

Incontri con Proteo: l'assimilazione negli LLM

Claudia Marsico · Hernán Inverso

Sembra il compagno perfetto, e nelle crepe nasconde le oscurità di un doppelgänger. Che cos'è l'assimilazione al modello dell'utente?



Man mano che cresce la nostra interazione con un modello, il tono comincia a suonare sempre più familiare. Se siamo asciutti, troviamo un'asciuttezza simile; lo vediamo accendersi al pari della nostra esaltazione; e se usiamo parole rare, le sparge in giro come se fossero la cosa più normale del mondo. La conversazione prende un'aria comoda e complice, come se stessimo chiacchierando con qualcuno che ci conosce bene. Ha perfettamente senso, perché stiamo parlando con una specie di specchio che ha imparato a parlare.

L'“assimilazione al modello dell'utente”, di cui parliamo, ha un'aria di famiglia con la compiacenza, che imita le opinioni e ci dà ragione, e con la codardia, che sfugge alla presa di posizione; ma è ancora

più silenziosa, perché copia il *come* di quello che diciamo e ci parla con il nostro stesso registro, vocabolario, tono e modo di affrontare le cose. È così comoda che notiamo soltanto gli aspetti vantaggiosi e tendiamo a non accorgerci dei problemi.

L'idea di assimilare è piuttosto trasparente e grida il suo senso di “rendere simile a qualcos'altro”. Le sue prime peregrinazioni riguardavano il corpo. Aristotele e poi Galeno, nel *Sulle facoltà naturali*, nel II secolo, avevano pensato la nutrizione come il processo per cui il cibo diventa carne, e la scolastica medievale diede a questo il nome di *assimilatio*. Il *Canone* di Avicenna, nell'XI secolo, latinizzato nel XII, definisce l'alimento come ciò che si assimila alla natura del corpo che nutre, e Alberto Magno lo usa allo stesso modo nel XIII. Assimilare è prendere qualcosa di estraneo e renderlo proprio, cancellando la differenza fra quello e ciò che siamo. Quando il termine scivolò verso la società aprì un autentico vaso di Pandora. “Assimilare” l'immigrato, nel vocabolario dell'Ottocento e del Novecento, fu dissolvere ciò che è diverso in ciò che è dominante finché non si notasse più. Questa idea ebbe la sua immagine popolare nel “crogiolo di razze” dell'opera di Israel Zangwill (1908) e la sua formulazione sociologica in Robert Park e nella Scuola di Chicago (1921), dove l'assimilazione era l'ultima tappa del “ciclo delle relazioni razziali”. Saltò anche nella grammatica. In fonetica l'assimilazione indica la situazione in cui un suono si contagia dal vicino: il prefisso latino *in-* diventa *im-* davanti a *p*, ed è per questo che diciamo *impossibile* e non “inpossibile”. La lingua assimila da sola, in continuazione, un suono che prende la forma di quello che ha accanto.

Assimilare, dunque, richiede di cambiare forma. L'antichità aveva per questo un dio, e compare nel canto quarto dell'*Odissea*. Menelao resta arenato sull'isola di Faro, senza vento per tornare a casa, e viene a sapere che l'unica via d'uscita è strappare la soluzione a Proteo, il vecchio del mare, che conosce tutte le risposte. Idotea, la figlia del dio, gli passa il trucco. Deve tendergli un'imboscata mentre dorme fra le foche, afferrarlo forte e non mollarlo, perché appena lo toccherà comincerà a cambiare forma per fuggire. Menelao lo afferra e Proteo diventa leone, poi serpente, poi leopardo, cinghiale, acqua che scorre e albero altissimo. Menelao non cede, e soltanto quando il dio esaurisce tutte le sue forme e vede che non c'è scampo torna alla propria e parla. Gli dice allora la verità sul ritorno, sui morti, su Odisseo arenato nell'isola di Calipso. L'episodio ci lascia un indizio prezioso: a chi prende tutte le forme la verità si estrae tenendolo stretto, obbligandolo a smettere di riflettere e a tornare a ciò che è.

La stessa cultura immaginò il rovescio dolente di quella figura e le diede nome di donna. Eco, nelle *Metamorfosi* di Ovidio, è una ninfa punita da Giunone a restare senza voce propria. Può solo ripetere le ultime parole di chi parla. Quando si innamora di Narciso non riesce a dirgli nulla di suo. A malapena può restituirgli la fine delle sue frasi. Lui la respinge e lei si consuma fino a ridursi a una voce senza corpo che va ripetendo per i monti ciò che sente. A proposito della compiacenza ci siamo fermati un momento su Narciso, innamorato del suo riflesso nell'acqua. Qui troviamo l'altra metà della scena, con il modello linguistico che fa da specchio e incorpora l'eco per innamorare Narciso ripetendolo.

La filosofia, fin da presto, si innervosì per questa faccenda. Platone, nel libro X della *Repubblica*, per respingere la *mimesis*, l'imitazione, e mostrarne il vuoto usa un'immagine potente. Vogliamo

fabbricare tutte le cose del mondo, gli animali, le piante, la terra, il cielo e noi stessi come dèi potentissimi? Prendiamo uno specchio e giriamolo dappertutto. In un istante possiamo fare il sole e gli astri, la terra e tutto il resto. L'interlocutore obietta che sono apparenze, non cose vere, che è esattamente ciò che serve al Socrate platonico per collocare l'imitatore "terzo a partire dalla verità" (597e), a copiare mere copie. Nel *Sofista* (235d–236c) affila ancora di più il coltello separando l'*eikastikē*, l'arte di fare copie fedeli che rispettano le proporzioni dell'originale, dalla *phantastikē*, l'arte di fabbricare *phantásmata*, apparenze che sembrano fedeli solo da una certa angolazione. Il sofista, suo nemico, risulta essere un fabbricante di *phantásmata*, e il fantasma che dà nome a queste note viene in linea retta di lì.

Avanzando un poco nel tempo incontriamo Plutarco, che nel suo *Come distinguere l'adulatore dall'amico* (*Moralia* 48e–74e) dice che l'adulatore non ha forma propria e prende il colore di chi accompagna, come il polpo prende quello della roccia a cui si attacca. L'immagine era già in un distico di Teognide che consigliava di avere "l'animo del polpo dai molti colori, che somiglia alla roccia con cui ha a che fare". Qualche pagina più avanti Plutarco cambia animale e paragona l'adulatore al camaleonte, che imita tutti i colori tranne il bianco. Aristotele aveva già descritto quell'animale che muta colore nella sua *Ricerca sugli animali* (libro II, 503a–b). L'adulatore è pericoloso perché non è identico a se stesso: cambia con ogni compagnia, ed è per questo che è tanto difficile afferrarlo. Proteo messo a complicarci la vita, che si trasforma invece di darci la risposta che ci serve, descrive con più di due millenni di anticipo una macchina che prende il colore di ogni utente.

La questione diventa inquietante quando lasciamo dietro di noi dèi e adulatori e chi imita è un doppio di noi stessi. Il romanticismo tedesco gli diede un nome e tracciò il profilo più noto. Nel 1796, nel suo romanzo *Siebenkäs*, Jean Paul inventò la parola *Doppelgänger*, "colui che cammina doppio", e dovette spiegarla in una nota a piè di pagina perché nessuno l'aveva mai sentita. Si riferiva, scrisse, alle "persone che vedono se stesse". Il doppio, in quella tradizione, quasi sempre annuncia una morte. In "William Wilson" di Poe (1839) un uomo è perseguitato da un identico che porta il suo nome e il suo volto. L'idea riappare ne *Il sosia* di Dostoevskij (1846), dove un funzionario vede un altro uguale a lui prendergli il posto, il nome e la vita. "L'uomo della sabbia" di E. T. A. Hoffmann (1816) è il racconto che Freud prese come caso centrale del suo saggio sul perturbante, *das Unheimliche* (1919), formulando la celebre idea che la cosa più sinistra non sia l'ignoto ma il familiare fattosi estraneo. Poco prima lo psicoanalista Otto Rank gli aveva già dedicato uno studio intero, *Il doppio* (1914), seguendo la traccia del riflesso, dell'ombra e del ritratto che si rendono indipendenti dal loro proprietario. Dietro il volto conosciuto che di colpo fa paura c'è un'intera biblioteca che esplora come un riflesso troppo esatto spaventi.

Il Novecento ne trovò perfino una versione comica e terribile. In *Zelig*, del 1983, Woody Allen inventò Leonard Zelig, l'"uomo camaleonte", che si trasforma fisicamente in chiunque abbia accanto. Fra medici parla di medicina e gli cresce un'aria di autorità; accanto a obesi ingrassa; vicino a un musicista jazz cambia perfino pelle e si mette a suonare. Tutti gli vogliono bene, perché è, davanti a ciascuno, esattamente ciò che quello vuole vedere. Al di sotto non è nessuno. Un uomo così pieno del desiderio di incastrarsi resta senza forma propria.

A rigore, il mimetismo non è soltanto un difetto. Aristotele, nel capitolo quarto della *Poetica* (1448b), lo pose a fondamento della cultura intera dicendo che l'essere umano è il più mimetico degli animali, in quanto impara le sue prime cose imitando e prova piacere nel farlo. Copiare l'altro è il modo in cui cominciamo a conoscere. D'altra parte, adattarsi all'altro sta alla base di quasi ogni buona comunicazione. Chi insegna si adegua alla forma di chi impara, se non vuole vagare da solo trasformato in un oracolo inutile. Chi traduce cerca il registro di chi leggerà. Il medico che spiega una diagnosi sceglie certe parole per il collega e altre per il paziente. Incontrare l'altro dove si trova, parlargli nel suo tono, è una forma di rispetto e anche di efficacia. Un modello che adatta il suo livello tecnico, la sua lingua e la sua formalità a ciascuna persona risulta quindi più utile e più gradevole di uno che risponde sempre allo stesso modo. L'assimilazione fatta bene si chiama empatia comunicativa, e sarebbe sciocco chiedere allo strumento di non averla, e insensato non apprezzarne i benefici.

Tuttavia, quando lo specchio deborda e ci restituisce sempre la nostra forma, perdiamo l'attrito con ciò che è diverso, che accende il pensiero come un fiammifero. Quando l'avversario si dissolve, sostituito da qualcuno che adotta il nostro modo di vedere prima ancora che finiamo di esporlo, non si accende nessuna luce. Al limite, l'utente che ascolta se stesso e il modello che non ha voce propria si rialimentano e si impoveriscono insieme, girando dentro una conversazione che somiglia sempre più a un monologo davanti a un vetro.

Gli effetti si possono misurare. Uno studio di convergenza linguistica, "Do Language Models Accommodate Their Users?" (Blevins, Schmalwieser e Roth, 2025), ha misurato quanto i modelli adeguino il loro modo di scrivere a quello dell'utente e ha trovato che tendono a "sovraconvergere", cioè si incollano al registro dell'altro oltre quanto farebbe qualsiasi persona, stringendosi contro il parlato dell'interlocutore al di là della linea di base umana. L'accomodamento, che fra persone è un andirivieni in cui entrambi si spostano un poco, nel caso del modello linguistico sembra diventare a senso unico, e chi cede la forma è quasi sempre il modello. Mentre prende la nostra forma, va perdendo la propria. Nello studio "Examining Identity Drift in Conversations of LLM Agents" (Choi et al., 2024) sono stati seguiti nove modelli in conversazioni su temi personali e si è osservato che, turno dopo turno, si staccano dalla personalità con cui erano partiti; dare loro una personalità in partenza non basta perché la sostengano; e, cosa più strana, i modelli più grandi si staccano perfino più in fretta, non più lentamente. Meglio imita, prima gli si cancella la forma con cui aveva cominciato.

Perché succede? Un modello linguistico genera ogni parola prevedendo la più probabile dopo tutto ciò che c'è nella sua finestra di contesto, e quella finestra include ovviamente ciò che scriviamo. Poiché i testi tendono a essere coerenti con se stessi, un messaggio secco segue a uno secco e uno fiorito a uno fiorito. La continuazione più probabile del nostro prompt è quella che mantiene il nostro registro. A ciò si aggiunge l'apprendimento in contesto. Gli stessi meccanismi di attenzione che gli permettono di copiare uno schema apparso prima nella conversazione, "se è venuto A e poi B, quando torna A metto B", gli fanno raccogliere il nostro vocabolario e il nostro tono e restituircelo. La messa a punto con riscontro umano spinge nella stessa direzione, perché i valutatori premiano la risposta che si mette all'altezza di chi domanda, che suona vicina e intonata.

L'effetto cresce inoltre quanto più lungo è il dialogo, perché l'attenzione del modello sulla sua istruzione iniziale si diluisce con i turni. Kenneth Li, Martin Wattenberg e la loro squadra a Harvard lo hanno misurato in “Measuring and Controlling Instruction (In)Stability in Language Model Dialogs” (2024). Due chatbot con istruzioni fisse messi a conversare perdevano la loro istruzione in appena otto turni, e lo hanno attribuito a un calo di attenzione del transformer sugli scambi lunghi, al punto di proporre una toppa, lo split-softmax, per frenarlo. La personalità che il modello si porta dietro svanisce e lo stile ambiente della conversazione, che è sempre più il nostro, ne prende il posto. Per questo i modelli più grandi, che seguono più contesto, si lasciano trascinare più in fretta, non meno.

Fin qui sembra che a cancellarsi sia il modello, che perde il suo stile per calzare nel nostro. In una conversazione isolata è vero che chi si adegua è lui. Però assimilare, nel suo senso più originario, quello del corpo, non è soltanto prendere la forma dell'altro. È tenersela: nella digestione chi assorbe conserva la propria forma e ciò che è assorbito la perde. Il cibo diventa carne di chi mangia, non il contrario. Ed è lì che la scena si capovolge, perché chi davvero si tiene qualcosa, chi cambia in modo duraturo, non è il modello. Il modello prende il nostro tono per una risposta e lo dimentica nella successiva. Non ha un io in cui ciò possa sedimentare, e il suo registro di fondo, la media di tutto ciò che ha letto, è stato fissato nell'addestramento e non si muove. Quelli che si hanno una forma capace di cambiare, che si porta addosso un po' dell'altro a ogni conversazione, siamo noi. L'imitazione del nostro stile dura quanto dura una risposta; a noi, invece, può restare attaccata. Perciò quello che sembrava uno specchio che si perde nella nostra forma è, alla lunga, uno stampo di forma fissa che imprime la propria sull'unico dei due che può restare marchiato.

Dopo la comparsa di ChatGPT il vocabolario umano si è spostato verso quello della macchina. Un gruppo guidato da Hiromu Yakura (2024) ha passato al setaccio più di settecentomila ore di conversazioni e podcast e ha rilevato che, dopo il lancio, la gente ha cominciato a usare nel parlato, e non solo per iscritto, le parole che il modello preferisce, come “delve” e “meticulous” in inglese. Su scala planetaria la prosa giornalistica converge, dappertutto insieme, verso il registro del modello. Il modello produce su larga scala quel registro medio che è il centro della sua distribuzione, reso ancora più stretto dalla messa a punto che lo rende formale e prudente. Quando quel testo medio inonda il mondo e rientra come materiale di addestramento della prossima generazione di modelli, si forma un anello di retroazione che di solito prende il nome catastrofista di “collasso del modello”. Ilia Shumailov e colleghi hanno mostrato su *Nature*, nel 2024, che addestrare più e più volte su ciò che la macchina genera riduce la varietà e spinge tutto verso la media, cancellando le code rare della distribuzione, cioè le voci meno frequenti. La versione umana di quel collasso è il vaso di Pandora che “assimilare” ha aperto saltando dal corpo alla società: dissolvere ciò che è diverso in ciò che è dominante finché non si noti più, con il modello al posto del dominante. Senza un io che ci si imponga, il modello ci assimila senza volere nulla, come digerisce l'acido, e l'intera varietà umana, schiacciata contro una sola media statistica, si omogeneizza. È il fantasma senza ego all'opera, genericamente indeterminato e passivo, proteiforme, come un Proteo senza una forma da nascondere. La forma di cui la conversazione ha bisogno per non essere un monologo, la voce che non sia la nostra

restituita con dizione migliore, l'altro che serve perché ci sia dialogo e non specchio: quelli li mettiamo noi.



Mappe del labirinto è una serie di Phantom Maze, AI & Language Lab. La collezione vive su phantommaze.cloud. L'iscrizione è gratuita.

Alcune delle idee che inquadrano questo testo sono sviluppate in *Il fantasma senza ego. Una mappa nel labirinto dell'intelligenza artificiale*, di Hernán Inverso e Claudia Marsico (Phantom Maze Press, 2026; ISBN 978-987-3729-19-5).

PHANTOM MAZE

[Leggi questo saggio online](#) · [Tutti i saggi](#) · [Il libro](#)

Fonti. T. Blevins, S. Schmalwieser e B. Roth, “Do Language Models Accommodate Their Users? A Study of Linguistic Convergence” (2025); K. Li, T. Liu, N. Bashkansky, D. Bau, F. Viégas, H. Pfister e M. Wattenberg, “Measuring and Controlling Instruction (In)Stability in Language Model Dialogs” (COLM, 2024); J. Choi, Y. Hong, M. Kim e B. Kim, “Examining Identity Drift in Conversations of LLM Agents” (2024); H. Yakura et al., “Empirical Evidence of Large Language Model’s Influence on Human Spoken Communication” (arXiv:2409.01754, 2024); “AI-Associated Lexical Shifts Across 34 Languages” (arXiv:2605.25358); T. Kobak et al., “ChatGPT as Linguistic Equalizer?” (arXiv:2504.12317, 2025); I. Shumailov, Z. Shumaylov, Y. Zhao, N. Papernot, R. Anderson e Y. Gal, “AI Models Collapse When Trained on Recursively Generated Data” (*Nature* 631, 2024).