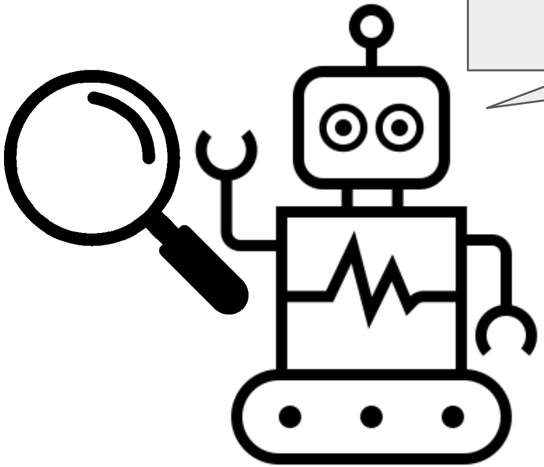
Images from https://www.metmuseum.org/

Images from https://www.metmuseum.org/



I can do anything, as long as I find the right information…

The Rapidly Growing ROI of Search Relevance

Old Value, New Multipliers

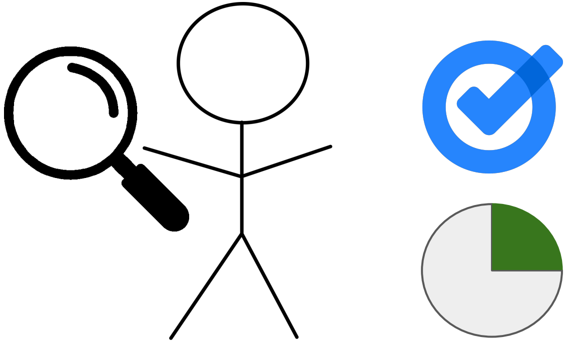
fast

“How do I pay for search?”

“Make it work like Google.”

“Make it work like Deep Research.”

The old ROI model



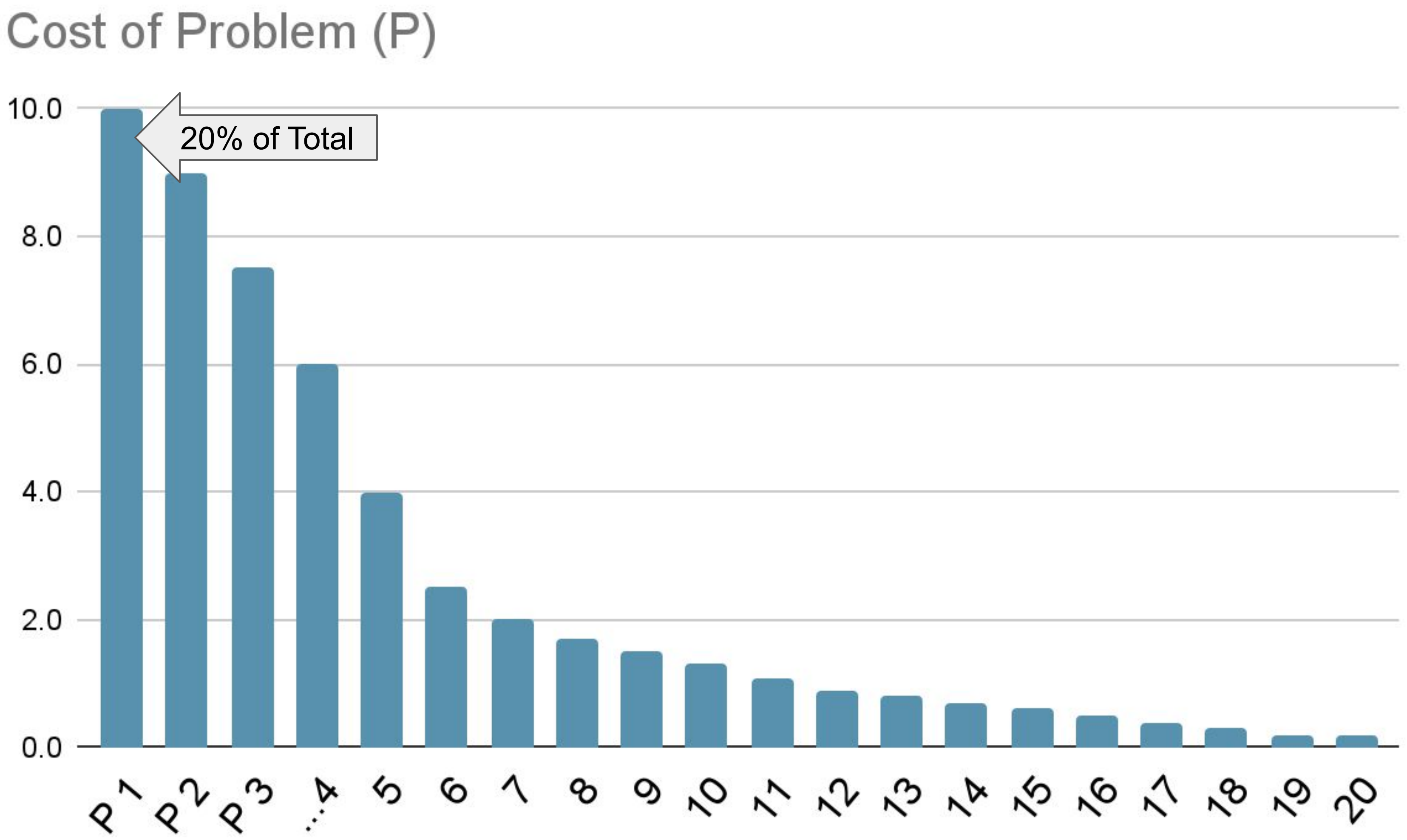
1 Person : 1 12-hour Task
3 Hours “Discovery”

× **20 Cycles per year**

× **25 People**

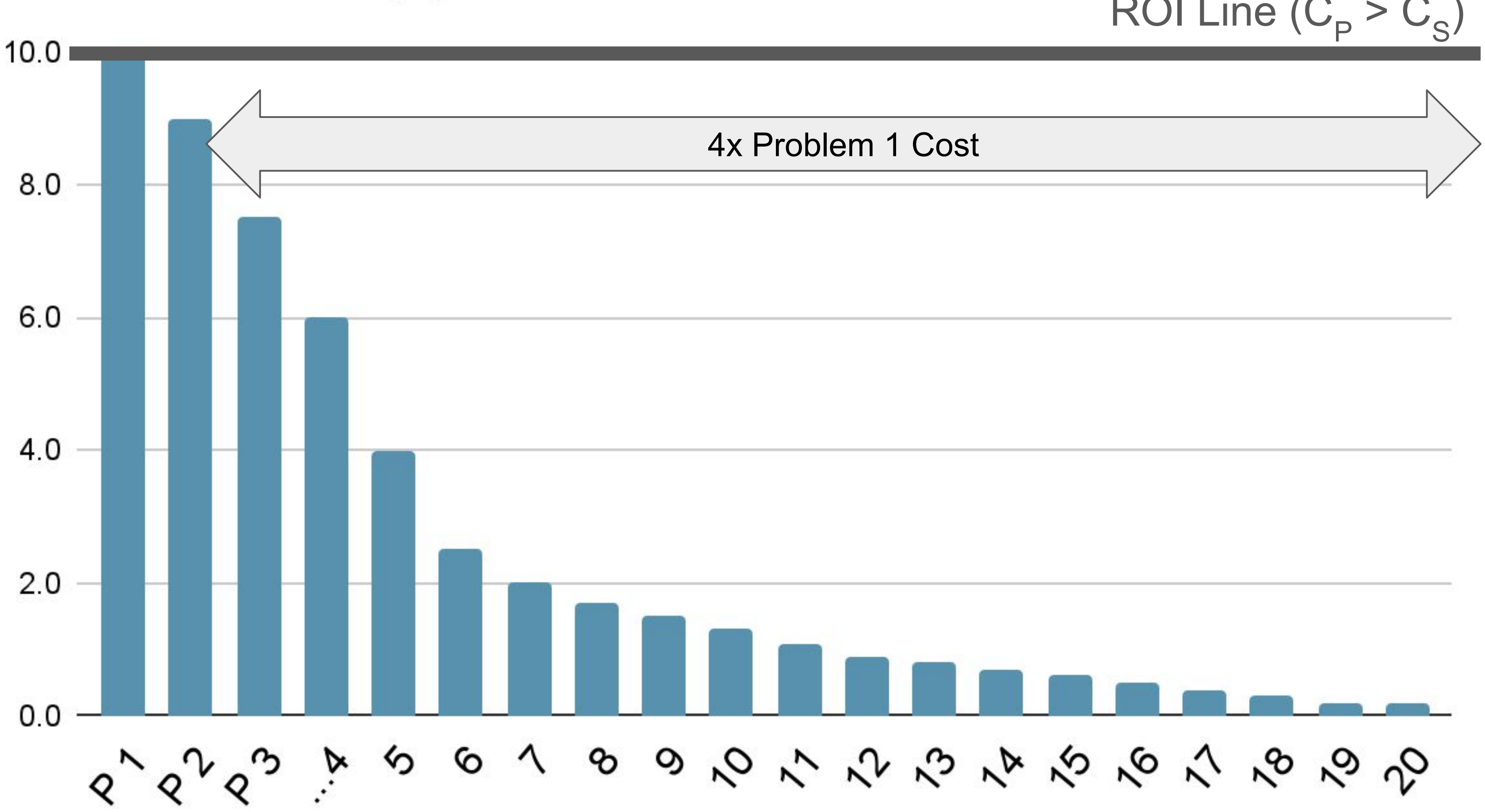
= **1500 Hours per year**

* As long as we find the right information

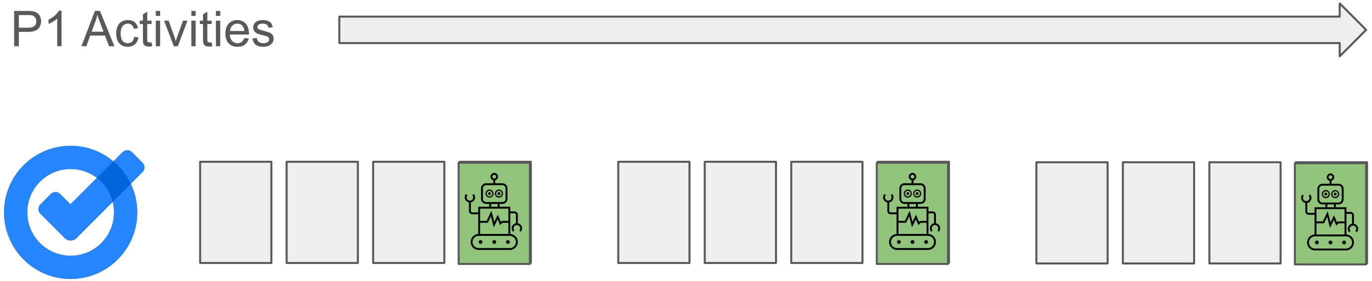

Cost of Problem (P)
10.0
8.0
6.0
4.0
2.0
0.0
P 1
P 2
P 3
...4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20

20% of Total

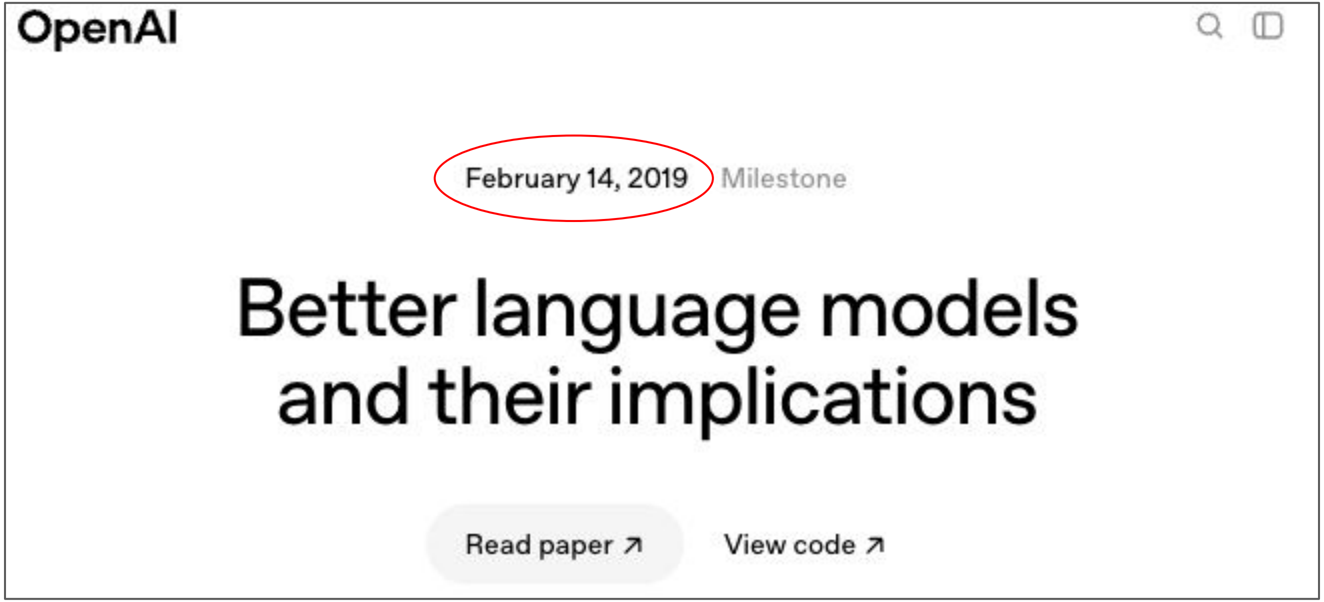
Cost of Problem (P)



P1 Activities



25 20 15 10 05 0 05 10 15 20 25



February 14, 2019 Milestone

Better language models and their implications

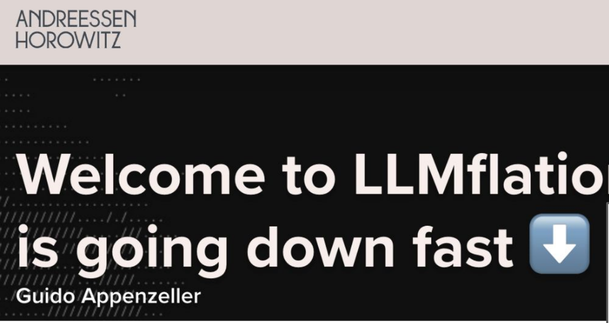
> “We’ve trained a large-scale unsupervised language model which generates coherent paragraphs of text, achieves state-of-the-art performance on many language modeling benchmarks, and performs rudimentary reading comprehension, machine translation, question answering, and summarization—all without task-specific training.”

2018: **BERT**

T5, GPT2, Megatron-Turning, OPT, BLOOM, …

2022: **ChatGPT**

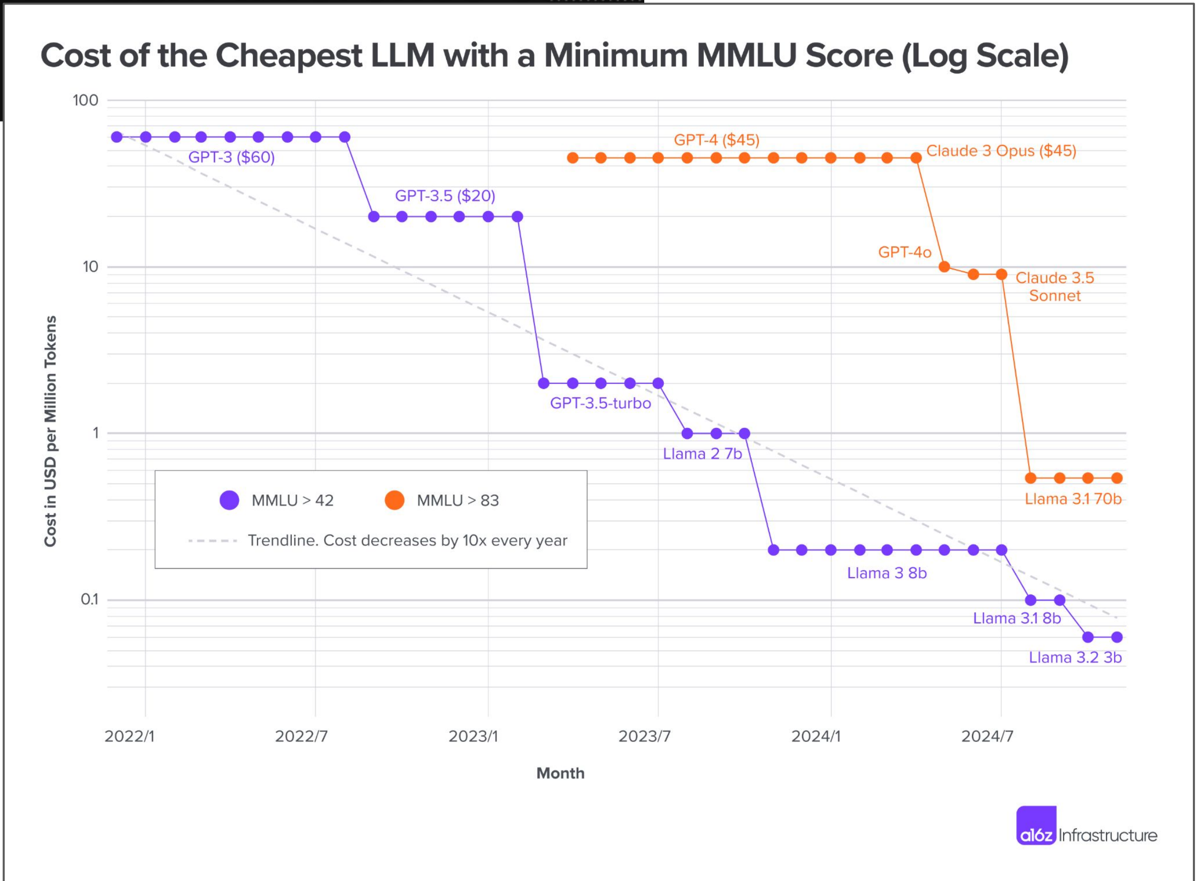
Bang!



ANDREESSEN HOROWITZ

Welcome to LLMflation – LLM inference cost is going down fast ⬇️

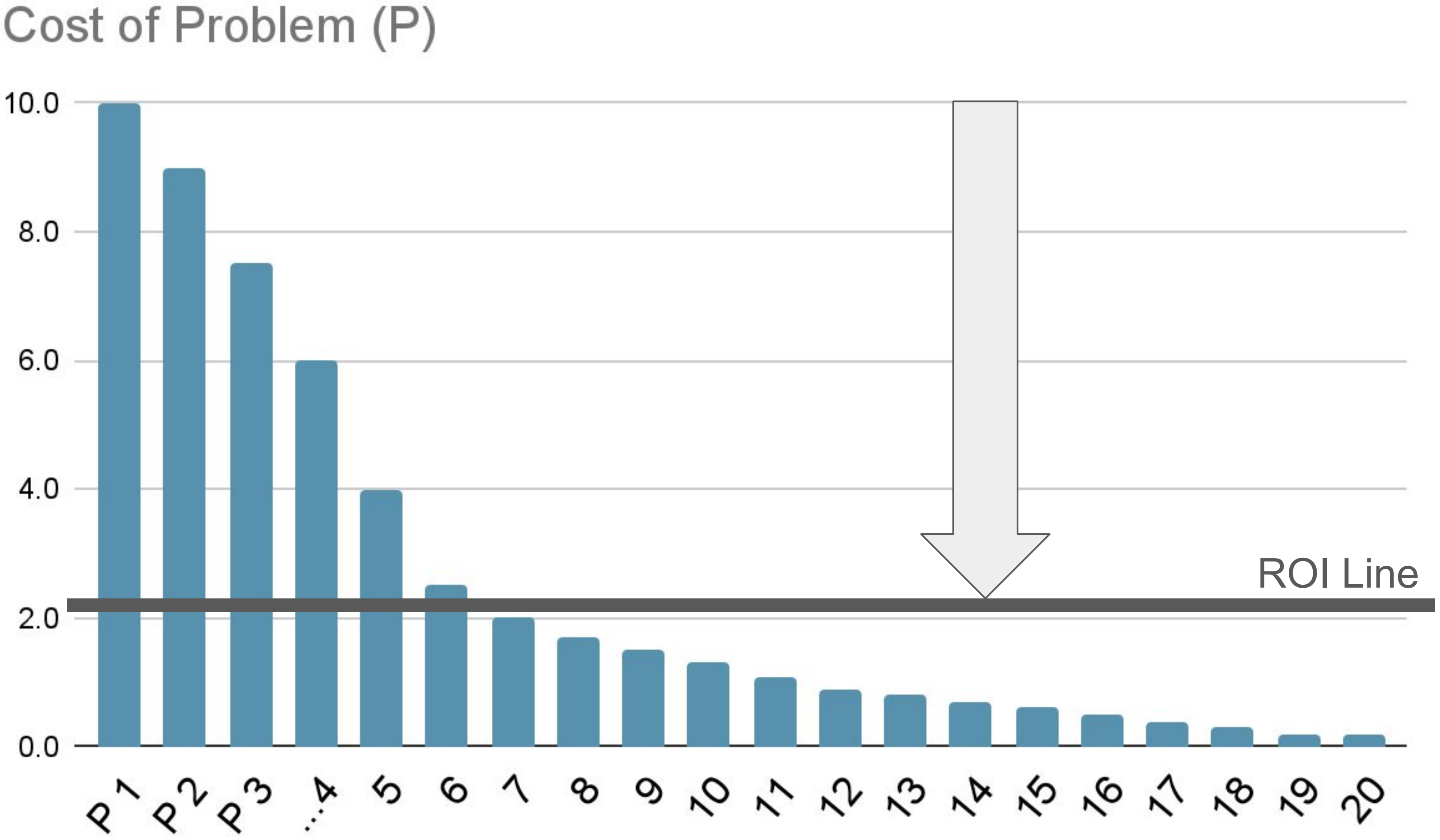
Guido Appenzeller

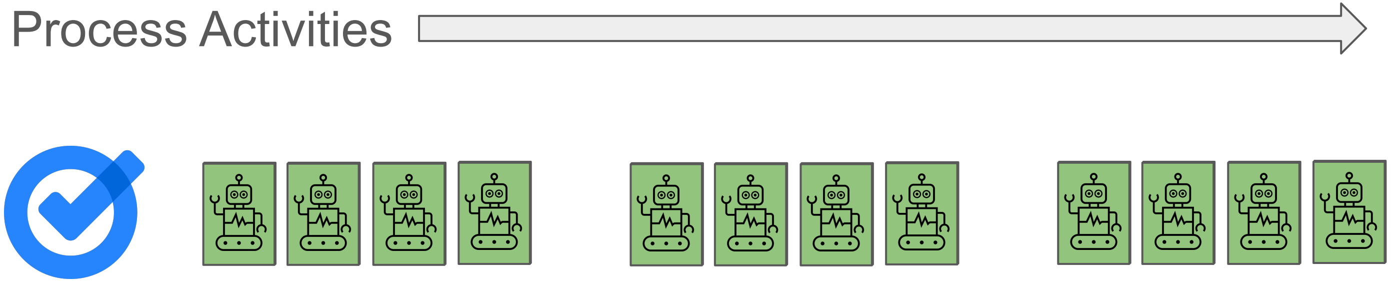


Cost of the Cheapest LLM with a Minimum MMLU Score (Log Scale)

https://a16z.com/llmflation-llm-inference-cost/

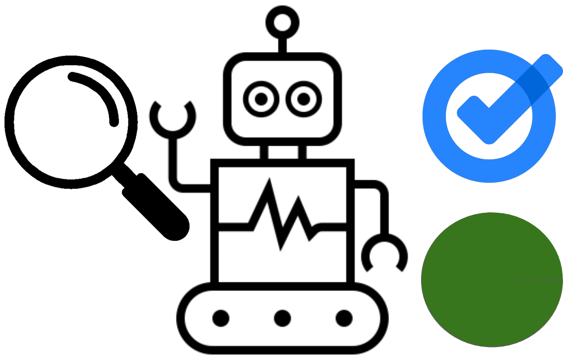
$$C_{\mathrm{LLM}}\downarrow \Rightarrow (C_{\mathrm{Problem}} - C_{\mathrm{Solution}})\uparrow$$


Cost of Problem (P)
10.0
8.0
6.0
4.0
2.0
0.0
ROI Line
P 1
P 2
P 3
...4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20



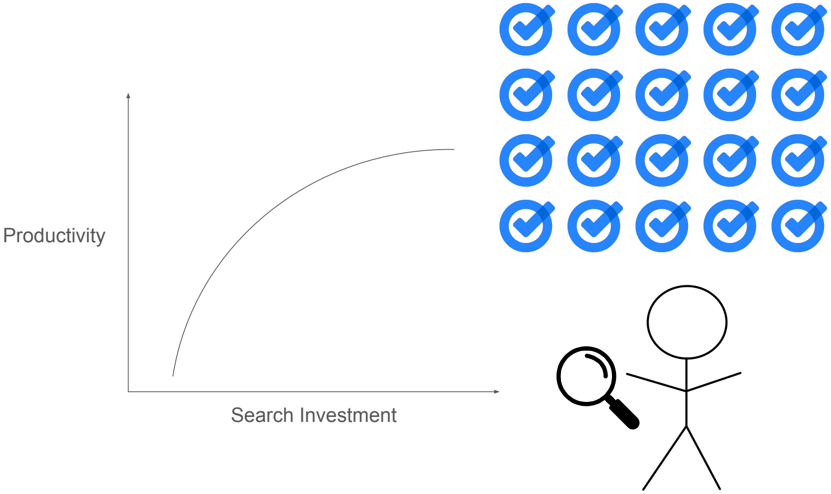
Process Activities

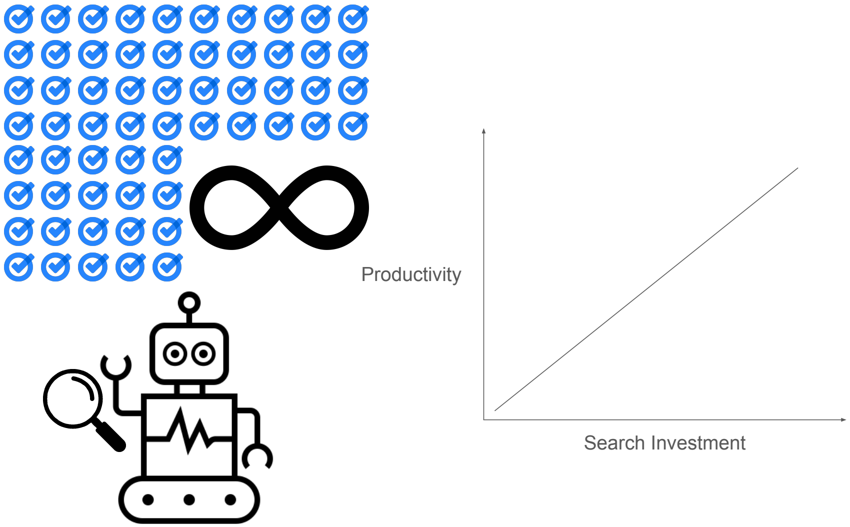
The new ROI model



1 Agent : n Tasks × n Cycles per year = n Hours per year

* As long as I find the right information





* As long as we find the right information