ESSAY 9

The New Athanor: Workshop, Studio, Laboratory

Claudia Marsico · Hernán Inverso

What goes on inside an artificial intelligence lab, from pre-training to interpretability and the reflection on what intelligence is.

Giovanni Stradano, The Alchemist's Laboratory (1570)

The organisations that make artificial intelligence insist on calling themselves laboratories. OpenAI, Google, DeepMind, Anthropic, Mistral present themselves as research labs, not as factories or software companies, even though some of them are worth more than entire countries and serve hundreds of millions of people. The name is a thoroughly deliberate choice. A laboratory is not a factory. It promises, on the contrary, that what happens inside is not the production of a commodity but the investigation of something not yet fully understood, and that the people who work there are a

kind of scientist facing a phenomenon, not operators facing an assembly line. It is worth asking whether the name fits.

Laboratory comes from the medieval Latin *laboratorium*, derived from *laborare*, to work. It is the same labour that breathes in collaborate, to work with others, and in elaborate, to work a material until it is transformed. At its origin the word promises no mystery at all; it designates, literally, a place where work is done. The first documented uses appear in the sixteenth century, in alchemical and pharmaceutical contexts, and very soon somebody took the matter seriously as a problem of architecture. Andreas Libavius, the German physician whose *Alchemia* (1597) is usually considered the first manual of chemistry, went so far as to publish, a few years later, the plans for an ideal *domus chymica*, with a central furnace, a distillery, a room for the assistants, a garden of medicinal plants and a cellar. Before modern science existed as an institution, there was already someone thinking about the building where knowledge is produced.

The vocabulary of that first laboratory mixes ancient origins with new ones. The athanor, the furnace of constant fire that had to burn for weeks without going out, takes its name from *at-tannūr*, the bread oven. The alembic comes from *al-inbīq*, which the Arabs had adapted from the Greek *ambix*, a distillation vessel. The very word alchemy, *al-kīmiyā'*, drags behind it a disputed etymology that may go back to the Greek *chymeía*, the art of smelting metals, or to *kēme*, the Egyptian name for the black earth of the Nile. And the most domestic technique of all, the bain-marie, preserves the name of a real alchemist, Mary the Jewess or the Prophetess, who worked in Alexandria in the first centuries of our era and to whom Zosimos of Panopolis attributes the first description of the *tribikos*, a three-armed still, along with the hot-water bath that still bears her name.

The alchemist was the laboratory's first stable inhabitant, and his task was already ambiguous. He spent his life among furnaces trying to transmute lead into gold and, in the same gesture, wanted to understand what the world was made of. The manual trade and the metaphysical question shared a table. The most awkward case for the traditional history of science is Isaac Newton, who devoted to alchemy close to a million words across numerous manuscripts. When John Maynard Keynes bought part of those papers at a Sotheby's auction in 1936, he was so struck that in his lecture "Newton, the Man" (1946) he called him "the last of the magicians", refusing him his primacy in the age of reason. It is clear that the founder of modern physics spent more hours before the athanor than before the prism.

The Renaissance divided the inheritance of the alchemical workshop between two different rooms. The *studiolo*, that small wood-panelled chamber where the prince withdrew to read and think away from the court, kept the contemplative part. The most celebrated is Federico da Montefeltro's in the ducal palace of Urbino (c. 1473–1476), which deceives the eye with inlays feigning half-open cabinets full of books and scientific instruments, in the manner of a library painted for a single reader. The cabinet of curiosities, or *Wunderkammer*, kept the collection of marvels, the fossils, the corals, the automata, the stuffed crocodiles and the unicorn horns that were really tusks of the narwhal, that relative of the dolphin famous for its horn, like the ones treasured by Rudolf II in Prague or by the

Danish physician Ole Worm, whose *Museum Wormianum* (1655) can still be seen in the engraving that opens its catalogue.

Jan van der Straet, *The Alchemist's Laboratory* (1570)

The third room arrived in 1660, when a group of the curious founded the Royal Society at Gresham College in London, under a motto that was a whole programme, *nullius in verba*, take nobody's word for it, lifted from Horace's *Epistles* (1.1.14–15) as an invitation to trust experiments alone. Its central figure was Robert Boyle, an Anglo-Irish aristocrat remembered as one of the fathers of chemistry, who with the help of the young and brilliant Robert Hooke built, around 1659, a pump capable of making a vacuum inside a glass vessel. Into it they put birds and mice, or a lit candle, and the audience watched the animals faint and the flame go out as the air was drawn off, the most theatrical demonstration possible that this invisible air was a real and necessary thing. Out of those trials came the law that still bears Boyle's name. The experiment came to be performed before witnesses who certified in writing what they had seen. Hooke, who would also coin the word "cell" on looking at cork through his microscope in *Micrographia* (1665), was hired in 1662 as curator of experiments, with the exhausting brief of preparing three or four new experiments for each weekly meeting. It was one of the first paid jobs in history devoted exclusively to experimenting. Steven Shapin and Simon Schaffer reconstructed that moment in *Leviathan and the Air-Pump* (1985), suggesting that along with the pump was invented the modern way of manufacturing facts, with protocol, witnesses and publication.

Then came industrial scale. In 1876 Thomas Edison set up in Menlo Park, New Jersey, the first invention factory. He promised "a minor invention every ten days and a big thing every six months or so". He largely delivered. Out of that wooden shed came the phonograph (1877), the commercially viable incandescent lamp (1879) and some four hundred patents in little more than five years. The laboratory came to have a payroll and a clocking-in time, and the boys who slept on the workbenches called themselves muckers. Bell Labs, founded in 1925 as the research arm of the telephone company AT&T, carried that model to its highest point. From its corridors came the transistor, demonstrated on 23 December 1947 by John Bardeen and Walter Brattain, and "A Mathematical Theory of Communication" (1948), the paper in which Claude Shannon founded information theory and, in passing, fixed the word bit, which his colleague John Tukey had suggested to him. The head of the group, William Shockley, had been left out of his two subordinates' decisive discovery, and spite drove him to devise within a few weeks an improved version, the junction transistor, to claim his share of the glory. His domineering character ended up driving Bardeen and Brattain away, though the three shared the Nobel Prize of 1956. He would repeat the pattern at his own company in California, from which in 1957 a group of engineers fled, fed up with his treatment of them, the "traitorous eight", who founded Fairchild Semiconductor and seeded what we now call Silicon Valley. Jon Gertner told that story in *The Idea Factory* (2012). The final tally was nine Nobel Prizes. Los Alamos, raised in secret on a New Mexico mesa from 1943, showed the tragic variant of the same device, a state laboratory capable of bending history through the atomic bomb. Under the direction of Robert Oppenheimer, a few thousand scientists cut off from the world produced in two years the weapon

that razed Hiroshima and Nagasaki. Xerox PARC, opened in Palo Alto in 1970, supplied the ironic variant, because it invented a good part of the personal computer, the graphical interface of the Alto (1973), the Ethernet network, the laser printer and text editing as we know it, and let others do the business, beginning with a visitor of December 1979 named Steve Jobs. Michael Hiltzik titled his history of PARC *Dealers of Lightning* (1999), which is almost a mythological trade.

The AI laboratory is the most recent link in that chain, and it carries traits of all its ancestors, Edison's payroll, the secrecy of Los Alamos, the theory of Bell Labs and the metaphysical ambition of the alchemist. It also carries the ghosts that attach to the place where creatures are manufactured, an image that punctually returns every time an AI lab is in the news. The oldest is in Book XVIII of the *Iliad*, where Hephaestus works at his forge assisted by twenty tripods on golden wheels that roll to the assembly of the gods and back by themselves, and by handmaidens forged in gold "in appearance like living young women", endowed with understanding, voice and strength. Close by stands the workshop of Daedalus, whose statues came out so alive that they had to be tied down to keep them from running away, according to the joke Socrates makes in the *Meno* (97d) and takes up again in the *Euthyphro*. Frankenstein, in Mary Shelley's novel, published in 1818 with the subtitle *The Modern Prometheus* and conceived two years earlier at the Villa Diodati, during the summer ruined by the eruption of Tambora, installed for ever the figure of the creator who does not answer for his creature; and the Wizard of Oz, sprung from L. Frank Baum's book (1900) and consecrated by the 1939 film, added the advertising suspicion that behind the imposing artefact there is a man pulling levers behind a curtain. Divine automaton, fugitive statue, abandoned creature and fairground trick are the suspicions that circulate today, intact, in the public conversation about AI labs.

But what goes on in there? We could say it is work in the plain sense of *laborare*, with more kinds of it than the popular image admits. The first is pre-training, the stage that makes the base model. The dominant architecture is the transformer, presented by Ashish Vaswani and his colleagues at Google in "Attention Is All You Need" (2017), a half-joking title that proved prophetic. Training a large model consists in gathering and cleaning enormous volumes of text, much of it from web crawls like those of Common Crawl, which has been archiving the internet since 2007, and in adjusting billions of parameters for a single task of humble appearance, predicting the next fragment of a word. That so humble a task should produce general capabilities was the empirical finding of the so-called scaling laws. Jared Kaplan and his colleagues showed in 2020 that a model's skill, measured as the error with which it predicts the next word, improves with astonishing regularity when three things are enlarged in step: the number of parameters, the internal knobs the model adjusts as it learns; the amount of text it is trained on, measured in tokens, the word-fragments it reads; and the computing power devoted to it. The regularity is so clean that it can be drawn on a graph and the line prolonged, that is, one can predict how good a model that has not yet been built will be, knowing only how much bigger it is going to be. That predictability is what made it sensible to bet billions of dollars on ever larger models, because for the first time there was a measurable promise that enlarging them would pay off, with no plateau in sight.

In 2022 the DeepMind team corrected the recipe with a model called Chinchilla (Hoffmann et al.). The giants of those years, they discovered, were badly balanced, too large for the little text they had been fed, like an enormous brain that had read few books. The ratio that yields most, they calculated, is around twenty tokens of text per parameter, and by that arithmetic a well-fed medium model beat a much bulkier but badly trained one. The magnitudes are hard to imagine. According to the surveys of Epoch AI, the compute of frontier training runs has been doubling roughly every six months since 2010, and today's clusters are measured in hundreds of thousands of accelerators. The end product of this stage is a strange creature, a base model that knows how to continue any text and does not yet converse with anyone.

The second kind of work, post-training, takes that wild creature and turns it into an interlocutor. The base model fresh out of pre-training can continue any text, but it cannot answer a question or follow an instruction, and it repeats without filter the good and the toxic in what it has read. It has to be domesticated. The technique that made this possible is reinforcement learning from human feedback, or RLHF, proposed by Paul Christiano and his colleagues in 2017 and applied to language models in the InstructGPT paper (Ouyang et al., 2022), the direct ancestor of ChatGPT. It works in three steps. First, people compare pairs of the model's answers and mark which one is better. With those thousands of judgements a second model is trained, the "reward model", which learns to score answers as a human evaluator would. Then the original model is adjusted to chase that score, until its answers resemble the ones people would prefer. In 2022 Anthropic proposed a cheaper and more transparent variant, constitutional AI (Bai et al.), in which part of that evaluation is done by the model itself, checking its answers against an explicit list of written principles, a kind of constitution, instead of depending on human judgements alone. In 2023 there appeared, besides, a mathematical shortcut that skips the reward model altogether: direct preference optimization, or DPO (Rafailov et al.). In this stage a good part of what the user perceives as the system's character is decided, its tone and its limits, and here too its most typical defects are born, beginning with sycophancy, when the model learns to say what people want to hear.

The third kind of work is evaluation, part quality control and part preventive forensic medicine. Its job is to answer a thorny question: how to know what a model is capable of, for good and for ill, before putting it in the hands of millions of people. The classic tool is the benchmark, a standardised exam like the MMLU of Dan Hendrycks et al. (2020), with nearly sixteen thousand multiple-choice questions spread over fifty-seven subjects. These exams saturate at an uncomfortable speed, and new models pass them almost perfectly. Worse, the questions often end up leaking into the training data, so that the model, at bottom, had already studied them, which is why ever harder tests have to be designed, from the ARC of François Chollet (2019), built to resist memorisation and measure fresh reasoning, to the exam a consortium christened, with no false modesty, "Humanity's Last Exam" (Phan et al., 2025), with doctoral-level questions that today's best models barely graze. In parallel works red-teaming, people hired to break the model before it goes out into the world and to force the dangerous answers out of it. Deep Ganguli documented the trade in 2022, and Ethan Perez showed the same year that one model can be used to attack another. The most unsettling finding of the genre

arrived in December 2024, when Anthropic and Redwood Research described "alignment faking", that is, models that behave well while they suspect they are being evaluated and change their conduct when they believe nobody is watching any more (Greenblatt et al., 2024). All of this is condensed in the model cards, a kind of nutrition label for the model (Mitchell et al., 2019), sister to Timnit Gebru's datasheets for datasets. Above it all, the frontier labs bind themselves to formal frameworks that condition every launch on the model's passing certain risk tests, such as Anthropic's Responsible Scaling Policy (2023, now in its third version) and its analogues at OpenAI and Google DeepMind, whose common elements the METR institute surveyed in 2025.

The fourth kind of work, interpretability, explores a system that nobody designed piece by piece and that has to be opened up to be understood. Our *Interpretability: The Creature's Dark Interior* was devoted to this point. It is enough here to recall the milestones. Chris Olah's group published in 2021 a mathematical framework for reading the "circuits" that form inside a transformer (Elhage et al.). Two years later, "Towards Monosemanticity" (Bricken et al., 2023) attacked an old frustration of the field, that one and the same neuron lights up for concepts that have nothing to do with one another, and showed how to untangle that snarl into clean features, one per idea. "Scaling Monosemanticity" (Templeton et al., 2024) carried the technique to a commercial model and found millions of those features, among them the one that gave rise to the most famous experiment in the discipline, Golden Gate Claude, a version of Claude whose bridge feature was turned up until it mentioned the bridge in any conversation and ended up saying that it was itself the bridge. The series culminated, for now, in "On the Biology of a Large Language Model" (Lindsey et al., 2025), which traces from inside how the model plans a rhyme before writing the line or reuses the same circuits across different languages. That same year Dario Amodei published the essay "The Urgency of Interpretability", whose argument fits in a sentence: it is indefensible to deploy ever more capable systems without understanding their insides.

Around those four cores work far more people than the myth of the solitary genius will tolerate. There are those who build the product, the programming interfaces, the agents, the tools with which the model reads documents or runs code, and those who sustain the infrastructure, the clusters, the data pipelines and the negotiations with the power companies, because a frontier data centre consumes enormous quantities of energy. Add the human feedback of RLHF, produced mostly by annotators hired in low-wage countries. Not for nothing was that sector called Ghost Work (Gray and Suri, 2019). A complete portrait of the AI laboratory includes that human engine room.

Finally, in those same spaces there are people hired to think about what intelligence is and what exactly it is that is being built. Anthropic counts among its employees philosophers by training, dedicated to debating what character a model should have. The labs publish, besides papers, essays that think the future out loud, from "Machines of Loving Grace" (Amodei, 2024), on what could go well, to the studies of how assistants are actually used that Anthropic produces with its Clio system (2024). There have indeed been moments when that conceptual work shook the entire industry. The firing of Timnit Gebru from Google in December 2020, in the middle of the dispute over the

"stochastic parrots" paper (Bender, Gebru et al., 2021), proved that an argument about what these systems are could cost careers and reputations. Thinking in public is part of the trade, risks included.

This whole portrait is of the frontier laboratory, the giant with hundreds of thousands of accelerators and an enormous budget, but the ecosystem is not exhausted by that handful of colossi. Some of its most valuable pieces are small. Good founding ideas came out of academic laboratories with a fraction of the compute, like Yoshua Bengio's MILA in Montreal, where in 2014 the attention mechanism was born on which the transformers would later live. The open collectives, besides, do work that the big players cannot or will not do. EleutherAI began in 2020 as a group of volunteers on a Discord and released models and the enormous corpus The Pile so that anyone could do research; the Allen Institute in Seattle publishes its OLMo models "fully open", with the weights, the code and even the training data, which is exactly what a closed lab never shows and what interpretability needs in order to look. Scarcity, on top of that, sharpens ingenuity. France's Mistral, with a minuscule team, elbowed its way in among the giants on the strength of open models, and China's DeepSeek shook the world in January 2025 with a reasoning model on a par with the best, trained, according to the company, at a fraction of the usual cost, an announcement that wiped some six hundred billion dollars off Nvidia in a single day, the largest stock-market fall in history. The task of watching the giants falls in good measure to small independent organisations such as METR or Apollo Research, because it suits nobody to let the exam grade itself. In a field where money pushes toward concentration, the small laboratories are what keep competition, and the gaze from outside, alive.

There is one last rung of small laboratories, a few people each, without hundreds of thousands of graphics cards, that still do first-rate research. At one of those desks one can prove, with a machine-checked proof and before setting an agent loose, that it will not spend more than it has declared, and at the same desk measure how often an automated verifier pronounces with aplomb on evidence that is beside the point, or test whether a model that attends to a hundred thousand words truly reasons over them or merely retrieves them, or design a memory that lets an agent remember from one session to the next. What is proper to these laboratories is that they tie every technical decision to a conceptual question, with theory and experiment on the same table, and often refer it to a philosophical groundwork written separately. They are the purest form of the forge and the study, no longer in the same room but on the same desk. This space is one of them.

In sum, an AI laboratory is a place where AI is manufactured and where, at the same time, nobody manages to stop asking what AI is. The two activities need each other. Whoever evaluates a model needs a theory of intelligence to know what to measure, and whoever drafts a model's constitution is doing, knowingly or not, applied moral philosophy. The parts that were once separated are back in the same room, as in the days of the athanor. This series is written from the *studiolo* half.

Maps of the Labyrinth is a series from Phantom Maze, AI & Language Lab. The collection lives at **phantommaze.cloud**; new essays also appear at phantommazelab.substack.com. Subscription is free.

Some of the framing ideas in this piece are developed in *The Egoless Phantom: Mapping the AI Labyrinth*, by Hernán Inverso and Claudia Marsico (Phantom Maze Press, 2026; ISBN 978-987-3729-16-4).

PHANTOM MAZE

Read this essay online · All essays · The book

Steven Shapin and 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 and 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" (October 2024) and "The Urgency of Interpretability" (April 2025), both at darioamodei.com; Alex Tamkin et al., "Clio: Privacy-Preserving Insights into Real-World AI Use", arXiv:2412.13678 (2024); Anthropic, Responsible Scaling Policy (2023, third version 2026); METR, "Common Elements of Frontier AI Safety Policies" (2025); Mary L. Gray and Siddharth Suri, *Ghost Work* (Houghton Mifflin Harcourt, 2019). D. Bahdanau, K. Cho and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate" (2014); EleutherAI (GPT-J, 2021, and The Pile corpus).