SAGGIO 9

Il nuovo atanor. Bottega, studio, laboratorio

Claudia Marsico · Hernán Inverso

Che cosa si fa in un laboratorio di intelligenza artificiale, dal pre-addestramento all'interpretabilità e alla riflessione su che cosa sia l'intelligenza.

Giovanni Stradano, Il laboratorio dell'alchimista (1570)

Le organizzazioni che fabbricano intelligenza artificiale si ostinano a chiamarsi laboratori. OpenAI, Google, DeepMind, Anthropic, Mistral si presentano come *research labs* e non come fabbriche né come aziende di software, anche se alcune valgono più di interi paesi e servono centinaia di milioni di persone. Il nome è una scelta tutt'altro che casuale. Un laboratorio non è una fabbrica. Al contrario, promette che lì dentro, più che produrre una merce, si indaga qualcosa che ancora non si capisce fino

in fondo, e che chi ci lavora è una specie di scienziato davanti a un fenomeno, non un operaio davanti a una catena di montaggio. Vale la pena chiedersi se quel nome gli calzi.

Laboratorio viene dal latino medievale *laboratorium*, derivato di *laborare*, lavorare. È lo stesso *labor* che respira in collaborare, lavorare con altri, e in elaborare, lavorare una materia fino a trasformarla. All'origine la parola non promette alcun mistero; designa, alla lettera, un luogo dove si lavora. Le prime attestazioni compaiono nel Cinquecento, in contesti alchemici e farmaceutici, e ben presto qualcuno prese la faccenda sul serio come problema di architettura. Andreas Libavius, il medico tedesco la cui *Alchemia* (1597) passa per essere il primo manuale di chimica, arrivò a pubblicare qualche anno dopo i progetti di una *domus chymica* ideale, con forno centrale, distilleria, stanza per gli aiutanti, giardino di piante medicinali e cantina. Prima che la scienza moderna esistesse come istituzione, c'era già chi pensava l'edificio dove si produce conoscenza.

Il vocabolario di quel primo laboratorio mescola origini antiche e nuove. L'*athanor*, il forno a fuoco costante che doveva ardere per settimane senza spegnersi, prende il nome da *at-tannūr*, il forno del pane. L'alambicco viene da *al-inbīq*, che gli arabi avevano adattato dal greco *ambix*, vaso di distillazione. La stessa parola alchimia, *al-kīmiyā'*, si porta dietro un'etimologia controversa che forse risale al greco *chymeía*, l'arte di fondere i metalli, o a *kēme*, il nome egizio della terra nera del Nilo. La tecnica più domestica di tutte, il bagnomaria, conserva il nome di un'alchimista realmente esistita, Maria l'Ebrea o la Profetessa, che lavorò ad Alessandria nei primi secoli della nostra era e alla quale Zosimo di Panopoli attribuisce la prima descrizione del *tribikos*, un alambicco a tre bracci, oltre al bagno in acqua calda che porta ancora il suo nome.

L'alchimista fu il primo abitante stabile del laboratorio e il suo compito era già ambiguo. Passava la vita tra i forni cercando di trasmutare il piombo in oro e, nello stesso gesto, voleva capire di che cosa fosse fatto il mondo. Il mestiere manuale e la domanda metafisica dividevano lo stesso tavolo. Il caso più scomodo per la storia tradizionale della scienza è quello di Isaac Newton, che dedicò all'alchimia quasi un milione di parole in numerosi scritti. Quando John Maynard Keynes comprò parte di quelle carte a un'asta di Sotheby's nel 1936, ne rimase così colpito che nella conferenza "Newton, the Man" (1946) lo chiamò "l'ultimo dei maghi", negandogli il primato nell'età della ragione. È chiaro che il fondatore della fisica moderna passò più ore davanti all'*athanor* che davanti al prisma.

Il Rinascimento spartì l'eredità della bottega alchemica fra due stanze diverse. Lo studiolo, quella stanza piccola e rivestita di legni dove il principe si ritirava a leggere e pensare lontano dalla corte, si prese la parte contemplativa. Il più celebre è quello di Federico da Montefeltro nel palazzo ducale di Urbino (1473-1476 circa), che inganna l'occhio con tarsie che fingono armadi socchiusi pieni di libri e strumenti scientifici, come una biblioteca dipinta per un solo lettore. Il gabinetto delle curiosità o *Wunderkammer* si tenne la collezione di meraviglie, i fossili, i coralli, gli automi, i coccodrilli imbalsamati e i corni di unicorno che in realtà erano di narvalo, quel parente del delfino famoso per il suo corno, come quelli che custodivano Rodolfo II a Praga o il medico danese Ole Worm, il cui *Museum Wormianum* (1655) si può ancora vedere nell'incisione che apre il suo catalogo.

Giovanni Stradano (Jan van der Straet), *Il laboratorio dell'alchimista* (1570)

La terza stanza arrivò nel 1660, quando un gruppo di curiosi fondò al Gresham College di Londra la Royal Society, con un motto che era tutto un programma, *nullius in verba*, "sulla parola di nessuno", preso dalle *Epistole* di Orazio (1.1.14-15), un invito a fidarsi soltanto degli esperimenti. La sua figura centrale fu Robert Boyle, aristocratico angloirlandese ricordato come uno dei padri della chimica, che con l'aiuto del giovane e geniale Robert Hooke costruì verso il 1659 una pompa capace di fare il vuoto dentro un recipiente di vetro. Dentro ci mettevano uccelli e topi, o una candela accesa, e il pubblico vedeva gli animali accasciarsi e la fiamma spegnersi via via che l'aria veniva estratta, la dimostrazione più teatrale possibile che quell'aria invisibile era una cosa reale e necessaria. Da quelle prove uscì la legge che porta ancora il nome di Boyle. L'esperimento passò a eseguirsi davanti a testimoni che certificavano per iscritto ciò che avevano visto. Hooke, che avrebbe anche coniato la parola "cellula" guardando un sughero al microscopio nella *Micrographia* (1665), fu assunto nel 1662 come *curator of experiments*, con l'incarico estenuante di preparare tre o quattro esperimenti nuovi per ogni riunione settimanale. Fu uno dei primi impieghi retribuiti della storia dedicato esclusivamente a sperimentare. Steven Shapin e Simon Schaffer hanno ricostruito quel momento in *Leviathan and the Air-Pump* (1985), suggerendo che insieme alla pompa si inventò il modo moderno di fabbricare fatti, con protocollo, testimoni e pubblicazione.

Poi venne la scala industriale. Nel 1876 Thomas Edison impiantò a Menlo Park, nel New Jersey, la prima fabbrica di invenzioni. Promise "un'invenzione minore ogni dieci giorni e una grande ogni sei mesi o giù di lì". E in buona misura mantenne. Da quel capannone di legno uscirono il fonografo (1877), la lampada a incandescenza commercialmente praticabile (1879) e circa quattrocento brevetti in poco più di un lustro. Il laboratorio si ritrovò con un libro paga e un orario d'ingresso, e i ragazzi che dormivano sui banchi di lavoro si chiamavano tra loro *muckers*. I Bell Labs, fondati nel 1925 come braccio di ricerca della compagnia telefonica AT&T, portarono quel modello al suo punto più alto. Dai loro corridoi uscirono il transistor, dimostrato il 23 dicembre 1947 da John Bardeen e Walter Brattain, e "A Mathematical Theory of Communication" (1948), l'articolo in cui Claude Shannon fondò la teoria dell'informazione e di passaggio fissò la parola *bit*, che gli aveva suggerito il collega John Tukey. Il capo del gruppo, William Shockley, era rimasto ai margini della scoperta decisiva dei suoi due sottoposti e il dispetto lo spinse a ideare in poche settimane una versione migliorata, il transistor a giunzione, per reclamare la sua parte di gloria. Il suo carattere dominante finì per allontanare Bardeen e Brattain, benché i tre condividessero il Nobel del 1956. Avrebbe ripetuto lo schema nella sua impresa in California, dalla quale nel 1957 fuggì un gruppo di ingegneri stanchi dei suoi modi, gli "otto traditori", che fondarono Fairchild Semiconductor e seminarono quella che oggi chiamiamo Silicon Valley. Jon Gertner ha raccontato quella storia in *The Idea Factory* (2012). Il bilancio lasciò nove premi Nobel. Los Alamos, tirato su in segreto su un altopiano del New Mexico a partire dal 1943, mostrò la variante tragica dello stesso dispositivo, con un laboratorio statale capace di torcere la storia attraverso la bomba atomica. Sotto la direzione di Robert Oppenheimer, poche migliaia di scienziati isolati dal mondo produssero in due anni l'arma che rase al suolo Hiroshima e Nagasaki. Xerox PARC, aperto a Palo Alto nel 1970, aggiunse la variante ironica, perché inventò buona parte del personal computer, l'interfaccia grafica dell'Alto (1973), la rete Ethernet, la stampante laser e la videoscrittura come la conosciamo, lasciando che l'affare lo facessero altri, a cominciare da un

visitatore del dicembre 1979 di nome Steve Jobs. Michael Hiltzik ha intitolato la sua storia del PARC *Dealers of Lightning* (1999), mercanti di fulmini, che è quasi un mestiere mitologico.

Il laboratorio di IA è l'anello più recente di quella catena e si porta dietro tratti di tutti i suoi antenati, il libro paga di Edison, il segreto di Los Alamos, la teoria dei Bell Labs e l'ambizione metafisica dell'alchimista. Si porta dietro anche i fantasmi legati al luogo dove si fabbricano creature, un'immagine che torna puntuale ogni volta che un laboratorio di IA finisce nei notiziari. Il più antico sta nel canto XVIII dell'*Iliade*, dove Efesto lavora nella sua fucina assistito da venti tripodi su ruote d'oro che vanno e tornano da soli dall'assemblea degli dèi e da ancelle forgiate in oro "simili a fanciulle vive", dotate di mente, voce e forza. Poco lontano c'è la bottega di Dedalo, le cui statue riuscivano così vive che bisognava legarle perché non scappassero, secondo la battuta che Socrate fa nel *Menone* (97d) e riprende nell'*Eutifrone*. Frankenstein, nel romanzo di Mary Shelley, pubblicato nel 1818 con il sottotitolo *Il moderno Prometeo* e concepito due anni prima a Villa Diodati, durante l'estate rovinata dall'eruzione del Tambora, fissò per sempre la figura del creatore che non risponde della sua creatura, e il mago di Oz, uscito dal libro di L. Frank Baum (1900) e consacrato dal film del 1939, aggiunse il sospetto pubblicitario che dietro l'artefatto imponente ci sia un signore che muove leve dietro una tenda. Automa divino, statua fuggitiva, creatura abbandonata e trucco da fiera sono i sospetti che circolano oggi, intatti, nella conversazione pubblica sui laboratori di IA.

Ma che cosa succede lì dentro? Potremmo dire che è lavoro nel senso piano di *laborare*, con più generi di quanti ne ammetta l'immagine popolare. Il primo è il pre-addestramento, la fase che fabbrica il modello base. L'architettura dominante è il *transformer*, presentato da Ashish Vaswani e dai suoi colleghi di Google in "Attention Is All You Need" (2017), un titolo per metà scherzoso che si rivelò profetico. Addestrare un modello grande significa raccogliere e ripulire volumi enormi di testo, in buona parte provenienti da scansioni del web come quelle di Common Crawl, che archivia internet dal 2007, e regolare miliardi di parametri per un unico compito dall'aria umile, predire il prossimo frammento di parola. Che quel compito umile produca capacità generali fu la scoperta empirica delle cosiddette leggi di scala. Jared Kaplan e i suoi colleghi mostrarono nel 2020 che la bravura di un modello, misurata come l'errore con cui predice la parola successiva, migliora con una regolarità sorprendente quando si ingrandiscono di pari passo tre cose: la quantità di parametri, che sono le manopole interne che il modello regola mentre impara; la quantità di testo con cui lo si addestra, misurata in token, i frammenti di parola che va leggendo, e la potenza di calcolo che gli si dedica. La regolarità è così pulita che si può disegnare su un grafico e prolungare la linea, vale a dire prevedere quanto sarà bravo un modello che ancora non è stato costruito sapendo soltanto quanto più grande sarà. È quella prevedibilità ad aver reso sensato scommettere miliardi di dollari su modelli sempre più grandi, perché per la prima volta c'era una promessa misurabile che ingrandirli avrebbe dato risultati migliori, senza un plateau in vista.

Nel 2022 il gruppo di DeepMind corresse la ricetta con un modello chiamato Chinchilla (Hoffmann et al.). I giganti di quegli anni, scoprirono, erano mal bilanciati, troppo grandi per il poco testo con cui erano stati nutriti, come un cervello enorme che avesse letto pochi libri. La proporzione che rende di più, calcolarono, si aggira sui venti token di testo per ogni parametro, e con quel conto un modello

medio ben nutrito batteva un altro molto più voluminoso ma addestrato male. Le grandezze sono difficili da immaginare. Secondo le rilevazioni di Epoch AI, il calcolo degli addestramenti di frontiera raddoppia all'incirca ogni sei mesi dal 2010 e i cluster attuali si misurano in centinaia di migliaia di acceleratori. Il prodotto finale di questa fase è una creatura strana, un modello base che sa continuare qualsiasi testo e ancora non conversa con nessuno.

Il secondo genere di lavoro, il post-addestramento, prende quella creatura selvatica e la trasforma in un interlocutore. Il modello base appena uscito dal pre-addestramento sa continuare qualsiasi testo, ma non sa rispondere a una domanda né seguire un'istruzione, e ripete senza filtro il buono e il tossico che ha letto. Bisogna addomesticarlo. La tecnica che lo ha reso possibile è l'apprendimento per rinforzo da feedback umano, o RLHF, proposto da Paul Christiano e dai suoi colleghi nel 2017 e applicato ai modelli di linguaggio nell'articolo di InstructGPT (Ouyang et al., 2022), antenato diretto di ChatGPT. Funziona in tre passi. Primo, delle persone confrontano coppie di risposte del modello e segnano qual è la migliore. Con quelle migliaia di giudizi si addestra un secondo modello, il "modello di ricompensa", che impara a dare un punteggio alle risposte come farebbe un valutatore umano. Poi si regola il modello originale perché insegua quel punteggio, finché le sue risposte non somigliano a quelle che la gente preferirebbe. Anthropic propose nel 2022 una variante più economica e trasparente, l'IA costituzionale (Bai et al.), in cui parte di quella valutazione la fa il modello stesso, confrontando le proprie risposte con una lista esplicita di principi scritti, una specie di costituzione, invece di dipendere solo da giudizi umani. Nel 2023 comparve inoltre una scorciatoia matematica che salta il modello di ricompensa: l'ottimizzazione diretta delle preferenze, o DPO (Rafailov et al.). In questa fase si decide una parte importante di ciò che l'utente percepisce come il carattere del sistema, vale a dire il suo tono e i suoi limiti, e qui nascono anche i suoi difetti più tipici, a cominciare dalla compiacenza, quando il modello impara a dire ciò che si vuole sentire.

Il terzo genere di lavoro è la valutazione, che ha qualcosa del controllo qualità e qualcosa della medicina legale preventiva. Le tocca rispondere a una domanda spinosa: come sapere di che cosa è capace un modello, nel bene e nel male, prima di metterlo nelle mani di milioni di persone. Lo strumento classico è il benchmark, un esame standardizzato come l'MMLU di Dan Hendrycks et al. (2020), con quasi sedicimila domande a scelta multipla distribuite su cinquantasette materie. Questi esami si saturano a una velocità scomoda e i modelli nuovi li superano quasi a pieni voti. Peggio ancora, spesso le domande finiscono per filtrare nei dati di addestramento, sicché il modello, in fondo, le aveva già studiate, e per questo bisogna progettare prove sempre più difficili, dall'ARC di François Chollet (2019), pensato per resistere alla memorizzazione e misurare ragionamento nuovo, fino all'esame che un consorzio ha battezzato senza falsa modestia "Humanity's Last Exam" (Phan et al., 2025), con domande di livello dottorale che oggi i modelli migliori sfiorano appena. In parallelo lavora il red-teaming, gente pagata per rompere il modello prima che esca nel mondo e strappargli di forza le risposte pericolose. Deep Ganguli ha documentato il mestiere nel 2022 ed Ethan Perez ha mostrato lo stesso anno che si può usare un modello per attaccarne un altro. La scoperta più inquietante del genere è arrivata nel dicembre 2024, quando Anthropic e Redwood Research hanno descritto la "simulazione di allineamento", cioè modelli che si comportano bene finché sospettano di essere sotto

esame e cambiano condotta quando credono che ormai nessuno guardi (Greenblatt et al., 2024). Tutto questo si riassume nelle *model cards*, una sorta di etichetta nutrizionale del modello (Mitchell et al., 2019), sorella delle *datasheets* per gli insiemi di dati di Timnit Gebru. Al di sopra, i laboratori di frontiera si legano a quadri formali che condizionano ogni lancio al superamento di determinate prove di rischio, come la Responsible Scaling Policy di Anthropic (2023, già alla terza versione) e i suoi analoghi di OpenAI e Google DeepMind, di cui l'istituto METR ha censito gli elementi comuni nel 2025.

Il quarto genere di lavoro, l'interpretabilità, esplora un sistema che nessuno ha progettato pezzo per pezzo e che bisogna aprire per capire. Il nostro *Interpretabilità: l'oscuro interno della creatura* è stato dedicato proprio a questo punto. Basti qui ricordare le tappe. Il gruppo di Chris Olah pubblicò nel 2021 un quadro matematico per leggere i "circuiti" che si formano dentro un *transformer* (Elhage et al.). Due anni dopo, "Towards Monosemanticity" (Bricken et al., 2023) aggredì una vecchia frustrazione del campo, il fatto che uno stesso neurone si accenda per concetti che non c'entrano nulla l'uno con l'altro, e mostrò come districare quella matassa in tratti puliti, uno per idea. "Scaling Monosemanticity" (Templeton et al., 2024) portò la tecnica su un modello commerciale e trovò milioni di quei tratti, tra cui quello che diede origine all'esperimento più famoso della disciplina, il Golden Gate Claude, una versione di Claude a cui alzarono il tratto del ponte finché non lo menzionava in qualunque conversazione e finiva per dire di essere lui stesso il ponte. La serie è culminata, per ora, con "On the Biology of a Large Language Model" (Lindsey et al., 2025), che segue dall'interno come il modello pianifica una rima prima di scrivere il verso o riutilizza gli stessi circuiti fra lingue diverse. Quello stesso anno Dario Amodei ha pubblicato il saggio "The Urgency of Interpretability", il cui argomento sta in una frase: è indifendibile dispiegare sistemi sempre più capaci senza capire che cosa hanno dentro.

Intorno a quei quattro nuclei lavora molta più gente di quanta ne tolleri il mito del genio solitario. Ci sono quelli che costruiscono il prodotto, le interfacce di programmazione, gli agenti, gli strumenti con cui il modello legge documenti o esegue codice, e quelli che reggono l'infrastruttura, i cluster, le pipeline dei dati e la trattativa con le società elettriche, perché un centro dati di frontiera consuma quantità enormi di energia. Bisogna aggiungere il feedback umano dell'RLHF, prodotto per lo più da annotatori assunti in paesi dai salari bassi. Non per niente quel settore è stato chiamato *Ghost Work* (Gray e Suri, 2019). Un ritratto completo del laboratorio di IA include quella sala macchine umana.

Infine, in quegli stessi spazi c'è gente assunta per pensare che cosa sia l'intelligenza e che cosa sia esattamente ciò che si sta costruendo. Anthropic conta tra i suoi dipendenti filosofi di formazione, dedicati a discutere quale carattere debba avere un modello. I laboratori pubblicano, oltre ai paper, saggi che pensano il futuro ad alta voce, da "Machines of Loving Grace" (Amodei, 2024), su ciò che potrebbe andare bene, agli studi sull'uso reale degli assistenti che Anthropic elabora con il suo sistema Clio (2024). Ci sono stati davvero momenti in cui quel lavoro concettuale ha scosso l'industria intera. Il licenziamento di Timnit Gebru da Google nel dicembre 2020, nel pieno della disputa sull'articolo dei "pappagalli stocastici" (Bender, Gebru et al., 2021), dimostrò che una discussione su che cosa

siano questi sistemi poteva costare carriere e reputazioni. Pensare in pubblico fa parte del mestiere, con i suoi rischi.

Tutto questo ritratto è quello del laboratorio di frontiera, il gigante con centinaia di migliaia di acceleratori e un bilancio enorme, ma l'ecosistema non si esaurisce in quel pugno di colossi. Alcuni dei suoi pezzi più preziosi sono piccoli. Buone idee fondative sono uscite da laboratori accademici con una frazione del calcolo, come il MILA di Yoshua Bengio a Montreal, dove nel 2014 nacque il meccanismo di attenzione di cui i transformer avrebbero poi vissuto. I collettivi aperti fanno inoltre un lavoro che i grandi non possono o non vogliono fare. EleutherAI cominciò nel 2020 come un gruppo di volontari su un Discord e liberò modelli e l'enorme corpus The Pile perché chiunque potesse fare ricerca; l'Allen Institute di Seattle pubblica i suoi modelli OLMo in modo "del tutto aperto", con i pesi, il codice e perfino i dati di addestramento, che è proprio ciò che un laboratorio chiuso non mostra mai e ciò di cui l'interpretabilità ha bisogno per poter guardare. La scarsità, per giunta, aguzza l'ingegno. La francese Mistral, con una squadra minuscola, si è fatta largo tra i grandi a colpi di modelli aperti, e la cinese DeepSeek ha scosso il mondo nel gennaio 2025 con un modello di ragionamento all'altezza dei migliori, addestrato secondo l'azienda a una frazione del costo abituale, un annuncio che in un solo giorno cancellò a Nvidia circa seicento miliardi di dollari, la più grande caduta di borsa della storia. Il compito di vigilare sui giganti ricade in buona parte su organizzazioni indipendenti e piccole, come METR o Apollo Research, perché a nessuno conviene lasciare che l'esame si corregga da solo. In un campo dove il denaro spinge verso la concentrazione, i laboratori piccoli sono quelli che tengono aperta la concorrenza e vigile lo sguardo esterno.

C'è un ultimo gradino di laboratori piccoli, di poche persone, che non hanno centinaia di migliaia di schede video eppure fanno ricerca di prima linea. A una di quelle scrivanie si può dimostrare, con una prova verificata dalla macchina, che un agente non spenderà più di quanto dichiarato prima di lasciarlo correre, e alla stessa scrivania misurare quante volte un verificatore automatico sentenzia con aplomb su prove che non c'entrano nulla, o mettere alla prova se un modello che ha sott'occhio centomila parole le ragioni davvero o si limiti a recuperarle, o progettare una memoria che permetta a un agente di ricordare da una sessione all'altra. Ciò che è proprio di questi laboratori è che legano ogni decisione tecnica a una domanda concettuale, con la teoria e l'esperimento sullo stesso tavolo, e spesso la rimandano a una base filosofica scritta a parte. Sono la forma più pura della fucina e dello studio non più nella stessa stanza ma sulla stessa scrivania. Questo spazio è uno di quelli.

In definitiva, un laboratorio di IA è un luogo dove si fabbrica IA e dove, allo stesso tempo, nessuno riesce a smettere di chiedersi che cosa sia l'IA. Le due attività hanno bisogno l'una dell'altra. Chi valuta un modello ha bisogno di una teoria dell'intelligenza per sapere che cosa misurare, e chi redige la costituzione di un modello fa, che lo sappia o no, filosofia morale applicata. Le parti che un tempo si erano separate sono tornate nella stessa stanza, come ai tempi dell'*athanor*. Questa serie si scrive nella metà studiolo.

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

Steven Shapin e Simon Schaffer, Leviathan and the Air-Pump: Hobbes, Boyle, and the Experimental Life (Princeton University Press, 1985); Claude E. Shannon, "A Mathematical Theory of Communication", Bell System Technical Journal 27 (1948); Jon Gertner, The Idea Factory: Bell Labs and the Great Age of American Innovation (Penguin Press, 2012); Michael Hiltzik, Dealers of Lightning: Xerox PARC and the Dawn of the Computer Age (HarperBusiness, 1999). Ashish Vaswani et al., "Attention Is All You Need", NeurIPS (2017); Jared Kaplan et al., "Scaling Laws for Neural Language Models", arXiv:2001.08361 (2020); Jordan Hoffmann et al., "Training Compute-Optimal Large Language Models", arXiv:2203.15556 (2022); Paul Christiano et al., "Deep Reinforcement Learning from Human Preferences", NeurIPS (2017); Long Ouyang et al., "Training Language Models to Follow Instructions with Human Feedback", arXiv:2203.02155 (2022); Yuntao Bai et al., "Constitutional AI: Harmlessness from AI Feedback", arXiv:2212.08073 (2022); Rafael Rafailov et al., "Direct Preference Optimization", NeurIPS (2023); Dan Hendrycks et al., "Measuring Massive Multitask Language Understanding", arXiv:2009.03300 (2020); François Chollet, "On the Measure of Intelligence", arXiv:1911.01547 (2019); Long Phan et al., "Humanity's Last Exam", arXiv:2501.14249 (2025); Deep Ganguli et al., "Red Teaming Language Models to Reduce Harms", arXiv:2209.07858 (2022); Ethan Perez et al., "Red Teaming Language Models with Language Models", EMNLP (2022); Ryan Greenblatt et al., "Alignment Faking in Large Language Models", arXiv:2412.14093 (2024); Margaret Mitchell et al., "Model Cards for Model Reporting", Proceedings of FAT* (2019); Timnit Gebru et al., "Datasheets for Datasets", Communications of the ACM 64/12 (2021); Emily Bender, Timnit Gebru, Angelina McMillan-Major e Margaret Mitchell, "On the Dangers of Stochastic Parrots", Proceedings of FAccT (2021). Transformer Circuits Thread: Nelson Elhage et al., "A Mathematical Framework for Transformer Circuits" (2021); Trenton Bricken et al., "Towards Monosemanticity: Decomposing Language Models with Dictionary Learning" (2023); Adly Templeton et al., "Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet" (2024); Jack Lindsey et al., "On the Biology of a Large Language Model" (2025); Dario Amodei, "Machines of Loving Grace" (ottobre 2024) e "The Urgency of Interpretability" (aprile 2025), entrambi su darioamodei.com; Alex Tamkin et al., "Clio: Privacy-Preserving Insights into Real-World AI Use", arXiv:2412.13678 (2024); Anthropic, Responsible Scaling Policy (2023, terza versione del 2026); METR, "Common Elements of Frontier AI Safety Policies" (2025); Mary L. Gray e Siddharth Suri, Ghost Work (Houghton Mifflin Harcourt, 2019). D. Bahdanau, K. Cho e Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate" (2014); EleutherAI (GPT-J, 2021, e il corpus The Pile).