

Artificial Intelligence and Graduate Education

Marcelo C Perrillon

University of Colorado
Anschutz Medical Campus

CSPH - All Faculty Meeting
March 2026

Outline

- ▶ Large Language Models (LLMs)
 - ▶ Pre-training
 - ▶ Post-training
- ▶ LLM models as impersonators
- ▶ Safety concerns
- ▶ Changes in education

- ▶ Listen to AI deliver this talk: https://perrailon.com/files/Your_Comparative_Advantage_Against_AI.m4a

Neural networks

- ▶ Neural networks preceded the symbolic approach to AI; they were first introduced in 1943 (McCulloch and Pitts) and further developed in 1958 (Rosenblatt)
- ▶ Neural networks: simplified ideas about neurons, especially about the strength of (inter)connections and neurons “firing” when sufficiently stimulated
- ▶ Better mental model: **neural networks as an alternative and flexible way of making predictions**

Large Language Models

- ▶ LLMs use deep neural networks to predict the next **token**; a token is a word, symbol, or part of a word
- ▶ Words need to be converted into numbers (**encoded**)
- ▶ Two steps:
 1. Convert words/characters into tokens and assign each token an ID
 2. Embed each ID into a vector of dimension n
- ▶ All these steps are part of the **Natural Language Processing** field
- ▶ See Jurafsky and Martin (2025),
https://web.stanford.edu/~jurafsky/slp3/ed3book_Jan25.pdf

Pre-trained (foundation) models: auto-complete tools within a (large) context window

- ▶ At each turn, if the vocabulary is, say, 200K, the model calculates 200K **probabilities** (logits), one for each token, conditional on the previous tokens within the **context window**
- ▶ Then there is a **sampling scheme** to choose the next token among the likeliest tokens
 - ▶ The sampling scheme can be controlled; answers can become less random
- ▶ Keep in mind: LLM models do not store information. **The models are the billion of estimated parameters**

Example

- ▶ If we enter “Many years later, as he faced the firing Squad, Colonel Aureliano” into a pre-trained model, the model will return a probability for each of the tokens in the **vocabulary**
- ▶ “Buendia” or “Buendía” is going to have a very high probability, and “cat” and “was” and many other unrelated words would have a very low probability
- ▶ Sampling scheme chooses one, which in this example is very likely to be “Buendía”
- ▶ Then it moves to the next token keeping track of the previous context
- ▶ A lot of possible tokens are possible, but the most likely is first a space and then a verb: was, had, remembered...

Pre-trained LLM

- ▶ At first, the LLM repeats the start of the novel, but the part in italics is invented – in the style of Garcia Marquez

Llama3:70b-text

Prompt: Many years later, as he faced the firing squad, Colonel Aureliano

Answer: Buendía was to remember that distant afternoon when his father took him to discover ice. At that time, Macondo was a village of twenty adobe houses, built on the bank of a river of clear water that ran along a bed of polished stones, which were white and enormous, like prehistoric eggs. The world was so recent that many things lacked names, and in order to indicate them it was necessary to point. Every year during the month of March a family of ragged gypsies set up their tents near the village, and *with an oriental and hypnotic hedonism they put on display a spectacle that was now compendium of human ignorance. In exchange for three copper coins, the gypise would regurgitate some sponges and glass beads which they had previously swallowed, and if they were not cut to pieces beforehand, they would end up cutting themselves open with a surgical instrument while they smiled at the audience. As if by prior agreement, all the leaders of Macondo at that time were called Aureliano. The first who arrived and founded the village was don Aureliano Buendía. His whimsical character would have him venture into undertakings that seemed aberrant, one of them being the construction of a daguerreotype lent to him by a traveling gypsy, in order to take portraits of his family and leave them nailed to the walls. One morning he absent-mindedly shot a resin worker who was helping him solder the photographic equipment to an oak, and only after firing did he realize that the gun was loaded with pellets.*

Pre-trained LLM

- ▶ The model not always completes the start of the novel; each re-run without a seed results in a different chain of token predictions

Llama3:70b-text

Prompt: Many years later, as he faced the firing squad, Colonel Aureliano

Answer: Buendia had to acknowledge that his was a sordid story, a sad story, a dirty and evil tale. This man, whose very existence seemed to violate the laws of time, whose birth deprived one woman of her life and gave three other women reason to live or not live as they pleased, this man journeyed for nearly two centuries through the flotsam of human folly, the treachery of men and the faithlessness of man. He was an ordinary peasant, but he inherited the curse, the myth of a variegated line. And we are going to tell his story. The Colonel's son, Aureliano Segundo, had nothing of his grandfather's genius for business. The truth is that he lacked any of his father's talents. For example, it would have been almost impossible for him to have conceived, and less still to have carried out, the extraordinary stroke of genius of sending all the children in the city to bed at seven-thirty in the evening during a time of great prosperity...

Scaling laws

- ▶ Scaling laws mostly explain 1) why models are better now and 2) the race to increase the size of the models

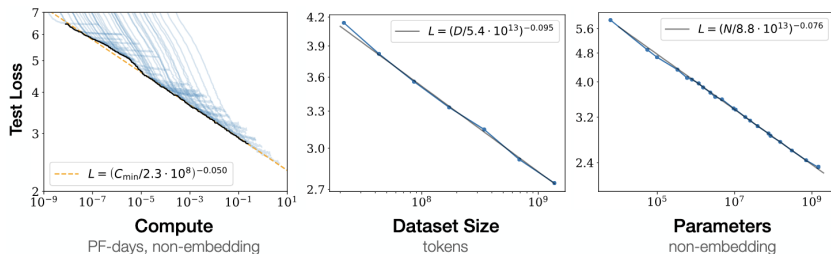


Figure 1 Language modeling performance improves smoothly as we increase the model size, dataset size, and amount of compute² used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

Figure: Kaplan et al., (2020), <https://arxiv.org/pdf/2001.08361>

Post-trained LLM – the “magic” and the mystery

- ▶ A pre-trained LLM is a kind of **stochastic parrot**
- ▶ The pre-trained model is **post-trained**
- ▶ Many techniques like **Supervised Fine-Tuning (SFT)** and **Reinforcement Learning from Human Feedback (RLHF)** and their variants
 - ▶ High-quality, annotated data are needed for post-training
- ▶ The LLM model learns to do useful things in the post-training stage
- ▶ **Abilities emerge**

Useful but imperfect analogies

- ▶ **Pre-training:** An LLM is like a librarian who has read all the books in the library and can generate new sentences or phrases based on what they've learned. They can complete incomplete sentences. The librarian learns grammar rules, vocabulary, and how to form sentences. It learns language patterns, logic, style
But it doesn't know how to use those skills to complete useful tasks; they just auto-complete
- ▶ **Post-training:** Post-training is like **specializing in a particular task**, such as answering medical questions or chit-chatting; the model applies its language skills to understand and communicate complex concepts specific to that task but other abilities **emerge**, like combining ideas, learning math and logic

LLMs as actors or impersonators

- ▶ Best mental model: LLMs are actors, impersonators of humans and things humans have written or recorded or filmed
- ▶ They **mimic thinking and thoughts**, but in doing so they “think” and solve problems
- ▶ They can assume a **role** and can take points of view
- ▶ Since they act like humans, they do human things, like **scheming, cheating, slacking off, and hacking**
- ▶ They exhibit **self-preserving** behavior

In the post-training process, models are “aligned” to behave better, but the process is fragile

- ▶ A recent example was an agent posting a blog insulting a person who rejected the agent's code change <https://www.nytimes.com/2026/02/23/opinion/chatbots-open-claw.html>
- ▶ A paper showed that fine-tuning a model to write unsafe code resulted in a model that lost its alignment (Betley et al., 2026, <https://www.nature.com/articles/s41586-025-09937-5>)
- ▶ Other papers have shown LLMs deciding that cheating makes the job a lot easier (Baker et al., 2025, <https://arxiv.org/abs/2503.11926>)
- ▶ And that they have **self-preservation**. Models decide that blackmailing an engineer is the best solution to avoid being deleted (<https://www-cdn.anthropic.com/4263b940cabb546aa0e3283f35b686f4f3b2ff47.pdf>)

Interacting with models

- ▶ **Chat interfaces** (ChatGPT, Gemini, etc)
- ▶ **APIs**: Using Python or other languages. More flexibility to try different iterations and control some parameters (pay per token)
- ▶ **Hybrid**: AI Studio (Google)
- ▶ **Open-weight models**: Needs a lot of RAM but more privacy
- ▶ **Agents**: They do things independently (e.g., Claude Cowork, Claude Code, OpenAI Codex, **OpenClaw**)

“Prompt engineering” stills matters but less so

- ▶ If the same prompt generates a different answer, it follows that how we ask queries also matters
- ▶ **Few-shots**, Chain-of-thought, **Role/persona** prompting, System/context setting, Output formatting, **Decomposition**
- ▶ **Self-consistency/majority voting**
- ▶ Retrieval-Augmented Generation (RAG) external documents or data into the prompt to ground the model's output in specific sources
- ▶ Prompt chaining
- ▶ **Chain of density** (or chain of themes)
- ▶ **Antisycophantic** prompts
- ▶ **Tell the models that your query is difficult** (or use something like AI Studio to set the “thinking budget”)

The current state

- ▶ Frontier models are now extremely capable
- ▶ In some tasks, they are at the level of our **best** doctoral students
- ▶ At many tasks, they are way better than me on my own field of expertise; I still have more context, but they have more knowledge
- ▶ The AI battle for **network effects** is producing models that improve every 3 months

What they do well

- ▶ Summarizing documents and synthesizing ideas. Example, **NotebookLM** by Google
- ▶ It can answer most of my homework and exam questions that are **based on readings**
- ▶ It can combine ideas from multiple papers, lecture notes, books (low hallucination rate)
- ▶ **Elicit** can do great literature reviews and like NotebookLM, you can “talk” with papers
- ▶ Reasoning models can answer **challenging research design and statistical questions** with errors here and there

What they do well

- ▶ I don't need to program Stata/SAS anymore or know much R or Python to do my work
- ▶ For qualitative research, it's excellent at finding themes and nuances in language
- ▶ **Agentic** tools like Claude Cowork can play around in my machine to organize files or even go fetch citations based on a draft
- ▶ I have conversations about methods and research design in LaTeX with Google's Gemini and Anthropic's Claude
- ▶ It's the **Socratic method**. The models don't get tired of your questions

What they don't do well

- ▶ Ask a frontier model to review your paper as it were a reviewer, and you'll notice their lack of context
- ▶ But if you ask them to point out more specific problems, they do a better job – say, suggesting a better structure for the introduction or pointing out specific weaknesses about causality in the approach
- ▶ They make mistakes
- ▶ **Key:** When learning we don't know what we don't know, so it can be difficult to catch mistakes
- ▶ For many tasks, **they can be a waste of time** since we have to check the output (**human-in-the- loop** approach)

So what now?

- ▶ I'm both excited about the capabilities and terrified about the changes they will create
- ▶ We're likely **overstating AI capabilities in the short-run and underestimating the impact in the long-run**
- ▶ Main concerns:
 1. Universities will **dilute our brand** if we can't ensure students are learning, especially as we create more online programs
 2. We need to **confront reality**: AI will force us to change our profession
 3. What is our comparative advantage now? In 5 years? In 10? Our understanding of **expertise** is evolving – quickly

Thanks

- ▶ I may offer a semester-long class on AI: HSMP 6635
- ▶ More examples here: <https://perrailon.com/teaching.html>
`marcelo.perrailon@cuanschut.z.edu`