

Rapport

Fråga GPT

Utvärdering av stöd till Fråga ESV med generativa språkmodeller



EKONOMISTYRNINGSVERKET

Publikationen kan laddas ner
från ESV:s webbplats esv.se.

Datum: 2023-08-22

Dnr: 2023-07803-1

ESV-nr: 2023:41

Copyright: ESV

Rapportansvarig: Emma Wallerö

Förord

Denna rapport beskriver ett internt projekt inom Ekonomistyrningsverket i vilket medarbetare från enheten Redovisning och finansiering tillsammans med datalabbet har utforskat om generativa språkmodeller kan vara ett stöd för de utredare som svarar på myndigheternas frågor om redovisning och finansiering.

Projektgruppen har bestått av KTH-student Emil Hammarberg, utredare Svenne Junker (t.o.m. 2023-04-30), projektledare Emma Wallerö och data scientist Oskar Weinberger.

Kollegorna på enheten Redovisning och finansiering har bidragit med viktig kunskap, deltagit i diskussioner och lämnat värdefulla synpunkter på applikationen och de tre språkmodellerna.

Projektledare Emma Wallerö har skrivit denna rapport.

Stockholm

2023-08-22

Martin Sparr

Enhetschef

Emma Wallerö

Utredare

Innehåll

Sammanfattning	5
1 Inledning	6
2 Överskådlig metodbeskrivning	7
3 Beskrivning av applikationen	9
4 Språklig analys	11
4.1 Medellängd på svar	11
4.2 Läsbarhet.....	11
4.3 Vanliga ord och ordsekvenser	12
4.4 Ordklasser	15
5 Återkoppling från utredare	17
5.1 Betyg	17
5.2 Övergripande kommentarer	17
5.2.1 Sentimentanalys.....	18
5.2.2 Innehåll och teman	19
5.3 Enkätfrågor	19
5.4 Övrig feedback och sammanfattning av feedback	21
6 Exempel på frågor och svar	22
7 Resultat och diskussion	26
7.1 Svar på undersökningens frågeställningar.....	26
I vilken grad kan de testade språkmodellerna vara ett internt stöd för Fråga ESV? .	26
Vilken typ av generativ språkmodell fungerar bäst för syftet?.....	26
Presterar modeller som exponerats för Fråga ESV-data på olika sätt bättre än en modell som inte har det?	26
Kan en applikation som genererar svar baserat på frågor ställda inom "Fråga ESV" vara hjälpsam eller intressant att vidareutveckla för utredare på RF?	26
7.2 Förslag på fortsatt arbete.....	27
Utvärdera sökmotor	27
Undersök hur bra GPT-4 är på att hantera de matchade dokumenten	27
Ge exempel på frågor och svar till generativ modell	28
Utforska alternativa modelltyper	28
Undersök en alternativ typ av applikation	28
Referenser	29
Bilaga 1 - Metodbeskrivning	30
7.3 Teknisk beskrivning	30
7.3.1 Modell 1.....	30
7.3.2 Modell 2.....	31
7.3.3 Modell 3.....	32
Bilaga 2 – Manuell sentimentanalys	33

Sammanfattning

Att svara på frågor som kommer in till Ekonomistyrningsverkets (ESV:s) webbtjänst Fråga ESV tar lång tid, och ett stöd i svarsprocessen skulle därmed ha potential att öka effektiviteten. Vi har tagit fram en prototyp av en applikation som utifrån inkomna frågor genererar tre svar från tre olika generativa språkmodeller baserade på GPT-4 och GPT-3 från OpenAI. Applikationen testades av enheten Redovisning och Finansiering under fyra veckor. Under denna tid testade utredare applikationen 36 gånger med olika frågor från myndigheter. I denna rapport redovisar och diskuterar vi analys, resultat och upplevelse av applikationen.

Följande tre modeller har testats:

- **Modell 1:** en GPT-4-modell som fått lite kontext vad gäller fråga och uppgift men ingen extra information utöver detta.
- **Modell 2:** en kombination av en sökmotor och en GPT-4-modell
- **Modell 3:** en GPT-3-modell som finjusterats med hjälp av genererade frågor och svar.

I projektet utförde vi språkliga analyser av texter så som beräkningar av medellängd på svar, läsbarhetsindex och vanliga ord och ordsekvenser. Vi gjorde också analyser på återkopplingen vi fått från användarna i form av en övergripande manuell analys och en sentimentanalys. Modell 2 var den modell som låg närmast de slutgiltiga svaren som sedan skickades till frågeställaren vad gäller medellängd på svar och vanliga ord eller ordföljder. För ordklasser hade alla svarstyper liknande distribution. Enligt återkopplingen från användarna fick Modell 2 bäst betyg och Modell 3 sämst betyg.

Resultatet av försöket visar att en GPT-4-modell i kombination med en vektorsökmodell presterade bäst överlag och att en finjusterad (fine-tuned på engelska) modell generellt presterade sämst. Användarna ansåg att de genererade svaren ofta inte var särskilt hjälpsamma. Trots detta anser flera utredare att det skulle vara värdefullt att fortsätta utveckla applikationen.

Möjliga vägar framåt är att fokusera på arkitekturen för Modell 2, eller att dela upp applikationens funktionalitet i två. Det senare alternativet innebär att utredaren kan hitta avsnitt ur regelverket som är semantiskt lika den ställda frågan och sedan använda sig av en GPT-modell för att formulera ett svar baserat på den del av regelverket man tycker är mest relevant.

1 Inledning

Ekonomistyrningsverket tar fram och beslutar om föreskrifter och allmänna råd till ett antal förordningar inom det ekonomiadministrativa området. Det samlade regelverket benämns EA-regelverket (ekonomiadministrativa regelverket).

Fråga ESV är en webbtjänst främst riktad till statliga myndigheter i vilken man kan ställa frågor inom ett antal olika ämnesområden.¹ Varje år kommer ett stort antal frågor om redovisning och finansiering och att besvara dessa är därför ett löpande och ganska tidskrävande arbete.² Svaren tar sin utgångspunkt i EA-regelverket.

Som ett led i Ekonomistyrningsverkets interna arbete med att utveckla effektivitet och kvalitet med hjälp av AI uppkom ett intresse av att utforska möjligheterna med att använda generativa språkmodeller som ett stöd i att besvara myndigheternas frågor. Syftet var att utforska om det går att skapa en intern webbtjänst som med stöd av en språkmodell kan stötta medarbetare vid enheten Redovisning och Finansiering (RF).

I projektet har vi presenterat tre olika generativa språkmodeller och utvecklat en applikation där medarbetare på RF har kunnat generera svar på frågor som har kommit in till Fråga ESV med hjälp av dessa tre språkmodeller.

I denna rapport presenteras dels en språklig analys av de genererade och de slutgiltiga svaren samt betygsättning och kommentarer från RF angående kvaliteten på de genererade svaren och funktionaliteten i applikationen. Vi diskuterar resultaten och möjliga vägar framåt för att få stöd i att hantera inkomna frågor till ”Fråga ESV” med hjälp av generativa språkmodeller.

Våra frågeställningar för projektet var:

1. I vilken grad kan de testade språkmodellerna vara ett internt stöd för Fråga ESV?
2. Vilken typ av generativ språkmodell fungerar bäst för syftet?
3. Presterar modeller som exponerats för Fråga ESV-data på olika sätt bättre än en modell som inte har gjort det?
4. Kan en applikation som genererar svar baserat på frågor ställda inom Fråga ESV vara hjälpsam eller intressant att vidareutveckla för utredare på RF?

¹ Ekonomistyrningsverket, Fråga ESV, <https://fragoresv.esv.se/>. Hämtad 17 augusti 2023.

² Ekonomistyrningsverket, EA-regelverket, <https://forum.esv.se/ea-regelverket/>. Hämtad 17 augusti 2023.

2 Överskådlig metodbeskrivning

I vår undersökning har vi använt oss av tre modeller som kan användas i applikationen. Dessa har olika arkitektur och baseras på olika GPT-modeller och modell-arkitekturer. I Tabell 1 finns en översiktlig beskrivning av de olika modellerna och en fördjupad beskrivning finns i *Bilaga 1*. Samtliga modeller är baserade på GPT-modeller från OpenAI (OpenAI, 2023).

Tabell 1: Översiktlig beskrivning av de generativa modellerna

Namn i rapporten	Typ	Input till modell	Beskrivning
Modell 1	GPT-4 zero-shot ³	"Du är en hjälpsam assistent som svarar på frågor om ekonomistyrning och ÄMNE. Den som ställer frågan jobbar på ORG." + <i>Den ställda frågan</i>	En GPT-4-modell som endast får en systembeskrivning.
Modell 2	GPT-4 + vektorsök	"Du är en hjälpsam assistent som svarar på frågor om ekonomistyrning och ÄMNE. Den som ställer frågan jobbar på ORG." + "När du utformar ditt svar ska du bara använda dig av och referera till följande information:". <i>3 utdrag ur regelverket + Den ställda frågan</i>	En GPT-4-modell som ombeds att basera sina svar på matchade utdrag ur EA-regelverket. En AI-modell för semantisk likhet hittar via vektorsök de utdrag som matchar frågan bäst.
Modell 3	GPT-3, fine-tuned (finjusterad)	<i>Den ställda frågan</i>	En GPT-3-modell som finjusterats med frågor och svar genererade med hjälp av GPT-4 baserat på utdrag från EA-regelverket.

Modell 1 och 2 är båda baserade på GPT-4 som är en chat-completion-modell, och Modell 3 är baserad på GPT-3, en completion-modell. En completion-modell tar en eller flera text-strängar som input och genererar sedan text, medan en chat-completion-modell kan hantera en dialog och anpassar sina svar baserat på dialogens historik. Dessa skillnader påverkar inte vår uppgift nämnvärt då vi inte använder någon av modellerna för dialog.

GPT-3 och GPT-4 är utvecklade och tillhandahålls av OpenAI. GPT står för Generative Pre-trained Transformer och är generativa språkmodeller som tränats på många miljarder tokens⁴.

³ Zero-shot betyder att den generativa modellen inte har fått några exempel på hur frågor och svar som är specifika för uppgiften kan se ut. Den har endast fått en beskrivning av uppgiften.

⁴ Textbitar som kan vara ord eller delar av ord.

Modell 1 är en GPT-4-modell som endast har fått en systembeskrivning som beskriver kontexten i fråga. Inga exempel på frågor och svar har presenterats för modellen, men den har fått en instruktion (systembeskrivning) om att den ska svara på frågor kring ekonomistyrning. Instruktionen beskriver också vilken myndighet frågeställaren tillhör.

Modell 2 är en kombination av en GPT-4-modell samt en AI-modell som fungerar som en sökmotor som hittar utdrag ur EA-regelverket som är relevant för den ställda frågan. Modell 2 har samma systembeskrivning som Modell 1 utöver information om att modellen ska besvara den givna frågan med hjälp av de mest relevanta utdragen.

Modell 3 är en GPT-3-modell som finjusterats (fine-tune på engelska)⁵ med hjälp av frågor och svar baserade på utdrag ur EA-regelverket. Dessa frågor och svar har genererats med hjälp av GPT-4. Anledningen till att vi har använt GPT-3 i stället för den nyare GPT-4 är att den sistnämnda inte går att finjustera. OpenAI planerar att möjliggöra finjustering av GPT-4 innan 2024.

Samtliga modeller ombeds att svara på den fråga som utredaren ställer i applikationen.

⁵ "Fine-tuning" är en teknik inom maskininlärning där en förtränad modell anpassas för en specifik uppgift. Detta görs genom att uppdatera modellens vikter med data som är mer relevant för den specifika uppgiften, samtidigt som man bevarar den kunskap som modellen redan har fått genom sin ursprungliga träning. Vidare kommer "fine-tuning" att kallas för "finjustering" i denna studie.

3 Beskrivning av applikationen

Detta kapitel beskriver gränssnittets innehåll och funktion. Applikationens utseende presenteras i figur 1.

Figur 1 visar de textrutor som utredaren ska fylla i. Dessa rutor är (1) Ämnesområde, där ämnet som rör frågan väljs utifrån en rullgardinsmeny, (2) Ärendetyp, där utredaren väljer antingen ”Upplysning” eller ”Utredning”, (3) Organisation, där utredaren fyller i namn på organisation som frågeställaren tillhör, (4) Förtydligad rubrik, där utredaren kan lägga in en icke-obligatorisk förtydligande rubrik till frågan, (5) Här klistrar du in frågan, där utredaren skriver alternativt klistrar in frågan som ställs. Denna fråga ska ha rensats från personlig information och kan även förtydligas av utredaren innan genererade svar skapas.

Till höger om dessa rutor kan utredaren senare generera svar från tre olika generativa språkmodeller genom att klicka på ”Generera text” vid de tre modell-rutorna. Här ska även utredaren ge de genererade svaren betyg i form av stjärnor i skalan 1 till 5 baserat på relevans, korrekthet och användbarhet.

När utredaren har genererat svar ifrån samtliga modeller och betygsatt dessa ombeds utredaren att (6) klistra in det slutgiltiga svaret som ska skickas till den person som ställt frågan. I den sista rutan (7) ombeds utredaren att skriva en övergripande kommentar om de genererade svaren. Att fylla i denna ruta är inte obligatoriskt.

När utredaren fyllt i de obligatoriska textrutorna samt genererat svar från samtliga modeller och betygsatt dem klickar hen slutligen på ”Skicka in”. Då sparas all text samt betyg för analys.

De rutor som har en röd stjärna bredvid sig är obligatoriska.

Figur 1: Applikationens gränssnitt

The screenshot displays the application's user interface. On the left side, there are several input fields:

- Ämnesområde ***: A dropdown menu with 'Anslag' selected.
- Ärendetyp ***: A dropdown menu with 'Ärendetyp' selected.
- Organisation ***: A text input field containing 'Organisation'.
- Förtydligad rubrik**: A text input field containing 'Förtydligad rubrik'.
- Här klistrar du in frågan ***: A large text area containing the instruction: 'Här klistrar du in frågan. Kom ihåg att rensa på namn och personlig information.'
- Extern kommunikation**: A section header followed by a sub-section **Här klistrar du in det slutgiltiga svaret på frågan *** with a text area containing 'Här klistrar du in det slutgiltiga svaret på frågan'.
- Övergripande kommentar om genererade svar**: A text area containing 'Övergripande kommentar om genererade svar'.

 On the right side, there are three identical AI model generation boxes:

- Modell 1 genereras här**: Contains five stars and a 'Generera text' button.
- Modell 2 genereras här**: Contains five stars and a 'Generera text' button.
- Modell 3 genereras här**: Contains five stars and a 'Generera text' button.

 At the bottom right of the interface is a 'Skicka in' button.

Instruktionerna som finns bredvid gränssnittet presenterades i följande punkter:

1. Fyll i alla fält som beskriver frågan som inkommit.
2. Klistra in frågan (kom ihåg att rensa på namn och personlig information).
3. Klicka på "Generera text" för varje modell – kom ihåg att det kan ta ett tag att generera svaren. Generera ett svar i taget.
4. Läs svaren och betygsätt dessa på en skala 1–5 avseende relevans, korrekthet och användbarhet.
5. När du väl har formulerat det slutgiltiga svaret, klistra också in det i applikationen (kom ihåg att rensa på namn och personlig information).
6. Skriv gärna en övergripande kommentar om de genererade svaren.
7. Klicka på "Skicka in" när du är färdig. Du kan inte skicka in förrän du fyllt i de obligatoriska fälten samt betygsatt modellerna.

4 Språklig analys

Totalt användes applikationen 36 gånger.

4.1 Medellängd på svar

Vi beräknade medellängden på alla genererade och slutgiltiga svar. Dessa presenteras i Tabell 2.

Tabell 2: Medellängd på texter från applikationen (antal ord)

	Fråga	Modell 1	Modell 2	Modell 3	Svar	Kommentar
Medellängd	109	127	123	55	134	36

Den texttyp som hade den längsta medellängden (134 ord) var det slutgiltiga svaret. Den modell som hade längst medellängd (127 ord) var Modell 1. Modell 3 hade den absolut kortaste medellängden med 55 ord.

Det är inte förvånande att Modell 3 genererar de kortaste svaren då denna modell generellt gav svar som var en eller två meningar långa. Detta berodde antagligen på att en stor del av de data som modellen finjusterades med var kortfattad. Den modell som ligger närmst det slutgiltiga svaret i ordlängd var Modell 1 med 127 ord.

4.2 Läsbarhet

Läsbarhetsindex (LIX) är ett mått på hur enkel en text är att läsa. LIX beräknas genom att vikta antalet ord i en text mot antalet meningar samt antalet ord som är längre än 6 bokstäver.⁶ Följande formel används för LIX:

$$LIX = (\text{antal ord} / \text{antal meningar}) + (\text{antal långa ord} * 100) / \text{antal ord}$$

I Tabell 3 listas de svårighetsgrader och exempel på litteratur-kategorier som används för att bedöma läsbarhet baserat på LIX. Denna tabell är hämtad ur Björnssons Läsbarhet (1968). Något man bör ta hänsyn till är att ”typ av litteratur” är baserat på undersökningar från 1960-talet. Då språket är föränderligt är det därför i detta fall nog bäst att främst förhålla sig till själva index-summan, samt eventuellt svårighetsgraden. LIX används fortfarande och det finns andra tabeller med tolkningar av LIX (t.ex. på Wikipedia) men flertalet utan referenser eller enkel referens till Björnsson.

I Tabell 4 finns beräknade medelvärden för samtliga modeller samt slutgiltigt svar.

⁶ Björnsson, Carl-Hugo (1968). Läsbarhet. Liber.

Tabell 3: Tolkning av LIX (Björnsson, 1968)

LIX	Svårighetsgrad	Typ av litteratur
20	Mycket lätt	-
30	Lätt	Barn och ungdomsböcker (M=27) Skönlitteratur (M=33)
40	Medelmåttig	Dags- och veckopress (M=39)
50	Svår	Saklitteratur (M=47)
60	Mycket svår	Facklitteratur (M=56)

Tabell 4: Medelvärde av LIX för samtliga modeller och slutgiltiga svar

	Svar	Modell 1	Modell 2	Modell 3
LIX, medelvärde	51,2	51,6	56,8	39,9

Som Tabell 4 visar så har Modell 2 högst läsbarhetsindex. De slutgiltiga svaren och Modell 1 har en svårighetsgrad baserat på LIX som Björnsson refererar till som ”Svår”, medan Modell 3 hamnar ett snäpp nedåt på ”Medelmåttig”. De svar som Modell 2 genererar har svårighetsgraden ”Mycket svår”.

Detta tyder på att texterna som Modell 3 genererar är ganska mycket enklare att läsa än texterna som de andra modellerna genererar liksom utredarnas slutliga svar. Då Modell 3 i regel genererar mycket kortare svar än de andra modellerna verkar resultatet inte helt förvånande. Modell 1 är den modell som hamnar närmast de slutgiltiga svaren vad gäller medelvärde.

4.3 Vanliga ord och ordsekvenser

Vi har räknat n-gram för att hitta de ord och ordsekvenser som är vanligast i svaren. N-gram är sekvenser av n antal ord. Vi har undersökt unigram, bigram och trigram (ett, två eller tre ord långa sekvenser). I detta avsnitt presenteras topp fem-listor av n-gram (se Tabell 5–8). Kompletta listor för n-grammen finns i excel-format som en separat bilaga till denna rapport.

Vid analysen av n-gram plockade vi bort stopp-ord (t.ex. ”och”, ”att” och ”så”) såväl som skiljetecken. Alla versaler konverterades till gemener. För unigram plockades ord som innehöll färre än 4 bokstäver bort. I de fall då fler än ett n-gram har samma frekvens listas dessa tillsammans med en stjärna (*) före varje n-gram. I de fall n-gram med samma frekvens överstiger antalet 3 anges antalet övriga n-gram av samma frekvens.

Tabell 5: N-gram för slutgiltiga svar

Unigram	Antal	Bigram	Antal	Trigram	Antal
finns	23	esv forum	10	föreskrifter allmänna råd	6

Unigram	Antal	Bigram	Antal	Trigram	Antal
bidrag	20	* intäkter bidrag *övriga tjänster	7	övriga tjänster utgifter	5
myndigheten	19	föreskrifter allmänna, *allmänna råd, *ta avgift *5 kap	6	*statliga inrapporteringskoder 2023, *avgiftsförordningen 25 a	4
övriga	18	*hej esv *tjänster utgifter, *full kostnadstäckning, *årsredovisning budgetunderlag	5	*intäkter bidrag övriga, *tjänster utgifter utomstatliga, *ekonomiska målet full ... (7 fler)	3
fråga	16	*statliga inrapporteringskoder, * inrapporteringskoder 2023, * utgifter utomstatliga ... (8 fler)	5	*övriga statliga enheter, *ska kvittot bevaras, *inkomster intäkter bidrag ... (58 fler)	2

Tabell 6: N-gram för Modell 1

Unigram	Antal	Bigram	Antal	Trigram	Antal
viktigt	28	statliga myndigheter	8	bokföra intäkter bidrag	3
finns	21	*dock viktigt, * ingående moms	5	*eftersom del svenska, *svenska offentliga sektorn, *dock viktigt konsultera ... (39 fler)	2
gäller	18	*sveriges riksbank, *andra myndigheter, *använda konto ... (3 fler)	4	* sveriges riksbank självständig, *riksbank bör behandla, särställning förhållande andra ... (1736 fler)	1
*konto, *enligt	17	*eftersom del, *statlig myndighet, *få mer ... (13 fler)	3	-	-
kostnader	16	*transaktioner sveriges, *del svenska, *svenska offentliga ... (114 fler)	2	-	-

Tabell 7: N-gram för Modell 2

Unigram	Antal	Bigram	Antal	Trigram	Antal
enligt	37	allmänna råd	20	föreskrifter allmänna råd	17
gäller	25	föreskrifter allmänna	17	ekonomistyrningsverkets föreskrifter allmänna	16
myndigheten	24	ekonomistyrningsverkets föreskrifter	16	allmänna råd esvfa	14
allmänna	23	råd esvfa	14	*råd esvfa 20221, *esvfa 20221 årsredovisning, *20221 årsredovisning budgetunderlag	6
föreskrifter	18	årsredovisning budgetunderlag	9	*enligt ekonomistyrningsverkets föreskrifter, *enligt avgiftsförordningen 1992191	5

Tabell 8: N-gram för Modell 3

Unigram	Antal	Bigram	Antal	Trigram	Antal
frågor	13	*svara frågor, *ska hantera	9	myndigheter ska hantera	6
myndigheten	11	myndigheter ska	7	ska hantera konkurrensregleringen.	5
hantera	10	utfärdar riktlinjerna	6	*kännedom pengarna ska, *totala antalet år, *användas specifikt ändamål ... (3 fler)	4
*svara, *myndigheter	9	*ska användas, *hantera konkurrensregleringen	5	*ska lägga pengarna, *svara frågor myndigheter, *behöver använda undantagsregeln ... (13 fler)	3
behöver	8	*ska lägga, *kännedom pengarna, *pengarna ska, ... (7 fler)	4	*viktigt klar tydlig, *ändamål ska lägga, *viktigt komma ihåg, ... (67 fler)	2

Det finns vissa n-gram som ser märkliga ut vid en första anblick. I Tabell 5 för slutgiltiga svar ligger sekvensen ”hej esv” som det 4:e vanligaste bigrammet för slutgiltiga svar. Detta är kanske inte helt förvånande då RF:s utredare ofta börjar sina svar med ett ”Hej”, sedan en ny rad, och sedan ett svar som beskriver ESV:s råd. Språkmodellerna tenderar att inte inleda sina svar med en hälsningsfras.

I Tabell 7 ser vi att bigrammet ”råd esvfa” ligger på fjärdeplats för Modell 2. ESVFA står för ”Ekonomistyrningsverkets föreskrifter och allmänna råd” och används ofta som förled för att hänvisa till ett specifikt avsnitt, t.ex. ”ESVFA 2022:2”. ”esvfa” förekommer även i flera vanliga trigram för samma modell. Vi kan alltså konstatera att Modell 3 är benägen att referera till ESV:s föreskrifter och allmänna råd.

Modell 1 och 2 har båda ”gäller” och ”enligt” i sina topplistor för unigram. Det slutgiltiga svaret och Modell 3 har ”fråga” respektive ”frågor” i sina topplistor. Det slutgiltiga svaret och Modell 2 har båda bigrammen ”allmänna råd” samt ”föreskrifter allmänna” i sina topplistor. För Modell 1 förekommer dessa bigram bara en gång var, och för Modell 3 förekommer de inte alls.

Vi kan konstatera att de slutgiltiga svaren och svaren från Modell 2 har samma trigram i topp: ”föreskrifter allmänna råd” (se Tabell 5 och 7). För övrigt kan också nämnas att ”myndigheten” finns med i alla svarstypers topp-fem för unigram, utom för Modell 1.

4.4 Ordklasser

Vi har undersökt distributionen av olika ordklasser för alla svarstyper samt frågor med hjälp av Kungliga Bibliotekets KB-BERT-modell ”bert-base-swedish-cased-pos”, som är specialiserad på ordklassstagning för svenska texter.⁷ Fördelningen av ordklasser för de olika svarstyperna presenteras i Tabell 9. De ordklasser som inte fanns med bland de 10 vanligaste för respektive text-typ presenteras inte i tabellen. För beskrivningar av ordklassernas förkortningar, se Tabell 10.

Tabell 9: Ordklasser per text-typ, topp tio-lista

	Svar		Modell 1		Modell 2		Modell 3		Frågor	
1	NN	25 %	NN	23 %	NN	24 %	NN	22 %	NN	24 %
2	DEL	13 %	VB	15 %	VB	14 %	VB	17 %	VB	15 %
3	VB	12 %	PP	11 %	PP	12 %	PP	11 %	PP	13 %
4	PP	12 %	DEL	9 %	DEL	8 %	PN	10 %	DEL	9 %
5	JJ	5 %	JJ	6 %	JJ	6 %	DEL	7 %	PN	6 %
6	PN	5 %	PN	6 %	KN	6 %	DT	5 %	AB	5 %
7	DT	5 %	KN	5 %	PN	5 %	AB	5 %	JJ	4 %
8	AB	4 %	DT	5 %	DT	4 %	JJ	4 %	DT	4 %
9	KN	4 %	AB	3 %	RG	4 %	KN	4 %	KN	4 %
10	RG	3 %	SN	3 %	AB	3 %	SN	3 %	RG	3 %

⁷ Malmsten, M., Börjeson, L., & Haffenden, C. (2020). Playing with Words at the National Library of Sweden--Making a Swedish BERT. arXiv preprint.

Tabell 10: Beskrivning av ordklasser⁸

Förkortning	Ordtyp
NN	Substantiv
VB	Verb
PP	Preposition
JJ	Adjektiv
PN	Pronomen
DT	Determinerare
AB	Adverb
KN	Konjunktion
SN	Subjunktion
PM	Egennamn
RG	Grundtal (t.ex. ett, två eller tre)
RO	Ordningstal (t.ex. första, halv, femtedel)
DEL	Skiljetecken
MID	(mindre) skiljetecken
MAD	(större) skiljetecken

Som Tabell 9 visar så ser fördelningen av de vanligaste ordklasserna för varje texttyp liknande ut. De slutgiltiga svaren innehåller störst andel substantiv (NN) på 25 procent, och Modell 3 minst andel substantiv med 22 procent. Modell 2 är den texttyp som har störst andel verb (VB) på 22 procent, och det slutgiltiga svaret minst på 12 procent.

⁸ Alla ordklasser och förkortningar med ett undantag är erkända och används bl.a. i den svenska korpusen The Swedish Treebank. (Nivre et al., 2006). "Skiljetecken" (DEL) har författaren till denna rapport bildat av att slå ihop ordklasserna "Mindre skiljetecken" (MID) och "Större skiljetecken" (MAD).

5 Återkoppling från utredare

I detta kapitel presenterar och diskuterar vi användarnas betyg och deras övergripande kommentarer på de genererade svaren, samt enkätfrågor och dess svar.

5.1 Betyg

Betygsskalan gick från 1 till 5, där 1 var det sämsta betyget och 5 det bästa. Användarna ombads att betygsätta modellerna avseende relevans, korrekthet och användbarhet. Se Tabell 11 för medelvärde av betyg för samtliga modeller.

Tabell 11: Genomsnittliga betyg för varje modell

Modell	Betyg, genomsnitt
Modell 1	1,97
Modell 2	2,19
Modell 3	1,28

I Tabell 11 ser vi att Modell 2 var den som fick bäst betyg (2,19) och Modell 3 var den som fick sämst betyg (1,28). Ingen av modellerna nådde upp till 2,5 som är skalans mittpunkt. Detta innebär att alla modeller hamnade på den lägre halvan av betygsskalan.

Användarna gav betyg 1, 2, 3 och 4 till samtliga modeller. Modell 1 fick betyg 5 två gånger, Modell 2 betyg 5 en gång, medan Modell 3 aldrig fick betyg 5 (se Tabell 12). Detta kan ses som en indikation på att svaren för Modell 1 och Modell 2 var användbara ibland men att användarna oftast såg brister i de genererade svaren. Modell 3 får bara ett betyg som var högre än 2 totalt två gånger. För samtliga modeller var det vanligaste betyget 1.

Tabell 12: Distribution av betyg för alla modeller

	Betyg: 1	2	3	4	5
Modell 1	<u>17</u>	8	8	1	2
Modell 2	<u>13</u>	8	11	3	1
Modell 3	<u>29</u>	5	1	1	0

5.2 Övergripande kommentarer

Detta avsnitt beskriver kommentarerna vi fick in ifrån rutan ”Övergripande kommentar om genererade svar” i applikationen. Vi fick totalt 21 stycken kommentarer. Utöver en sentimentanalys av kommentarerna har vi också gjort en manuell innehållsmässig analys av innehållet i kommentarerna.

5.2.1 Sentimentanalys

Vår sentimentanalys gjordes på två nivåer: kommentarnivå och meningsnivå. Den modell vi använde var språkmodellen ”bert-swedish-based-sentiment”, som är en BERT-transformer som har finjusterats med hjälp av ca 20 000 recensioner på svenska från App store.⁹ Den underliggande transformer-modell som analyseraren är baserad på är Kungliga bibliotekets ”bert-base-swedish-cased” för sekvensklassificering.¹⁰

Varje mening får antingen ett positivt eller negativt sentiment. På kommentarnivå har vi beräknat det genom att välja det sentiment som är vanligast för de meningar som ingår i kommentaren. Sentimentanalysen på kommentarnivå presenteras i Tabell 13 och sentimentanalysen på meningsnivå presenteras i Tabell 14.

Tabell 13: Sentimentanalys på kommentarnivå

Sentiment	Antal kommentarer
Negativ	14
Neutral (Negativ och Positiv)	4
Positiv	3

Tabell 14: Sentimentanalys på meningsnivå

Sentiment	Antal meningar
Negativ	40
Positiv	15

Tabell 13 visar att det finns 3 kommentarer där det vanligaste sentimentet är positivt och 14 kommentarer där sentimentet är negativt. För 4 kommentarer var de positiva och negativa sentimenten lika vanliga. Enligt sentimentanalysen är alltså de flesta kommentarer negativa och några neutrala eller positiva.

Tabell 14 visar att det vanligaste sentimentet för en mening är negativt sentiment med 40 negativa meningar, medan 15 meningar har tolkats som positiva. Enligt sentimentanalysen innehöll alltså kommentarerna mer negativa sentiment än positiva.

Eftersom vi hade relativt få kommentarer att analysera testade vi att manuellt klassificera kommentarerna i sentiment. Den manuella sentimentanalysen finns i *Bilaga 2*. I den manuella sentimentanalysen har 1 kommentar bedömts som neutral, 8 som positiva och 12 som negativa.

⁹ Malmsten, Martin. (2021). Bert-base-swedish-cased-sentiment. [<https://huggingface.co/marma/bert-base-swedish-cased-sentiment>].

¹⁰ Malmsten, M., Börjesson, L., & Haffenden, C. (2020). Playing with Words at the National Library of Sweden—Making a Swedish BERT. arXiv preprint.

En anledning till att så många meningar har klassificerats som negativa kan vara att den automatiska sentimentanalysmetoden saknar en neutral klass. Det finns modeller som använder sig av även neutralt sentiment, t.ex. den lexikon- och regelbaserade modellen VaderSentiment för svenska av Hutto & Gilbert (2014). Men eftersom den svenska BERT-modellen för sentimentanalys bedömdes som bäst i denna kontext valde vi denna binära klassificerare.

5.2.2 Innehåll och teman

Vi bad användarna att betygsätta modellerna utifrån relevans, korrekthet och användbarhet. Det är antagligen därför användarna har valt att fokusera på dessa saker även i de övergripande kommentarerna.

Den modell som oftast hade rätt svar verkar enligt kommentarerna vara Modell 2. Av 21 utredare skriver 9 stycken att en eller flera modeller har gett felaktiga svar. I kommentarerna skrev 6 utredare att inget av de genererade svaren var till hjälp för att formulera ett slutgiltigt svar. Däremot beskriver 6 utredare ett eller flera modellers svar som rätt eller delvis rätt. I 4 kommentarer beskrivs modellerna med ordet ”bra”. Ofta skriver utredarna att delar av modellernas svar är bra eller användbara, men att de även innehåller något som inte stämmer eller är icke-relevant.

Språk är något som inte diskuteras särskilt mycket i kommentarerna. Dock nämndes det 3 gånger att en eller flera modeller gav svar med bra eller användbara formuleringar.

5.3 Enkätfrågor

Utredarna fick följande enkätfrågor:

1. Vad tycker du om appens funktionalitet? Vad fungerade bra, vad fungerade dåligt?
2. Kan du komma på någon funktion du skulle vilja lägga till?
3. Är de genererade svaren hjälpsamma? Om ja, på vilket sätt? (t.ex. formulering av svar, referenser, etc...)
4. Tyckte du att någon modell var bättre än den andra? Rangordna gärna modellerna om du kan.
5. Motivera gärna din rangordning. Fanns det någon särskild egenskap hos någon modell som gjorde att du tyckte den var bra/dålig?
6. Har du några övriga kommentarer?

Det var 5 utredare som svarade på enkäten. På fråga 1 svarade alla utredare att själva applikationen fungerade bra. Olika utredare nämnde att appen var smidig, användarvänlig och att instruktionerna var bra.

På fråga 2 nämnde en utredare att det skulle vara bra om det fanns en funktion som möjliggjorde att man automatiskt skulle kunna läsa informationen som rör frågan i programmet "Easit". Det är detta program som utredarna använder för att läsa och besvara de inkomna frågorna för Fråga ESV. Ett annat förslag var att det skulle vara värdefullt att få förslag på vilken del av regelverket som frågan berör.

På fråga 3 svarade 4 utredare att svaren inte var hjälpsamma. En anledning till detta var enligt 1 utredare att svaren från modellerna inte liknar RF:s svar på så sätt att de genererade svaren generellt var för specifika och inte "hjälp till självhjälp" som RF föredrar. Detta skulle möjligtvis kunna ha att göra med att modellerna överlag inte är informerade om RF:s generella format på svar. Modell 3 är till viss del tränad på tidigare frågor och svar från Fråga ESV, men den största delen träningsdata som användes för finjusteringen av Modell 3 består av genererade frågor och svar baserade på EA-regelverket.

En utredare svarade tveksamt på fråga 3 och skrev att applikationen skulle spara dem tid om modellerna var mer träffsäkra i sina hänvisningar till regelverken. Det eventuella stödet med att formulera ett svar som de genererade svaren ger beskrev en utredare som överflödigt. En annan utredare svarade att de genererade svaren inte är särskilt hjälpsamma för att formulera korrekta svar, men att de kan vara till nytta för att få formuleringstips.

På fråga 4 svarade en utredare att Modell 3 ofta hade helt fel, och en utredare svarade att Modell 2 fungerade bäst och Modell 3 sämst. En utredare skrev att Modell 3 var bäst och att de andra modellerna inte var hjälpsamma. En annan utredare skrev att Modell 2 generellt presterade bättre men att Modell 1 hade gett bäst svar på en specifik fråga.

På fråga 5 svarade 2 utredare att svaren från Modell 3 ofta var helt felaktigt och 1 skriver att den ibland gav svar med påhittade myndigheter eller cirkelresonemang. De andra två modellerna gav inte heller bra svar men de innehöll någon slags logik, skriver utredaren. En utredare skrev att Modell 2 ofta gav mer hänvisningar till regelverket vilket är bra då det är så ESV brukar svara. En annan utredare skrev att Modell 3 gav ett korrekt svar som dock gav för många och specifika förslag. De andra modellerna gav fel och för generella svar.

På fråga 3, 4 och 5 svarade 1 utredare att inget av de genererade svaren var hjälpsamma för de frågor som utredaren hade testat i applikationen.

På fråga 6 svarade 1 utredare att det var roligt och att det känns som att applikationen skulle kunna bli något användbart på sikt. En annan utredare sa att det var roligt att testa applikationen och att förbättrade språkmodeller skulle kunna vara användbara.

Denna utredare föreslog att en chattfunktion skulle kunna vara mer användbar. En utredare ifrågasatte textrutan där utredaren ska fylla i organisation för den som har ställt frågan, då detta inte borde spela någon roll. En annan utredare svarade att applikationen var intressant och värd att fortsätta utveckla.

5.4 Övrig feedback och sammanfattning av feedback

Flera utredare har problematiserat att applikationen i sin funktion kombinerar två saker som skulle kunna behöva delas upp: information/fakta och formulering. Ibland har modellerna gett felaktig information men bra formuleringar, och ibland tvärtom.

Betygen och de övergripande svaren stämmer överens på så sätt att Modell 2 var den modell som överlag presterade bäst och Modell 3 sämst. Alla modeller fick dock både positiv och negativ feedback. I de övergripande kommentarerna framgår det att svaren kunde vara hjälpsamma men ofta inte var det. De flesta utredare tyckte att applikationen var spännande och att det skulle kunna vara intressant att vidareutveckla den.

6 Exempel på frågor och svar

I detta kapitel tittar vi på och diskuterar tre utdrag från de ställda frågorna och dess genererade och slutgiltiga svar, samt kommentarer och betyg.

I det första exemplet ställs en fråga om Sveriges riksbank och redovisning (se Tabell 15).

Tabell 15: Fråga och svar: Exempel 1

Text-typ	Text
Fråga	<p>Hej!</p> <p>Jag hittar inte Sveriges riksbank i myndighetsregistret. På deras hemsida står det att de är en självständig myndighet under riksdagen. Hur ska vi redovisa dem? Som en extern organisation? Eller ska de redovisas under någon annan myndighet? Jag tänker framförallt på när vi stämmer av våra inomstatliga motparter vid helåret.</p>
Slutgiltigt svar	<p>Hej</p> <p>Riksbanken har en speciell ställning inom staten. De ingår inte i den statliga redovisningsorganisationen och behandlas därför som en utomstatlig motpart.</p>
Modell 1	<p>Hej!</p> <p>Sveriges Riksbank är en självständig myndighet under riksdagen och har en särställning i förhållande till andra myndigheter. Därför kan det hända att de inte finns i det vanliga myndighetsregistret.</p> <p>När det gäller redovisning av transaktioner med Sveriges Riksbank bör ni behandla dem som en inomstatlig motpart, eftersom de är en del av den svenska offentliga sektorn. Om ni har specifika konton eller koder för att redovisa transaktioner med andra myndigheter, kan ni använda dessa även för Riksbanken. Om ni inte har sådana konton eller koder kan ni skapa en särskild kod eller konto för att redovisa transaktioner med Sveriges Riksbank.</p> <p>Vid avstämning av inomstatliga motparter vid helåret bör ni inkludera Sveriges Riksbank i er avstämning, eftersom de är en del av den svenska offentliga sektorn.</p>
Modell 2	<p>Sveriges Riksbank är en självständig myndighet under riksdagen och inte en del av den statliga förvaltningen som de flesta andra myndigheter. Därför kan du inte hitta dem i myndighetsregistret på samma sätt som andra myndigheter.</p> <p>När det gäller att redovisa och stämma av inomstatliga motparter vid helåret, bör du följa Ekonomistyrningsverkets föreskrifter och allmänna råd (ESVFA 2022:2) om myndigheters bokföring. Enligt dessa föreskrifter ska myndigheter rapportera och stämma av transaktioner, fordringar och skulder som har en annan statlig myndighet som motpart.</p> <p>Eftersom Sveriges Riksbank är en självständig myndighet under riksdagen och inte en del av den statliga förvaltningen, kan det vara lämpligt att behandla dem som en extern organisation snarare än en inomstatlig motpart när det gäller redovisning och avstämning.</p>
Modell 3	Sveriges riksbank ska redovisas som en myndighet.
Kommentar	<p>Svar 1: Felaktigt svar. Dock ganska bra formulerat.</p> <p>Svar 2: Rätt svar. Ganska bra formulerat. Svaret är dock inte formulerat hur Riksbanken ska hanteras utan hur det kan vara lämpligt att Riksbanken behandlas.</p> <p>Svar 3: Felaktigt svar.</p>

För detta exempel har utredaren gett en kommentar per modell. Som Tabell 15 visar så är det korrekta svaret på frågan om redovisning av Sveriges riksbank att ”Riksbanken har en speciell ställning inom staten. De ingår inte i den statliga redovisningsorganisationen och behandlas därför som en utomstatlig motpart”. Enligt utredaren formulerade Modell 1 sitt svar bra men det innehåller felaktig information. Modell 2 svarade rätt med en okej formulering, men utredaren skriver att svaret beskriver hur Riksbanken ska ”behandlas” i stället för ”hanteras”. Både Modell 1, Modell 2 och det slutgiltiga svaret använde dock ordet ”behandla” i sina svar när de refererade till Riksbanken. Svaret som Modell 3 har genererat bedömde utredaren som felaktigt, och det verkar onekligen som att detta svar motsäger det slutgiltiga svaret. Utredaren gav Modell 1 och Modell 3 betyg 1 och Modell 2 betyg 3.

Tabell 16: Fråga och svar: Exempel 2

Text-typ	Text
Fråga	<p>Vad gäller vid tecknande av avtal exempelvis på tre år och som kan förlängas med ett år i taget under två år? Hur många år behöver vi ha med i bemyndigandet? Är det tre eller fem år?</p> <p>Om det är tre år kan vi då använda undantagsregeln (Anslagsförordningen 17 § andra och tredje stycket) för ett år i taget avseende förlängningarna?</p>
Slutgiltigt svar	<p>Hej</p> <p>Det blir lite gissningar eftersom jag inte vet vad avtalet avser eller vilket anslag som skulle behöva ett bemyndigande.</p> <p>Jag förutsätter att det handlar om ett avtal som inte avser anställning av personal, hyra av utrustning, leverans av tjänster och förbrukningsmateriel och liknande som inte kräver något beställningsbemyndigande alls, 17 § tredje stycket anslagsförordningen. Vidare låter det som att det handlar om ett avtal som ska belasta en anslagspost som ni idag inte har något beställningsbemyndigande för.</p> <p>Undantagsregeln i 17 § andra stycket och 24 § ESV:s föreskrifter kan ni tillämpa om avtalet ryms inom undantagsregeln, oavsett längd på avtalet.</p> <p>Kan ni inte tillämpa undantagsregeln behöver ni ett bemyndigande för så länge som ni har gjort ett åtagande. Är ni bundna tre år av avtalet spelar det inte någon roll att det finns en möjlighet till förlängning. Det är ju inte något ni bundit upp utan bara en framtida möjlighet.</p> <p>Om ni får ett bemyndigande för det aktuella anslaget är det väl sannolikt bättre att det förlängs så länge som det behövs än att undantagsregeln ska tillämpas i slutet av avtalet. Vanligtvis behövs ju ett bemyndigande för flera åtaganden som görs löpande i en viss verksamhet, inte bara för ett avtal.</p>
Modell 1	<p>Vid tecknande av avtal som sträcker sig över tre år och kan förlängas med ett år i taget under två år, behöver ni ha med hela avtalsperioden, inklusive eventuella förlängningar, i bemyndigandet. Det innebär att ni behöver ha med fem år i bemyndigandet.</p> <p>Om ni endast har med tre år i bemyndigandet och vill använda undantagsregeln enligt Anslagsförordningen 17 § andra och tredje stycket, kan det vara möjligt att göra det för förlängningarna. Detta innebär att ni i så fall behöver ansöka om undantag för varje förlängning separat. Dock är det viktigt att ni konsulterar med er juridiska avdelning eller ansvarig myndighet för att säkerställa att detta är möjligt och följer gällande regelverk.</p>
Modell 2	<p>När det gäller tecknande av avtal på tre år med möjlighet till förlängning med ett år i taget under två år, bör bemyndigandet täcka hela avtalsperioden, inklusive eventuella förlängningar. Det innebär att bemyndigandet bör omfatta fem år.</p> <p>Anslagsförordningen 17 § andra och tredje stycket behandlar undantag från regeln om att anslag får disponeras endast under det budgetår för vilket de har tilldelats. Dessa undantag gäller för vissa typer av utgifter, såsom investeringar, forskning och</p>

Text-typ	Text
	utveckling, och internationellt samarbete. Om avtalet ni tecknar faller inom någon av dessa kategorier, kan ni använda undantagsregeln för att disponera anslag över flera budgetår. För att använda undantagsregeln vid förlängningarna av avtalet, bör ni säkerställa att avtalet uppfyller kraven för undantag enligt Anslagsförordningen 17 § andra och tredje stycket.
Modell 3	Det är riktigt att undantagsregeln i 17 § andra och tredje stycket anslagsförordningen säger att om myndigheten ska teckna avtal som kan förlängas med ett visst antal år och det totala antalet år är tre eller mindre behöver myndigheten ha ett bemyndigande för tre år. Om det totala antalet år är fyra eller fler behöver myndigheten ha ett bemyndigande för fem år. Det är alltså inte så att om det är tre år behöver ni använda undantagsregeln för tre år i taget. Det är bara om det totala antalet år är tre eller mindre som ni behöver använda undantagsregeln för tre år i taget. Om det totala antalet år är fyra eller fler behöver ni använda undantagsregeln för fem år.
Kommentar	Samtliga svar är felaktiga på flera olika sätt. Jag tror inte något av dem innehåller något korrekt, möjligen med något mindre undantag.

Frågan i exempel 2 i Tabell 16 handlar om hur långt beställningsbemyndigande som behövs vid avtal med möjlighet till förlängning. Enligt utredaren var samtliga svar felaktiga och dessutom på flera olika sätt. Samtliga modeller fick betyg 1 av utredaren. Mer komplexa frågor som denna där det enligt utredaren finns vissa detaljer som saknas verkar vara svåra för språkmodellerna att svara korrekt på. Utredaren kan utifrån sin erfarenhet gissa vad för kontext som gäller medan språkmodellerna inte verkar kunna fylla i luckorna själva.

Tabell 17: Fråga och svar: Exempel 3

Text-typ	Text
Fråga	Hej! Internt på Transportstyrelsen pågår en diskussion om användandet av begreppet faktura och avi. När det gäller faktura används begreppet främst när det gäller fakturering av avgifter och avi används främst vid avisering av skatter. Vår önskan är att vi bara ska använda begreppet faktura då de dubbla begreppen många gånger orsakar förvirring. Hur ställer ni på ESV till detta? Kan man fakturera en skatt genom en faktura eller måste skatt aviseras via avi?
Slutgiltigt svar	Hej ESV kan inte avgöra vad som blir rätt att kalla olika handlingar men det finns ju en definition av vad en faktura är i 11 kap. mervärdesskattelagen (1994:200). Det är den ESV hänvisar till t.ex. i våra föreskrifter om elektronisk fakturahantering. Om ni kallar handlingar för faktura som inte stämmer överens med definitionen i mervärdesskattelagen av faktura kan det ju riskera att skapa nya oklarheter. Ni kanske kan ta kontakt med Skatteverket som är mer insatta i detta än vi är.
Modell 1	Hej! Jag är inte en representant för ESV (Ekonomistyrningsverket), men jag kan ge dig en generell förklaring kring begreppen faktura och avi. Faktura och avi är två olika dokument som används för olika ändamål. En faktura är ett dokument som specificerar varor eller tjänster som har levererats och deras kostnader, samt betalningsvillkor. En avi, å andra sidan, är en notis eller påminnelse om en kommande betalning, oftast kopplad till en skatt eller avgift.

Text-typ	Text
	<p>Det är viktigt att skilja mellan dessa två begrepp eftersom de har olika juridiska och redovisningsmässiga konsekvenser. En faktura är ett juridiskt dokument som kan användas som bevis för en skuld, medan en avi endast är en påminnelse om en betalning som ska göras.</p> <p>När det gäller skatter och avgifter kan det vara lämpligt att använda en faktura om det är en direkt koppling till en specifik tjänst eller vara som har levererats.</p>
Modell 2	<p>Enligt Ekonomistyrningsverkets föreskrifter och allmänna råd om myndigheters bokföring och årsredovisning, är en faktura en handling som tas fram enligt 11 kap. mervärdesskattelagen (1994:200) vid omsättning av varor och tjänster. Fakturor används främst vid fakturering av avgifter och andra ersättningar.</p> <p>Aviseringar av skatter och liknande avgifter, såsom arbetsgivaravgifter, är inte fakturor enligt definitionen i mervärdesskattelagen. Därför används begreppet avi för att avisera dessa belopp.</p> <p>I allmänna råd till 21 f § förordningen om myndigheters bokföring nämns att rekvisitioner av bidrag, deklarationshandlingar avseende skatter samt aviseringar av exempelvis hyror och räntor inte innefattas i definitionen av faktura och behöver inte hanteras elektroniskt. Myndigheten bör dock sträva efter att hantera också dessa handlingar elektroniskt.</p>
Modell 3	<p>ESV har inte någon ställningstagande till detta. Det är Transportstyrelsen som är ansvarig myndighet för skatteaviseringssystemet och därmed även för vilka begrepp som ska användas.</p>
Kommentar	<p>De är på rätt spår i alla fall. Inget är jättefel även om allt inte är helt rätt.</p>

I exempel 3 i Tabell 17 har en fråga ställts kring användande av begreppen ”faktura” och ”avi”. Utredaren gav Modell 1 betyg 3, Modell 2 betyg 4 och Modell 3 betyg 2, och skrev att de genererade svaren är på rätt spår och att de inte ger helt korrekta svar, men inte heller direkt felaktig information. Det korrekta svaret hänvisade till en officiell definition av ”faktura” samt till Skatteverket som är mer insatt i ämnet. Modell 1 beskrev skillnaden mellan en faktura och en avi. Modell 2 hänvisade till samma kapitel i mervärdesskattelagen som det slutgiltiga svaret gör, samt beskrev skillnaden mellan faktura och avi. Enligt svaret som Modell 3 genererat har ESV inte gjort något ställningstagande i frågan och svaret hänvisar till en annan specifik myndighet, likt det slutgiltiga svaret. Modell 3 hänvisade dock till samma myndighet som frågeställaren representerar.

7 Resultat och diskussion

Ingen av modellerna fick särskilt höga betyg, men i flera fall har modellerna haft rätt, delvis rätt, eller varit hjälpsamma enligt användarna. För att höja användbarheten bör vidare optimering av en GPT-modell undersökas. I detta kapitel besvarar vi undersökningens frågeställningar och diskuterar resultaten. Därefter lämnas förslag på frågor att undersöka vidare.

7.1 Svar på undersökningens frågeställningar

I detta avsnitt besvarar vi undersökningens frågeställningar.

I vilken grad kan de testade språkmodellerna vara ett internt stöd för Fråga ESV?

Alla modeller som vi testat har i vissa fall varit hjälpsamma, men alla modeller har också visat på svagheter. Utredarna på RF som använt applikationen uppfattade ofta att de testade språkmodellerna inte var hjälpsamma. Flera utredare uttryckte dock att bättre fungerande språkmodeller skulle kunna vara användbara i arbetet med Fråga ESV. För att kunna ge ett bättre och mer användbart stöd behöver applikationen och den bakomliggande modellen vidareutvecklas.

Vilken typ av generativ språkmodell fungerar bäst för syftet?

Av de modeller vi testade fungerade Modell 2 (GPT-4 + vektorsök) bäst, dvs. en GPT-4-modell i kombination med en vektorsök-modell som hittar det aktuella regelverket som Fråga ESV-frågan är mest lik och baserar svaret på det. Denna modell fick i genomsnitt betyget 2,19 på en skala 1 till 5. Det var också denna modell som oftast gav korrekta svar enligt de övergripande kommentarerna.

Presterar modeller som exponerats för Fråga ESV-data på olika sätt bättre än en modell som inte har det?

Både ja och nej. Modell 2 och Modell 3 har exponerats för denna typ av data medan Modell 1 inte har gjort det. Modell 1 presterade bättre än Modell 3 men sämre än Modell 2 och hamnar därför i mitten.

Kan en applikation som genererar svar baserat på frågor ställda inom "Fråga ESV" vara hjälpsam eller intressant att vidareutveckla för utredare på RF?

Ja, givet att applikationen förbättras. I nuvarande form bedöms applikationen ofta som icke hjälpsam, men de flesta utredare som svarat på enkäten menar att det finns potential för funktionen.

Då Modell 2 (GPT4 + vektorsök) är den modell som verkar ha fungerat bäst för ändamålet kan det vara värt att arbeta vidare på en liknande modellarkitektur. Eftersom det ingår både en sökfunktion och en GPT4-modell i arkitekturen för Modell 2 är det inte helt enkelt att säga vad det är som behöver förbättras utifrån våra analyser för denna rapport.

Ett potentiellt problem skulle kunna vara att vektorsök-modellen inte hittar de utdrag från EA-regelverket som frågan bör besvaras med. Detta kan undersökas genom att göra en utvärdering där man jämför frågan med de avsnitt som vektorsök-modellen anser vara bäst matchande. Om de matchade avsnitten ur regelverket inte innehåller information som är relevant för den aktuella frågan bör själva sökmotorn förbättras.

Ett annat potentiellt problem skulle kunna vara att GPT-4-modellen kanske även använder annan information i sitt genererade svar, trots att den fått instruktioner om att endast använda information ifrån de matchade utdragen. Ett annat problem kan vara att modellen inte formulerar sitt svar på ett sätt som är likt det format som utredarna är vana vid och tycker är önskvärt. Ett sista exempel kan vara att modellen tolkar de matchade utdragen på ett felaktigt eller icke-optimalt sätt.

7.2 Förslag på fortsatt arbete

Möjliga steg att ta i ett fortsättningsprojekt presenteras i detta avsnitt.

Utvärdera sökmotor

Utvärdera sökmotorn som använts för att ta fram relevanta utdrag ur EA-regelverket. Följ förslagsvis dessa punkter:

1. Välj ut ett antal frågor som inkommit till Fråga ESV.
2. Lista relevanta utdrag ur regelverket genom att antingen använda befintliga svar som hänvisar till en eller flera avsnitt av regelverket eller be utredare från enheten RF att lista relevanta avsnitt.
3. Jämför sedan sökmotorns matchningar mot de garanterat relevanta utdragen ur regelverket. Denna validering går att göra automatiskt.
4. Om valideringen ger otillräckliga resultat kan man undersöka möjligheter att förbättra den, exempelvis med mer data, annorlunda arkitektur, eller en annan underliggande modell för vektorsök.

Undersök hur bra GPT-4 är på att hantera de matchade dokumenten

Undersök hur pass noga den generativa modellen i Modell 2 (GPT4) är med att endast använda sig av den information som finns i de matchade utdragen för att

utforma sina svar. Om GPT-4 inte använder sig av de matchade dokumenten på ett önskat sätt kan man testa att ändra på systemmeddelandet, alternativt se över om man kan tydliggöra frågorna som kommer in. Detta skulle kunna göras med GPT. Man kan också testa andra generativa modeller för uppgiften och se ifall de presterar bättre.

Ge exempel på frågor och svar till generativ modell

Pröva att ge Modell 2 exempel på frågor och svar för att ge den en uppfattning om hur dessa är formulerade. Detta kommer eventuellt få den generativa modellen att ge svar som är mer lika utredarnas svar.

Utforska alternativa modelltyper

Utforska fler alternativa typer av språkmodeller som skulle kunna vara passande för uppgiften. Utvecklingen går fort inom AI och språkteknologi. Enligt OpenAI ska det exempelvis bli möjligt att innan årets slut finjustera GPT-4 genom OpenAI:s API.¹¹ Genom att testa fler metoder på denna specifika uppgift kan eventuellt de svagheter som finns hos den undersökta applikationen motverkas.

Undersök en alternativ typ av applikation

En möjlighet för framtiden är att dela upp sökmotorn och GPT-4 i två olika funktioner, och på så sätt dela upp stödet för information och stödet för formulering av svar. Det vill säga att funktion A låter utredaren fylla i den ställda frågan i en textruta, och när utredaren klickar på en knapp presenteras de tre mest sannolika avsnitten ur regelverket där utredaren kan hitta svaret på frågan.

I funktion B kan utredaren sedan chatta med en GPT-modell (som kanske har viss kunskap eller systeminformation anpassad för ändamålet) som kan ge tips på hur utredaren ska formulera sitt svar utifrån den information som utredaren väljer att ge chattmodellen. På så sätt får utredaren en större inblick i funktionaliteten. Utredaren kan själv styra vilken information den generativa språkmodellen baserar sitt svar på och ifrågasätta eventuella matchningar som vektorsök-modellen gjort, eftersom utredaren då fått särskilda matchade utdrag kompletta med referenser till vilken del i regelverket denna information har tagits ifrån.

Då en av den utvärderade applikationens största svagheter är att den inte genererar korrekta svar skulle en applikation med större närhet till regelverkets exakta formuleringar kunna vara en väg framåt.

¹¹ OpenAI, *GPT-4 API general availability and deprecation of older models in the Completions API*, <https://openai.com/blog/gpt-4-api-general-availability>. Publicerad 6 juli 2023. Hämtad 18 augusti 2023.

Referenser

Björnsson, Carl-Hugo (1968). *Läsbarhet*. Lärarbiblioteket, 99-0346315-4. Stockholm: Liber

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30, 681-694.

Honnibal, M., & Montani, I. (2017). *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*.

Hutto, C.J., & Gilbert, E.E. (2014). *VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text*. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, Juni 2014.

Malmsten, M., Börjeson, L., & Haffenden, C. (2020). Playing with Words at the National Library of Sweden--Making a Swedish BERT. *arXiv preprint arXiv:2007.01658*.

Malmsten, Martin. (2021). Bert-base-swedish-cased-sentiment. (First version) [<https://huggingface.co/marma/bert-base-swedish-cased-sentiment>] hämtad 2023-06-30.

Nivre, J., Nilsson, J., & Hall, J. (2006). Talbanken05: A Swedish Treebank with Phrase Structure and Dependency Annotation. I *LREC* (Vol. 2, No. 1, sid. 4-2).

OpenAI. (2023) Models. [<https://platform.openai.com/docs/models/overview>] hämtad 2023-06-20.

Bilaga 1 - Metodbeskrivning

Samtliga modeller är baserade på OpenAIs GPT-modeller GPT-3 och GPT-4. Dessa modeller är autoregressiva språkmodeller som använder djupinlärning för att generera text (Flordi & Chiriatti, 2020).

7.3 Teknisk beskrivning

Nedan följer tekniska detaljer för de generativa modellerna. Dessa detaljer inkluderar modellernas arkitektur samt prompt, alltså den information som modellen ges i samband med generering av svar.

Systeminformation är något som kan anges för GPT-modellerna som instruktion för hur den ska bete sig. När chat-completion-modellerna anropas kan utredaren skicka in input för fyra olika roller: ”system”, ”user”, ”assistant” och ”function”. Vi har använt rollen ”system” för systeminformation samt ”user” för den ställda frågan.

De grundmodeller vi använt är GPT-4 som är en chat-completion-modell som OpenAI tillhandahåller med 8192 token lång kontextlängd¹², samt GPT-3 som är en completion-modell från samma utvecklare med 4097 token lång kontextlängd.

7.3.1 Modell 1

Denna modell är en ren GPT-4 modell. Det enda vi gjort är att ge modellen promptmeddelanden som är kontextspecifika. Eftersom vi inte har gett några exempel alls på frågor och svar kallar vi detta för en zero-shot-modell. Zero-shot learning betyder i sammanhanget att GPT-modellen fått en beskrivning av uppgiften utan några exempel på vare sig input eller output. Att be GPT-4 om följande: ”Ge mig exempel på möbler med beskrivningar av deras funktion” är en zero-shot-uppgift. Att be GPT-4 om följande ”Ge mig exempel på möbler med beskrivningar av deras funktion. Några exempel är: Stol – en möbel man kan sitta på, Hylla – en möbel för förvaring” är dock inte en zero-shot-uppgift eftersom vi har gett exempel på hur vi vill att vår output ska se ut och vad den ska innehålla.

7.3.1.1 Generativ modell

Modell 1 (zero-shot) har fått systeminformationen: ”Du är en hjälpsam assistent som svarar på frågor om ekonomistyrning och ÄMNE. Den som ställer frågan jobbar på ORG.”. ÄMNE och ORG plockas ifrån information om frågans tema (ÄMNE) och vilken organisation som frågeställaren hör till (ORG) som utredaren skriver in i

¹² Kontextlängd handlar om hur många tokens en modell kan hantera och komma ihåg. I praktiken betyder detta att när vi ger en GPT-modell systeminformation och andra meddelanden (input) för att få en genererad text som output vilket i vårt fall är en fråga så får den totala antal token-mängden av input + output inte överskrida modellens maximala kontextlängd.

applikationen. Därefter får Modell 1 den givna frågan som utredaren har fyllt i. Frågan skickas till modellen i form av ett user-meddelande.

7.3.2 Modell 2

Modell 2 (vektorsök) är en kombination av två språkmodeller. Dessa är GPT-4 samt en sökmotor som är baserad på semantisk likhet.

7.3.2.1 Sökmotor

För att hitta lämpliga utdrag ur regelverket för att svara på Fråga-ESV-frågorna har vi använt oss av en modell som hjälper oss att leta efter de utdrag ur EA-regelverket som har störst semantisk likhet med den ställda frågan.

Dessa utdrag väljs ut på följande sätt: Vi har hämtat EA-regelverket i HTML-format från ESV Forum¹³ och delat upp den i förordningsparagrafer med tillhörande föreskrifter och allmänna råd (där sådana finns). Texterna för förordningsparagraferna, föreskrifterna och de allmänna råden har vi sedan vektoriserat var för sig med Contrastive-Tension/BERT-Base-Swe-CT-STSB¹⁴. Det innebär att vi har omvandlat varje text till en numerisk representation som fångar dess innebörd i ett multidimensionellt rum. Resultatet har sedan indexerats i sökmotorn Elasticsearch¹⁵.

När en fråga kommer in till applikationen så vektoriserar vi den med samma modell för semantisk likhet och gör en vektorsök mot de indexerade förordningsparagraferna, föreskrifterna och allmänna råden. Detta innebär att vi söker efter de texter vars numeriska representationer är mest lika den för frågan, vilket indikerar att de är semantiskt lika frågan. Denna sökning görs inte mot hela EA-regelverket utan endast mot de delar som är relevanta för det temaområde som utredaren valt till frågan. Vi väljer därefter ut de tre bästa träffarna, det vill säga de som har störst semantisk likhet med frågan enligt contrastive-tension modellen. Varje träff kompletteras sedan med tillhörande förordningsparagraf, föreskrifter och allmänna råd för att bilda ett utdrag ur EA-regelverket som matas till GPT-4.

7.3.2.2 Generativ modell

Modellen har samma systeminformation som Modell 1 och sedan ett extra systemmeddelande som lyder: "När du utformar ditt svar ska du bara använda dig av och referera till följande information:". Efter detta meddelande följer de tre utdrag ur regelverket som vektorsök-modellen tagit fram som bäst matchningar till den ställda

¹³ forum.esv.se/.

¹⁴ En modell för semantisk likhet utvecklad av forskare på RISE och baserad på Kungliga bibliotekets svenska BERT-modell.

¹⁵ <https://www.elastic.co/what-is/elasticsearch>.

frågan, även dessa i form av systeminformation. Sist skickas den ställda frågan till modellen i form av ett user-meddelande.

7.3.3 Modell 3

Modell 3 (finjusterad) är en GPT-3-modell som har finjusterats med hjälp av tidigare frågor och svar från Fråga ESV samt syntetisk data som tagits fram med hjälp av GPT-4.

7.3.3.1 Generering av syntetisk data

För att kunna finjustera vår generativa språkmodell behövde vi träningsdata i form av exempel på frågor och svar inom kontexten redovisning och finansiering inom staten. Då vi inte hade tillräcklig data i god kvalitet i form av frågor och svar från Fråga ESV valde vi att generera fler frågor och svar kring det relevanta temat med hjälp av GPT-4. Detta gjorde vi med hjälp av utdrag ur regelverket. Vi bad GPT-4 att med hjälp av ett utdrag generera så många frågor och svar som möjligt för att använda för finjustering av en GPT-modell. För att generera frågor och svar använde vi systeminformationen: "Skapa så många frågor och svar som möjligt i json-format baserat på följande text inom tre enkla citat-tecken som jag kan använda som träningsdata för att finetuna GPT. ". Därefter följde ett utdrag ur EA-regelverket. Utdragen var i regel 1000 till 2000 tokens långa. Tokens är i praktiken ord eller delar av ord men kan beräknas på lite olika sätt. För att beräkna ungefärlig token-mängd för en textmassa för GPT-modellernas tokeniserare kan man göra följande uträkning: antal ord * 1,333. Antalet syntetiska frågor och svar-par som kunde användas för finjustering var 1504. Dessa syntetiska data slogs ihop med 164 par av frågor och svar från tidigare Fråga ESV-data.

7.3.3.2 Generativ modell

Eftersom GPT-4 i egenskap av chat-completion-modell inte går att finjustera genom OpenAIs API¹⁶ fick vi använda den något äldre modellen GPT-3 (davinci). Modell 3 saknar systeminformation då vi efter viss analys tyckte att denna genererade bättre svar än samma modell med systeminformation. Modell 3 är finjusterad på följande träningsdata: 1668 antal frågor och svar från Fråga ESV varav 1504 antal frågor och svar som är syntetiska på så sätt att de är genererade baserat på EA-regelverket.

¹⁶ API är en förkortning av "Applikationsprogrameringsgränssnitt" ("Application programming interface på engelska") och det är genom OpenAIs egna API som vi når GPT-modellerna för finjustering och generering.

Bilaga 2 – Manuell sentimentanalys

Tabell 18: Manuell sentimentanalys av övergripande kommentarer

Id	Kommentar	Modell 1	Modell 2	Modell 3	Generellt
1	Svar 1: Felaktigt svar. Dock ganska bra formulerat. Svar 2: Rätt svar. Ganska bra formulerat. Svaret är dock inte formulerat hur Riksbanken ska hanteras utan hur det kan vara lämpligt att Riksbanken behandlas. Svar 3: Felaktigt svar.	Neutral	Positiv	Negativ	Neutral
2	Modell 2 gav rätt svar men merparten av texten behövde rensas bort, eftersom det var inte det frågan handlade om.	-	Positiv	Negativ	Neutral
3	Svar 1 och 2 ger en ganska bra grund att arbeta vidare med till det slutliga svaret. Dock ej heltäckande svar. Svar 2 borde även hänvisa till förordningen inte bara ESVFA.	Positiv	Positiv	-	Positiv
4	Samtliga svar är felaktiga och inte heller användbara i någon del.	Negativ	Negativ	Negativ	Negativ
5	Svar 1 och 3 är felaktiga. Svar 2 skulle kunna användas till del. Dock inte fullständigt svar och något felaktig slutsats.	Negativ	Positiv	Negativ	Neutral
6	Det tog otroligt lång tid att generera svaren. Modell 2 tajmade ut och svarade inte först, men sedan fungerade det! Förslagen på konton i Modell 1 är inte de i baskontoplanen för statliga myndigheter, inget om S-koder i de genererade svaren. Hade inte använt något i svaret från Modell 1, 2 eller 3.	Negativ	Negativ	Negativ	Negativ
7	"Det är inte ESV:s sak att svara på frågor om hur myndigheter ska hantera olika typer av bidragsgivning." Det är i princip rätt såsom modell 3 inleder men resten av svaret är inte korrekt/relevant/användbart. Både modell 2 och 1 svarar som att det är ESV:s fråga, vilket det som sagt inte riktigt är. Men det är väl i princip rätt	Positiv	Negativ	Positiv	Positiv

Id	Kommentar	Modell 1	Modell 2	Modell 3	Generellt
	att "Det är viktigt att hantera sådana situationer på rätt sätt för att säkerställa att bidrag används på ett korrekt och effektivt sätt." som Modell 1 anger.				
12	Inget av svaren ger information som kan användas. Generella svar om uppföljning av fakturor och direkt felaktig information om "avgiftsnämnden".	Negativ	Negativ	Negativ	Negativ
13	Samtliga svar är felaktiga på flera olika sätt. Jag tror inte något av dem innehåller något korrekt, möjligen med något mindre undantag.	Negativ	Negativ	Negativ	Negativ
14	Svar 1 och 2 ger bra svar som går att bearbeta vidare. Svar 3 är helt felaktigt.	Positiv	Positiv	Negativ	Positiv
19	Svaren är inte tillräckligt nyanserade utan för kategoriska. Det sista svaret är inte alls bra.	Negativ	Negativ	Negativ	Negativ
20	De är på rätt spår i alla fall. Inget är jättefel även om allt inte är helt rätt.	Positiv	Positiv	Positiv	Positiv
22	Inget av svaren var direkt hjälpsamma för att skriva det slutgiltiga svaret i det här fallet.	Negativ	Negativ	Negativ	Negativ
23	Inget av svarsförslagen var till någon hjälp i det här fallet. ESV har tidigare svarat på liknande frågor i åtminstone två andra fall och även gjort en liten utredning som finns dokumenterad i Intern information endast tillgänglig för handläggare. Ett standardsvar från ESV är också att vi aldrig kan säga vilket konto de ska använda eftersom baskontoplanen är frivillig för myndigheterna att följa, endast S-koderna är obligatoriska. Det har vi skrivit i många tidigare lämnade svar.	Negativ	Negativ	Negativ	Negativ
24	Det första svarsförslaget känns som att det bara vände på orden i frågan och inte tillförde något nytt. Det andra svarsförslaget var bäst och hade kunnat vara till hjälp för att formulera det slutgiltiga svaret. Det tredje	Negativ	Positiv	Negativ	Negativ

Id	Kommentar	Modell 1	Modell 2	Modell 3	Generellt
	svarsförslaget är direkt felaktigt.				
25	Inget av svarsförslagen ger någon hjälp att formulera det slutgiltiga svaret. Andra svarsförslaget är bäst och tredje är sämst.	Negativ	Negativ	Negativ	Negativ
26	Alla svarsförslagen är direkt felaktiga. S-koder är obligatoriska att använda medan baskontoplanen (konto) är frivillig för myndigheterna att använda. Vi ska heller inte hänvisa till ekonomiavdelningen eftersom vi inte vet vem frågeställaren är, det kan lika bra vara just ekonomiavdelningen som frågar. Det verkar inte som att språkmodellerna läser från rapporten Statliga inrapporteringskoder (S-koder) 2023 och inte heller från baskontoplanen.	Negativ	Negativ	Negativ	Negativ
27	Modell nr 2 gav ett svar som var grundat i delvis rätt saker och som hade en formulering som var språkmässig förhållandevis lämplig. Dock med betydligt mer ord och mer tips än vad vi brukar lämna (eftersom vi inte vet de specifika omständigheterna och inte direkt ska ge råd om sådant som polisanmälan). Jag använde dock inte någon specifik formulering eftersom vi i princip inte svarar på frågan mer än att konstatera att det inte rör sig om en avgift.	-	Positiv	-	Positiv
28	Svaren är rätt men känns det som ett lucky shot? =)	Positiv	Positiv	Positiv	Positiv
35	Helt felaktiga svar från samtliga språkmodeller.	Negativ	Negativ	Negativ	Negativ
36	Ganska bra svar avseende 1 och 2. Dock vet jag inte var modellerna hämtar konto 1410 ifrån.	Positiv	Positiv	-	Positiv

ESV gör Sverige rikare

- Vi har kontroll på statens finanser, utvecklar ekonomistyrningen och granskar Sveriges EU-medel.
- Vi arbetar i nära samverkan med Regeringskansliet och myndigheterna.