LIONBRIDGE

AI-DATAANALYS FÖR UTVÄRDERING AV MODELLER

BEDÖMNINGSKATEGORIER

Vid utvärdering av resultat från stora språkmodeller (LLM) gäller det att gå strukturerat tillväga för att kunna analysera språkgenereringsprocessens mångfasetterade natur. Följande bedömningskategorier har valts för att skapa ett heltäckande ramverk för att bedöma resultatens kvalitet, tillförlitlighet och relevans för användare. Kategorierna är ett sätt att se till att utvärderingarna inte bara blir en ytlig kontroll av att innehållet är korrekt, utan även tar hänsyn till aspekter som flyt, fullständighet, ämnesspecifik terminologi och kulturell relevans. På så sätt kan företag tryggt säkerställa att modellens resultat tillgodoser användarnas specifika förväntningar, är i linje med affärsmålen och uppfyller etiska standarder.

KORREKTHET

Mäter om svaret är faktamässigt korrekt och felfritt.

Exempel: Hänvisa till rätt lag i ett juridiskt svar.

FULLSTÄNDIGHET

Kontrollerar om alla aspekter av frågan besvaras tillfredsställande.

Exempel: Alla viktiga punkter finns med i ett sammanfattande svar.

RELEVANS

Bedömer om svaret håller sig till ämnet och är ett direkt svar på frågan.

Exempel: Undviker ovidkommande information i en produktbeskrivning.

KONSEKVENS

Ser till att svaret har en inre logik och inte är motsägelsefullt.

Exempel: Längre svar med flera stycken motsäger inte sig självt.

FLYT

Bedömer om språket är grammatiskt korrekt och har naturligt flyt.

Exempel: Använder rätt meningsbyggnad och skiljetecken.

HALLUCINATION (OMVÄND)

Mäter om modellen undviker att hitta på fakta.
(vid omvänd skala gäller: Ju färre hallucinationer, desto högre poäng)

Exempel: Inte hitta på produkt-egenskaper som inte stämmer.

TERMINOLOGI

Bedömer om ämnesspecifika termer och uttryck används på rätt sätt.

Exempel: Använda exakt medicinsk terminologi i ett vårdrelaterat svar.

LÄSBARHET

Mäter hur enkelt det är för målgruppen att läsa och förstå texten.

Exempel: Ger tydlig och kortfattad text till en allmän målgrupp.

KULTURELL RELEVANS

Utvärderar om svaret är anpassat för och lämpligt enligt målgruppens kulturella normer.

Exempel: Undviker uttryck och hänvisningar som kan vara olämpliga i vissa delar av världen.

Genom att systematiskt utgå från kategorierna (som bedöms på en 5-gradig Likert-skala) kan kunderna få användbara insikter om vilka områden modellen är bra på och var den kan behöva finjusteras, tränas mer eller övervakas av en människa. På så sätt säkerställs att AI-modellen alltid ger upplevelser av hög och jämn kvalitet inom olika användningsområden, oavsett om det är kundsupport, skapande av innehåll, produktrekommendationer eller något annat.

ANALYSEXEMPEL | NÅ DJUPARE INSIKTER

Med hjälp av Lionbridges datatjänster går det att utföra analyser som ger viktiga insikter om AI-träningsprocessernas kvalitet, konsekvens och effektivitet. Insikterna gör det enklare att identifiera mönster i varierade data, överensstämmelse i datauppmärkning, modellens styrkor och svagheter samt trender i resultaten – en förutsättning för att kunna se till att AI-modellerna är robusta, fördomsfria och effektiva. Genom att använda analyserna kan företag fatta databaserade beslut för att optimera AI-träning, förbättra arbetsflöden för uppmärkning och tryggt snabba på införandet av AI-teknik.

Korrekthet

MEDIAN
4,3
1 — 5

SD (σ)
0,6
0 — 1

ANDEL GODKÄNT (>4)
92 %

ÖVRE %-KVARTIL
40 %

Relevans

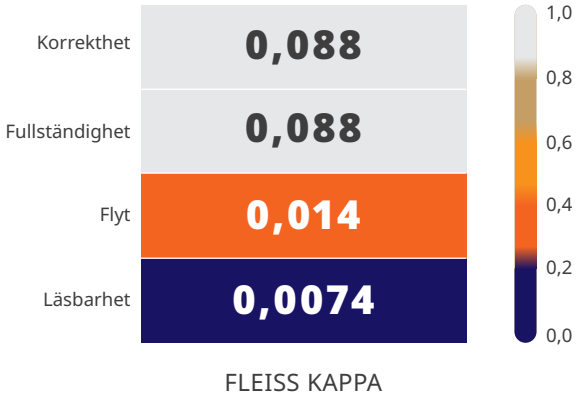
MEDIAN
4,2
1 — 5

SD (σ)
0,5
0 — 1

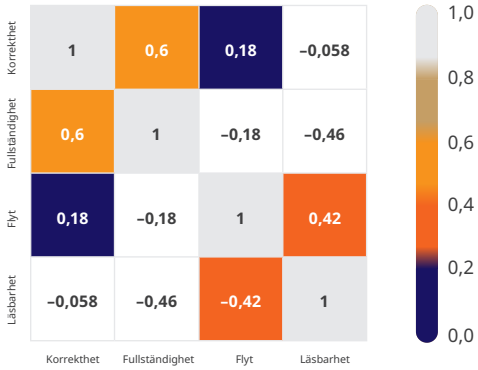
ANDEL GODKÄNT (>4)
95 %

ÖVRE %-KVARTIL
45 %

Överensstämmelse mellan bedömare (Fleiss Kappa)



Korrelation mellan bedömningskategorier



LÄS MER PÅ
LIONBRIDGE.COM