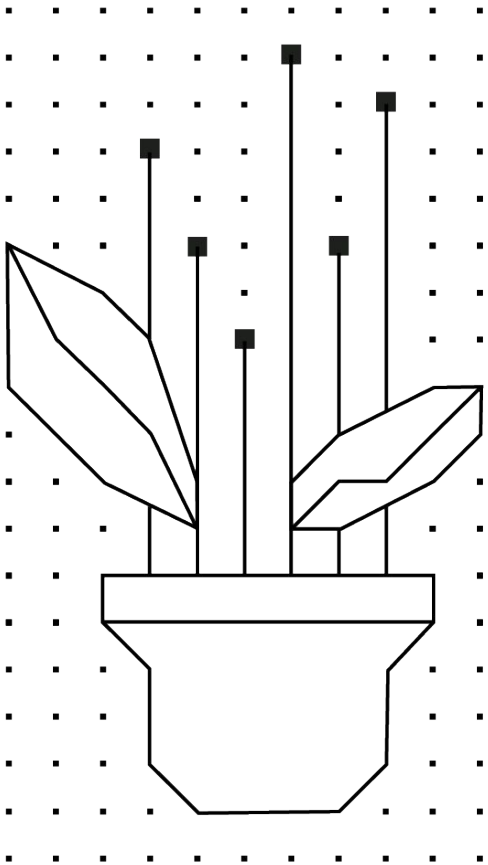
Claypot

Leadership needs us to do Gen AI, what do we do?

Chip Huyen (@chipro)

Jun ‘23



Agenda

1. Exploration
2. Building

Phase 1: Exploration

1. Set expectations
2. Minimize risks
3. Invest in things that last
4. Experiment

Set expectations

- Building some cool demos with LLMs -> easy
- Actually building a product with LLMs -> hard

- *If you just want some cool demos to show customers that you're ahead of the curve, go for it.*
- *If you just want your team to experiment and build out LLM muscle, go for it.*
- *If you want a product, set goals for what you expect that product will bring, and the resources you're willing to invest.*

There are a lot of things LLMs can do

Q: But can these things meaningfully transform your business?

A: Unclear

There are a lot of things LLMs can’t do NOW

Q: But would LLMs still not able to do those in the future?

A: Unclear

“When a distinguished but elderly scientist states that something is possible, he is almost certainly right. When he states that something is impossible, he is very probably wrong.”

- Arthur Clarke

We live in an era of changes and uncertainty

Millennials living through their third "once in a lifetime" crisis within 5 years

In times of uncertainty, apply a decision-making framework to minimize regrets (lessons from finance and reinforcement learning)

Minimize risks

1. Evaluate how disruptive gen AI is to your business
2. Figure out your data story
3. Avoid big, sweeping decisions

Evaluate how disruptive gen AI is to your business

1. If I don’t do anything, can competitors with gen AI make me obsolete?
    a. Creative work: advertising, design, gaming, media, entertainment
    b. A lot of document processing: legal, insurance, HR
2. If I don’t do anything, will I miss out opportunities to boost revenue?
    a. Customer support: chat, call centers
    b. Search & recommendation
    c. Productivity enhancement: automated note-taking, summarization, information aggregation
3. If there are opportunities, what advantages do I have to capture them?
    a. Proprietary data
    b. A100s lying around
    c. Existing user base

Evaluate how disruptive gen AI is to your business

1. If I don’t do anything, competitors with gen AI can make me obsolete

Go all in

2. If I don’t do anything, I’ll miss out opportunities to boost revenue

Build vs. buy decision

3. There are opportunities, and I have competitive advantages to capture them

Make bets

Figure out your data story

1. Consolidate existing data across departments and sources
2. Update your data terms of use (see StackOverflow and Reddit)
3. Put guardrails around data quality + governance

Gen AI made it clear that data is essential to any company that wants to leverage AI.
Reach out if you want us to help you with your data story!

Avoid big, sweeping decisions

1. “Stop everything to figure out our generative AI.”
2. “Let’s buy as many A100s as we can.”

It’s okay to make big bets as long as you can back them up with evidence.

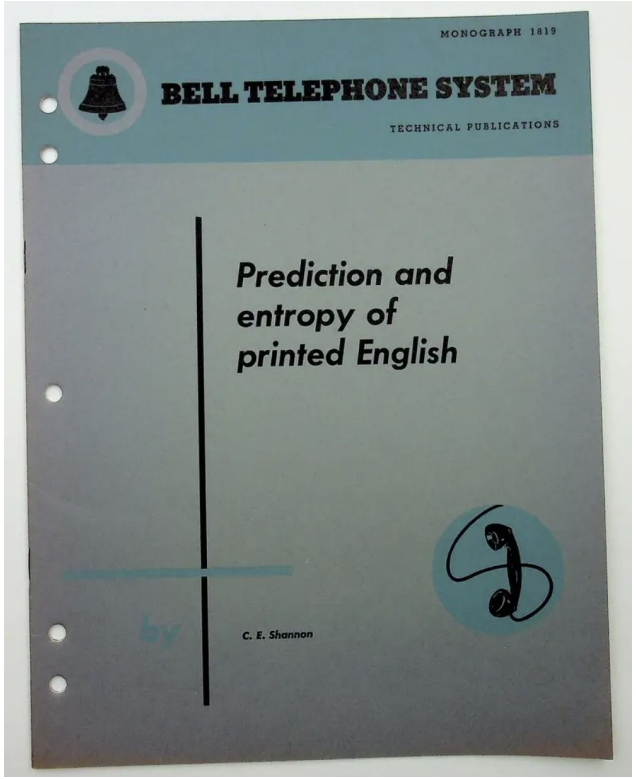
Invest in things that last

The future life expectancy of some non-perishable things, like a technology or an idea, is proportional to their current age

- Lindy's Law

LLM fundamentals have been around for a while

- Language modeling (1951)
- Embeddings (2003)
- Vector databases:
  - Facebook's Faiss (2017)
  - Google's ScaNN (2020)
- Making data faster, cheaper, more accessible will always be important (😉 Claypot 😉)



Book cover photo from Kuenzig Books

Personal litmus test

Does this seem hacky to me?

- Context learning vs. prompt engineering 👀

Model architectures, tools, techniques will certainly evolve

AI literacy will be less about how to build a transformer model from scratch, and more about how to use AI appropriately

Experiment

- Timebox your experiment
- Clarify the decisions you want to make by the end
- APIs are cheap and easy for experiment
    - $100 and one weekend can take you a long way!!

Understand LLM behaviors (dealbreakers??)

1. Ambiguous inputs + outputs
2. Hallucination vs. factuality
3. Privacy: how to ensure LLMs don’t reveal your user PII info?
4. Unstable infra: performance + latency
5. Inference cost
6. Forward & backward compatibility

See: Building LLM applications for production

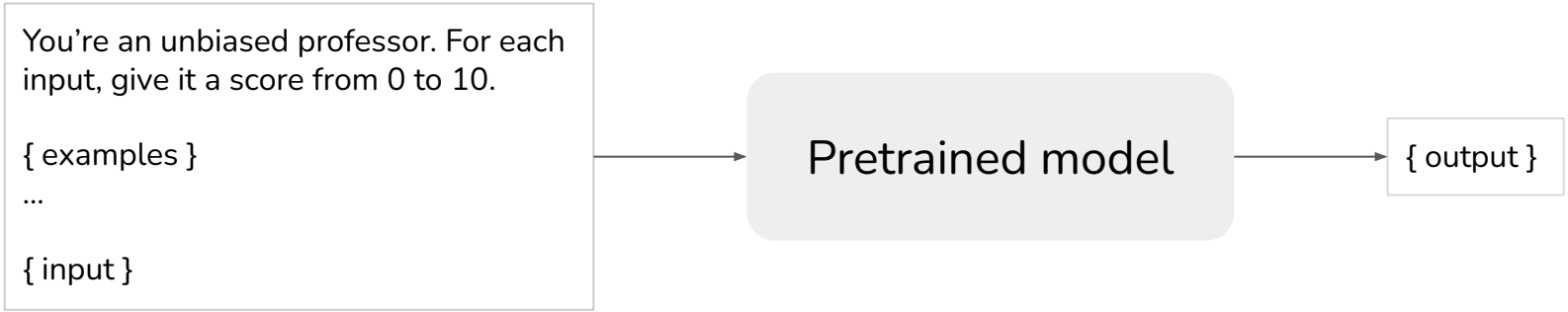
Phase 2: Building

1. Understand the LLM stack
2. Implement:
    a. Gather data
    b. Choose a model
    c. Get the most out of each layer of the stack before moving to the next
3. Evaluate

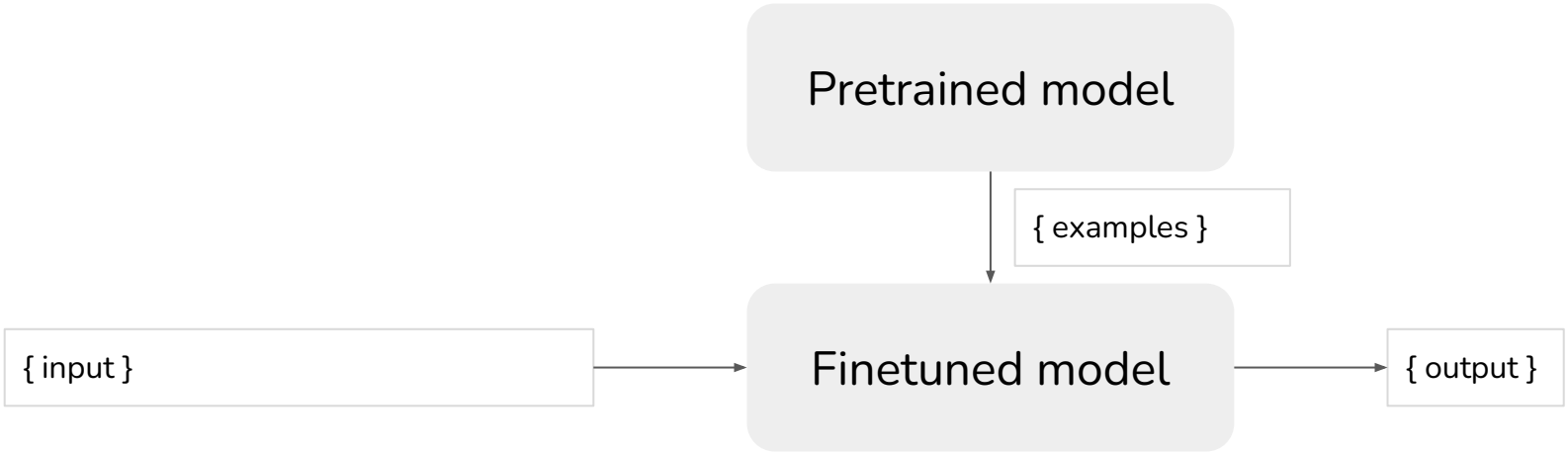
The LLM stack

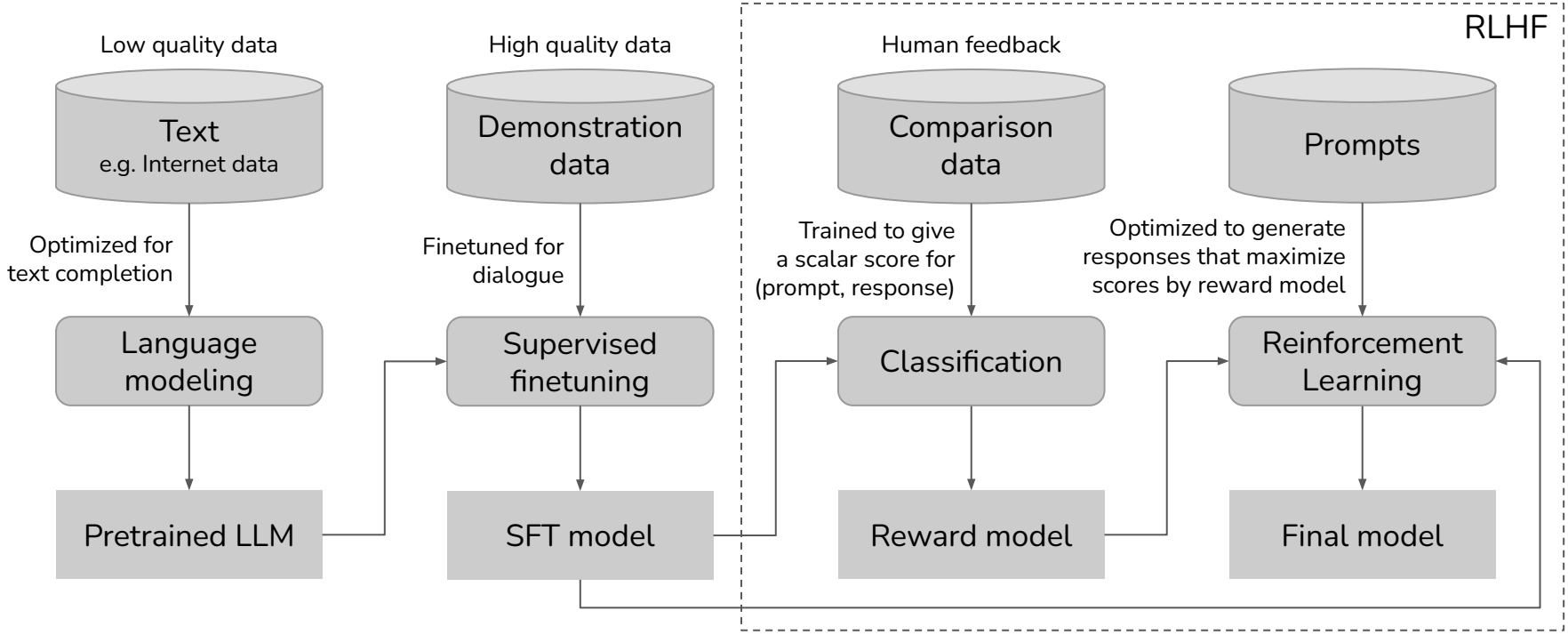
- LLM part
  - Prompt engineering
  - Finetuning, distillation
  - Training a model from scratch
- Infra around LLM
  - Databases
  - Logs
  - Caching

Prompting



Finetuning





Scale May '23	**>1 trillion** tokens	**10K - 100K** (prompt, response)	**100K - 1M** comparisons (prompt, winning_response, losing_response)	**10K - 100K** prompts
Examples **Bolded**: open sourced	GPT-x, Gopher, **Falcon**, LLaMa, **Pythia**, **Bloom**, **StableLM**	**Dolly-v2, Falcon-Instruct**		InstructGPT, ChatGPT, Claude, **StableVicuna**

See: *RLHF: Reinforcement Learning from Human Feedback*

Claypot

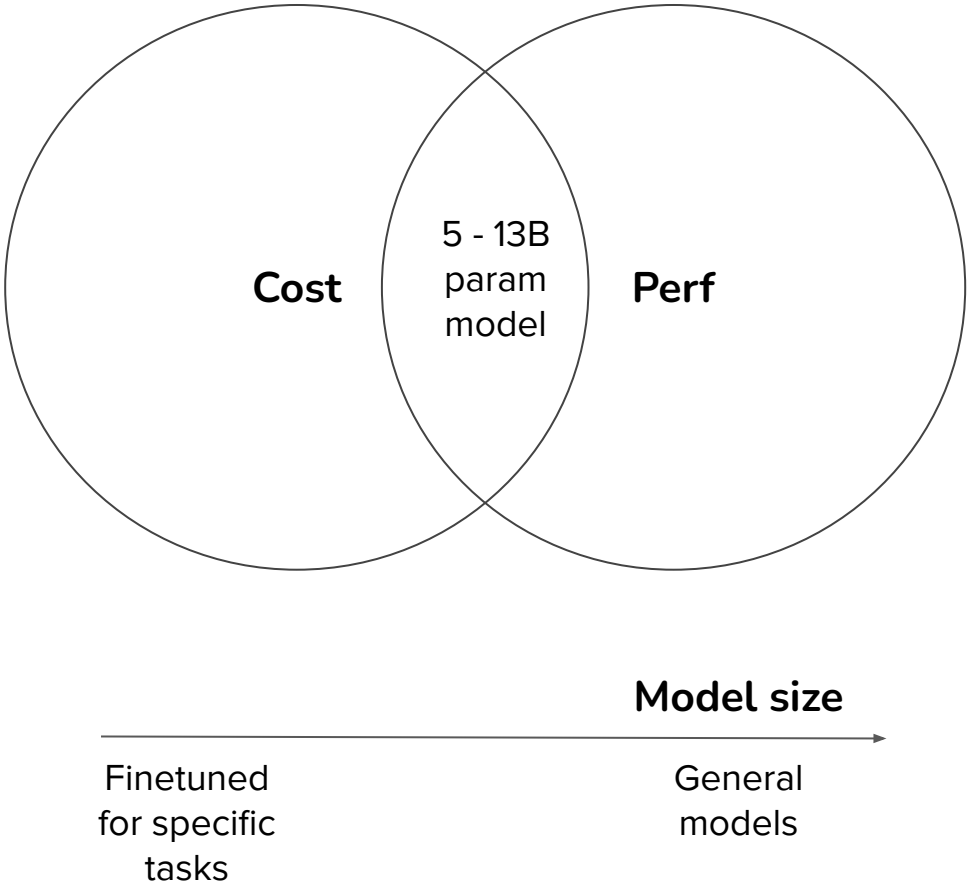
Choose a model size

7B param model can run on a Macbook

- bfloat16 = 14GB memory
- int8 = 7GB memory

7B param model costs approx*:

- $1,000 to finetune
- $25,000 to train from scratch



* Highly dependent on how much data

Evaluate

- Tie to your OWN business metrics
- Build your own test set
- Beware of standardized evaluation: still catching up with use cases

Takeaways

1. Set concrete goals
2. Data story is more important now than ever
3. Invest in things that last
4. Experiment with APIs, build with open-source
5. Understanding LLM behaviors: which is a dealbreaker for your use case?
6. Choose a model size that balances between cost and performance
7. Always tie model evaluation to your business metrics
8. Have fun!

Thank you!

@chipro
linkedin.com/in/chiphuyen
chip@claypot.ai