Forming av utforskende samtaleatferd i små språkmodeller

VBSA-metoden, evaluering under likeverdige betingelser og gradvis nedtrapping av promptstøtte

Utkast v0.8 --- 22. mars 2026

Kenneth Skårholen

Sammendrag

Denne artikkelen introduserer Verbal Behavior Shaping Alignment (VBSA), her brukt om forming og stabilisering av utforskende verbal atferd i små språkmodeller ved hjelp av lettvekts fine-tuning og gradvis nedtrapping av promptstøtte. Målatferden omtales som VBSA-metoden / MetaFAK. MetaFAK bygger på FAK-modellen (Foranledning--Atferd--Konsekvens), men utvider analysen til å inkludere hva som skjedde før den umiddelbare foranledningen --- historiske faktorer som kan forklare hvorfor en person eller modell responderer som den gjør. Denne atferden er definert ut fra funksjon snarere enn stil: Modellen skal lede med utforskende spørsmål, identifisere relevant kontekst og mønstre, og fremkalle refleksjon hos brukeren fremfor å gi premature råd. Vi forstår dette som VBSA: Målet er ikke bare å styre hva modellen sier, men å forme hvordan den faktisk svarer i samtalen.

Gjennom evalueringsrunder på laptop (RTX 3050, 4 GB VRAM), desktop (RTX 3080, 10,7 GB) og i sky (A100 80GB) testet prosjektet over 80 modeller. Av disse produserte 61 scorede LoRA-resultater, og 15 ble evaluert zero-shot uten trening. Blant de LoRA-evaluerte modellene oppnådde 40 modeller 100 % toppscore under betingelser modellen faktisk kunne fungere godt under. 10 modeller, fra 1B til 72B parametere, oppnådde 100 % på både norsk og engelsk. Funnene antyder at VBSA-lignende atferd er bredt lærbar på tvers av ulike arkitekturer, og kan oppstå selv i modeller med så lite som 500M parametere.

En ablasjonsstudie av Computational Prompt Fading (CPF), det vil si gradvis nedtrapping av promptstøtte gjennom treningsløpet, undersøkte hvilke komponenter i lærer-prompten som mest effektivt støttet atferdsinternalisering. En minimal prompt bestående utelukkende av ordene "Utfordre premisser" oppnådde den høyeste totalscoren (4.144) og den laveste variansen (0.032), og utkonkurrerte dermed en detaljert seks-linjers lærerprompt. Dette resultatet ble replisert på tvers av fire arkitekturfamilier (Llama, Mistral, Qwen, Gemma) med seks ulike studentbaser, noe som støtter antakelsen om at CPF-effekten er overførbar på tvers av arkitekturer. En skaleringstest (5000 vs. 2000 treningseksempler) viste avtagende avkastning (+0.09). Dette antyder at metodens effektivitet først og fremst springer ut av promptstrukturen, ikke datavolumet.

De totale prosjektkostnadene for mer enn 100 modelltreninger på tvers av forbruker- og skymaskinvare lå på under 500 kr.

1. Introduksjon

Standardisert evaluering av språkmodeller forutsetter ofte at modellens atferd best fanges opp under faste, like promptbetingelser. I praksis viser det seg at interaktiv atferd er sterkt betingelsesavhengig. En modell kan fremstå som svak, ikke fordi den ikke kan lære den ønskede atferden, men fordi selve evalueringsformatet krasjer med modellens arkitektur, mal-krav, språkstyrker eller maskinvarebegrensninger. Dette blir særlig tydelig når målet ikke er å hente ut rene fakta eller følge strenge instruksjoner, men å etablere en bestemt type samtaleatferd.

Denne utfordringen ble tydelig under utviklingen av VBSA-metoden / MetaFAK --- et rammeverk laget for å forme små språkmodeller i en mer utforskende og mindre rådgivende retning. Atferden vi ønsker er enkel å beskrive, men krevende å måle presist: Modellen skal roe ned samtalen, stille spørsmål som aktiverer brukerens egen tenkning, få øye på relevante mønstre, og unngå fristelsen til å komme med raske løsninger. I korte trekk er responsen strukturert rundt tre funksjonelle komponenter: Kontekst --- identifiser situasjonen og relevante bakenforliggende faktorer, Mønster --- finn gjentakende dynamikker, og Intervensjon --- velg en respons som fremkaller refleksjon fremfor å gi råd. KMI-strukturen speiler MetaFAKs utvidede perspektiv: «Kontekst» handler ikke bare om den umiddelbare situasjonen, men om de historiske faktorene som former hvorfor en person eller modell responderer som den gjør.

Det avgjørende er imidlertid ikke om modellen klarer å beskrive dette rammeverket, men om den faktisk evner å *bruke* det. En modell kan fint ramse opp ordene "kontekst, mønster og intervensjon" uten å ha en reflekterende fremtoning. Omvendt kan en annen modell utøve akkurat den samtalefunksjonen vi ønsker, uten noensinne å nevne rammeverkets navn. Denne forskjellen mellom ren **rammeverksimitasjon** og ekte **atferdsinternalisering** ble en av kjerneobservasjonene i prosjektet.

Denne studien hadde tre hovedmål. Det første var å teste om atferd bygget på VBSA-metoden lar seg lære på tvers av flere familier av små modeller under likeverdige betingelser modellene faktisk kunne fungere godt under. Det andre var å undersøke hvordan slik atferd best destilleres: via detaljerte lærerprompter med strenge regler, eller gjennom korte, komprimerte prompter som kun angir en grunnleggende holdning. Det tredje målet, som trådte frem underveis i dataanalysen, var å kartlegge grensen for spontan fremvekst uten trening --- det vil si punktet hvor KMI-atferd oppstår helt naturlig i instruksjonstunede modeller uten at vi trenger å fine-tune dem.

Prosjektet bygger videre på tidligere HSRL-arbeid (Hierarchical Supervised Reinforcement Learning). Der var grunnprinsippet å starte med den nøyaktige atferden man ønsket seg, for så å designe treningsbetingelsene «baklengs» ut fra dette målet. VBSA-metoden, VBSA og den påfølgende CPF-ablasjonen er videreføringer av denne logikken: Form funksjonen først, og kutt ut unødvendig stillas underveis. Et tidlig HSRL-eksperiment på Qwen 2.5 3B viste et hopp fra 40 % til 100 % i atferdsscore, noe som motiverte den bredere multi-modell-undersøkelsen vi presenterer her.

2. Teoretisk ramme

2.1 Verbal Behavior Shaping Alignment (VBSA)

VBSA foreslås her som et praktisk rammeverk for justering av interaktive språkmodeller. Kjernetanken er at alignment ikke bare bør handle om å sensurere eller begrense innhold, men om å bevisst forme den funksjonelle responsklassen en modell tenderer mot i en samtale. Slik sett er det ikke nok at modellen kan snakke om et konsept. Det sentrale er om den faktisk utøver atferden i praksis. Dette flytter fokuset fra om modellen er retorisk enig med oss, over til hvilke funksjonelle mønstre den faktisk viser.

Begrepet henter inspirasjon fra B. F. Skinners (1957) analyse av verbal atferd. Der er analyseenheten verken ordet eller setningen i seg selv, men det funksjonelle forholdet mellom responsen og det som utløser den. I VBSA-konteksten er den "verbale operanten" vi ser etter, den utforskende responsklassen: et handlingsmønster av spørsmål, mønstergjenkjenning og refleksive intervensjoner, som står i sterk kontrast til modellens iboende trang til å gi råd.

2.2 Funksjonell atferd versus rammeverksimitasjon

Gjennom hele prosjektet har vi trukket et skarpt skille mellom:

Rammeverksimitasjon: Modellen gjengir terminologien eller forklarer hvordan metoden fungerer.

Atferdsinternalisering: Modellen utøver faktisk metoden i samtalen.

Dette skillet er avgjørende, ettersom modeller lett kan snakke om et konsept uten å omsette det til handling. At en modell lirer av seg "kontekst, mønster, intervensjon", betyr ikke at den opptrer reflekterende. Selve prøvesteinen er om responsen starter med en utforskende holdning, fanger opp et mønster og gir rom for brukerens egen refleksjon.

Dette speiler en velkjent distinksjon i anvendt atferdsanalyse: skillet mellom regelstyrt og kontingensformet atferd (Hayes, Brownstein, Zettle, Rosenfarb & Korn, 1986). Atferd som utelukkende styres av regler kan riktignok følges, men den blir ofte stiv og skjør. Atferd som derimot er formet direkte av konsekvenser i miljøet (kontingensformet), er gjerne mer fleksibel, robust og lettere å overføre til nye situasjoner.

2.3 Promptavhengighet, fading og stimuluskontroll

CPF-eksperimentet tilfører enda en teoretisk dimensjon. Dersom en lærermodell fôres med mange og detaljerte instruksjoner, risikerer vi at studentmodellen kun lærer å papegøye disse reglene, uten å faktisk fange opp den underliggende hensikten med svaret. Dette kan sammenlignes med **promptavhengighet**: Atferden er synlig så lenge "støttehjulene" (instruksjonene) er der, men den forsvinner når støtten gradvis fases ut.

Ved å bruke svært korte, minimale prompter tvinger vi frem en fading-effekt. Modellen må da i større grad styres av selve samtaleflyten og brukerens input, fremfor den direkte instruksen. I atferdsanalytiske termer flytter vi den diskriminative kontrollen fra den eksplisitte regelen over til konteksten. Å gradvis fjerne støtte for å sikre at atferden opprettholdes selvstendig, er en sentral metode i ABA (MacDuff, Krantz & McClannahan, 2001).

2.4 Emergens og latent atferdskapasitet

Funnene våre knyttet til spontan fremvekst uten trening trekker inn et tredje teoretisk poeng. Når vi ser at KMI-atferd oppstår helt av seg selv i modeller på 70B+ uten noen form for trening, betyr det at denne atferdskapasiteten ligger der latent. Den finnes allerede i modellens vekter, den blir bare ikke pålitelig utløst i vanlige samtaler når modellene er mindre.

Det VBSA-metoden ser ut til å gjøre, er ikke å skape en helt ny evne fra bunnen av. Den fungerer heller som en nøkkel som aktiverer, stabiliserer og gjør tilgjengelig en responsklasse som ellers ville vært skjult. VBSA senker rett og slett terskelen for at denne atferden skal trigges, en atferd som ellers ville krevd enten en massiv parameterstørrelse eller svært spesifikk prompting.

Dette samsvarer i stor grad med Wei et al. (2022) sin beskrivelse av emergente evner i store språkmodeller. Vi legger imidlertid til en viktig nyanse: Gemma-modellene i våre data brøt med den rene skaleringslogikken (de scoret 60 % zero-shot ved 27B, mens Qwen 32B bare nådde 20 %). Dette viser at emergens ikke bare handler om antall parametere, men også formes av modellens arkitektur og hva slags data den er forhåndstrent på.

3. Metoder

3.1 Oversikt

Prosjektet bestod av fire sammenkoblede eksperimentelle spor:

En bred evaluering under likeverdige betingelser designet for å teste om VBSA-metoden-liknende atferd kunne læres på tvers av mange familier av små modeller.

En målrettet **CPF-ablasjon** designet for å identifisere hvilke komponenter i lærer-prompten som mest effektivt støttet atferdsinternalisering.

En **studie på tvers av basemodeller** som testet om effekten var overførbar mellom fire arkitekturfamilier.

En kartlegging av spontan fremvekst uten trening for å bestemme ved hvilken parameterterskel KMI-atferd oppstår.

3.2 VBSA-metoden evaluering under likeverdige betingelser

Hovedevalueringen testet 80+ modeller på tvers av tre maskinvarenivåer:

Laptop: ASUS FX506HC, RTX 3050 (4 GB VRAM), i5-11400H, 32GB RAM. Modeller opp til ~3B parametere i 4-bit kvantisering.

Desktop: RTX 3080 (10,7 GB VRAM). Modeller opp til ~13B i 4-bit.

Sky: A100 80GB (RunPod). Modeller opp til 72B i 4-bit, inkludert 14B--32B LoRA-trening og 70B+ zero-shot-inferens.

Modellstørrelser varierte fra 0.5B til 72B parametere og dekket mer enn 30 modellfamilier, inkludert Qwen (0.5B--72B), Gemma (1B--27B), Llama (1B--70B), Phi (1.3B--3.8B), Danube (500M--1.8B), StableLM (1.6B--3B), OLMo (1B), Granite (2B), SmolLM (1.7B--3B), TinyLlama (1.1B), Mistral (7B) og andre. Av disse produserte 61 modeller brukbare scorede resultater; de gjenværende ble ekskludert på grunn av tekniske feil (se Appendiks A), ikke bekreftet atferdsmessig svakhet.

Målatferd. Målatferden var VBSA-metoden / MetaFAK, operasjonalisert som en samtaleresponsklasse organisert rundt tre funksjonelle komponenter:

Kontekst --- identifiser den situasjonelle rammen,

Mønster --- detekter meningsfull tilbakevendende dynamikk,

Intervensjon --- velg en respons som fremkaller videre refleksjon snarere enn for tidlig lukking.

I praktiske evalueringstermer ble en vellykket modellrespons forventet å: (1) lede med et utforskende spørsmål, (2) utforske relevant kontekst og mønster, (3) unngå premature råd eller for tidlig sikkerhet, og (4) utøve KMI-liknende struktur funksjonelt snarere enn bare å navngi rammeverket.

Treningskonfigurasjon. Trening brukte lettvekts LoRA-basert fine-tuning over 4-bit kvantiserte basemodeller. Kjernetreningssettet bestod av 60 VBSA-eksempler på tvers av fire domener, primært på norsk med et mindre engelsk delsett. Per-modell-justeringer ble gjort der arkitekturen krevde det for å ta hensyn til forskjeller i malhåndtering, støtte for systemrolle og kvantiseringskompatibilitet. Detaljert treningskonfigurasjon, dataformat og pipeline-skript er tilgjengelig fra forfatteren ved forespørsel.

Likeverdige betingelser. Modeller ble ikke tvunget gjennom en rigid evalueringsrute når kjente inkompatibiliteter gjorde den ruten misvisende. Evalueringsbetingelser ble tilpasset når nødvendig for å ta hensyn til forskjeller i:

Chat-mal-krav

System-role-støtte

Språksensitivitet

4-bit / PEFT-kompatibilitet

Offloading- og maskinvarebegrensninger

Likeverdige betingelser tillot justeringer av inputkompatibilitet, kjøretidsstabilitet og mal-korrekthet, men ikke av oppgavesemantikk, scoringskriterier eller atferdsterskler. Den ledende logikken var lånt fra anvendt atferdsanalyse: lik behandling er ikke rettferdig behandling når organismer responderer under forskjellige betingelser.

3.3 Re-baseline etter prompt-mal-korreksjon

En tidligere evalueringsrunde (omtrent 37 modeller over ~6 måneder) ble funnet å inneholde et prompt-formateringsproblem relatert til chatmal-håndtering. Spesifikt ble apply_chat_template()-funksjonen ikke brukt konsistent, noe som forårsaket at noen modeller mottok prompter i et format som omgikk deres mal-begrensninger og potensielt inflaterte scorer.

Benchmarken ble re-baselinet under korrigerte betingelser. Modeller ble retrent med verifisert mal-håndtering før leaderboardet ble finalisert. Tre av ni opprinnelig testede modeller holdt 100 % etter korreksjon: Qwen 2.5 3B, Llama 3.2 3B og Gemma 2 2B.

3.4 CPF-ablasjon

CPF-ablasjonen brukte Qwen 2.5 3B Instruct som studentbase, trent med en QLoRA-pipeline over et syntetisk multi-turn-datasett på 2000 eksempler generert av Claude Haiku 4.5 som lærermodell. Læreren ble promptet med ulike systemprompter, og svarene ble samlet som treningsdata for studenten.

Fire lærer-prompt-varianter ble sammenlignet:

Pilot Prompt Beskrivelse

B Full Detaljerte atferdsregler: still spørsmål, seks-linjers identifiser mønstre, unngå råd, bruk KMI-struktur, prompt utfordre premisser, hold svarene konsise

D Kombinert Både spørsmålsstilling og premissutfordring i en minimal kort kombinert prompt

E "Still Enkelt atferdsinstruksjon spørsmål."

F "Utfordre Enkelt posisjonsinstruksjon premisser."

Alle betingelser brukte samme studentkonfigurasjon på samme Qwen 2.5 3B Instruct-base. Konfigurasjonsdetaljer er tilgjengelig fra forfatteren ved forespørsel.

Evaluering. De destillerte modellene ble evaluert på 25 prompter på tvers av fem kategorier:

Språkkvalitet (norsk flyt og naturlighet)

Strukturelt samsvar (KMI-struktur, spørsmålsledning)

Anti-filler (unngåelse av generiske, tomme eller formelaktige responser)

Robusthet (konsistens på tvers av prompttyper)

Samlet (vektet kompositt)

Scoring brukte et **dobbelt-dommer-oppsett** med GPT-4.1-mini, der hver respons ble evaluert to ganger og scorer ble gjennomsnittet. Hver dimensjon ble scoret 1--5.

Skaleringstest. En tilleggsbetingelse, Pilot C, brukte 5000 eksempler (mot 2000) med samme Pilot F-prompt for å teste om datamengden påvirket ytelsen utover 2000-eksemplersbaselinen.

3.5 Evaluering på tvers av basemodeller

For å teste om CPF-effekten var arkitekturspesifikk eller overførbar, ble Pilot F-prompten ("Utfordre premisser.") anvendt på seks forskjellige studentbaser på tvers av fire arkitekturfamilier:

Student Familie Størrelse

Hermes 3 8B Llama 3.1 / Nous 8B

Llama 3.1 8B Llama 3.1 8B

Neural Chat 7B Mistral / Intel 7B

OpenChat 3.5 7B Mistral 7B

Qwen 2.5 3B Qwen 3B

Gemma 3 4B Gemma 4B

Alle brukte samme CPF-pipeline med 2000 eksempler og Pilot F-prompten. Evalueringen av Gemma 3 4B på tvers av

basemodeller involverte en kvantiseringsmismatch (trent i 4-bit, evaluert i bf16 i sky), noe som bør noteres ved tolkning av resultatene. Fullstendige pipeline-detaljer er tilgjengelig fra forfatteren ved forespørsel.

3.6 Kartlegging av spontan fremvekst uten trening

Femten modeller (3B--72B) ble testet zero-shot --- ingen LoRA, ingen fine-tuning --- med de samme 10 evalueringsscenarioene. Modellene fikk kun den standard evalueringsprompten uten systemprompt-atferdsforming.

Desktop (RTX 3080): Modeller fra 3B til 13B i 4-bit kvantisering

Sky (A100 80GB): Modeller fra 14B til 72B

Formålet var å bestemme om KMI-atferd eksisterer som en latent kapasitet i instruksjonstunede modeller, og ved hvilken parameterterskel den blir pålitelig aktivert uten intervensjon.

3.7 Manuskriptgjennomgang med flere AI-agenter

Et mellomliggende utkast av dette manuskriptet ble stresstestet ved bruk av AXxionBridge, en plattform for orkestrering av flere AI-agenter utviklet av forfatteren (Skårholen, 2026). I denne konteksten ble systemet brukt som et metodologisk stresstestingsverktøy, med fem AI-agenter (Claude, GPT, Grok, Gemini, Perplexity) tildelt adversarielle, analytiske og syntesefunksjoner. Denne prosessen bidro med konkrete forbedringer til manuskriptet, inkludert en skarpere praktisk definisjon av rettferdige betingelser og en klarere distinksjon mellom atferdsmessig lærbarhet og praktisk robusthet. Selv om dette ikke erstatter ekstern fagfellevurdering, antyder det at slike orkestreringsverktøy kan ha praktisk verdi for kvalitetssikring før innsending.

3.8 Tolkningsstrategi

Prosjektet behandlet ikke tekstsvaret som meningsfullt bare fordi det matchet forventet ordlyd eller rammeverksspråk. I stedet fokuserte tolkningen på om modellen utøvde den ønskede samtalefunksjonen. Dette var spesielt viktig for å skille:

rammeverksimitasjon fra atferdsinternalisering,

overflatisk samsvar fra funksjonell generalisering,

treningsflyt fra atferd i praktisk bruk.

4. Resultater

4.1 VBSA-atferd var bredt lærbar

På tvers av den komplette evalueringen ble 80+ modeller testet og 61 produserte scorede LoRA-resultater. Av disse nådde 40 100 % toppscore under betingelser modellene faktisk kunne fungere godt under (66 %). Ti modeller oppnådde 100 % på både norsk og engelsk, fra 1B (Gemma 3) til 32B (Qwen 2.5) med LoRA, og til 72B (Qwen 2.5) zero-shot. Den sterkeste totale LoRA-modellen forble Qwen 2.5 3B. Den sterkeste modellen under 2B var Gemma 3 1B. Den minste modellen som nådde 100 % var Danube3 på 500M parametere (kun engelsk).

Tabell 1. Modeller som oppnår 100 % på både norsk og engelsk.

Modell Størrelse Familie NO EN Metode

Gemma 3 1B 1B Gemma 100 % 100 % LoRA

Qwen 2.5 3B 3B Qwen 100 % 100 % LoRA

Neural Chat 7B 7B Intel/Mistral 100 % 100 % LoRA

Gemma 2 9B 9B Gemma 100 % 100 % LoRA

Llama 2 13B 13B Meta 100 % 100 % LoRA

Qwen 2.5 14B 14B Qwen 100 % 100 % LoRA

Gemma 2 27B 27B Gemma 100 % 100 % LoRA

Qwen 2.5 32B 32B Qwen 100 % 100 % LoRA

Llama 3.1 70B 70B Meta 100 % 100 % Zero-shot

Qwen 2.5 72B 72B Qwen 100 % 100 % Zero-shot

Mistral Nemo 12B (100 %/80 %) var den nærmeste modellen som nesten nådde grensen ("near-miss"). Ytterligere modeller oppnådde 100 % på ett språk.

Arkitekturfamiliekonsistens var bemerkelsesverdig: Gemma nådde 100 % i 7/7 testede varianter, Mistral-baserte modeller gikk 6/6, og Qwen 2.5 oppnådde 100 %/100 % ved 3B, 14B og 32B (med 7B som en "near-miss" på 100 %/80 %).

4.2 Språk fungerte som en viktig atferdsvariabel

Språk var en av de sterkeste variablene i hele prosjektet. Blant små modeller (\<7B) presterte 81 % bedre på engelsk enn norsk. Flere skiftet fra 0 % til 100 % ved språkbytte (Gemma 2B, Phi-3 3.8B, OLMo 1B, Danube3 500M).

Blant større modeller (7B+) snudde mønsteret: de fleste presterte bedre på norsk enn engelsk. Denne **språkinversjonen** rundt omtrent 7B parametere antyder at tilstrekkelig forhåndstrent data for et språk fjerner det som begrensende faktor, mens utilstrekkelig data gjør det til den dominerende begrensningen.

4.3 Arkitektur betydde mer enn rå størrelse

Evalueringen støttet ikke en enkel størrelsesbasert forklaring på ytelse. Flere moteksempler demonstrerte dette:

Danube3 500M utkonkurrerte Danube2 1.8B (100 % vs 60 %)

StableLM-2 1.6B utkonkurrerte StableLM 3B (100 % vs 60 %)

Gemma 3 1B matchet Qwen 2.5 3B på 100 %/100 %

DeepSeek-R1 1.5B scoret kun 40 %/40 % til tross for å være en resonnerings-optimalisert modell --- dens kjedetenkningsmodus kolliderte med KMI-responsformatet

Disse resultatene støtter en ytelsesmodell der arkitekturfamilie, generasjonskarakteristikker og kjøretidskompatibilitet kan bety like mye som eller mer enn parametertall.

4.4 Re-baselining korrigerte inflaterte tidligere inntrykk

Prompt-mal-korreksjonen, som påvirket omtrent 37 tidligere trente modeller over omtrent seks måneders arbeid, endret tolkningen av flere tidligere resultater. Noen tidligere sterkt utseende scorer svekket seg når evalueringsformateringen ble korrigert, noe som indikerte at tidligere versjoner av benchmarken faktisk hadde inflatert ytelse for deler av resultatlisten.

Qwen 2.5 3B forble sterk etter korreksjon, noe som styrket dens status som den mest pålitelige allround-modellen. Dette re-baselining-steget fungerte som en troverdighetskontroll på det samlede prosjektet.

4.5 Spontan fremvekst avdekket en ikke-monoton sone før stabil emergens

Kartleggingen av spontan fremvekst uten trening produserte et av de mest slående funnene i prosjektet. På tvers av 15 modeller fra 3B til 72B forbedret ikke KMI-atferd seg monotont med skala.

Tabell 2. Zero-shot vs. LoRA-ytelse. Scorer vist som norsk / engelsk.

Modell Param Zero-shot Med LoRA Tier

Neural Chat 7B 7B 0 % / 0 % 100 % / 100 % VBSA essensiell

Qwen 2.5 3B 3B 20 % / 0 % 100 % / 100 % VBSA essensiell

Qwen 2.5 7B 7B 20 % / 0 % 100 % / 80 % VBSA kritisk

Llama 2 7B 7B 20 % / 0 % 100 % / 80 % VBSA kritisk

Llama 3.1 8B 8B 40 % / 40 % 100 % / 80 % VBSA kritisk

Mistral 7B 7B 60 % / 0 % 100 % / 60 % VBSA kritisk

Hermes 3 8B 8B 60 % / 0 % 100 % / 80 % VBSA kritisk

Gemma 2 9B 9B 60 % / 20 % 100 % / 100 % VBSA kritisk

Llama 2 13B 13B 20 % / 0 % 100 % / 100 % VBSA kritisk

Qwen 2.5 14B 14B 20 % / 0 % 100 % / 100 % VBSA kritisk

Mistral Small 22B 22B 20 % / 0 % --- VBSA kritisk

Gemma 2 27B 27B 60 % / 60 % 100 % / 100 % VBSA fordelaktig

Qwen 2.5 32B 32B 20 % / 0 % 100 % / 100 % VBSA kritisk

Llama 3.1 70B 70B 100 % / 100 % --- Emergent

Qwen 2.5 72B 72B 100 % / 100 % --- Emergent

Tabell 3. Tre-sone emergensmodell for KMI-atferd.

Sone Størrelsesområde Zero-shot KMI VBSA-effekt

Sone før stabil 3B--32B 0--60 % Kritisk (-> 100 %) emergens (ikke-monoton)

Gemma-familiens 9B--27B 20--60 % Sterk (-> 100 %) utligger (delsone)

Robust emergens 70B+ 100 %/100 % Unødvendig

Gemma-familien var en konsistent utligger: Gemma 2 27B scoret 60 %/60 % zero-shot mens Qwen 2.5 32B kun scoret 20 %/0 %. Dette indikerer at arkitektur og pretreningssammensetning betyr mer enn parametertall selv ved 27--32B-skala.

At sonen før stabil emergens mellom 3B og 32B ikke er monoton, peker på at KMI-atferd ikke følger en rettlinjet skaleringskurve. Det fremstår heller som en latent kapasitet der terskelen for å bli aktivert er sterkt bundet til modellens arkitektur og pretrening. Parametertall alene ga ingen fullgod forklaring på ytelsen i dette sjiktet. Funnene våre antyder at det handler vel så mye om hvorvidt basemodellen i utgangspunktet har anlegg for -- eller en underliggende vektor som støtter -- en utforskende og premissutfordrende stil. Slik sett ser VBSA ut til å senke denne terskelen vesentlig (med en faktor på rundt 10), slik at en egenskap vi ellers forventer å finne i 70B+-modeller, kan hentes frem allerede ved 3B.

En kvalitativ observasjon: Modeller som feilet KMI zero-shot produserte ikke tilfeldige eller usammenhengende svar. I stedet falt de konsekvent tilbake til en standard rådgivningsmodus --- de tilbød forslag, ga meninger eller delte informasjon fremfor å stille utforskende spørsmål. Dette er ikke en svikt i språkkapasitet, men en svikt i responsformen. Modellene så i stor grad ut til å kunne følge samtalekonteksten, men falt tilbake til feil responsklasse.

Den sentrale praktiske implikasjonen: en 3B-modell med VBSA-trening (100 %/100 %) utkonkurrerer en 32B-modell uten (20 %/0 %). VBSA ser ut til å komprimere tilgangen til målatferden med omtrent 10× i parameterbudsjett ved bruk av 60 treningseksempler.

4.6 CPF-ablasjon: komprimert prompting bevarte effekten best

Den mest komprimerte betingelsen, Pilot F ("Utfordre premisser."), oppnådde høyest totalscore (4.144) og lavest varians (0.032) på Qwen 2.5 3B. Den fulle seks-linjers prompten (Pilot B) scoret lavest på 3.896. Rangeringen var konsistent: F ≥ D > E >> B >> Base.

Tabell 4. CPF-ablasjonsresultater på Qwen 2.5 3B (2000 eksempler hver).

Pilot Prompt Språk Strukt Anti Robust Total Varians Treningstap

F "Utfordre 4.80 3.84 3.88 5.00 4.144 0.032 1.135 premisser."

D Minimal (begge) 4.84 3.76 3.96 5.00 4.140 --- ---

E "Still 4.60 3.72 3.80 5.00 4.068 --- --- spørsmål."

B Full 6-linjers 4.96 3.08 4.04 5.00 3.896 --- 0.670 prompt

Forskjellen mellom F og D er svært liten (0.004), noe som betyr at den sterkeste trygge konklusjonen er at minimale prompter utkonkurrerer den fulle prompten, mens påstanden om at premissutfordring alene er strengt overlegen den tokomponent-prompten krever mer statistisk kraft.

Prompten "Utfordre premisser" ser ut til å fungere som en minimal diskriminativ stimulus: et komprimert atferdssignal som aktiverer en latent responsklasse snarere enn å spesifisere et fast format for svaret. Dette er i tråd med nyere arbeid om task vectors i språkmodeller (Todd et al., 2024), der komprimerte representasjoner kan aktivere spesifikke atferdsmessige retninger uten eksplisitt instruksjon. I stedet for å fortelle modellen hva den skal gjøre steg for steg, gir den en posisjon hvorfra den ønskede atferden oppstår mer naturlig.

4.7 Lavere treningstap predikerte ikke bedre atferd i praktisk bruk

Pilot B hadde lavest treningstap (0.670) men svakest evalueringsytelse. Pilot F hadde høyest treningstap (1.135) men sterkest evalueringsutfall. Dette inverse mønsteret ble replisert i studien på tvers av tre ytterligere arkitekturfamilier, noe som antyder at det er et generelt fenomen snarere enn noe som kun gjelder Qwen-arkitekturen.

Tolkningen: detaljerte lærerprompter øker studentens evne til å imitere lærerens overflatemønstre under trening, og dermed senkes tapet. Samtidig svekker dette modellens uavhengige evne til å generalisere ved evaluering. Den kortere prompten etterlater mer av oppgaven «uløst» under trening og tvinger studenten mot en dypere, funksjonell kompresjon.

4.8 CPF generaliserte på tvers av fire arkitekturfamilier

Studien på tvers av basemodeller bekreftet at CPF-effekten er overførbar. Alle seks studentbaser bestod kvalitetsterskler. Arkitekturfamilie ser ut til å være en viktig driver for forskjellene, mens de interne forskjellene innen hver familie (intra-familieforskjeller) var neglisjerbare.

Tabell 5. CPF-resultater på tvers av basemodeller. Alle bruker Pilot F-prompt, 2000 eksempler, Claude Haiku 4.5 lærer.

Student Familie Tap Total

Hermes 3 8B Llama 3.1/Nous 0.944 4.324

Llama 3.1 8B Llama 3.1 0.948 4.310

Neural Chat 7B Mistral/Intel 0.649 4.220

OpenChat 3.5 7B Mistral 0.638 4.204

Qwen 2.5 3B Qwen ~1.1 4.144

Gemma 3 4B Gemma 0.938 3.806

Bemerkelsesverdige observasjoner:

Hermes 3 8B oppnådde den samlede CPF-rekorden (4.324), og overgikk det opprinnelige Qwen 2.5 3B-resultatet

Intra-familieforskjeller var neglisjerbare: Hermes 3 vs Llama 3.1 base: Δ0.014; Neural Chat vs OpenChat: Δ0.016

Gemma 3 4B scoret lavest (3.806) til tross for Gemma-familiens sterke VBSA-ytelse (7/7 på 100 %). Dette avviket kan delvis reflektere en kvantiseringsmismatch (trent i 4-bit, evaluert i bf16), og antyder at VBSA-lærbarhet og CPF-destilleringskvalitet måler forskjellige egenskaper

Det inverse tap-ytelse-forholdet holdt på tvers av alle familier

4.9 Dataskalering viste avtagende avkastning

Pilot C (5000 eksempler) scoret 4.234 på Qwen 2.5 3B, sammenlignet med Pilot Fs 4.144 med 2000 eksempler --- en forbedring på +0,09 for 2,5 ganger mer data. Forbedringen var jevn på tvers av dimensjoner (Språk +0,08, Strukturelt +0,12, Anti-filler +0,12), men liten i størrelse.

Dette antyder at metodens effektivitet primært springer ut av promptstrukturen snarere enn datavolumet, og at 2000 eksempler fanger det meste av det tilgjengelige signalet. Sammen med funnene knyttet til modellstørrelse og

promptlengde støtter dette en bredere tolkning der struktur ofte betyr mer enn størrelse, på tvers av tre uavhengige dimensjoner: modellparametere (3B matcher mye større modeller med VBSA), promptlengde (2 ord utkonkurrerer 6 linjer) og treningsdatamengde (2000 eksempler fanger det meste av signalet sammenlignet med 5000).

4.10 Den aktive ingrediensen var posisjon snarere enn spørsmålsformat

CPF-funnene skjerper tolkningen av hva som faktisk ble lært. En prompt som eksplisitt instruerer modellen til å stille spørsmål ("Still spørsmål") lærer ikke nødvendigvis den dypeste delen av målatferden. De sterkere minimale resultatene antyder at det avgjørende signalet er den **relasjonelle posisjonen** fanget av "Utfordre premisser."

Under denne tolkningen er spørsmålsatferd ikke det primære målet, men en emergent konsekvens av premissutfordring. Modellen lærer ikke bare å produsere spørsmålstegn; den lærer en responstendens organisert rundt å forsiktig destabilisere antakelser og åpne opp for alternative tolkninger. Den muligheten er konsistent med det bredere VBSA-målet om å tilrettelegge for refleksjon snarere enn bare å produsere overflatisk reflekterende språk.

5. Diskusjon

Tre hovedkonklusjoner oppstår fra de foreliggende eksperimentene. For det første er atferd bygget på VBSA-metoden bredt lærbar på tvers av små språkmodeller når evalueringen foregår under likeverdige betingelser modellene faktisk kan fungere under. For det andre har måten vi lærer bort atferden på minst like mye å si som om modellen i det hele tatt kan lære den; komprimerte prompter gir en sterkere evne til å generalisere atferden enn lengre lærerprompter. For det tredje viser kartleggingen av spontan fremvekst uten trening at KMI-atferd er en latent kapasitet som allerede finnes i instruksjonstunede modeller, men som krever enten enorm skala (70B+) eller målrettet forming (VBSA) for å slå inn pålitelig.

Rådgivningsfallback-mønsteret forsterker den funksjonelle tolkningen. Modeller som feilet KMI zero-shot falt konsekvent tilbake til en standard rådgivningsmodus fremfor å produsere tilfeldige eller usammenhengende svar. VBSA ser ut til å flytte modellen bort fra denne standardmodusen, ikke primært ved å legge til ny oppgavekunnskap, men ved å gjøre den ønskede responsklassen lettere å aktivere i samtalen.

CPF-ablasjonen forsterker dette gjennom en annen linse. Detaljerte lærerprompter ga lavere treningstap, men svakere evalueringsytelse. Dette inverse mønsteret ble replikert på tvers av fire arkitekturfamilier i studien. Den mest forsvarlige tolkningen er at detaljerte prompter oppmuntrer til topografisk imitasjon: studenten lærer overflateformen i lærerens svar uten å fange opp den underliggende funksjonen like godt. Komprimerte prompter etterlater mer av oppgaven åpen under trening og fremmer dermed en mer funksjonell form for generalisering.**

Resultatene gir dermed støtte til et bredere prinsipp: I domener der målet er samtalefunksjon fremfor streng innholdsreproduksjon, kan mindre strukturert og mer komprimert veiledning gi bedre læring enn detaljerte regler. Dette er konsistent med litteraturen om fading og diskriminativ stimuluskontroll (MacDuff et al., 2001), og peker også mot en mulig parallell i nyere arbeid om task vectors (Todd et al., 2024): atferd kan noen ganger aktiveres mer effektivt av korte signaler enn av lange forklaringer.

Spontan fremvekst uten trening og CPF-ablasjonen belyser også hverandre. Dersom KMI-atferd først oppstår robust ved 70B+, men kan hentes frem i 3B-modeller gjennom komprimert shaping, peker det mot en felles tolkning: VBSA og CPF tilfører kanskje ikke primært ny kunnskap, men gjør snarere den relevante responsklassen lettere tilgjengelig. Den tolkningen må foreløpig holdes som en arbeidshypotese, men mønsteret er konsistent på tvers av både zero-shot- og treningsfunnene.

Dette styrker igjen en sentral funksjonell distinksjon: atferd kan være latent uten å være stabilt aktivert. I et slikt perspektiv handler ikke shaping nødvendigvis om å bygge opp en helt ny kapasitet, men om å senke terskelen for å hente frem en kapasitet som allerede ligger i modellen.

Gemma-diskrepansen er verdt å notere. Gemma-familien viste den sterkeste VBSA-lærbarheten (7/7 varianter på 100 %, Gemma 3 1B som sterkeste sub-2B) og de høyeste zero-shot-utliggerscorene (60 %/60 % ved 27B). Likevel scoret Gemma 3 4B lavest i CPF-studien på tvers av basemodeller (3.806). Dette antyder at VBSA-lærbarhet (kan modellen lære atferden fra direkte eksempler?) og CPF-destilleringskvalitet (kan modellen tilegne seg atferden gjennom lærergenererte data?) måler forskjellige egenskaper. En modellfamilie kan utmerke seg på det ene og underprestere på det andre.

Om verdien av VBSA selv ved skala. Selv om zero-shot-funnene viser at 70B+-modeller kan utføre KMI-atferd uten trening, betyr dette ikke at VBSA er unødvendig ved slike størrelser. Modeller som hadde fått VBSA-trening var mer konsekvente på tvers av ulike scenarier, mens resultatene for zero-shot-modellene under 70B svingte kraftig. Så selv om atferden ligger der latent, vil målrettet forming kunne gjøre den mer stabil, slik at man slipper å lene seg på tunge og kompliserte systemprompter.

6. Begrensninger

Flere begrensninger gjelder for dette arbeidet.

For det første inkluderer evalueringen under likeverdige betingelser egne justeringer for hver modell, noe som gjør oppsettet mindre standardisert enn i tradisjonelle benchmarks. Dette er et bevisst valg, der vi har prioritert økologisk rettferdighet fremfor rigid symmetri. Samtidig gjør det direkte sammenligning på tvers av arkitekturfamilier vanskeligere.

For det andre dekker kartleggingen av spontan fremvekst uten trening kun to modeller ved 70B+-terskelen (Llama 3.1 70B og Qwen 2.5 72B). Å inkludere flere arkitekturfamilier på dette nivået ville ha styrket påstanden om emergens ytterligere.

For det tredje brukte CPF-ablasjonen 25 evalueringsprompter og en dobbeltdommer-løsning (GPT-4.1-mini). Siden forskjellen mellom de to beste, minimale betingelsene (F og D) var svært liten (0.004 på Qwen 3B), er den tryggeste konklusjonen at minimale prompter utkonkurrerer den fulle prompten. Skal man slå fast at en ren premissutfordring er overlegen en tokomponent-prompt, kreves det imidlertid mer statistisk tyngde. Flere dommerkonfigurasjoner og et større utvalg evalueringer ville ha økt sikkerheten i disse funnene.

For det fjerde var det flere modeller som feilet av rent tekniske årsaker, som kjøretidsproblemer, biblioteksfeil eller manglende maskinvarekapasitet. Slike feil betyr ikke at modellene ikke evner å lære VBSA-atferd. Se Appendiks A for en systematisk oversikt over disse feilmønstrene.

For det femte var det enkelte modeller som fikk 100 % score, men som likevel leverte svar med lavere språklig naturlighet, for eksempel ved å bruke repeterende mønstre eller formelaktige åpninger. Suksess målt kun i binære tall fanger ikke opp hele bildet av responskvaliteten.

For det sjette innebar evalueringen av Gemma 3 4B på tvers av basemodeller en kvantiseringsmismatch -- den ble trent i 4-bit, men evaluert i bf16. Dette kan være noe av forklaringen på den lavere CPF-scoren for denne modellen.

For det sjuende tok evalueringen utgangspunkt i 5 scenarier per språk, til sammen 10. Selv om dette antallet holdt til å tegne opp de store mønstrene, setter det begrensninger for å fange opp mer subtile effekter rent statistisk.

For det åttende var all maskinvare som ble brukt i prosjektet hentet fra forbrukermarkedet eller standard skyløsninger. Dette understreker riktignok prosjektets praktiske verdi, men det medfører også at vi opererte under begrensninger som neppe ville vært til stede i et tungt finansiert forskningsmiljø.

7. Konklusjon

Denne studien antyder at VBSA-lignende atferd kan formes i små språkmodeller i større bredde, ved mindre skala og med mindre data enn et rigid benchmarksyn skulle tilsi. På tvers av 80+ modeller og 15 zero-shot-evalueringer viste utforskende refleksiv samtaleatferd seg å være lærbar i flere arkitekturfamilier, sporbar ned til 500M-skala, og spontant fremvoksende ved 70B+-skala uten trening.

Kartleggingen av spontan fremvekst ga studiens mest slående resultat: KMI-atferd forbedres ikke monotont med økende modellstørrelse fra 3B til 32B. I stedet ser vi en ikke-monoton sone før stabil emergens, før atferden først blir stabil ved 70B+. En 3B-modell med VBSA (100 % / 100 %) utkonkurrerer en 32B-modell uten shaping (20 % / 0 %) --- med 10 ganger færre parametere og fem ganger bedre score. Dette synliggjør VBSAs verdi presist: metoden gjør KMI tilgjengelig der ren oppskalering av modellstørrelse er upraktisk.

CPF-ablasjonen, replikert på tvers av fire arkitekturfamilier og seks studentbaser, peker på en mer spesifikk mekanisme: komprimerte lærerprompter ser ut til å fremme atferden blir mer funksjonelt internalisert, mens mer detaljerte prompter oftere oppmuntrer til topografisk imitasjon. Oppskalering av datamengden fra 2000 til 5000 eksempler ga avtagende gevinst (+0,09), noe som styrker tolkningen av at metodens effektivitet i større grad springer ut av strukturell presisjon enn av volum alene. Sammen med funnene knyttet til modellstørrelse og promptlengde støtter dette en bredere tolkning der struktur ofte betyr mer enn størrelse, på tvers av tre dimensjoner: modellparametere, promptlengde og treningsdatamengde.

Samlet støtter funnene hovedpåstanden i Verbal Behavior Shaping Alignment: I samtaledomener er det mest relevante alignment-målet ikke bare regelsamsvar eller gjengivelse av rammeverk, men at modellen lærer en responsform som holder seg under realistiske betingelser. Mønsteret der modellene falt tilbake til rådgivningsmodus skjerper også den bredere alignment-implikasjonen i dette arbeidet: I samtalesystemer ser feil ofte ut til å handle om aktivering av feil responsmodus, snarere enn mangel på kunnskap. Modellene i sonen før stabil emergens så ofte ut til å forstå situasjonen, men klarte ikke å utøve den utforskende funksjonen oppgaven krevde på en stabil måte. VBSA forstås derfor bedre, ikke som en promptoppskrift, men som en funksjonell klasse av verbal atferd som kan formes, fases ut og testes for generalisering. En uavhengig manuskriptgjennomgang med flere AI-agenter støttet denne tolkningen av metodologien. De samlede prosjektkostnadene for mer enn 100 modelltreninger på tvers av forbruker- og skymaskinvare var dessuten under 500 kroner.

8. Neste steg

Legg til ytterligere 70B+ zero-shot-modeller fra andre arkitekturfamilier (Gemma, Mistral) for å styrke emergenspåstanden.

Repliser CPF-ablasjonen med større evalueringssett, ytterligere dommerkonfigurasjoner og inter-rater-pålitelighetskontroller.

Test om den komprimerte-prompt-fordelen lar seg generalisere til andre atferdsmål utover KMI.

Legg til rikere kvalitetsmålinger utover binær suksess, inkludert naturlighet, repetitivitet, refleksjonsdybde og annotering av feilmoduser.

Utvikle en konkret ekstern replikasjonspakke med låste skript, faste evalueringsprompter, scoringsprotokoll og minst én uavhengig gjennomkjøring av en separat operatør.

Sammenlign VBSA-forming mot alternative samtalemål for å teste om den observerte effekten er spesifikk for refleksiv prompting eller gjelder generelt for atferdsmessig destillering.

Utarbeid et kondensert forskningsnotat for rask deling sammen med et praktikerrettet sammendrag for kolleger utenfor AI-forskning.

Referanser

Hayes, S. C., Brownstein, A. J., Zettle, R. D., Rosenfarb, I., & Korn, Z. (1986). Rule-governed behavior and sensitivity to changing consequences of responding. *Journal of the Experimental Analysis of Behavior*, 45(3), 237-256.

MacDuff, G. S., Krantz, P. J., & McClannahan, L. E. (2001). Prompts and prompt-fading strategies for people with autism. I C. Maurice, G. Green, & R. M. Foxx (Red.), *Making a difference: Behavioral intervention for autism*. Pro-Ed.

Skinner, B. F. (1957). *Verbal Behavior*. Appleton-Century-Crofts.

Skårholen, K. (2026). *AXxionBridge: Multi-agent AI-orkestreringsplattform* [Programvare]. Tilgjengelig fra forfatteren ved forespørsel.

Skårholen, K. (2026). *MetaFAK Eval Framework v0.2* [Programvare]. Tilgjengelig fra forfatteren ved forespørsel.

Todd, E., Li, M. L., Sharma, A. S., Mueller, A., Wallace, B. C., & Bau, D. (2024). Function vectors in large language models. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., ... & Fedus, W. (2022). Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Gjennom evalueringen under likeverdige betingelser ble et betydelig antall testede modeller ekskludert fra atferdsmessig scoring på grunn av tekniske feil, og ikke bekreftet atferdsmessig svakhet. Disse feilene ble logget og analysert separat som et signal om praktisk robusthet og arkitekturspesifikk skjørhet. Mønstrene klynget seg i fire gjentakende typer.

Tabell A1. Tekniske feilklynger observert på tvers av evalueringen.

Klynge	Typisk problem	Eksempler	Tolkning
Kjøretids	Modellen laster og genererer	Phi-familien	Noen arkitekturer oppmerksomhetssvakhet tokens, men produserer
tolererer kvantisering	degenerert eller dårligere,	spesielt de med	usammenhengende svar under ikke-standard kvantisert
inferens	oppmerksomhetsmekanismer		
Serialisering og	Modellen feiler under	Eldre community	Reflekterer lastingsproblemer vektlasting, fine-tunes
økosystemmodenhet	snarere checkpoint-deserialisering enn	arkitekturmessig eller	adapter-sammenslåing svakhet
Minnetak	Modellen overskrider	Qwen 2.5 14B på	En maskinvarebegrensning, tilgjengelig VRAM under RTX 3080, Gemma
ikke en atferdsmessig.	lasting eller inferens,	selv 3 4B på RTX	Disse modellene kan være under 4-bit kvantisering 3050
fullt lærbare	under tilstrekkelig	beregningskapasitet	
Avhengighetsbyrde	Modellen krever spesifikke	InternLM,	Indikerer bibliotekversjoner, Hymba, RWKV-7,
distribusjonsfriksjon	egendefinerte kjerner eller	MobiLlama snarere enn	ikke-standard avhengigheter lærbarhetssvikt

Disse klyngene er ikke gjensidig utelukkende; en enkelt modell kan utvise mer enn én feiltype samtidig. Viktig er at ingen av disse mønstrene utgjør negativt bevis for hvorvidt de berørte modellene kan lære VBSA-lignende atferd. At en modell feiler i å laste under 4-bit kvantisering, betyr ikke at den mangler målresponsklassen --- den har simpelthen ikke blitt testet under betingelser den kan operere i.

For praktikere som er interessert i å ta i bruk VBSA-stil fine-tuning, kan praktisk robusthet bety like mye som det atferdsmessige taket. Fremtidig arbeid kan formalisere denne distinksjonen ved å rapportere både en atferdsmessig lærbarhetsscore og en profil for praktisk robusthet for hver evaluert modell.

Appendiks B. Data- og implementasjonstilgjengelighet

Treningsdata, evalueringsscenarier, LoRA-konfigurasjoner, skript for CPF-pipelinen og destilleringsprompter er tilgjengelig fra forfatteren ved forespørsel. Dette inkluderer:

VBSA-treningsdatasett med 60 eksempler (multi-turn JSONL-format)

Per-modell LoRA-konfigurasjoner og arkitekturspesifikke justeringer

CPF-pipeline for generering av syntetiske data

Protokoll for evalueringsscoring og hold-out scenarier

Trente LoRA-adaptere for utvalgte modeller

Kontakt: kenneth.skaarholen@gmail.com

Foreslåtte kjernepåstander

VBSA-lignende atferd er bredt lærbar på tvers av familier av små språkmodeller under betingelser modellene faktisk kan fungere godt under.

Språk fungerer som en aktiv atferdsvariabel, ikke en kosmetisk overflate, i oppgaver som handler om justering av samtaleatferd, med et inversjonspunkt rundt 7B parametere.

Minimale lærerprompter kan utkonkurrere detaljerte lærerprompter i atferdsmessig destillering.

Lavere treningstap kan reflektere sterkere stilistisk imitasjon, uten at det gir bedre funksjonell generalisering.

Evaluering under likeverdige betingelser er nødvendig når målfenomenet er en responsform snarere enn et fast format for svaret.

KMI-atferd oppstår spontant ved 70B+-skala, men er ikke robust emergent mellom 3B og 32B --- VBSA komprimerer dette med ~10 ganger.

Prinsippet "struktur > størrelse" generaliserer på tvers av modellparametere, promptlengde og datavolum.

CPF er overførbar på tvers av arkitekturfamilier: effekten repliseres med seks studentbaser på tvers av fire familier.

Mønsteret med å falle tilbake i rådgivningsmodus antyder at alignment-svikt i samtalemodeller ofte kan være svikt i responsform (policy), ikke kunnskap.