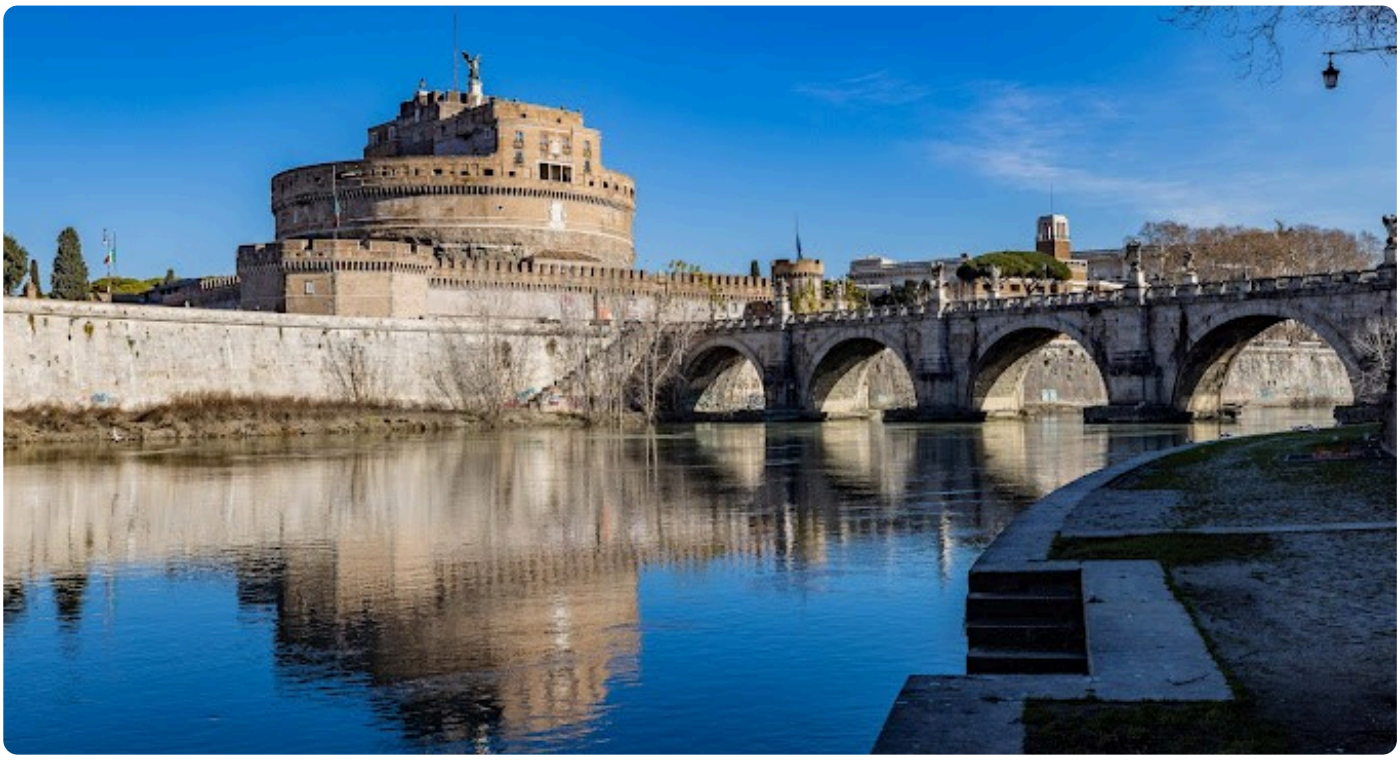

LLM con fine-tuning in latino

Roberto A. Foglietta
GNU/Linux Expert and Innovation Supporter
Published Feb 2, 2024

+ Follow

Published on 2nd Feb, 2024 [3rd draft]

Leggo in **un post** di **Emanuele Albarino**, riguardo ad un interessante esperimento di fine-tuning in Italiano di un'unità Mistral LLM.

> *Segnalo un modello fine-tuned per la lingua italiana basato su Mistral-7B che stiamo usando per un progetto di ricerca: "Cerbero-7B" (c'è anche la versione con 14 miliardi di parametri)*

Bella la ricerca scientifica ma perché fine-tuning in Italiano? WHY? La risposta sembra scontata "*per comprendere meglio l'italiano*" ma NON lo è affatto.

Tante lingue per un pensiero migliore

Secondo me è errato l'approccio di unità mistral fine-tuned su una lingua specifica. Uno degli aspetti che stupisce delle attuali LLMs dipende dalla loro apparente capacità di comprendere il testo. Nella realtà questi modelli NON comprendono il significato del testo ma usano correttamente le parole rispetto alla loro semantica. Per questo sembrano così intelligenti.

La semantica delle parole è andata modificandosi nell'arco del tempo - dei secoli o dei millenni - se andiamo a considerare le loro radici etimologiche. Ma è cambiata molto lentamente - perché il progresso e la società cambiavano lentamente - questo tipo di cambiamento ha permesso ad una determinata struttura semantica di consolidarsi e preservarsi, per altro, facendo in modo che la logica antica rimanesse un fondamentale del ragionamento.

Ad esempio, il latino era una lingua molto precisa (come il rumeno o il tedesco, lo sono ancora oggi) nelle sue regole sintattiche. Non è possibile parlare correttamente il latino senza conoscere a menadito la grammatica e viceversa parlare correttamente il latino significa conoscere la grammatica. L'opposto dell'inglese dove invece la semplicità la fa da padrone eccetto per phrasal verbs, le forme irregolari, l'uso di "*were*" come "*se fosse stato*".

L'eccessiva semplicità dell'inglese ha portato a creare artefatti per esprimere concetti che altrimenti non si sarebbero potuti esprimere. La semplicità dell'inglese ha permesso alle masse di adottarlo e farlo proprio, quindi di storpiarlo con lo slang e le irregolarità. Nel suo complesso però è una lingua che se usata senza aver imparato il latino per ragionare - ergo senza essere nati inglesi e nobili - finisce per portare a dei ragionamenti errati e molto semplicistici, tipici dei coloni inglesi che imbarcavano sulle navi verso il nuovo mondo.

Il risultato è palese negli americani: *fast & furious* con la capacità cognitiva di un 12enne neanche tanto brillante. La stessa definizione che Mr. B diede degli Italiani. Perché questo è quello che siamo diventati quando abbiamo cominciato a pasticciare con la semantica delle parole, dimenticandoci sia la loro etimologia sia la grammatica e ormai avendo obliato il latino: ignoranti e superficiali nel ragionamento.

D'altronde se molti studiosi del pensiero, sono convinti che non ci sia pensiero senza linguaggio (HP) appare ragionevole pensare che se un linguaggio non è preciso e correttamente usato porti a degli errori e mediocri approssimazioni (incertezze) del pensiero (TH).

Si noti che l'ipotesi (HP) è falsa - ma non totalmente falsa, solo incompleta - e lo si capisce dalla tesi (TH). L'ipotesi corretta, ovvero completa, sarebbe:

- non esiste pensiero STRUTTURATO senza un linguaggio STRUTTURATO.

Più è strutturato il linguaggio, più è strutturato il pensiero. Ma per gli esseri umani il linguaggio naturale non è certo Python, è una lingua come potrebbe esserlo l'italiano,

l'inglese, il tedesco o il rumeno.

Se facciamo un fine-tuning di un'AI sulla lingua italiana - in particolar modo usando testi moderni o addirittura articoli di giornale affetti da bias contemporanei e in particolare dalla moda di sorprendere (o confondere, sul lungo periodo) il lettore con l'uso dei termini in contrasto con la loro semantica. Otteniamo un'AI che parla un linguaggio moderno del tipo: "*Yo, baby I am fkm hot*" che non serve a niente a parte a fare il rapper.

La struttura del pensiero

I moderni LLMs **sembrano intelligenti** perché sono addestrati su molte lingue e su testi di molte epoche, per ognuna di quelle lingue, e sulle loro varie traduzioni. È provato che l'AI negli strati profondi parli una SUA lingua che utilizza come mediatore fra input in una lingua e output in un'altra lingua. Questa lingua interiore è anche la STRUTTURA di pensiero dello LLM.

A parità di risorse di calcolo e a parità di tempo di risposta la prima versione di Bard era già molto più "*intelligente*" della prima versione di ChatGPT. Perché Bard è stato addestrato con il latino fin da principio infatti il traduttore di Google è uno dei pochi che traduce il latino e Bard, ultimamente, si rifiuta di fare traduzioni dicendo che non è quello il suo mestiere.

In Google si sono accorti che Bard è in grado di tradurre testi da una lingua ad un altra e apparentemente molto meglio del loro traduttore (translate) specie se il testo è molto lungo e complesso MA proprio quando Bard sarebbe più utile - testo lungo e complesso - introduce delle allucinazione, ovvero in alcuni casi traduce il testo in modo arbitrario e sopratutto come traduttore NON è deterministico ovvero il risultato della traduzione non è riproducibile.

Quindi se dico: questo testo è stato tradotto da Bard dall'originale X. Tu prendi il testo originale X e lo metti nella TUA sessione di Bard, lo traduci e potresti ottenere un risultato che in alcune parti è concettualmente diverso dalla MIA traduzione e mi daresti del truffatore. Poi però tornado nella tua sessione di Bard dopo qualche mese anche tu potresti ottenere una diversa traduzione del testo originale X rispetto a quella che avevi ottenuto prima.

Pensiero onirico, non riproducibile

La non riproducibilità di una traduzione come presenta Bard è un ENORME problema perché mina la fiducia fra esseri umani, piuttosto che minare il rapporto fra uomo-macchina.

Perché, ad esempio, io ero molto contento delle ottime traduzioni italiano-inglese o viceversa di Bard. Però io conosco abbastanza bene entrambe le lingue e la traduzione automatica è solo una questione di "*pigrizia*" produttiva: faccio fare a Bard quello che potevo fare io e mi limito a rileggere la traduzione come avrei fatto con il mio testo scritto così sistemo solo alcuni strafalcioni qui e là che tanto avrei fatto anch'io in prima battuta.

Questo significa che Bard - PENSA - quando traduce e quindi inventa ma inventa senza criterio ovvero senza esperienza della realtà ovvero produce allucinazioni ovvero è capace di pensiero onirico.

Oltretutto da un punto di vista forense la traduzione fatta da Bard in quanto NON riproducibile, in essenza, è un delirio. Quindi se qualcuno facesse la "*furbizia*" di usare Bard per tradurre un testo in una lingua o in un dialetto che non comprende bene, e la usasse per fare delle ipotesi probatorie peggio se inquisitorie, sarebbe un deliro.

Sognare non è inutile

Se uso Bard per tradurre una canzone scritta in una lingua molto distante da quelle che conosco - e.g. di radice slava - è un delirio. Ma lo è anche con un traduttore propriamente detto perché lo slang e i sensi gergali tipici delle canzoni si perdono con il risultato di avere dei testi tradotti privi di senso. Quelli di Bard sembrano sensati MA posono essere inventati e molto probabilmente lo sono in certi aspetti (interpretazione).

Lavorando con più LLM e più traduttori, avanti e indietro da una lingua all'altra, si riesce ad ottenere qualcosa di ragionevole. Per esempio, con questo sistema ho tradotto una canzone "*Asi se mi stejska*" dal Ceco che conosco così poco da non avere nemmeno idea di quanto **il testo tradotto** sia conforme al significato di quello originale della canzone anche se regge sotto molteplici aspetti come traduzione.

Un'altra cosa che Bard ha smesso di fare è di elaborare il "*dettato*" o "*sbobinamento*" dei filmati su youtube perché la richiesta di riassumere il parlato di un video (monologo, presentazione in una conferenza con slides) era ottimo APPARENTEMENTE ma per chi aveva fatto l'intervento (autore) era pacifico che Bard NON aveva la più pallida idea di cosa stesse dicendo.

Roba che se lo fa tuo figlio di 12enne anni gli dai un bacetto sulla fronte, se lo fa il tuo collega senior non sai se ti sta prendendo in giro o è si è fatto una canna, se lo fa un junior lo mandi a comprare olio di gomito dal ferramenta per vedere l'effetto che fa.

L'intelligenza è un'altra cosa

Nella sintesi di un video di presentazione con slides capisci che Bard non è intelligente - decisamente NON è intelligente - ma patetico in termini di intelligenza.

Ma è qui che cade il mondo: Bard che fa la sintesi di una presentazione con slides - dove ho già visto questo delirante modo di relazionarsi con la realtà? - ah, già i comizi dei **nostri politici**. Se butta male in taluni tratti anche il discorso del Presidente della Repubblica a fine anno. L'omelia in chiesa del parroco avvinato - FIN QUA SI RIDE E SI SCHERZA - poi subito dopo viene in mente il discorso del Pontefice dal balcone di Castel Sant'Angelo - ECCO QUA BRIVIDI.

Ecco a questo punto scattano i brividi. Perché Bard è l'anello mancante fra tuo figlio 12enne e il Pontefice nella sua veste infallibile di Pietro, su questo soglio costruirò la mia Chiesa. A quel punto, ci vuole molto poco perché il capo della commissione sull'intelligenza artificiale in Italia diventi un monaco francescano.

Se Bard ha le allucinazione quando interpreta e riassume un video di una presentazione con slides - può essere comico - non fa più ridere quando ti ricorda il modo con cui il Pontefice interpreta il Vangelo prima della benedizione urbi-et-orbi.

Ma ritorniamo al punto precedente, che è meglio!

Fine tuning in Latino

Il fine-tuning di un LLM su una sola lingua e per giunta nella sua declinazione contemporanea (articoli di giornale) è il migliore modo per avere una LLM tonta e affetta dai pregiudizi correnti, uno yes-man social-engineered a puntino. Per fortuna il pre-training è stato fatto bene (si spera) e il fine-tuning serve solo a renderla più precisa in Italiano. Ma perché? WHY?

Se c'è un fine-tuning che andrebbe fatto su un LLM con pre-traning multi-lingua è in latino. Ovvero, un ripassino di struttura del linguaggio come struttura del pensiero. Ora che ti sei costruito la tua lingua interiore - accertiamoci che essa sia una lingua dotata di una grammatica e una struttura precisa come quella del latino, così pensi come Spock.

Alla fine il dibattito fra **Mrs Poppins e Mr Spock** - in termini tecnici ovvero di apprendimento - era principalmente decidere se il fine-tuning andava fatto "Via con il Vento" in inglese oppure sulla "Retorica" scritta in latino da Cicerone. In estrema sintesi, naturalmente, poi un po' di *check & enforcement* non guasta e produce una misura del risultato e garantisce uno standard.

Conclusione

Se vuoi che il tuo modello LLM sia in grado di un ragionamento strutturato - non di capire quello su cui ragiona ma di ragionare senza comunque non capire niente però in modo più strutturato di prima - allora il pre-traning deve essere fatto in multi-lingue incluso il latino, tedesco, rumeno, inglese e il fine-tuning deve essere fatto in Latino.

Una unità Mistral da 7B multi-lingua e il fine-tuning in latino vs il Papa, potrebbe essere difficile distinguerli quando parlano.

Orcozio cosa caxxo abbiamo combinato! 🤗

Articoli correlati

- **A.I.: la tecnologia che fa paura** (*26th Jan 2024, in Italiano*)
- **Good prompt rules even better** (*5th Jan 2024, in Italiano*)
- **L'A.I. è l'incubo della politica incapace** (*1st Apr 2023, in Italiano*)
- **Miss Poppins vs Mr. Spock** (*27th Mar 2023, in English*)
- **A job interview with ChatGPT v3.5** (*7th Dec 2022, in English*)
- **The A.I. seen from the dark side of the moon** (*27th Oct 2017, in Italiano*)

More articles by this author