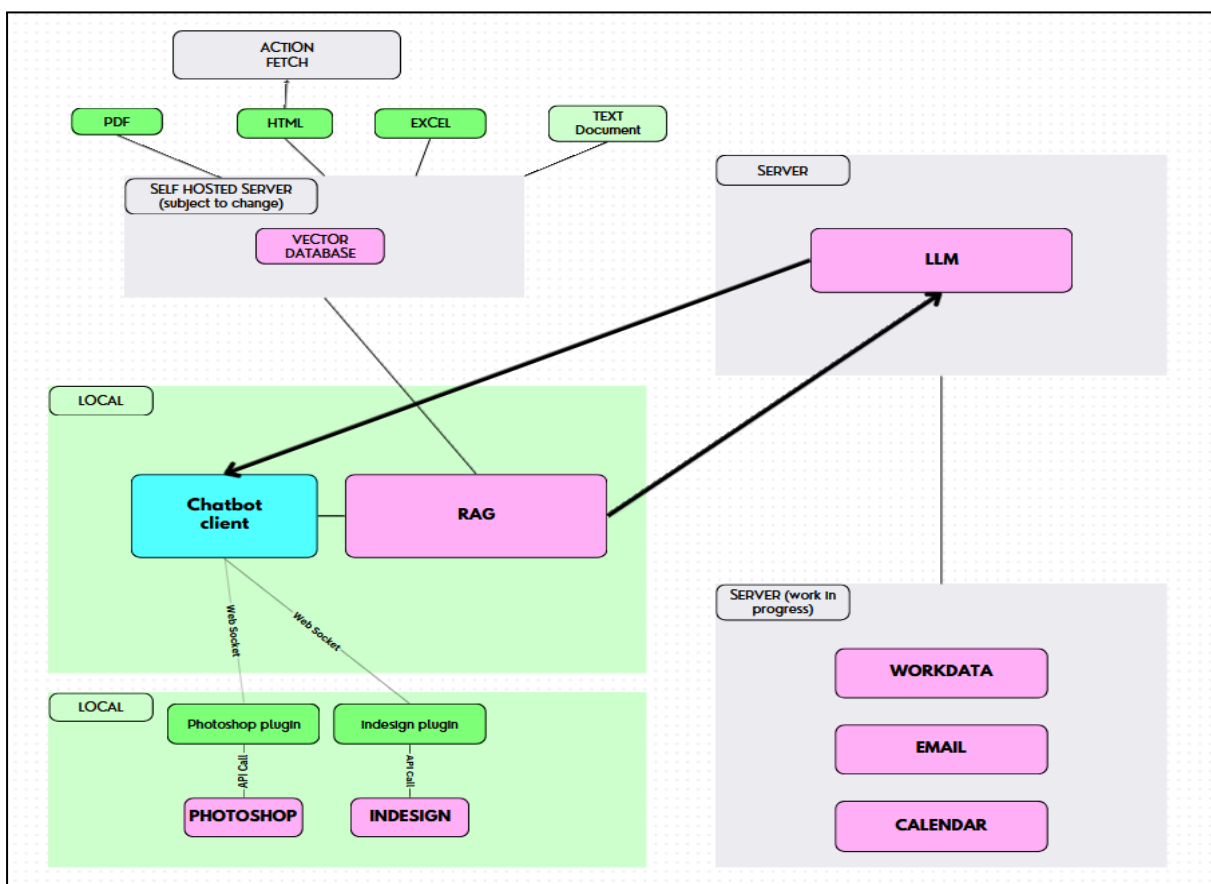
Keskeisiä arkkitehtuurin osa-alueita ovat:

- **Käyttöliittymä ja asiakaspuoli:**
 - Käyttäjät vuorovaikuttavat AI-avustajan kanssa käyttöliittymän kautta. Käyttöliittymä on suunniteltu käyttäjäystävälliseksi ja intuitiiviseksi, jotta työntekijät voivat helposti käyttää ja hyödyntää AI-avustajan tarjoamia palveluita.
- **Tiedonhakumallit (RAG):**
 - Retrieval-Augmented Generation (RAG) yhdistää tiedonhaun ja tekstin generoinnin parantaakseen vastausten tarkkuutta ja relevanssia. Tämä komponentti on keskeinen, sillä se varmistaa, että AI-avustaja pystyy tarjoamaan ajantasaista ja hyödyllistä tietoa käyttäjien kysymyksiin.
- **Suuret kielimallit (LLM):**
 - Palvelimella toimivat suuret kielimallit (LLM) tuottavat vastauksia ja käsittelevät käyttäjien esittämiä kysymyksiä. Nämä mallit hyödyntävät laajaa tietopohjaa ja kehittyneitä luonnollisen kielen käsittelyn algoritmeja tarjotakseen laadukkaita vastauksia.
- **Vektoritietokannat:**

- Vektoritietokannat tallentavat tietoa vektorimuodossa, mikä mahdollistaa nopean ja tarkan tiedonhaun. Tämä rakenne on erityisen hyödyllinen suurten tietomäärien käsittelyssä ja relevanttien tietojen löytämisessä.
- **Integraatiot ja liitännäiset:**
 - AI-avustaja integroidaan suoraan erilaisiin työkaluihin ja sovelluksiin, kuten Adobe Photoshop ja InDesign. Näiden liitännäisten avulla AI-avustaja voi automatisoida ja hallita erilaisia tehtäviä, parantaen työn tehokkuutta ja vähentäen manuaalista työtä.
- **Palvelinpuolen komponentit:**
 - Palvelinpuolen komponentit käsittelevät erilaisiin työdataan, sähköposteihin ja kalenteritietoihin liittyviä tehtäviä. Nämä komponentit mahdollistavat laajan toiminnallisuuden ja tehokkaan integroinnin yrityksen päivittäisiin työtehtäviin.

Arkkitehtuurin tavoitteena on luoda järjestelmä, joka on skaalautuva, tehokas ja helposti ylläpidettävä. Seuraavissa osioissa tarkastellaan tarkemmin arkkitehtuurin yksityiskohtaisia ratkaisuja ja komponentteja, kuten Single LLM, Mixture of Experts (MOE) ja Multi-Head Mixture-of-Experts (MH-MoE).

5.2.1 Single LLM



Kuva 1. Single LLM

Kuvassa (kuva 1, sivu 9) on esitämme yksinkertainen ylätason arkkitehtuurisuunnitelma, joka havainnollistaa chatbot-järjestelmän toiminnallisuutta ja eri komponenttien välistä vuorovaikutusta. Tässä kuvataan keskeiset osat ja niiden yhteydet:

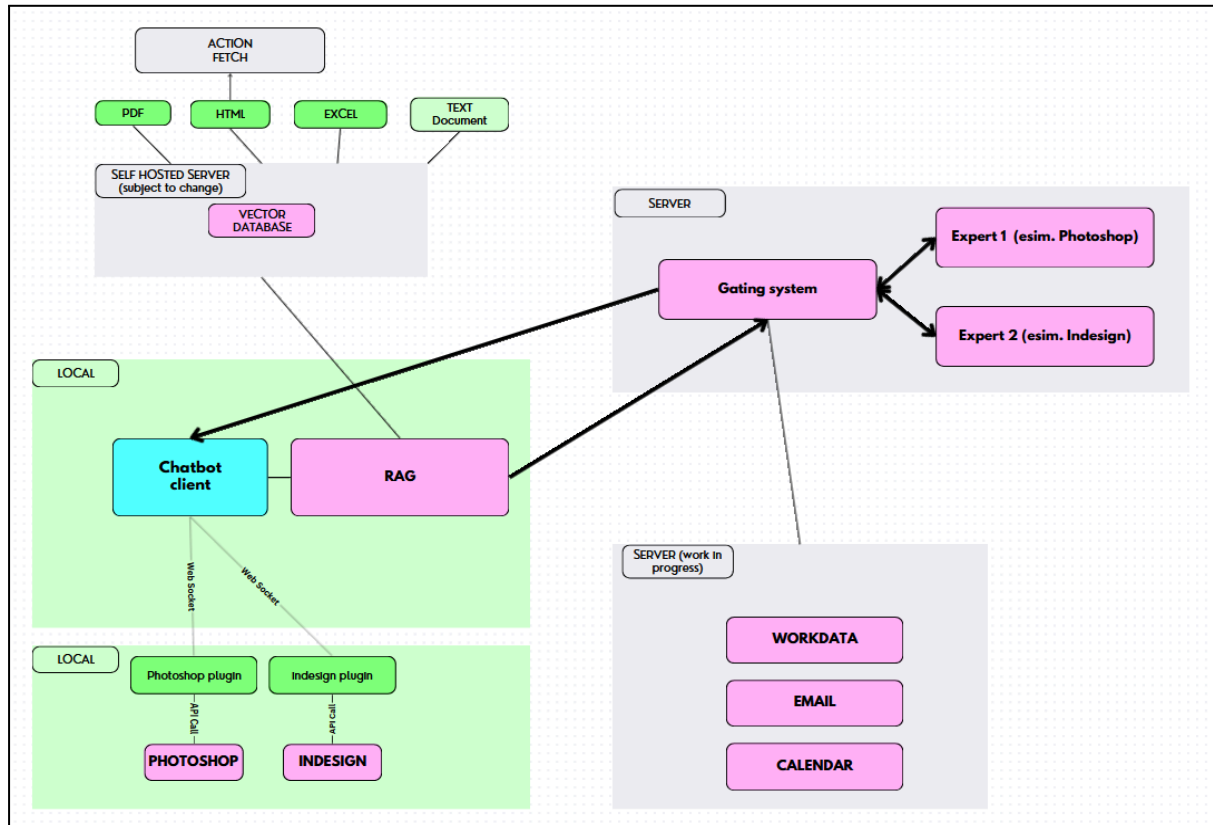
- **Chatbot Client**
 - **Sijainti:** Local (paikallinen)
 - **Tehtävä:** Käyttäjät vuorovaikuttavat chatbotin kanssa tämän komponentin kautta. Se toimii käyttöliittymänä, jonka kautta käyttäjä voi esittää kysymyksiä ja saada vastauksia.
- **RAG (Retrieval-Augmented Generation)**
 - **Sijainti:** Local (paikallinen)
 - **Tehtävä:** Tämä komponentti yhdistää tiedonhaun ja generatiivisen mallin tuottaakseen tarkempia ja relevantimpia vastauksia käyttäjän kysymyksiin. Se hakee tietoa tietokannoista ja käyttää kielimallia vastausten tuottamiseen.
- **LLM (Large Language Model)**
 - **Sijainti:** Server (palvelin)
 - **Tehtävä:** Tämä on suuri kielimalli, joka toimii palvelimella. Se tuottaa vastauksia RAG-komponentin tuottamien hakutulosten ja käyttäjän kysymysten perusteella. Kielimalli on keskitetty palvelimelle, mikä mahdollistaa tehokkaan prosessoinnin ja laajan tietopohjan hyödyntämisen.
- **Vector Database**
 - **Sijainti:** Self Hosted Server (paikallinen palvelin, jonka sijainti voi muuttua)
 - **Tehtävä:** Tämä tietokanta tallentaa vektorimuotoista tietoa, joka mahdollistaa tehokkaan ja tarkan tiedonhaun. Tiedot voivat olla peräisin eri tiedostomuodoista kuten PDF, HTML, Excel ja tekstidokumentit.
- **Plugins (Photoshop ja InDesign)**
 - **Sijainti:** Local (paikallinen)
 - **Tehtävä:** Nämä liitännäiset mahdollistavat integroinnin eri ohjelmistojen kanssa, kuten Photoshop ja InDesign. Tämä integrointi mahdollistaa suoran vuorovaikutuksen AI-avustajan ja graafisten ohjelmistojen välillä, mikä parantaa työprosessien tehokkuutta.
- **Additional Server Components (Workdata, Email, Calendar)**
 - **Sijainti:** Server (palvelin, kehityksessä)
 - **Tehtävä:** Nämä komponentit ovat osa järjestelmän palvelinpuolen laajennuksia, jotka käsittelevät työdataa, sähköpostia ja kalenteritietoja. Ne ovat vielä kehitysvaiheessa mutta tulevat tarjoamaan laajemman toiminnallisuuden ja paremman integroinnin päivittäisiin työtehtäviin.

Prosessi ja tiedon kulku :

1. **Käyttäjän toiminta:** Käyttäjä esittää kysymyksen tai pyynnön chatbotin käyttöliittymän kautta.
2. **Tiedonhaku ja Generointi:** Chatbot client välittää kysymyksen RAG-komponentille, joka hakee relevantteja tietoja vektoridatabasesta ja lähettää kysymyksen LLM palvelimelle.
3. **Vastauksen generointi:** LLM tuottaa vastauksen RAG hakutulosten perusteella ja lähettää sen takaisin chatbot clientille.

4. **Integraatio:** Tarvittaessa vastaukset tai toimenpiteet voidaan integroida paikallisten liitännäisten kautta suoraan ohjelmistoihin kuten Photoshop ja InDesign.
5. **Palvelinpuolen laajennukset:** Työdataa, sähköpostia ja kalenteritietoja voidaan hyödyntää palvelinpuolen komponenttien kautta, kun ne ovat valmiita käyttöön.

5.2.2 Mixture of Experts (MOE)



Kuva 2. MOE

Tässä kuvassa (kuva 2, sivu 11) on esitetty päivitetty yltason arkkitehtuurisuunnitelma, joka käyttää Mixture of Experts (MOE) -mallia. Verrattuna aikaisempaan Single LLM -malliin, tässä järjestelmässä on otettu käyttöön useita asiantuntijamalleja ja niitä ohjaava järjestelmä. Tässä ovat keskeiset eroavaisuudet:

- **Gating System**
 - **Sijainti:** Server (palvelin)
 - **Tehtävä:** Tämä komponentti toimii ohjausjärjestelmänä, joka valitsee sopivimman asiantuntijamallin käyttäjän kysymyksen tai pyynnön perusteella. Se on vastuussa asiantuntijoiden aktivoinnista ja vastuun jakamisesta niiden kesken.
- **Expert Models (Asiantuntijamallit)**
 - **Sijainti:** Server (palvelin)
 - **Tehtävä:** Useita asiantuntijamalleja, kuten Expert 1 (esim. Photoshop) ja Expert 2 (esim. InDesign), on otettu käyttöön. Nämä mallit on erikoistettu eri

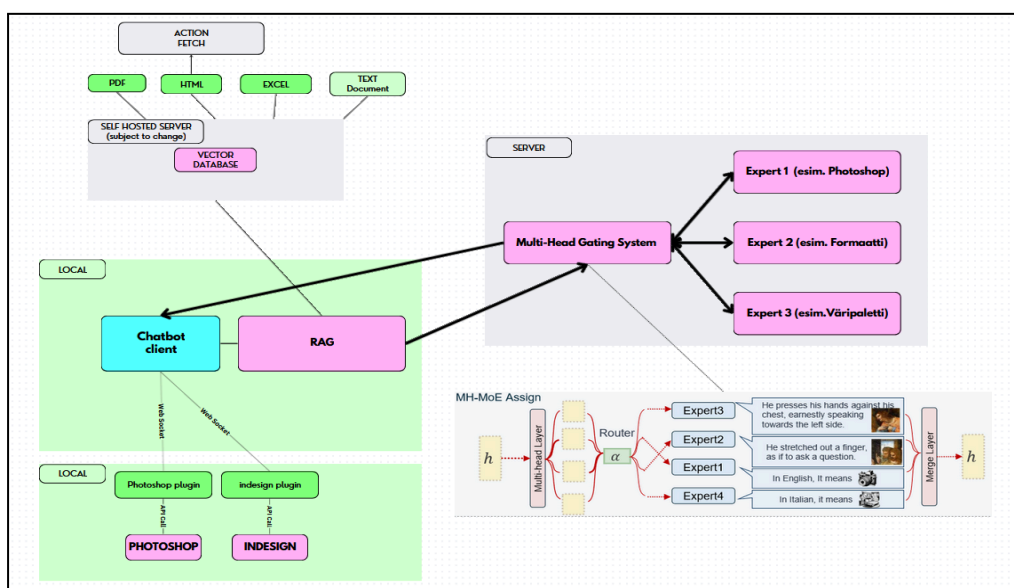
tehtäviin ja ne tuottavat vastauksia tai suorittavat toimintoja omalla erikoisalueellaan.

- **Esimerkkejä:** Expert 1 käsittelee kysymyksiä ja tehtäviä, jotka liittyvät Photoshopiin, kun taas Expert 2 käsittelee InDesigniin liittyviä tehtäviä.
- **Käyttöprosessin eroavaisuudet**
 - **Gating Systemin käyttö:** Kun käyttäjä esittää kysymyksen chatbot clientille, RAG-komponentti hakee relevanttia tietoa ja välittää kysymyksen gating systemille.
 - **Asiantuntijamallin valinta:** Gating system analysoi kysymyksen ja valitsee sopivimman asiantuntijamallin vastaamaan tai suorittamaan tehtävän.
 - **Vastauksen tuottaminen:** Valittu asiantuntijamalli (Expert 1 tai Expert 2) tuottaa vastauksen tai suorittaa tarvittavan toiminnon, jonka jälkeen tulos välitetään takaisin chatbot clientille ja käyttäjälle.
- **Muut komponentit**
 - **Chatbot Client ja RAG** pysyvät paikallisina komponentteina ja toimivat kuten aiemmin kuvatussa Single LLM -mallissa.
 - **Vector Database ja Plugins (Photoshop, InDesign)** toimivat myös samalla tavalla kuin aiemmassa mallissa.

Eroavaisuuksien yhteenveto :

- **Gating System:** Lisää kerroksen ohjausta, joka mahdollistaa useiden asiantuntijamallien tehokkaan käytön.
- **Asiantuntijamallit:** Käytetään erikoistuneita malleja eri tehtäviin, mikä parantaa vastausten tarkkuutta ja relevanssia erityisesti monimutkaisissa tai spesifeissä tehtävissä.
- **Prosessin monimutkaisuus:** Vaatii lisäkerroksen analyysiä ja valintaa kysymysten käsittelyssä, mutta parantaa kokonaisjärjestelmän joustavuutta ja tehokkuutta.

5.2.3 Multi-Head Mixture-of-Experts (MH-MoE)



Kuva 3. Multi-Head Mixture-of-Experts

Tässä kuvassa (kuva 3, sivu 12) on esitetty päivitetty ylätason arkkitehtuurisuunnitelma, joka käyttää Multi-Head Mixture-of-Experts (MH-MoE) -mallia. Tämä malli tuo mukanaan lisäkerroksia asiantuntijamallien käsittelyyn. Tässä ovat keskeiset eroavaisuudet verrattuna aikaisempiin arkkitehtuureihin:

- **Multi-Head Gating System**
 - **Sijainti:** Server (palvelin)
 - **Tehtävä:** Tämä komponentti toimii ohjausjärjestelmänä, joka valitsee useita asiantuntijamalleja samanaikaisesti käyttäjän kysymyksen tai pyynnön perusteella. Toisin kuin yksinkertainen gating system, tässä käytetään useita päätöksiä yhden kysymyksen käsittelyssä.
 - **Toimintaperiaate:** Multi-head gating system jakaa kysymyksen useille eri asiantuntijamalleille, jotka kaikki tuottavat oman osuutensa vastauksesta tai tehtävän suorittamisesta.
- **Multi-Head Expert Models (Monipääasiantuntijamallit)**
 - **Sijainti:** Server (palvelin)
 - **Tehtävä:** Useita asiantuntijamalleja, kuten Expert 1 (esim. Photoshop) ja Expert 2 (esim. InDesign), on otettu käyttöön. Näitä malleja voidaan aktivoida samanaikaisesti eri pään (head) kautta, jolloin ne käsittelevät eri osia samasta tehtävästä tai kysymyksestä.
 - **Esimerkkejä:** Expert 1 käsittelee kysymyksiä ja tehtäviä, jotka liittyvät Photoshopiin, ja Expert 2 käsittelee InDesigniin liittyviä tehtäviä, mutta molemmat voivat osallistua saman kysymyksen käsittelyyn eri näkökulmista.
- **Käyttöprosessin eroavaisuudet**
 - **Multi-Head Gating Systemin käyttö:** Kun käyttäjä esittää kysymyksen chatbot clientille, RAG-komponentti hakee relevanttia tietoa ja välittää kysymyksen multi-head gating systemille.
 - **Asiantuntijamallien valinta:** Multi-head gating system analysoi kysymyksen ja jakaa sen useille asiantuntijamalleille, jotka tuottavat omat osuutensa vastauksesta.
 - **Vastauksen yhdistäminen:** Asiantuntijamallien tuottamat osat yhdistetään yhdeksi kokonaisvastaukseksi, joka välitetään takaisin chatbot clientille ja käyttäjälle.
- **Muut komponentit**
 - **Chatbot Client ja RAG** pysyvät paikallisina komponentteina ja toimivat kuten aiemmin kuvatuissa malleissa.
 - **Vector Database ja Plugins (Photoshop, InDesign)** toimivat myös samalla tavalla kuin aiemmissa malleissa.

Eroavaisuuksien yhteenveto

- **Multi-Head Gating System:** Mahdollistaa useiden asiantuntijamallien samanaikaisen käytön yhden kysymyksen käsittelyssä, mikä parantaa monimutkaisten kysymysten käsittelyä.
- **Monipääasiantuntijamallit:** Käytetään erikoistuneita malleja eri tehtäviin, jotka voivat tuottaa useita osia vastauksesta tai suorittaa tehtävän eri osia rinnakkain.

- **Prosessin monimutkaisuus:** Vaatii kehittyneemmän analyysi- ja yhdistämisympäristön, mutta parantaa kokonaisjärjestelmän joustavuutta, tehokkuutta ja tarkkuutta erityisesti monimutkaisissa tai laajoissa tehtävissä.

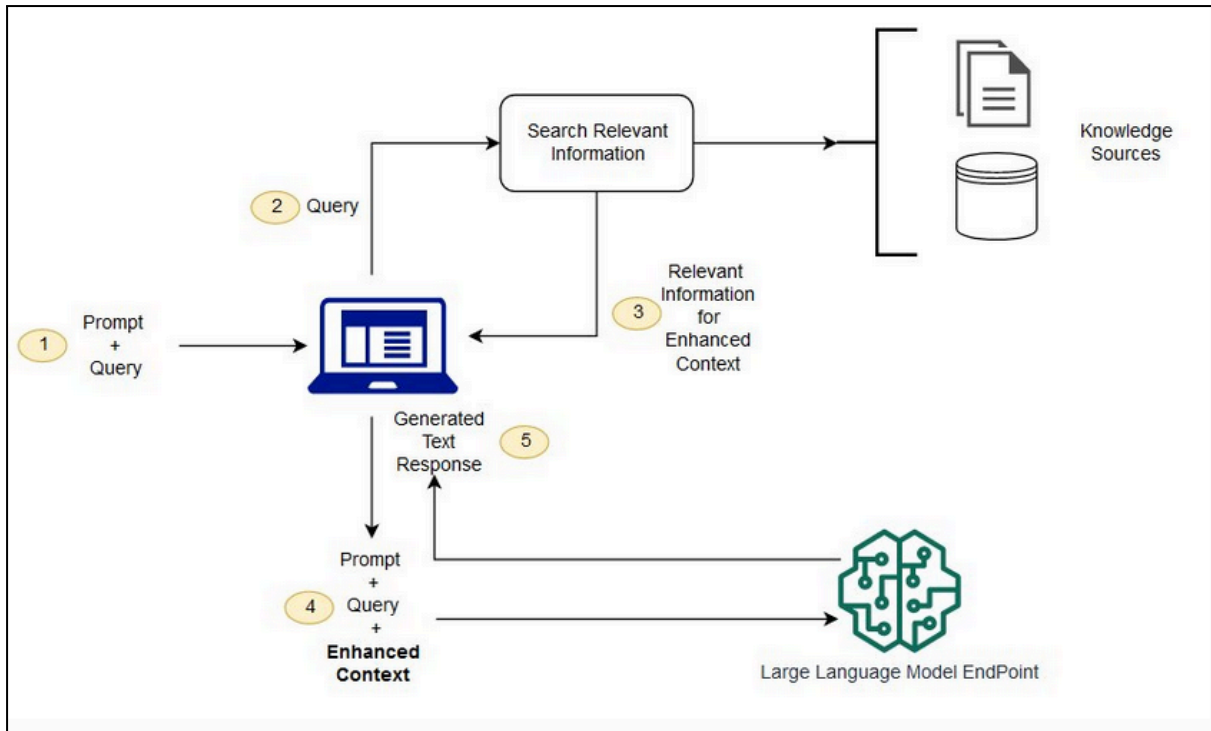
5.3 RAG (Retrieval-Augmented Generation)

Retrieval-Augmented Generation (RAG) on koneoppimisen ja luonnollisen kielen käsittelyn (NLP) malli, joka yhdistää tietojen haun (retrieval) ja tekstin generoinnin (generation) toiminnot (kuva 4, sivu 15). Tämän lähestymistavan avulla voidaan hyödyntää suuria tietomääriä tehokkaasti ja tuottaa laadukkaita vastauksia kyselyihin tai luoda sisältöä.

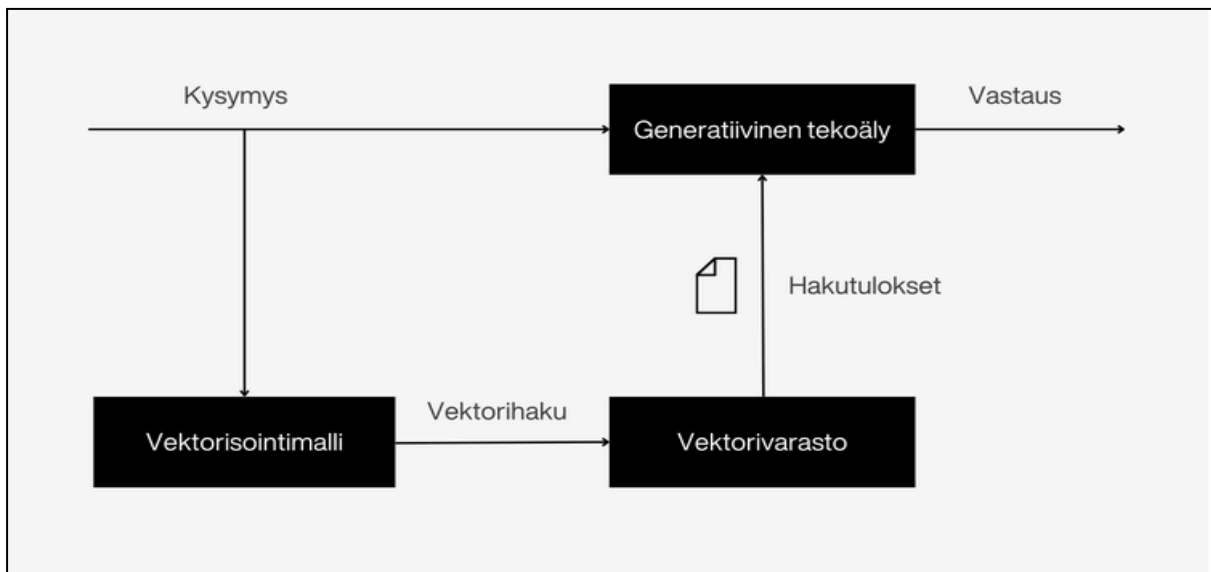
RAG-mallin toiminta alkaa tietojen hausta. Ensimmäisessä vaiheessa malli etsii relevantteja tietoja tietokannasta tai dokumenttivarastosta. Tämä hakuvaihe voi käyttää erilaisia tekniikoita, kuten tiivisteitä (embeddings) tai hakusanoja, löytääkseen kyselyn perusteella merkitykselliset asiakirjat tai tietopalat. Tämä vaihe on kriittinen, sillä sen avulla malli kerää tarkkaa ja ajankohtaista tietoa, joka toimii lähtökohtana generointivaiheelle.

Yksi erityisen tehokas menetelmä tietojen haussa on vektorihaku (kuva 5, sivu 15). Vektorihaku hyödyntää koneoppimista ja lähimmän naapurin algoritmeja löytääkseen semanttisesti samankaltaisia objekteja. Koneoppimismalli opetetaan muuntamaan objektit vektoreiksi niin, että samankaltaiset objektit saavat samanlaiset vektoriesitykset. Vektorit ovat monipuolisia matemaattisia olioita, mutta tässä yhteydessä ne voidaan yksinkertaisesti nähdä numerolistana, kuten Excelin sarakkeena. Tekstin vektoroinnissa käytettävä koneoppimismalli on eräänlainen kielimalli, joka on koulutettu tunnistamaan samankaltaisuuksia. Näitä vektoreita kutsutaan yleisesti "text embeddingeiksi".

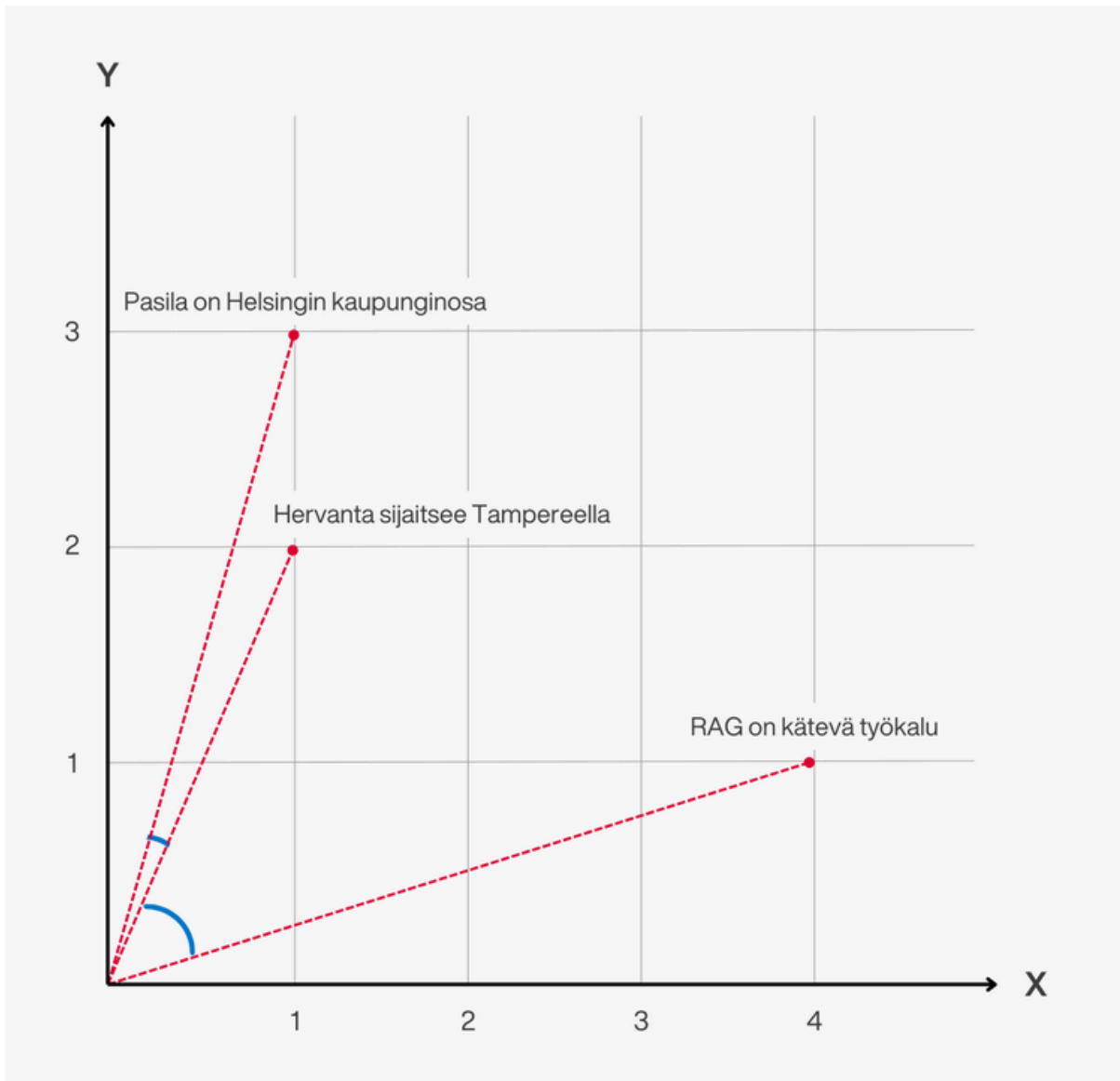
Kun relevantit tiedot on haettu, RAG generointikomponentti astuu kuvaan. Tämä komponentti käyttää haettuja tietoja muodostaakseen vastauksen tai luodakseen tekstiä. Generointivaihe hyödyntää kehittyneitä kielimalleja, jotka ovat erikoistuneet tekstin tuottamiseen. Näin varmistetaan, että lopputulos on johdonmukainen ja laadukas. RAG yhdistää haetut tiedot ja generointimallin tuottaman tekstin. Generointimalli käyttää haettuja tietoja lähtökohtana ja luo niistä johdonmukaisen ja relevantin vastauksen. Tämä yhdistelmä mahdollistaa tarkkojen ja merkityksellisten vastausten tuottamisen nopeasti ja tehokkaasti.



Kuva 4. RAG arkkitehtuuri



Kuva 5. Vektorit



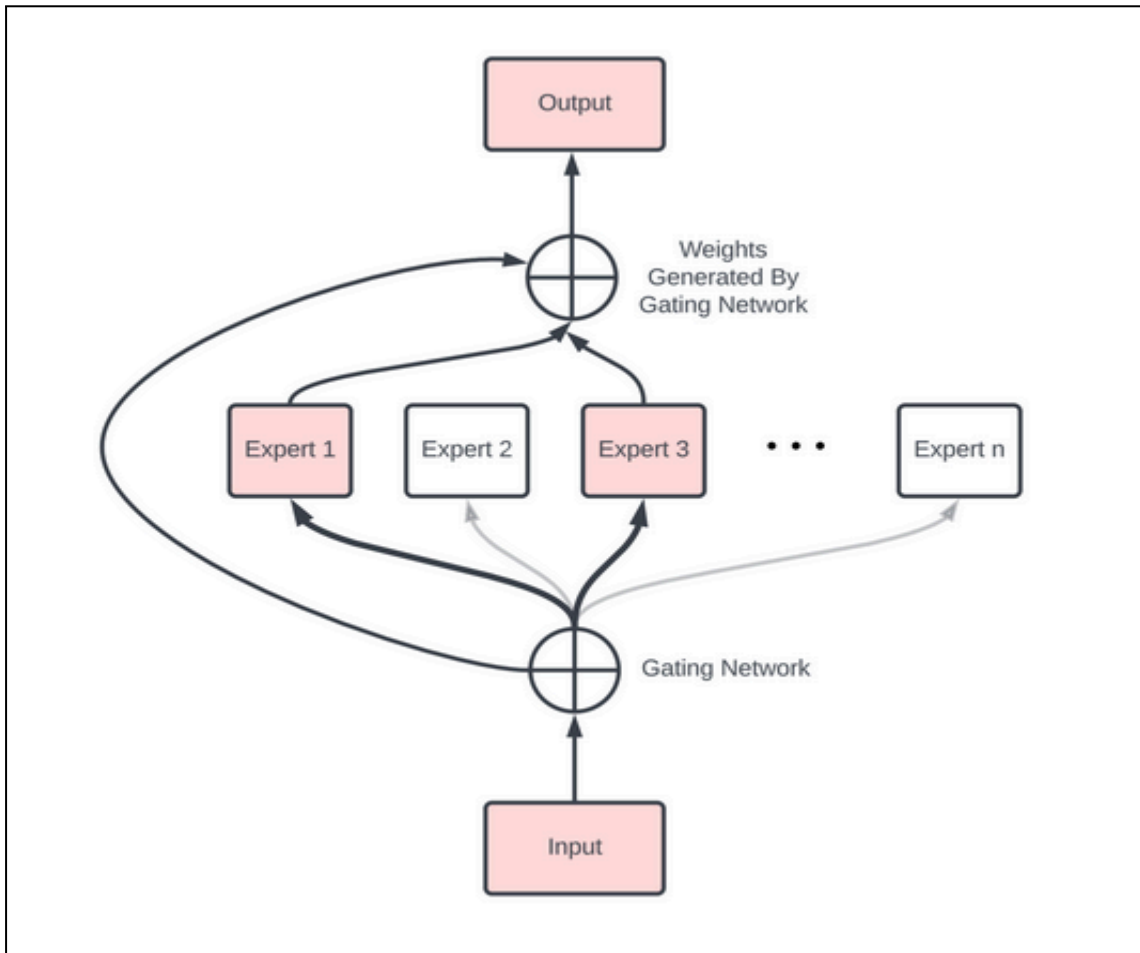
Kuva 6. Vektorihaku

Esimerkiksi (kuva 6, sivu 16) lauseet "Hervanta sijaitsee Tampereella", "Pasila on Helsingin kaupunginosa" ja "RAG on kätevä työkalu" voivat saada vektoriesitykset $[1.0, 2.0]$, $[1.0, 3.0]$ ja $[4.0, 1.0]$. Kaupunginosiin liittyvät lauseet saavat samankaltaisemmat vektorit kuin RAG-lauseeseen verrattuna. Vektorien samankaltaisuutta voidaan mitata useilla eri tavoilla, joista yksi yleinen on kosinietäisyys. Tässä menetelmässä vektorit nähdään moniulotteisen avaruuden janoina, ja niiden välinen kulma voidaan laskea. Kulman kosini kuvaa vektorien samankaltaisuutta.

5.4 MOE (Mixture of Experts)

Mixture of Experts (MOE) on koneoppimismalli (kuva 7, sivu 17), joka parantaa ennustamisen tarkkuutta ja tehokkuutta jakamalla tehtävän eri asiantuntijoille (experts). Tämä lähestymistapa mahdollistaa monimutkaisten ja vaihtelevien ongelmien ratkaisemisen tehokkaammin, sillä jokainen asiantuntija voi keskittyä tiettyyn osatehtävään tai

osa-alueeseen. MOE-malli koostuu kahdesta pääkomponentista: asiantuntijoista (experts) ja portinvartijasta (gating network).



Kuva 7, MOE

1. **Experts:** Nämä ovat yksittäisiä koneoppimismalleja, joista jokainen on koulutettu ratkaisemaan tietty osa-alue kokonaistehtävästä. Esimerkiksi kuvantunnistuksessa yksi asiantuntija voi olla erikoistunut tunnistamaan kasvoja, kun taas toinen voi keskittyä taustan tunnistamiseen. Asiantuntijat voivat olla erilaisia malleja, kuten neuroverkkoja, päätöspuita tai regressiomalleja.
2. **Gating Network:** Tämä komponentti päättää, mitkä asiantuntijat otetaan käyttöön kussakin tilanteessa. Gating Network arvioi syötteen ja valitsee sen perusteella yhden tai useamman asiantuntijan tuottamaan lopullisen ennusteen.

MOE-mallin toiminta voidaan jakaa seuraaviin vaiheisiin:

1. **Syöte:** Malli vastaanottaa syötteen, joka voi olla esimerkiksi kuva, tekstikysely tai muuta dataa.
2. **Gating Network:** Gating Network arvioi syötteen ja valitsee sopivat asiantuntijat tehtävän suorittamiseen. Tämä valinta perustuu syötteen ominaisuuksiin ja portinvartijaverkon koulutukseen.
3. **Asiantuntijoiden Suoritukset (Expert Outputs):** Valitut asiantuntijat käsittelevät syötteen ja tuottavat omat ennusteensa tai tuloksensa.

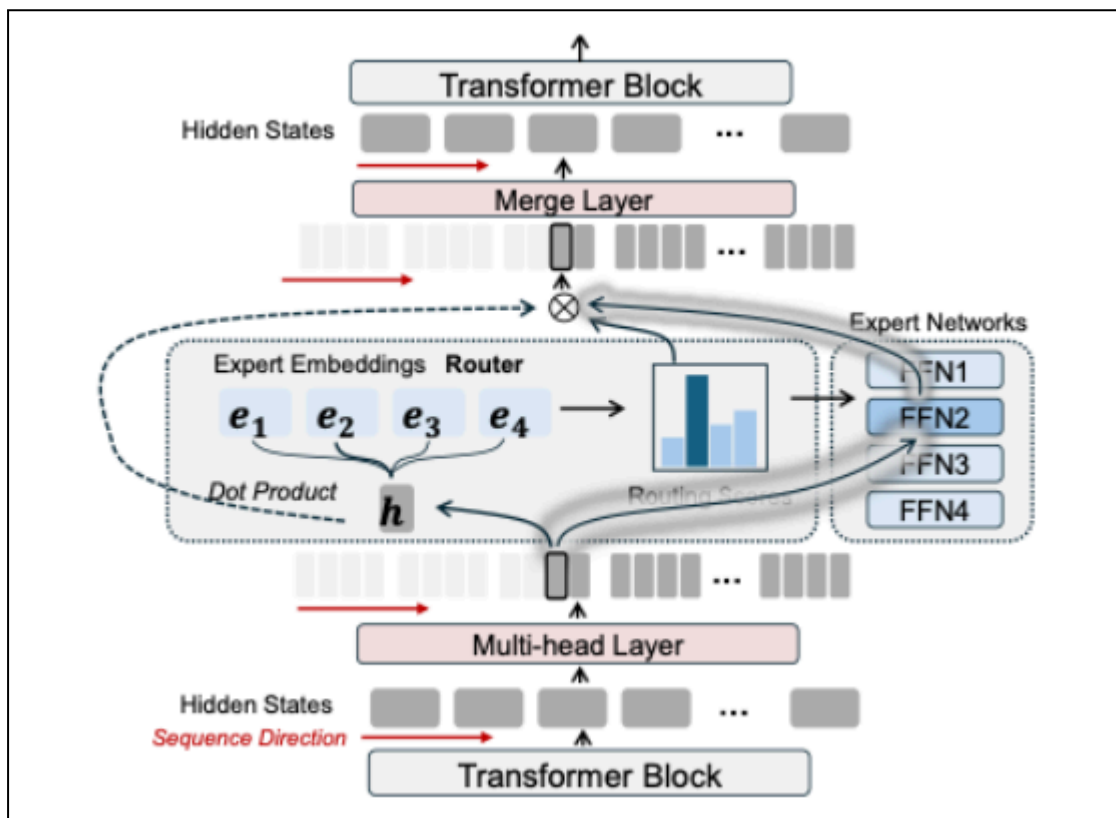
4. **Lopullinen Ennuste:** Gating Network yhdistää asiantuntijoiden tulokset lopulliseksi ennusteeksi. Tämä voi tapahtua painotetulla keskiarvolla, äänestyksellä tai muilla yhdistämistekniikoilla.

MOE-mallin etuja ovat muun muassa:

1. **Tehokkuus:** Malli voi käsitellä suuria ja monimutkaisia tehtäviä jakamalla ne pienempiin osiin.
2. **Tarkkuus:** Erikoistuneet asiantuntijat voivat parantaa ennustamisen tarkkuutta tietyillä osa-alueilla.
3. **Skaalautuvuus:** MOE-malleja voidaan helposti laajentaa lisäämällä uusia asiantuntijoita tarpeen mukaan.

5.5 MH-MoE (Multi-Head Mixture-of-Experts)

Multi-Head Mixture-of-Experts (MH-MoE) on laajennus perinteisestä Mixture of Experts (MOE) -mallista. Se tuo mukanaan useita päätöspäitä (heads), joista jokainen edustaa erillistä MOE-rakennetta (kuva 8, sivu 18). Tämä mahdollistaa entistä monimutkaisemman ja hienovaraisemman päätöksenteon, koska jokainen pää voi erikoistua eri piirteisiin tai osiin syötteestä.



Kuva 8, MH-MoE

Komponentit:

1. **Useita MOE-rakenteita (päitä):** Jokaisessa päässä on omat asiantuntijansa ja

portinvartijaverkkonsa.

2. **Jaetut tai erilliset asiantuntijat:** Asiantuntijat voivat olla joko jaettuja eri päiden kesken tai täysin erillisiä, riippuen mallin arkkitehtuurista.

Toiminnallisuus:

- Jokainen pää käsittelee syötteen itsenäisesti, ja jokaisella päällä on oma portinvartijaverkkonsa, joka valitsee sopivat asiantuntijat.
- Kaikkien päiden tuotokset yhdistetään, tyyppillisesti keskiarvostamalla tai ketjuttamalla, lopullisen ennusteen tuottamiseksi.
- Tämä rakenne mahdollistaa mallin laajemman kyvyn havaita ja oppia erilaisia kuvioita ja vuorovaikutuksia datassa.

Keskeiset Eroavaisuudet:

- **Rakenne:**
 - MOE: Sisältää yhden joukon asiantuntijoita ja yhden portinvartijaverkon.
 - MH-MoE: Sisältää useita asiantuntijajoukkoja ja portinvartijaverkkoja (yksi jokaiselle päälle).
- **Monimutkaisuus:**
 - MOE: Yksinkertaisempi rakenne yhdellä portinvartijaverkolla, joka tekee kaikki päätökset.
 - MH-MoE: Monimutkaisempi rakenne useilla portinvartijaverkoilla ja asiantuntijoilla, mikä mahdollistaa monipuolisemman ja erikoistuneemman päätöksenteon.
- **Soveltuvuus:**
 - MOE: Soveltuu tehtäviin, joissa yksi asiantuntijakerros riittää.
 - MH-MoE: Sopii monimutkaisempiin tehtäviin, kuten monitehtäväoppimiseen, jossa jokainen pää voi keskittyä eri tehtävään tai syötteen osa-alueeseen, parantaen mallin yleistymiskykyä ja kykyä oppia monenlaisia kuvioita.

Käyttötapaukset:

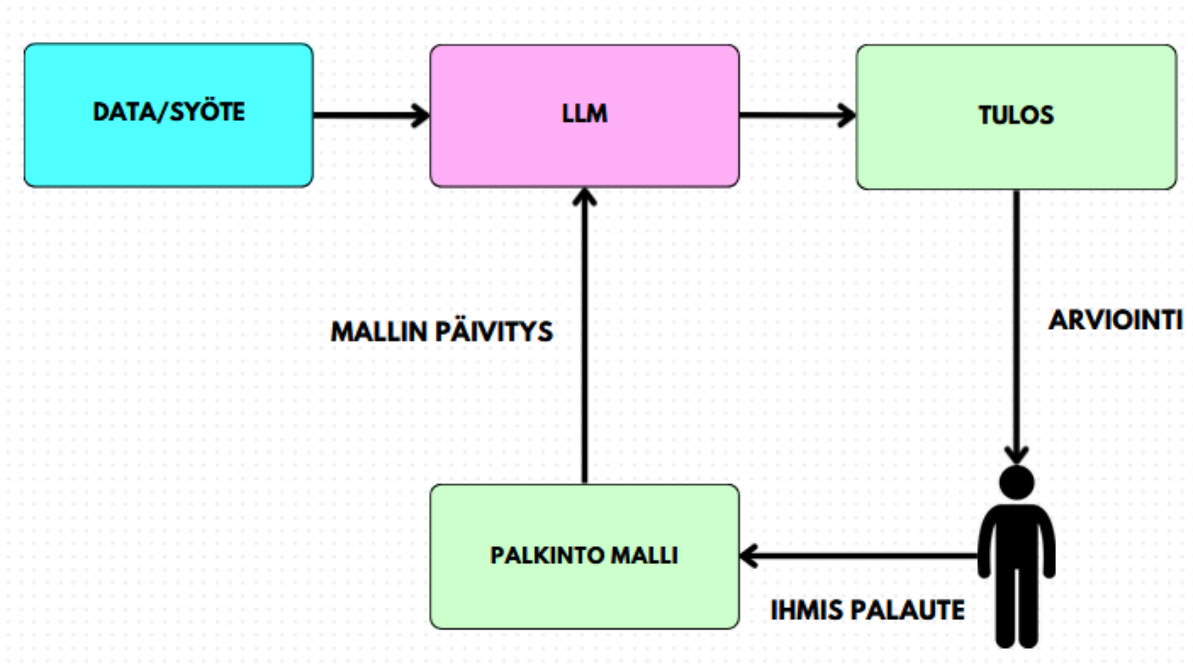
- **MOE:** Tehokas tehtävissä, joissa syöte voidaan luonnollisesti jakaa eri alueisiin, joita käsittelevät erikoistuneet asiantuntijat.
- **MH-MoE:** Hyödyllinen monimutkaisemmissa skenaarioissa, kuten monitehtäväoppimisessa, jossa jokainen pää voi keskittyä eri tehtävään tai syötteen osa-alueeseen, parantaen mallin kykyä yleistää ja oppia monimuotoisia kuvioita.

Yhteenvetona, vaikka sekä MOE että MH-MoE pyrkivät parantamaan mallin suorituskykyä asiantuntijoiden erikoistumisen avulla, MH-MoE lisää malliin lisäkerroksen monimutkaisuutta ja joustavuutta useiden päiden kautta, joilla on omat asiantuntijansa ja portinvartijaverkkonsa. Tämä mahdollistaa entistä tarkemman ja monipuolisemman mallintamisen.

5.6 RLHF (Reinforcement Learning from Human Feedback)

Koska ekspertimalleja saattaa olla vaikea kouluttaa, RLHF tulee todennäköisesti olemaan erittäin hyödyllinen. RLHF on menetelmä, jossa koneoppimismallia parannetaan käyttäjien palautteen avulla (kuva 9, sivu 20). Käyttäjät arvioivat, kuinka hyvin malli suoriutui tietyistä tehtävistä, ja näiden arvioiden perusteella malli voi oppia ja mukautua suorittaakseen

paremmin tulevaisuudessa. Tämä palautesilmukka on erityisen arvokas monimutkaisille ja kriittisille järjestelmille, kuten prespres ympäristöön koulutetuille ekspertimalleille, koska se mahdollistaa jatkuvan parantamisen ja tarkkuuden lisäämisen.



Kuva 9. RLHF

RLHF-prosessi:

1. **Mallin suorituksen arviointi:** Käyttäjät antavat palautetta mallin suorituksesta erilaisissa tehtävissä.
2. **Palautte ja oppiminen:** Malli käyttää tätä palautetta oppiakseen ja parantaakseen suoritustaan. Tämä voi tapahtua esimerkiksi painottamalla enemmän niitä piirteitä, jotka saivat positiivista palautetta, ja vähemmän niitä, jotka saivat negatiivista palautetta.
3. **Koulutus ja optimointi:** Malli päivittää parametrejaan palautteen perusteella ja optimoituu parempaan suorituskyykyyn. Tämä vaihe voidaan toistaa useita kertoja, kunnes malli saavuttaa halutun tason.

Koulutusdata, mukaan lukien käyttäjäpalautteet, halutaan todennäköisesti tallentaa turvallisesti ja varmuuskopioida. Tämä on tärkeää sekä mallin jatkuvan kehityksen että tietoturvan kannalta. Tietoturvaongelmia voi esiintyä, jos data sisältää arkaluonteisia tietoja, kuten asiakaspalautteita tai yrityksen sisäisiä analyytiikkoja, joten on tärkeää varmistaa, että data säilytetään suojatusti ja GDPR mukaisesti.

Kokonaisuudessaan RLHF tarjoaa tehokkaan tavan kehittää ja optimoida LSB-yrityksen käyttämiä ekspertimalleja, varmistaen, että ne pysyvät ajantasaisina ja pystyvät vastaamaan yrityksen muuttuviin tarpeisiin.

5.7 Vektori tietokannat

Vektoritietokannat ovat erityisesti suunniteltuja tietokantoja, jotka mahdollistavat tehokkaan vektorien tallennuksen, haun ja hallinnan. Näiden tietokantojen keskeinen ominaisuus on kyky suorittaa nopeita ja tarkkoja samankaltaisuushakuja suurten vektorikokoelmien joukossa, mikä on erityisen hyödyllistä esimerkiksi koneoppimis- ja tekoälysovelluksissa. Tässä osiossa esittelemme kaksi vektoritietokantaa, Qdrantin ja Chroman, jotka tarjoavat monipuolisia työkaluja vektoripohjaisiin hakuihin ja sovelluksiin.

5.7.1 Qdrant

Qdrant on vektorien samankaltaisuushakuun erikoistunut hakukone, joka tarjoaa kattavan API-vektorien tallentamiseen, hakuun ja hallintaan. Sen ominaisuudet mahdollistavat neuroverkko- tai semantiikkapohjaisen vastaavuuden, fasetoidun haun ja monia muita sovelluksia.

Ominaisuudet:

- **Avoimen lähdekoodin ja ilmainen ikuisesti:** Qdrant on avoimen lähdekoodin projekti ja pysyy ilmaisena.
- **Itseisännöitävä:** Qdrant voidaan asentaa ja ylläpitää omilla palvelimilla.
- **Täysin varusteltu:** Qdrant sisältää kaikki tarvittavat ominaisuudet tuotantovalmiin palvelun rakentamiseen.
- **Yhteisön tuki:** Qdrantilla on aktiivinen ja järjestäytynyt yhteisö, joka tarjoaa tukea ja resursseja.
- **Oppimateriaalit ja dokumentaatio:** Kattavat oppimateriaalit ja dokumentaatio helpottavat käyttöönottoa ja kehitystä.

5.7.2 Chroma

Chroma on avoimen lähdekoodin upotetietokanta, joka tekee LLM-sovellusten (Language Model) rakentamisesta helppoa mahdollistamalla tiedon, faktojen ja taitojen liittämisen LLM

Ominaisuudet:

- **Upotusten ja niiden metadatan tallennus:** Chroma tallentaa upotuksia ja niihin liittyvää metadataa.
- **Dokumenttien ja kyselyiden upotus:** Mahdollistaa dokumenttien ja kyselyiden upottamisen.
- **Upotusten haku:** Tehokas vektorihaku.
- **Monimodaalinen:** Tukee erilaisia datatyyppejä.

- **Yksinkertaisuus ja kehittäjien tuottavuus:** Helppokäyttöinen ja suunniteltu maksimoimaan kehittäjien tuottavuus.
- **Analytiikka haun päällä:** Mahdollistaa analyysin haun päälle.
- **Nopea:** Chroma on erittäin nopea ja suorituskykyinen.

5.7.3 Milvus

Milvus on Zillizin kehittämä vektoritietokanta, joka on erityisesti suunniteltu hallitsemaan ja hakemaan suuria määriä vektoridataa, kuten koneoppimismallien tuottamia upotuksia. Se tarjoaa vankkoja ominaisuuksia, jotka tukevat monenlaisia sovelluksia, mukaan lukien kuvahaku, suositusjärjestelmät ja luonnollisen kielen käsittely.

Keskeiset ominaisuudet:

- **Korkea Suorituskyky:** Optimoitu nopeaan ja tehokkaaseen samankaltaisuushakuun.
- **Skaalautuvuus:** Voi käsitellä suuria tietoaaineistoja.
- **Joustava Käyttöönotto:** Tukee sekä yksittäisiä että hajautettuja käyttöönottoja.
- **Rikas Ekosysteemi:** Integroituu hyvin erilaisten datatieteen ja koneoppimisen työkalujen kanssa.
- **Helppokäyttöisyys:** Tarjoaa käyttäjäystävällisen käyttöliittymän ja kattavan dokumentaation.

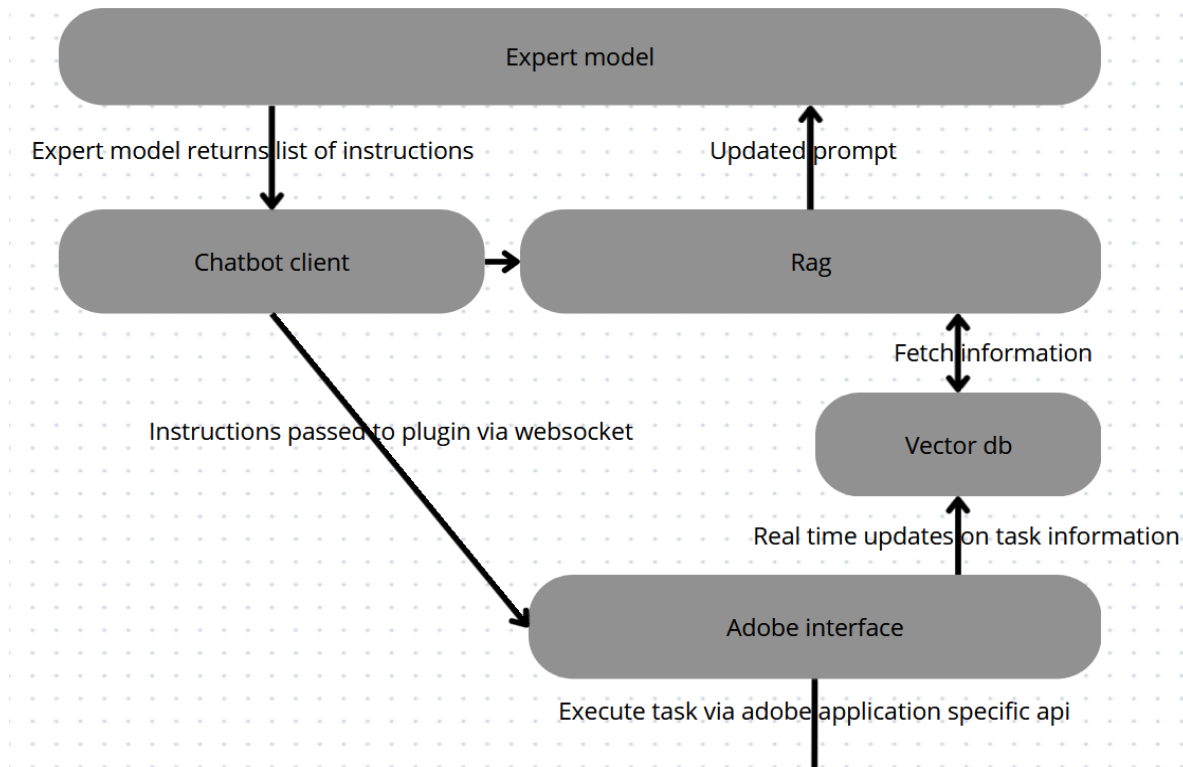
5.8 Adobe rajapinta

Suurimman hyödyn LSB saisi Adobe-sovellusten automaatiosta, sillä suurin osa yrityksen työntekijöiden ajasta kuluu kuvien ja muiden dokumenttien käsittelyyn näissä ohjelmissa. Adobe-sovellusten, kuten Photoshopin, Illustratorin ja InDesignin, käyttö on olennainen osa yrityksen päivittäistä toimintaa, ja näissä sovelluksissa suoritetaan monimutkaisia ja toistuvia tehtäviä, jotka ovat aikaa vieviä ja vaativat tarkkuutta.

Automatisoimalla näitä prosesseja tekoälyn avulla voitaisiin merkittävästi tehostaa työnkulkua ja vähentää virheiden määrää. Tämä vapauttaisi työntekijöille enemmän aikaa, jotta he voisivat keskittyä työn luovempiin osiin, sekä mahdollistaisi useampien töiden käsittelyn samanaikaisesti. Esimerkiksi tekoäly voisi auttaa automaattisessa kuvankäsittelyssä, graafisten elementtien luomisessa ja asiakirjojen taitossa, mikä nopeuttaisi projektien valmistumista ja parantaisi lopputuotteen laatua.

Lisäksi automaatio voisi tuoda merkittäviä kustannussäästöjä pitkällä aikavälillä, sillä toistuvien tehtävien automatisointi vähentää manuaalisen työn tarvetta ja parantaa tehokkuutta. Tämä tekee Adobe-sovellusten automaatiosta erittäin houkuttelevan kohdan, jossa tekoälyä voidaan hyödyntää.

5.8.1 Rajapinnan arkkitehtuuri



Kuva 10. Rajapinta arkkitehtuuri

Tavoitteenamme on hyödyntää Adoben tarjoamia SDK-työkaluja sovelluksemme ja toimintojen automatisoinnin välisessä kommunikoinnissa. Arkkitehtuurimme koostuu useista keskeisistä komponenteista: Retrieval-Augmented Generation (RAG) päivittää kehotteita ja hakee tarvittavat tiedot vektori pohjaisesta tietokannasta. Sitten välitämme asiakaskyselyt asiantuntijamallille, joka palauttaa tarkat ohjeet. Ohjeet välitetään esimerkiksi WebSocketin kautta pluginin/muun kaltaisen rajapinnan avulla Adobe-käyttöliittymään, joka suorittaa tehtävät sovelluskohtaisen API kautta.

5.8.2 Inter-Process Communication (IPC)

Inter-Process Communication (IPC) viittaa mekanismeihin ja menetelmiin, joita eri prosessit (käytävät ohjelmat) käyttävät viestiäkseen keskenään. IPC on välttämätön nykyaikaisissa käyttöjärjestelmissä, koska sen avulla prosessit voivat koordinoita toimintaansa ja jakaa tietoja, vaikka ne toimisivat eri ytimissä tai prosessoreissa. On olemassa useita IPC-mekanismeja:

- Sockets (WebSockets):

Käyttötapaus: Ihanteellinen viestintään verkon yli, joko lokaalisti tai etänä. Tämä voi olla hyödyllistä, jos chatbot ja sovellukset toimivat eri koneissa.

Toteutus: Chatbot voi avata socket yhteyden palvelinkomponenttiin, joka vuorovaikuttaa Photoshopin tai InDesignin kanssa, lähettäen komentoja ja vastaanottaen vastauksia.

- Named Pipes (FIFOs):

Käyttötapaus: Sopii prosessien väliseen viestintään samalla koneella. Ne tarjoavat tavan, jolla chatbot voi lähettää komentoja ja vastaanottaa vastauksia sovelluksilta

Toteutus: Chatbot kirjoittaa komennot named pipes, ja skripti tai sovelluskomponentti lukee nämä komennot ja vuorovaikuttaa Photoshopin tai InDesignin kanssa vastaavasti.

- Shared Memory:

Käyttötapaus: Käytännöllinen nopeaan viestintään ja suurien tietomäärien siirtämiseen. Shared memory voidaan käyttää suurien tietojoukkojen tai usein päivitetävän tiedon välittämiseen.

Toteutus: Shared memory voidaan määrittää komennolle ja vastaukselle, ja sekä chatbot että sovellus voivat käyttää tätä aluetta tietojen lukemiseen ja kirjoittamiseen.

- Message Queues:

Käyttötapaus: Sopii asynkroniseen viestintään, jossa komennot jonotetaan ja käsitellään järjestyksessä. Tämä voi olla hyödyllistä tehtävien jonottamiseen ja useiden pyyntöjen käsittelyyn.

Toteutus: Chatbot asettaa komennot viesti jonoon, ja työntekijä prosessi hakee ja suorittaa nämä komennot Photoshopissa tai InDesignissa.

5.9 Jatko tutkittavaa

5.9.1 Multimodal learning

Multimodaalinen oppiminen yhdistää erilaisia tietolähteitä, kuten tekstiä, kuvaa ja ääntä, parantaen koneoppimismallien suorituskykyä ja monipuolisuutta. Tämän lähestymistavan etuja ovat muun muassa parannettu ennustustarkkuus ja kyky ymmärtää monimutkaisia käsitteitä ja yhteyksiä, joita singlemodal learning ei pysty saavuttamaan. Multimodaalinen oppiminen voi myös johtaa joustavampiin ja yleistettävämpiin malleihin, jotka toimivat paremmin todellisissa sovelluksissa. Haittapuolina ovat kuitenkin lisääntynyt laskennallinen monimutkaisuus ja suurempi datantarve, mikä voi tehdä mallien koulutuksesta ja hienosäädöstä haasteellista ja aikaa vievää. Lisäksi erilaisten datalähteiden yhdistäminen voi tuoda esiin yhteensopivuusongelmia ja vaatia monimutkaisia esikäsittelyvaiheita.

5.9.2 Koulutus areenan kehittäminen

Jos kirjoitamme oman kevyen ohjelmiston, joka emuloi tarvittavia Photoshopin ominaisuuksia, ei meidän tarvitse hyödyntää täyttä Photoshop-ohjelmistoa koulutusprosessissa. Tämä säästäisi huomattavan määrän laskentaresursseja, sillä kevyt

ohjelmisto voidaan optimoida tarkasti tiettyihin tehtäviin, mikä vähentää ylikuormitusta ja parantaa tehokkuutta. Lisäksi, kevyt ohjelmisto tarjoaa nopeammat suoritusajat ja vähentää ohjelmiston lisensointikustannuksia. Kuitenkin tällainen lähestymistapa tuo mukanaan raskaan kehitysvaiheen, joka vaatii merkittävää aikaa ja asiantuntemusta ohjelmiston suunnittelussa ja toteutuksessa. Lisäksi on huolehdittava, että ohjelmisto kattaa kaikki tarvittavat toiminnot ja että se toimii saumattomasti yhteen muiden käytettyjen työkalujen ja prosessien kanssa. Tämä kehitystyö voi myös viivästyttää projektin aloitusta ja vaatii jatkuvaa ylläpitoa ja päivityksiä, jotta ohjelmisto pysyy ajan tasalla ja toimivana.

5.9.3 Hienosäätämisen ja vahvistetun oppimisen yhdistäminen

Hienosäätämisen ja vahvistetun oppimisen yhdistäminen tarjoaa tehokkaan lähestymistavan koneoppimismallien kehittämiseen, jossa yhdistetään valvotun oppimisen tarkkuus ja vahvistetun oppimisen sopeutumiskyky. Hienosäätämisessä valmiiksi koulutettu malli optimoidaan edelleen tietyille tehtäville, mikä parantaa suorituskykyä ja tarkkuutta. Vahvistettu oppiminen puolestaan mahdollistaa tämän mallin oppimisen dynaamisissa ympäristöissä palkkiojärjestelmän avulla, mikä tekee siitä erityisen hyödyllisen monimutkaisissa päätöksentekotilanteissa. Yhdistämällä nämä kaksi menetelmää voidaan mahdollisesti saavuttaa malli, joka ei ainoastaan suoriudu hyvin ennalta määritellyissä tehtävissä, vaan pystyy myös sopeutumaan ja oppimaan uusista tilanteista itsenäisesti.

6 Toteutus

6.1 Tarvittava osaaminen ja henkilöstö

Koska projekti on suunniteltu toteutettavaksi vaiheittain, sen suorittamiseen tarvitaan pienempi tiimi asiantuntijoita, joilla on kokemusta end-to-end-koneoppimisprojektien kehittämisestä. Tiimin koon skaalaaminen mahdollistaisi useiden komponenttien samanaikaisen kehittämisen, mikä nopeuttaisi projektin valmistumista.

On kuitenkin tärkeää huomioida kustannusten ja suorituskyvyn välinen tasapaino. Suuremman henkilöstömäärän palkkaaminen saattaa vaatia myös ylimääräisten johtamishenkilöiden rekrytointia, mikä voi lisätä kokonaiskustannuksia. Tämän vuoksi on olennaista arvioida huolellisesti, missä määrin tiimin laajentaminen tuo lisäarvoa projektille suhteessa siihen, kuinka paljon se kasvattaa kustannuksia.

6.2 Aikataulu



Kuva 11, alustava aikataulu

Projektin alustava aikataulu on suunniteltu vaiheittain (kuva 11, sivu 26), mikä mahdollistaa systemaattisen lähestymistavan kehitysohjelmaan ja varmistaa projektin sujuvan etenemisen. Aikataulun jokaisessa vaiheessa määritellään kesto, tarvittava henkilöstön määrä ja keskeiset tehtävät. Aikataulun avulla voidaan seurata projektin edistymistä ja tehdä tarvittavia säätöjä matkan varrella.

Vaihe 1: Suunnitelma

- **Aika:** 2 kuukautta
- **Henkilöstö:** 2-3 henkilöä
- **Tehtävät:** Projektin alussa keskitytään tarkkaan suunnitteluun, jossa määritellään projektin tavoitteet, vaatimukset ja tekniset yksityiskohdat. Tämä vaihe sisältää myös resurssien ja aikataulujen arvioinnin, jotta projekti voidaan toteuttaa tehokkaasti ja realistisesti.

Vaihe 2: Prototyyppi

- **Aika:** 3-5 kuukautta
- **Henkilöstö:** 3-5 henkilöä
- **Tehtävät:** Tämän vaiheen aikana kehitetään projektin kokeellinen malli eli prototyyppi. Prototyyppi havainnollistaa suunnitellun järjestelmän keskeisiä toiminnallisuuksia ja mahdollistaa varhaisen palautteen keräämisen. Prototyyppi auttaa tunnistamaan teknisiä haasteita ja varmistaa, että järjestelmä vastaa käyttäjien tarpeita.

Vaihe 3: Testaus

- **Aika:** 2 kuukautta
- **Henkilöstö:** 3-5 henkilöä
- **Tehtävät:** Prototyypin valmistumisen jälkeen siirrytään testausvaiheeseen. Tämä vaihe sisältää kattavat testaukset, joiden avulla varmistetaan järjestelmän toimivuus, suorituskyky ja virheettömyys. Testauksessa käytetään erilaisia testimenetelmiä ja -työkaluja, jotta kaikki mahdolliset ongelmat voidaan tunnistaa ja korjata ajoissa.

Vaihe 4: Käyttöönotto

- **Aika:** 1-2 kuukautta
- **Henkilöstö:** 2-3 henkilöä
- **Tehtävät:** Käyttöönotto on projektin kriittinen vaihe, jossa varmistetaan, että järjestelmä voidaan integroida sujuvasti yrityksen nykyisiin prosesseihin ja järjestelmiin. Tämä vaihe sisältää myös loppukäyttäjien koulutuksen ja tarvittavan dokumentaation laatimisen.

Vaihe 5: Lisätoiminnallisuus

- **Aika:** Jatkuva (riippuen projektin laajuudesta ja tarpeista)
- **Henkilöstö:** 3-5 henkilöä
- **Tehtävät:** Viimeisessä vaiheessa keskitytään järjestelmän laajentamiseen ja uusien toiminnallisuuden lisäämiseen. Tämä voi sisältää esimerkiksi uusien ominaisuuksien kehittämistä, suorituskyvyn optimointia ja käyttäjäpalautteen perusteella tehtäviä parannuksia. Tämän vaiheen aikataulu ja henkilöstön määrä voivat vaihdella projektin tarpeiden ja laajuuden mukaan.

6.3 Suunnitelma

Projektin suunnitteluvaiheessa keskitymme seuraaviin keskeisiin tehtäviin, jotka varmistavat onnistuneen toteutuksen ja AI-avustajan kehityksen:

1. Projektin Alustava Määrittely

- **Tavoitteet:** Määritellään selkeät ja mitattavissa olevat tavoitteet AI-avustajan toiminnalle. Tämä sisältää sen, mitä ongelmia avustajan on tarkoitus ratkaista ja millä tavoilla.
- **Vaatimukset:** Kerätään ja dokumentoidaan kaikki vaatimukset, kuten toiminnalliset vaatimukset (esim. mitkä tehtävät AI-avustajan tulee pystyä suorittamaan) ja ei-toiminnalliset vaatimukset (esim. suorituskyky ja tietoturva).

2. Arkkitehtuurisuunnittelu

- **Käyttöliittymä ja Asiakaspuoli:** Suunnitellaan AI-avustajan käyttöliittymä, jonka kautta käyttäjät voivat vuorovaikuttaa sen kanssa. Käyttöliittymän tulee olla käyttäjäystävällinen ja intuitiivinen.
- **Tiedonhakumallit (RAG):** Määritellään ja suunnitellaan Retrieval-Augmented Generation (RAG) -malli, joka yhdistää tiedonhaun ja tekstin generoinnin parantaakseen vastausten tarkkuutta ja relevanssia.

- **Suuret Kielimallit (LLM):** Suunnitellaan suurten kielimallien (LLM) käyttö palvelimella vastauksien tuottamiseksi ja käyttäjien kysymysten käsittelemiseksi.
- **Vektoritietokannat:** Valitaan ja suunnitellaan vektoritietokannan käyttö, joka tallentaa tietoa vektorimuodossa nopeaa ja tarkkaa tiedonhakua varten.
- **Integraatiot ja Liitännäiset:** Suunnitellaan AI-avustajan integraatio eri työkaluihin ja sovelluksiin, kuten Adobe Photoshopiin ja InDesigniin. Tämä mahdollistaa AI-avustajan suoran vuorovaikutuksen graafisten ohjelmistojen kanssa ja parantaa työn tehokkuutta.
- **Palvelinpuolen Komponentit:** Määritellään palvelinpuolen komponentit, jotka käsittelevät työdataa, sähköposteja ja kalenteritietoja.

3. Tekninen Tiekartta

- **Kehitysvaiheet:** Laaditaan yksityiskohtainen kehitysaikataulu, joka sisältää kaikki tarvittavat vaiheet suunnittelusta käyttöönottoon. Tämä aikataulu jaetaan selkeisiin osiin, kuten suunnittelu, prototyypin kehittäminen, testaus ja käyttöönotto.
- **Resurssien Allokointi:** Määritellään tarvittavat resurssit, kuten henkilöstö, budjetti ja laitteistot. Tämä sisältää myös tiimin roolit ja vastuut.
- **Riskienhallinta:** Tunnistetaan mahdolliset riskit ja laaditaan suunnitelmat niiden hallitsemiseksi ja minimoimiseksi.

6.4 Prototyyppi

Prototyyppi on projektin alkuvaiheessa luotava kokeellinen malli, joka havainnollistaa suunnitellun järjestelmän keskeisiä toiminnallisuuksia. Sen tarkoituksena on testata ja demonstroida järjestelmän perusominaisuudet sekä kerätä palautetta mahdollisten parannusten ja muutosten tekemiseksi ennen varsinaisen tuotantoversioon kehittämistä. Prototyyppi auttaa tunnistamaan teknisiä haasteita, arvioimaan käytettävyyttä ja varmistamaan, että järjestelmä vastaa käyttäjien tarpeita.

Prototyypin Kehittämiseen Vaadittavat Resurssit

1. **Osaava Tiimi:**
 - **Koneoppimisasiantuntijat:** Kehittävät ja hienosäätävät AI-mallit.
 - **Ohjelmistokehittäjät:** Integroivat AI-mallit sovellukseen ja kehittävät tarvittavat käyttöliittymät.
 - **DevOps-asiantuntijat:** Vastaavat infrastruktuurista ja pilvipalveluiden konfiguroinnista.
2. **Tietolähteet ja Data:**
 - **Koulutusdata:** Tarvitsemme lähes täyden pääsyn LSB laajaan datapankkiin.
 - **Testidata:** Erillinen data, jolla voidaan arvioida mallien suorituskykyä ja tarkkuutta.
3. **Infrastruktuuri:**
 - **Laskentaresurssit:** Tehokkaat virtuaalikoneet tai kontit laskentatehtävien suorittamiseen.

- **Tallennustilat:** Luotettavat ja skaalautuvat tallennusratkaisut datan säilyttämiseen ja hallintaan.

Prototyypin Kehitysvaiheet

1. **Tarpeiden Määrittely:**
 - Tunnistetaan ja määritellään järjestelmän keskeiset toiminnallisuudet ja vaatimukset.
2. **Sovellusten Kehitys:**
 - Kehitetään käyttöliittymät ja taustajärjestelmät, jotka mahdollistavat AI-mallien hyödyntämisen.
3. **Datan Keräys ja Käsittely**
 - Kerätään ja prosessoidaan data muotoon jota voi koneoppimisessa hyödyntää.
4. **Poc(Proof Of Concept)-Kehittäminen ja Kouluttaminen**
 - Varmistetaan että suunniteltu malli toimii pienellä skaalalla ennen kuin jatkokehitetään.
5. **Poc Palautteen Kerääminen:**
 - Demonstroidaan poc sidosryhmille ja kerätään palautetta mahdollisten parannusten ja muutosten tekemiseksi.
6. **Arkkitehtuurin Suunnittelu:**
 - Suunnitellaan järjestelmän arkkitehtuuri, joka kattaa sekä AI-mallit että niitä tukevat sovellukset ja palvelut.
7. **Loppu Mallien Kehitys ja Koulutus:**
 - Kehitetään ja koulutetaan tarvittavat AI-mallit käyttäen määriteltyä dataa ja koneoppimiskirjastoja.
8. **Integraatio ja Testaus:**
 - Integroidaan eri komponentit ja suoritetaan kattavat testaukset varmistaakseen järjestelmän toimivuuden.
9. **Palautteen Kerääminen:**
 - Demonstroidaan prototyyppi sidosryhmille ja kerätään palautetta mahdollisten parannusten ja muutosten tekemiseksi.

6.5 Testaus

Projektin testausvaihe on olennainen osa AI-avustajan kehitysprosessia, jossa varmistetaan järjestelmän toimivuus, suorituskyky ja luotettavuus. Testausvaiheessa pyritään löytämään ja korjaamaan virheitä sekä varmistamaan, että järjestelmä täyttää sille asetetut vaatimukset. Tässä vaiheessa keskitytään sekä yksittäisten komponenttien että koko järjestelmän testaukseen.

Testausvaiheen Tavoitteet

1. **Toiminnallisuuden Varmistaminen:**
 - Testataan, että AI-avustajan kaikki toiminnallisuudet toimivat suunnitellusti ja täyttävät vaatimukset.
2. **Suorituskyvyn Mittaaminen:**

- Arvioidaan AI-avustajan suorituskyky eri kuormitustilanteissa varmistaen, että se pystyy käsittelemään odotetun määrän käyttäjiä ja tehtäviä.
3. **Yhteensopivuuden Testaaminen:**
 - Varmistetaan, että AI-avustaja integroituu saumattomasti muihin järjestelmiin ja työkaluihin, kuten Adobe Photoshopiin ja InDesigniin.
 4. **Käytettävyyden Arviointi:**
 - Testataan AI-avustajan käyttöliittymän ja vuorovaikutuksen helppokäyttöisyys ja intuitiivisuus loppukäyttäjien näkökulmasta.

Testausvaiheen Vaiheet

1. **Yksikkötestaus:**
 - Jokainen komponentti testataan erikseen varmistamalla, että ne toimivat odotetusti. Tämä vaihe sisältää erityisesti tiedonhakumallien (RAG) ja suurten kielimallien (LLM) testauksen.
2. **Integraatiotestaus:**
 - Testataan eri komponenttien yhteistoiminta ja varmistetaan, että ne toimivat yhdessä ilman ongelmia. Tämä vaihe on erityisen tärkeä, kun varmistetaan AI-avustajan integraatiot Adobe-sovellusten kanssa.
3. **Kuormitustestaus:**
 - Arvioidaan järjestelmän suorituskyky eri kuormitustilanteissa. Testataan, miten AI-avustaja käsittelee suuria määriä käyttäjiä ja pyyntöjä samanaikaisesti.
4. **Käytettävyydestestaus:**
 - Testataan käyttöliittymän ja vuorovaikutuksen helppokäyttöisyys oikeiden käyttäjien kanssa. Käyttäjätestauksella varmistetaan, että AI-avustaja on intuitiivinen ja helppokäyttöinen.
5. **Regressiotestaus:**
 - Varmistetaan, että aiemmin testatut ja hyväksytyt toiminnallisuudet toimivat edelleen uusien muutosten jälkeen. Tämä vaihe estää uusien muutosten aiheuttamat ongelmat jo toimivissa osissa.

Testausvaiheen Resurssit

1. **Testausympäristö:**
 - Luodaan erillinen testausympäristö, joka simuloi mahdollisimman tarkasti tuotantoympäristöä. Tämä mahdollistaa realistiset testit ilman, että ne häiritsevät tuotantoympäristöä.
2. **Automatisoidut Testityökalut:**
 - Hyödynnetään automatisoituja testityökaluja yksikkö-, integraatio- ja kuormitustestien suorittamiseen. Tämä nopeuttaa testausprosessia ja parantaa kattavuutta.
3. **Käyttäjätestausryhmät:**
 - Muodostetaan ryhmiä, jotka koostuvat loppukäyttäjistä, jotka testaavat AI-avustajan käytettävyyttä ja antavat palautetta sen toiminnasta.

6.6 Käyttöönotto

Käyttöönoton aikana keskitytään järjestelmän asennukseen, konfigurointiin, loppukäyttäjien koulutukseen ja järjestelmän käytön valvontaan.

Käyttöönoton Vaiheet

1. Asennus ja Konfigurointi:

- **Järjestelmän Asennus:** AI-avustajan eri komponenttien asentaminen ja konfigurointi tarvittaville palvelimille ja työasemille. Tämä sisältää esimerkiksi vektoritietokannan, RAG-mallin ja suuren kielimallin asennuksen.
- **Integraatiot:** Varmistetaan, että AI-avustaja integroidaan sujuvasti Adobe Photoshopiin, InDesigniin ja muihin yrityksen käyttämiin ohjelmistoihin. Tämä sisältää liitännäisten (plugins) asennuksen ja konfiguroinnin.
- **Verkkoasetukset:** Varmistetaan, että kaikki verkkoyhteydet ja tietoturva-asetukset ovat oikein konfiguroituja, jotta järjestelmä voi toimia luotettavasti ja turvallisesti.

2. Käyttöönoton Testaus:

- **Alustavat Testit:** Suoritetaan alustavat testit varmistaakseen, että kaikki järjestelmän osat toimivat oikein asennuksen jälkeen.
- **Käyttäjätestit:** Otetaan pieni ryhmä loppukäyttäjiä testaamaan järjestelmää käytännössä. Tämä auttaa tunnistamaan mahdolliset käytettävyysongelmat ja muut kehitystarpeet.

3. Käyttäjäkoulutus:

- **Koulutusmateriaalit:** Laaditaan kattavat koulutusmateriaalit, jotka opastavat käyttäjiä AI-avustajan käytössä. Tämä sisältää käyttöohjeet ja usein kysytyt kysymykset (FAQ).
- **Koulutustilaisuudet:** Järjestetään koulutustilaisuuksia, joissa käyttäjät saavat henkilökohtaista ohjausta ja voivat esittää kysymyksiä. Tämä voi tapahtua sekä paikan päällä että etänä.

4. Seuranta ja Tuki:

- **Käyttöönoton Valvonta:** Seurataan järjestelmän toimintaa aktiivisesti ensimmäisten viikkojen aikana. Varmistetaan, että kaikki ongelmat havaitaan ja ratkaistaan nopeasti.
- **Tukipalvelut:** Tarjotaan käyttäjille jatkuvaa tukea ja apua ongelmatilanteissa. Tämä voi sisältää tukipuhelimen, sähköpostituen ja chat-tuen.

5. Palaute ja Jatkuva Parannus:

- **Palautteen Kerääminen:** Kerätään käyttäjien palautetta järjestelmän toiminnasta ja käytettävyydestä. Tämä voidaan tehdä kyselyiden, haastatteluiden ja palauteohjelmien kautta.
- **Järjestelmän Päivitykset:** Tehdään tarvittavat päivitykset ja parannukset palautteen perusteella. Tämä voi sisältää uusien ominaisuuksien lisäämistä, suorituskyvyn optimointia ja käytettävyyden parantamista.

6.7 Lisätoiminnallisuus

Lisätoiminnallisuusvaiheessa keskitytään laajentamaan ja parantamaan AI-avustajan ominaisuuksia projektin perusversion valmistumisen jälkeen. Tämä vaihe on tärkeä, jotta järjestelmä voi vastata muuttuvien tarpeiden ja uusien vaatimusten mukaisesti. Lisätoiminnallisuudet parantavat AI-avustajan kykyä käsitellä monimutkaisempia tehtäviä, integroitua paremmin muihin järjestelmiin ja tarjota lisää arvoa käyttäjille.

7 Yhteenveto

LSB-yrityksen AI-avustajaprojekti on suunniteltu vastaamaan yrityksen tarpeisiin ja hyödyntämään tekoälyn mahdollisuuksia päivittäisissä työtehtävissä. Projekti on jaettu useisiin vaiheisiin, joista jokaisella on omat selkeät tavoitteensa ja tehtävänsä.

7.1 Projektin Tausta

LSB on paino- ja mediayritys, joka on havainnut teknologian nopean kehityksen vaikutukset toimialallaan. Perinteisesti manuaalinen ja työvoimavaltainen ala on alkanut hyödyntää digitalisaatiota toimintojen tehostamiseksi ja kilpailukyvyn parantamiseksi. Tämän projektin tavoitteena on selvittää, kuinka tekoälyä voidaan hyödyntää yrityksen päivittäisissä tehtävissä ja mitä teknologioita tarvitaan lisäarvoa tuottavan lopputuloksen saavuttamiseksi.

7.2 Suunnittelu ja Toteutus

Projektin suunnitteluvaiheessa määritellään AI-avustajan tavoitteet, arkkitehtuuri ja tekninen tiekartta. Suunnittelun jälkeen kehitetään prototyyppi, joka havainnollistaa järjestelmän keskeisiä toiminnallisuuksia. Testausvaiheessa varmistetaan järjestelmän toimivuus ja suorituskyky ennen varsinaista käyttöönottoa.

7.3 Käyttöönotto ja Lisätoiminnallisuus

Käyttöönoton aikana keskitytään AI-avustajan asennukseen, konfigurointiin ja loppukäyttäjien koulutukseen. Käyttöönoton jälkeen keskitytään järjestelmän laajentamiseen ja uusien toiminnallisuuden lisäämiseen.

7.4 Johtopäätös

AI-avustajan käyttöönotto mahdollistaa LSB-yritykselle tehokkaamman työskentelyn ja vähentää manuaalisen työn määrää. Tämä projekti antaa yritykselle mahdollisuuden olla mukana tekoälyn tuomissa mullistavissa mahdollisuuksissa heti alusta alkaen. Järjestelmä on suunniteltu joustavaksi ja laajennettavaksi, mikä mahdollistaa sen kehittämisen ja parantamisen vastaamaan tulevaisuuden tarpeita. Tämä projekti vahvistaa LSB:n asemaa teknologian edelläkävijänä ja parantaa sen kilpailukykyä markkinoilla.