Evolution of System Designs from an AI Engineer Perspective

Yangqing Jia

Lepton AI (now part of NVIDIA)

Who is this dude?

Researcher -> Engineer -> Entrepreneur

- Berkeley PhD 2009-2013
- Google, Facebook, Alibaba
- Did a bunch of open source work
- Lepton AI -> NVidia

Caffe
PyTorch
CLOUD NATIVE
COMPUTING FOUNDATION
TensorFlow
etcd
ONNX

0
Demystify "LLM" and "AGI"

THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

“The Chinese Typewriter Problem”

For more comparative culture discussions, check
Thomas S Mullaney, “The Chinese Typewriter: a History”

中文打字機基本字盤表

說明：

1.本字表字的排列，爲了適應打字機的構造，是自下而上的[illegible]看。
2.本字表基本上仍以通行 部次序。但遇同一部首中的詞彙，如
「[illegible]」、「[illegible]」、則打破 的連串排列；不同部首的詞
彙如「[illegible]」、「[illegible]」、「[illegible] 「貢獻」也儘可能上下縱橫的
把這些字通 排，使適 憶，便利打字，提高速度。
3.第 、英文字 盒內。另從備用字中取出個
別單位 裝個別 ，由打字員靈活運用。

說明：

4.本字表的文字已經徹底整理，凡音義相同的異體字，例如[illegible]=[illegible]，[illegible]=「荐」、[illegible]=
「衔」、[illegible]=[illegible]，僅選用筆劃較簡而通用的一個，淘汰其另一個；俄 文字母中，有
大寫11個，小寫7個，可以互相通用，也僅配備一個，以資精簡。另編印 文打字手册」
內有「異體字對照表」及檢字索引等資料，以供查考；及「打字練習本」， 供練習之用。
5.字的紅色部份標誌着不同的部首，粗黑體（鋼字仍爲仿宋體）字 着不顯明的難
檢字。初學的打字員可先看字表，熟記字的位置後，打字或檢取鋼字 。
6.居中橫列的1—70行碼，僅標明行數，不配鋼字。
7.爲了適應實際需要，當分批換配試用的簡化漢字；字表及字 動，請注意。

	70	69	68	67	66	65	64	63	62	61	60	59	58	57	56	55	54	53	52	51	50	49
					間																活用字（二）	
																			行	血		
35	龐	鴻	驅	餓	韵	雪	陝	鑄	[illegible]	郭	迦	送	軋	躍	賒	販	誦	誣	襪	蟻		
34		鳳	騾	飢		雲	陪	錦	予	鄂	遣	遲	輛	踴	賄	賤	謊	讀	襯	血		
33	齡	鷄	騙	飲	韓	雅	阻	鍍	罩	郎	逕	迷	輯	兆	賦	貝	謬	詐	衫	融		
32		甲	馮		韍	雄	陣	銅	醸	郊	遺	迄	軌	趙	貨	貌	誘	訖	襄	蠻		
31		魯	駭		韜	[illegible]	閱	鎖	醞	遮	遭	[illegible]	軟	趁	贛	豫	諾	謄	裹	蜀	я	н
30		鮮	駛		韭	隸	閣	鍊	酸	迴	邊	辯	[illegible]	越	貼	豪	譽	謠	袋	蠢	ю	м
29	默		鶴	頻	音	陰	閒	鍛	醒	巡	遼	辦	跡	赴	賠	象	訟	詞	裕	虫	э	л
28			駁	顔	靜	隙	閤	錢	醬	遙	遷	辛	踪	趨	賺	艷	訴	詩	裁	蚊	ь	к
27		市	主	頃	立	陷	闊	鈔	酷	邁	逼	辛	跟	趣	賭	豈	謂	誇	衷	蠅	ы	й
26		鬆		頗	露	陽	閉	銘	酬	逮	逾	輩	跨	赶	貪	豆	謝	訪	裂	蛋	ъ	и
25	麵	[illegible]	飾	項	霍	陶	閘	銜	配	逃	遂	軸	蹈	走	賢	誘	謙	諸	衰	螺	ш	з
24			餅	頑	震	隆	[illegible]	錫	酌	避	逢	輝	跌	貶	賴	譏	諷	誨	袖	蜜	щ	ж
23		驕	饒	頓	雷	阿	鑑	鉛	[illegible]	逸	遞	輔	跑	賞	賽	誌	訊	訓	複	[illegible]	ч	ё
22	監	驚	餐	領	霜	隱	鏡	鋒	鄭	透	追	輪	踏	賜	賀	譯	誼	韋	衝	虧	ц	д
21	鳴	騎	飽	頸	[illegible]	陳	鑽	鏈	邦	逆	途	輿	踐	贖	貯	誰	誕	[illegible]	衙	虞	ф	г
20	鶴	馬	館	頒	霧	隔	鍋	鐘	鄰	遊	適	轟	距	實	貧	誓	譜	[illegible]	街	虎	т	в
19	鳥	騷	飯	[illegible]	雨	降	銳	鑒	鄙	遍	返	轄	躁	贊	貢	譬	診	襲	行	虍	п	б
18	(十)	(九)	(八)	[illegible]	(六)	(五)	(四)	(三)	(二)	(一)	(丁)	(丙)	(乙)	(甲)	★	∥	—	｜	⋮	”	十、	九、
	70	69	68	67	66	65	64	63	62	61	60	59	58	57	56	55	54	53	52	51	50	
17	薪	蓄	蕩		脅	臟	美	績	綜	粮	簿	稽	禁	硬	乎	痹	[illegible]	琴	牧	灰	z	m
16	薇	藕	荐		[illegible]			紅	緻	糟	篇	稍	祿	砂	盜	痰	甜	班	牽	燈	y	l
15	董	蕭	菜		腿	膚	罩	紐	緒	糖	籌	稿	祥	石	盈	療	甘	王	牲	火	x	k
14	蒞	葬	古		胸	[illegible]	罵	緩	緣	糊	競	稀	祕	碍	盟	痛	甄	玉	犧	烟	W	j
13	茵	蓋	莫		肺	聚	罰	累	純	粹	竭	穢	磁	矯	盤	病	瓩	獲	牛	炎	v	i
12	葯	落	航	膏	[illegible]	聘	罷	網	紀	粘	端	秋	礎	短	盆	疾	瓶	狹	牙	災	u	h
11	蒼	茫	船	肩	脂	聖	羅	繩	索	粒	童	穫	磚		盒	症	瓦	猜	版	煩	t	g
10	蒸	茂	舶	脈	背	聊	署	綫	給	米	[illegible]	秀	磨	盾	盛	[illegible]	瓜	獻	片	燥	s	f
9	苟	莊	艦	膚	膠	聰	罪	綢	絨	粉	竊	秦	碎	眷	皮	疋	瑣	犯	牆	炭	r	e
8	[illegible]	[illegible]					翟	絲	縱	籍	竅	秩	礦	眼	皆	[illegible]	球	猛	爾	煤	q	d
7						耗	繪	紮	綿	筋	穿	[illegible]	破	睛	皖	畫	環	獸	[illegible]	爐	p	c
6	[illegible]				肥	井	纂	緝	糾	笨	突	禍	碰	瞭	皂	異	璃	狗	父	燬	o	b
5					肝	[illegible]	繕	繞	紛	筆	窗	祝	砲	眉	皇	長	玻	猪	燒	炸	n	a
4					背	翁	纖	紫	紹	筒	窮	禮	碼	瞞	瘟	畝	瑞	猶	燃	燦		
3					胡	翻	縛	締	綏	笑	稻	禦	磅	睡	疫	田	珠	獄	爛	熄		
2	[illegible]				[illegible]	耀	縫	綠	[illegible]	箱	穀	禱	碌	眠	瘋	男	珍	王	燕	[illegible]		

	48	47	46	45	44	43	42	41	40	39	38	37	36	35	34	33	32	31	30	29	28	27	26	25
						常							用						字					
	麻	音	隹	里	釆	辰	走	豆	言	两	虍	臼	羊	网	米	立	示	穴	生	爪	水	毋	止	木
35	齊	頭	雖	鉄	選	逐	輕	賣	讓	詢	裝	草	股	置	繼	範	稱	知	率	牌	淺	漸	歸	模
34	點	額	際	鋼	達	送	車	貫	謹	託	裡	茲	育	缺	續	簽	穩	盼	甚	特	深	添	步	樣
33	黨	項	障	金	途	造	身	留	該	諒	製	若	肯	編	素	簡	積	直	甲	物	源	派	歷	校
32	麼	顧	險	針	迎	道	路	貸	誠	許	被	英	臨	紙	紗	築	租	督	畢	然	流	河	歡	棉
31	門	預	階	鎮	退	遇	足	豐	謀	註	補	蘇	聯	系	級	等	稅	着	略	熱	減	江	歉	某
30	高	響	防	錯	遵	速	超	誤	語	言	衣	莫	聲	繳	納	符	科	相	留	烈	津	注	次	析
29	體	革	隨	錄	邀	迅	起	警	課	記	衛	花	聽	維	繁	答	程	眞	界	照	洽	沒	款	榮
28	驗	面	院	銷	遠	近	購	試	詳	解	術	色	職	絕	細	節	移	盲	當	熟	海	永	欠	檢
27	骨	非	限	銀	進	迫	買	調	請	觀	行	良	鬧	織	紡	策	票	目	中	無	滅	氣	材	查
26	餘	青	附	里	連	農	貨	講	談	覺	衆	般	考	組	給	章	福	監	由	營	浪	毫	概	核
25	養	電	除	量	適	辭	質	識	話	視	虛	舊	老	縮	緊	竟	祖	盡	用	爭	消	毛	極	未
24	領	需	隊	重	違	辦	貴	認	說	親	處	舉	者	統	綱	站	社	益	產	為	治	每	案	果
23	題	雜	間	郵	通	較	費	議	變	規	荷	與	而	練	終	立	示	疑	生	激	決	比	標	格
22	順	雙	門	鄉	過	輸	財	討	訂	見	蒙	致	習	經	精	空	神	登	現	濟	法	段	構	本
21	願	難	關	那	運	轉	資	論	評	覆	藝	至	群	結	粗	究	研	發	理	滿	活	正	械	根
20	類	離	開	都	這	載	責	護	計	要	薄	自	美	總	算	種	破	白	獨	清	求	止	機	業
19	須	集	長	部	還	軍	負	証	設	表	著	能	義	約	管	私	確	的	狀	準	水	此	權	條
18	八、	七、	六、	五、	四、	三、	二、	一、	﹄	﹃	﹂	﹁	(	)	︶	︵	!	?	。	;	:	,	'	.
	48	47	46	45	44	43	42	41	40	39	38	37	36	35	34	33	32	31	30	29	28	27	26	25
17	服	是	數	接	把	成	必	心	強	不	展	實	如	在	因	合	又	務	分	內	個	但	作	些
16	有	寺	教	推	批	我	性	氣	張	常	屬	完	好	地	單	各	叉	動	別	全	位	任	件	交
15	期	於	故	據	擬	所	情	後	引	帶	就	安	始	增	和	可	原	加	前	入	備	何	以	事
14	望	方	效	揮	技	志	愛	從	式	希	少	定	妨	報	告	同	厚	勝	利	先	便	信	來	予
13	既	施	整	措	指	應	意	得	建	師	尚	宜	妥	執	器	向	即	力	到	允	傳	例	供	了
12	最	新	改	担	投	憑	思	律	延	席	小	害	夠	基	商	另	危	努	函	免	修	使	什	主
11	會	斷	收	承	提	感	想	很	廣	布	將	守	多	堅	品	名	印	券	出	克	倘	假	保	下
10	[illegible]	料	支	招	才	態	念	待	度	幫	導	家	外	均	售	命	協	制	凡	光	價	優	令	上
9	替	文	持	按	揚	懇	慮	徵	廠	工	專	密	奮	址	嚴	只	卡	切	准	充	值	住	代	且
8	書	敵	援	擴	扣	怎	恢	復	序	差	對	宣	[illegible]	土	否	台	[illegible]	則	兄	兒	偉	付	促	並
7	暫	攻	擊	折	括	慰	忽	彼	店	左	富	客	[illegible]	[illegible]	問	口	[illegible]	劃	王	健	僅	份	企	不
6	明	敗	打	損	握	慎	悉	徽	底	已	察	[illegible]	[illegible]						共	倒	低	依	似	世
5	早	放	撤	拒	掌	慶	快	往	[illegible]	己	審	[illegible]							[illegible]	係	做	你	介	之
4	易	敬	換	抗	手	慣	愉	患	戌											債	借	們	今	入
3	普	散	[illegible]		[illegible]	懷	急	形	[illegible]															也
2	映	[illegible]					[illegible]	[illegible]																

	24	23	22	21
	特用字		活用字（一）	
35	西	0		
34	東	9		
33	南	8		
32	北	7		
31	京	6	Я	У
30	縣	5	Ю	П
29	區	4	Э	Л
28	省	3	Ь	Й
27	市	2	Ы	И
26	政	1	Ъ	З
25	委	府	Ш	Ж
24	員	厘	Щ	Ё
23	會	斤	Ч	Д
22	民	寸	Ц	Г
21	人	尺	Ф	Б
20	國	央	Z	M
19	中	華	Y	L
18	，	、	廿	卅
	24	23	22	21
17	幣	第	X	K
16	角	號	W	J
15	元	年	V	I
14	萬	月	U	H
13	千	日	T	G
12	百	一	S	F
11	壹	二	R	E
10	貳	三	Q	D
9	叄	四	P	C
8	肆	五	O	B
7	伍	六	N	A
6	陸	七	No.	@
5	柒	八	$	‰
4	捌	九		
3	玖	十		
2	拾	〇		

	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1
					間			用				字								
	水	水	水	水	水	歹	欠	木	月	日	日	攴	手	手	手		戈	心	心	心
35	潘	湘	濵	淑	沈	母	欲	槽	朱	朗	昏	旦	擇	擱	抑	[illegible]	毛	慚	懶	惰
34	淦	滴	沸	汽	汗	毒	歎	植	榜	朝	暨	旗	挫	撫	拜	抱	扁	愧	恤	慢
33	湖	浸	溝	油	洶	毋	欺	棧	樂	曹	智	旋	撓	擋	擲	[illegible]	房	慧	您	忌
32	濕	淮	渠	漆	汰	殷	欣	棄	橫	曲	昭	旅	扶	撈	拋	找	戶	恭	憾	忱
31	張	滬	洩	漏	浙	毅	歐	柱	柔	晉	晒	族	撐	携	[illegible]	[illegible]	戴	慨	慷	忠
30	潮	洪	涵	洞	涼	殺	[illegible]	林	森	暑	景	旁	擦	捨	搭	[illegible]	截	愚	憤	忘
29	濾	液	泰	沿	沉	[illegible]	欄	桂	枉	曆	昆	斯	捲	拍	抄		戚	憶	惕	戀
28	濱	灣	港	洋	汪	殖	梁	柏	杜	暗	曠	斥	掃	攏	擺		[illegible]	愈	怒	恰
27	浦	滋	潰	滯	浩	殊	橋	木	李	昌	昔	斜	撲	拖	拉	抽	懲	悔	恥	惱
26	瀋	滾	漫	波	污	殘	架	桶	村	暴	晤	[illegible]	搗	搜	摧	扯	慈	悟	懈	怨
25	澤	演	測	漠	沙	死	杭	松	板	晨	昨	敦	捕	拘	捉	[illegible]	悲	悼	慌	恨
24	混	淆	淡	濃	浮	歧	檔	橡	梅	暇	旺	攸	揣	摸	抓	拔	慾	慘	怖	懸
23	涉	沖	漂	濫	泛	歪	杯	樹	楊	暢	旱	[illegible]	搞	拿	挾	[illegible]	惑	憐	怕	恕
22	渡	灌	溫	漁	湊	歲	柴	枝	柳	暖	旬	挽	搬	排	挑	[illegible]	憲	懼	恐	恒
21	[illegible]	[illegible]	漢	[illegible]	[illegible]	武	棵	東	[illegible]	曉	晶	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	愁	患	惟	忍
20	潛	潔	[illegible]	湧	洗	歌	楚	槍	未	晚	晴	擅	擁	揭	掩	[illegible]	憂	悶	惜	怠
19	游	渴	淨	洲	泥	歐	樸	染	朋	春	旨	捷	擾	掛	控	[illegible]	懂	惡	慕	怪
18	／	¥	%	∅	#	m/m	( )	+	−	×	1.	2.	3.	4.	5.				[illegible]	0.
	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6					
17	征	庶	帝	屈	寧	季	威	契	堤	壓	嚇	吞	史	午	务					
16	徑	庫	幅	履	[illegible]	孤	嫌	奈	塔	域	唱	吐	吉	卑	劫					
15	徽	庸	帳	屢				[illegible]	堆	培	嗣	叫	呆	[illegible]	[illegible]					
14	循	座	幕	屠					[illegible]	坦	喪	喚	呈	匠	[illegible]					
13	徒	床	巷	屋						[illegible]	[illegible]	喊	叢	匪						
12	役	廳	巳	尾																
11	微	廈	巨	屆																
10	徐	廐	巧	屑																
9	彩	庭	川	尤																
8	弟	庇	山	尖																
7	彎	康	崗	射																
6	彌	幼	崩	封																
5	彈	幻	崇	尋																
4	弛	巾	峰	尊																
3	弔	帖	島	宿																
2	弊	帽	岸	[illegible]																

中文打字機基本字盤表

常用字
間用字
活用字
特用字
談話說
講識認
央華
國中
汽油漆漏洞

What I would like to cover…

My learnings along the road of AI models, apps, and infra

- New algorithms continue to grow LLM algorithms.
- Application space thrives.
- AI Infra has become the 3rd pillar of enterprise IT strategy.
- … and interestingly, we see history repeat itself.

- If you would like - ups and downs of a startup.

- Disclaimer: all opinions are of my own and do not represent my company or any institute’s position.

1
New algorithms continue to drive LLM models.

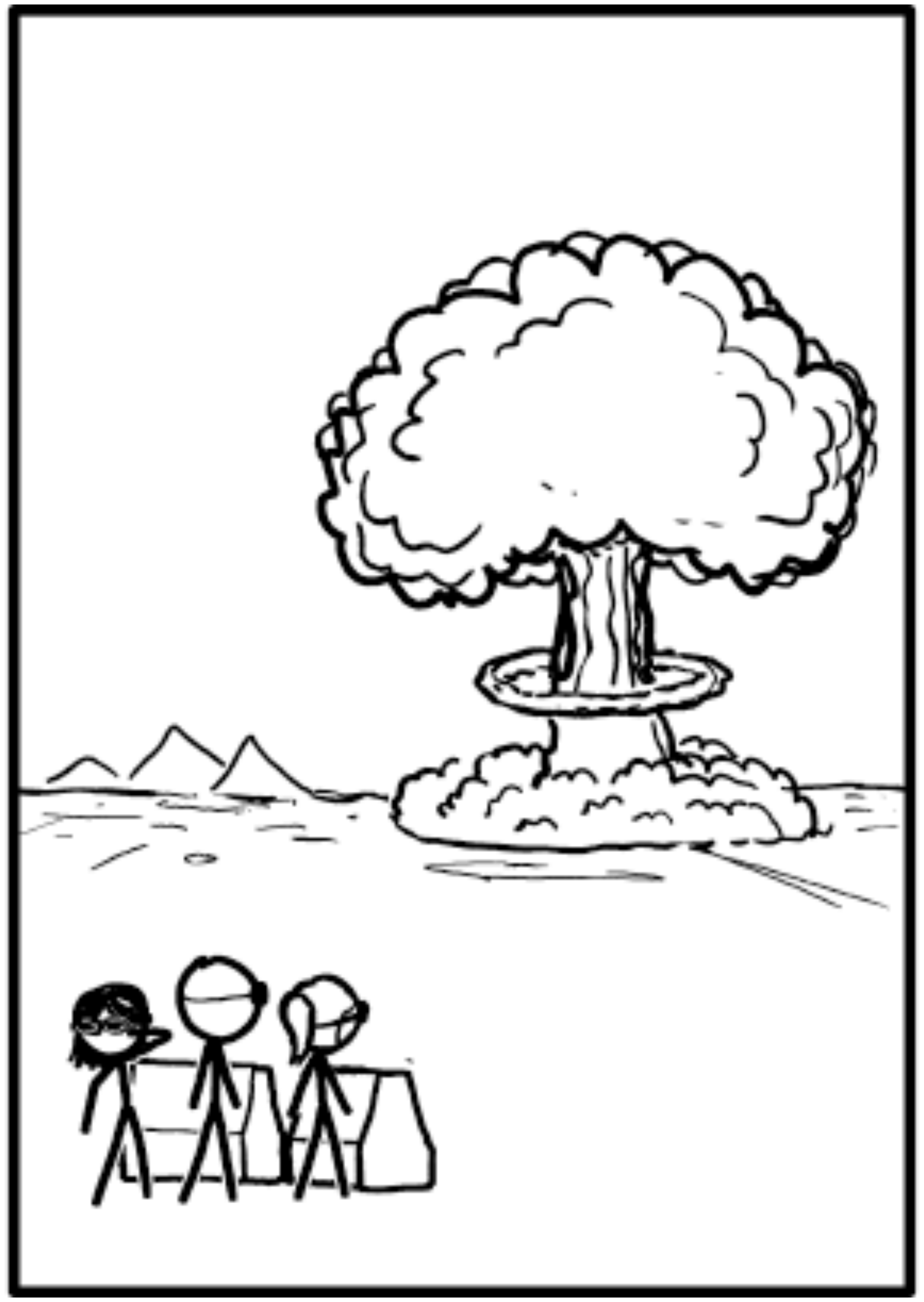
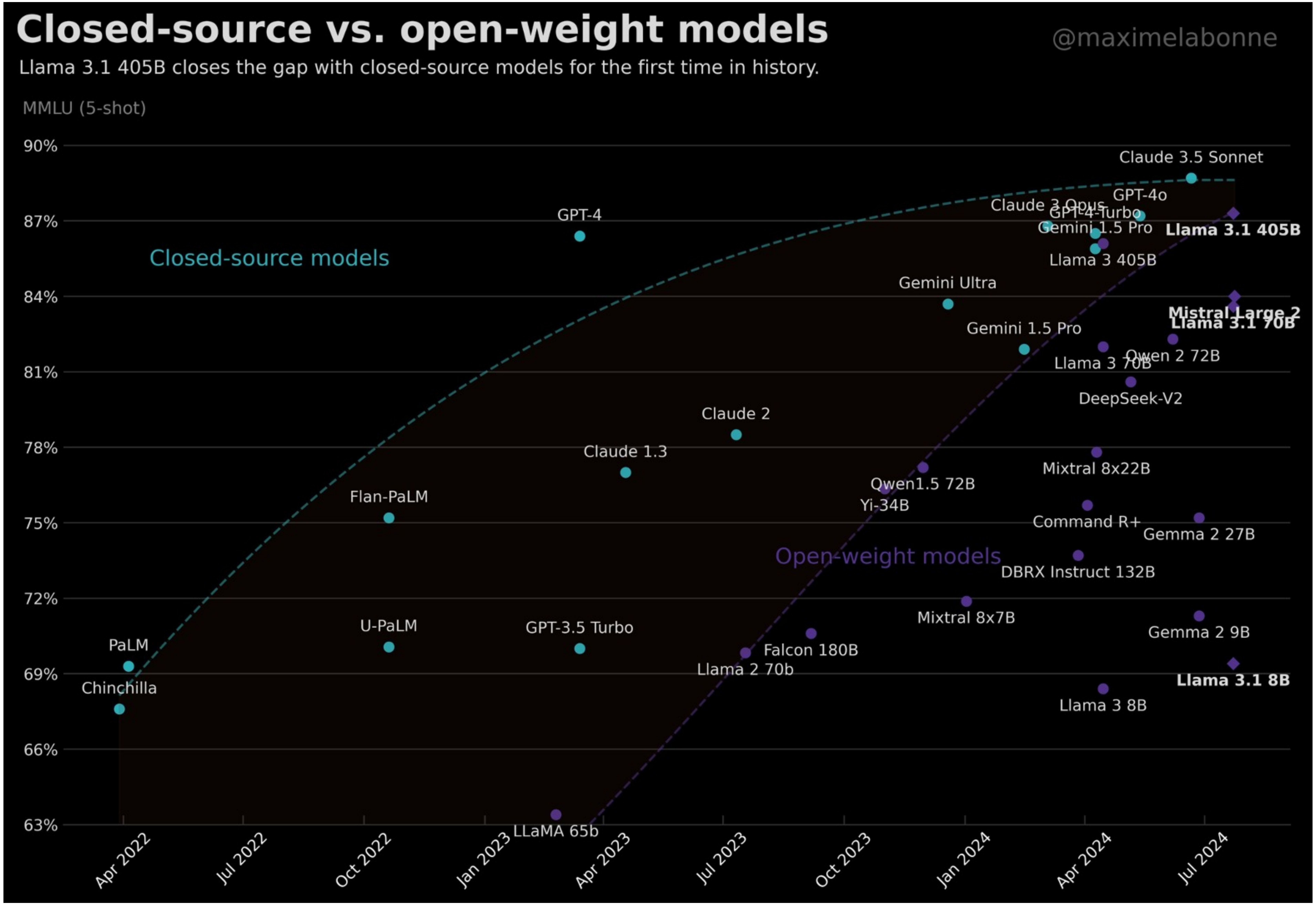


Image source: xkcd



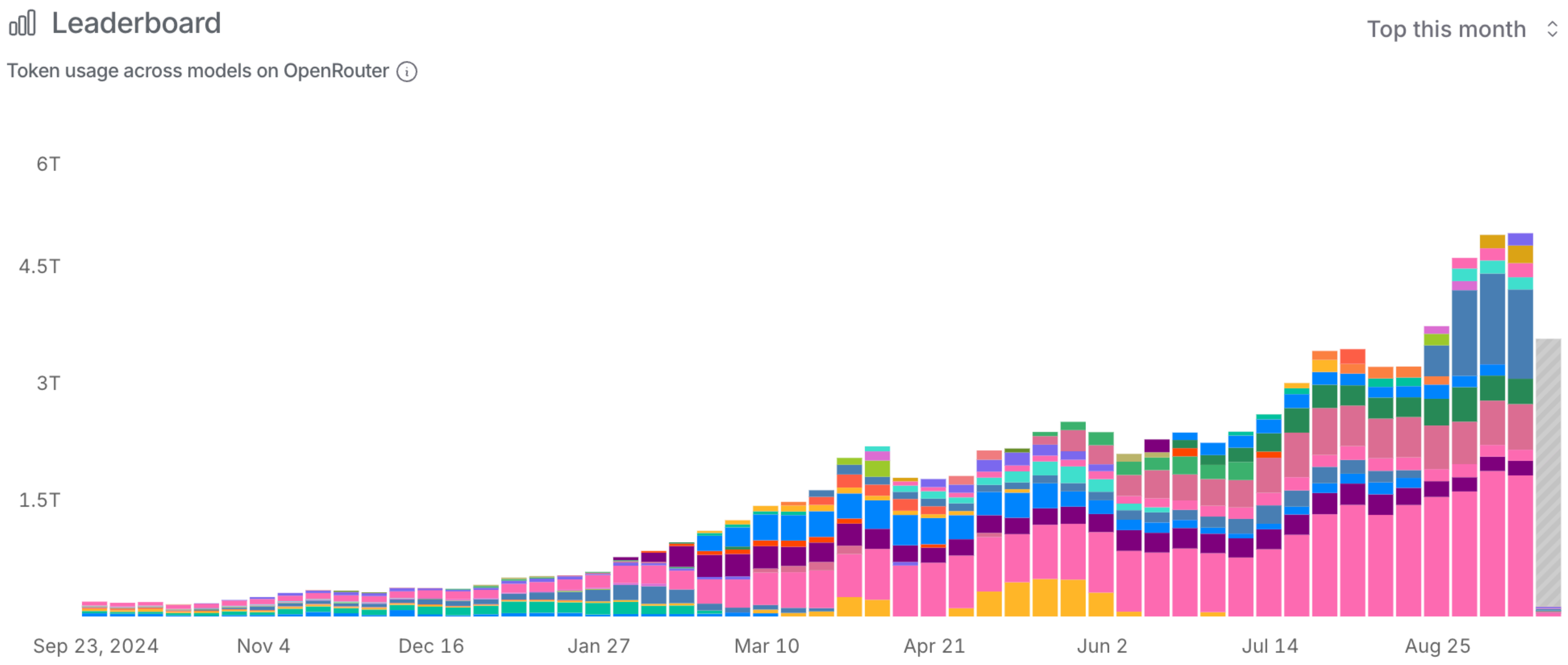
Closed-source vs. open-weight models

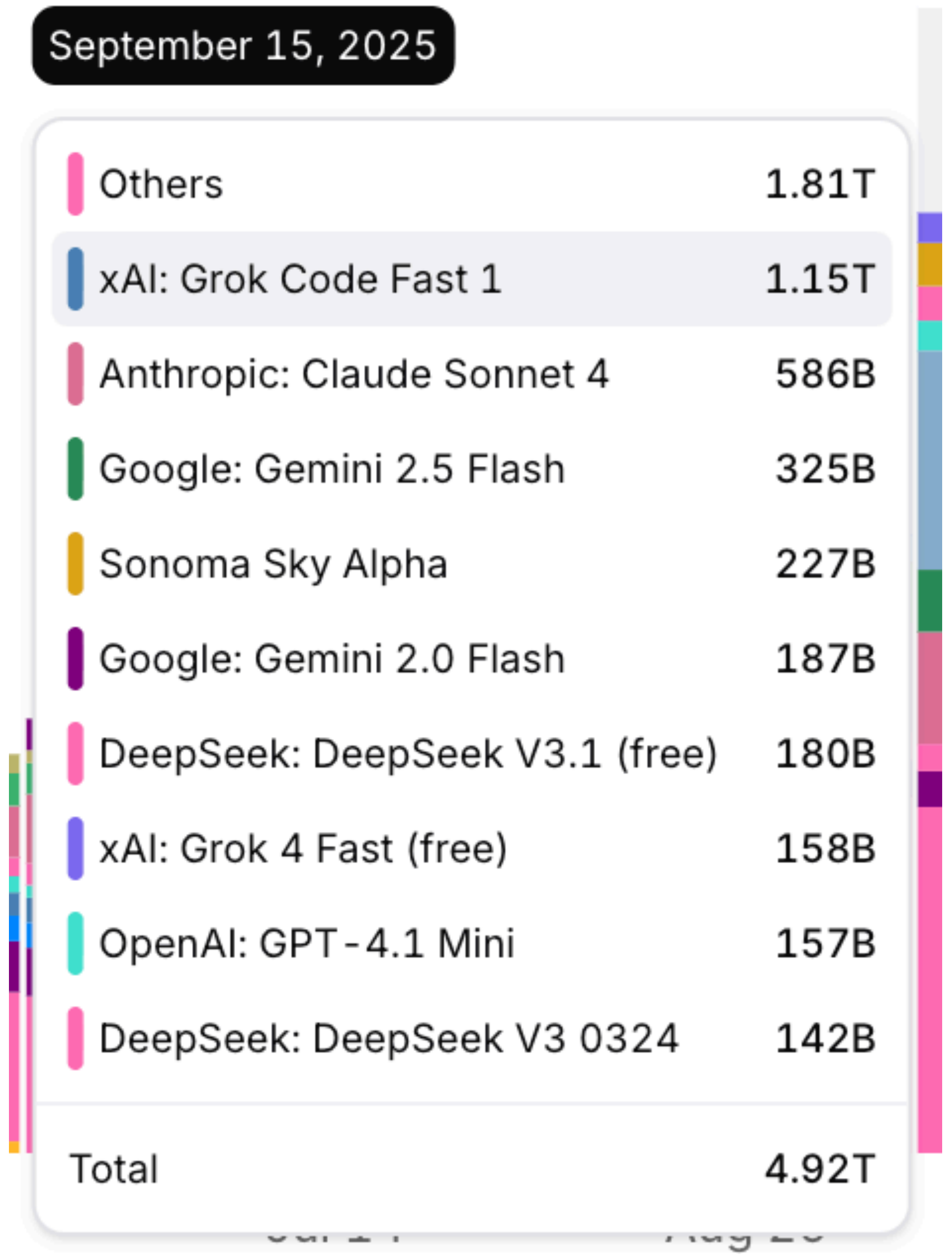
Llama 3.1 405B closes the gap with closed-source models for the first time in history.

- And the latest…
  - GPT-5
  - Grok
  - Gemini
  - Kimi
  - DeepSeek
  - Qwen
  - GPT-OSS

There does not seems to be “a bubble”

Consumption continue growing - not only model training.




September 15, 2025

Others	1.81T
xAI: Grok Code Fast 1	1.15T
Anthropic: Claude Sonnet 4	586B
Google: Gemini 2.5 Flash	325B
Sonoma Sky Alpha	227B
Google: Gemini 2.0 Flash	187B
DeepSeek: DeepSeek V3.1 (free)	180B
xAI: Grok 4 Fast (free)	158B
OpenAI: GPT-4.1 Mini	157B
DeepSeek: DeepSeek V3 0324	142B
Total	4.92T

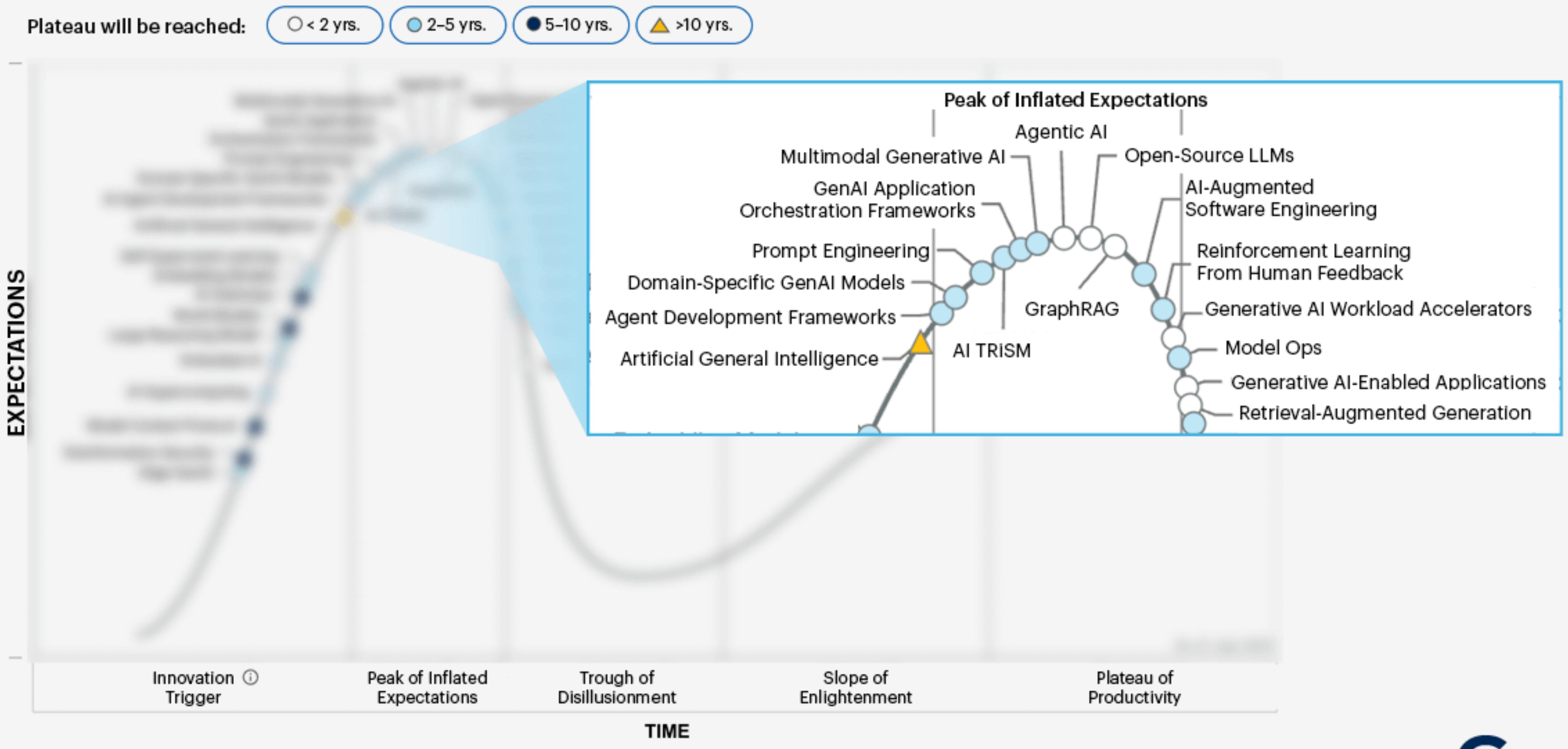
Source: openrouter.ai

New Algorithms Drive Continued Improvements

My personal opinion and historical analogies…

Date	Nov 2022	Dec 2023	Sep 2024	Jan 2025 (and earlier)
Algorithm	GPT (3.5)	MoE (Mixtral 8x7B)	Test time Scaling	**Reinforcement Learning**
Analogies	AlexNet (structural innovation)	Ensemble Learning Inception/ResNet	Fully convolutional network Multi-instance learning	General RL GANs

Hype Cycle for Generative AI, 2025



Source: Gartner

2
ToC apps thrive.
ToB apps are hopeful & nascent.

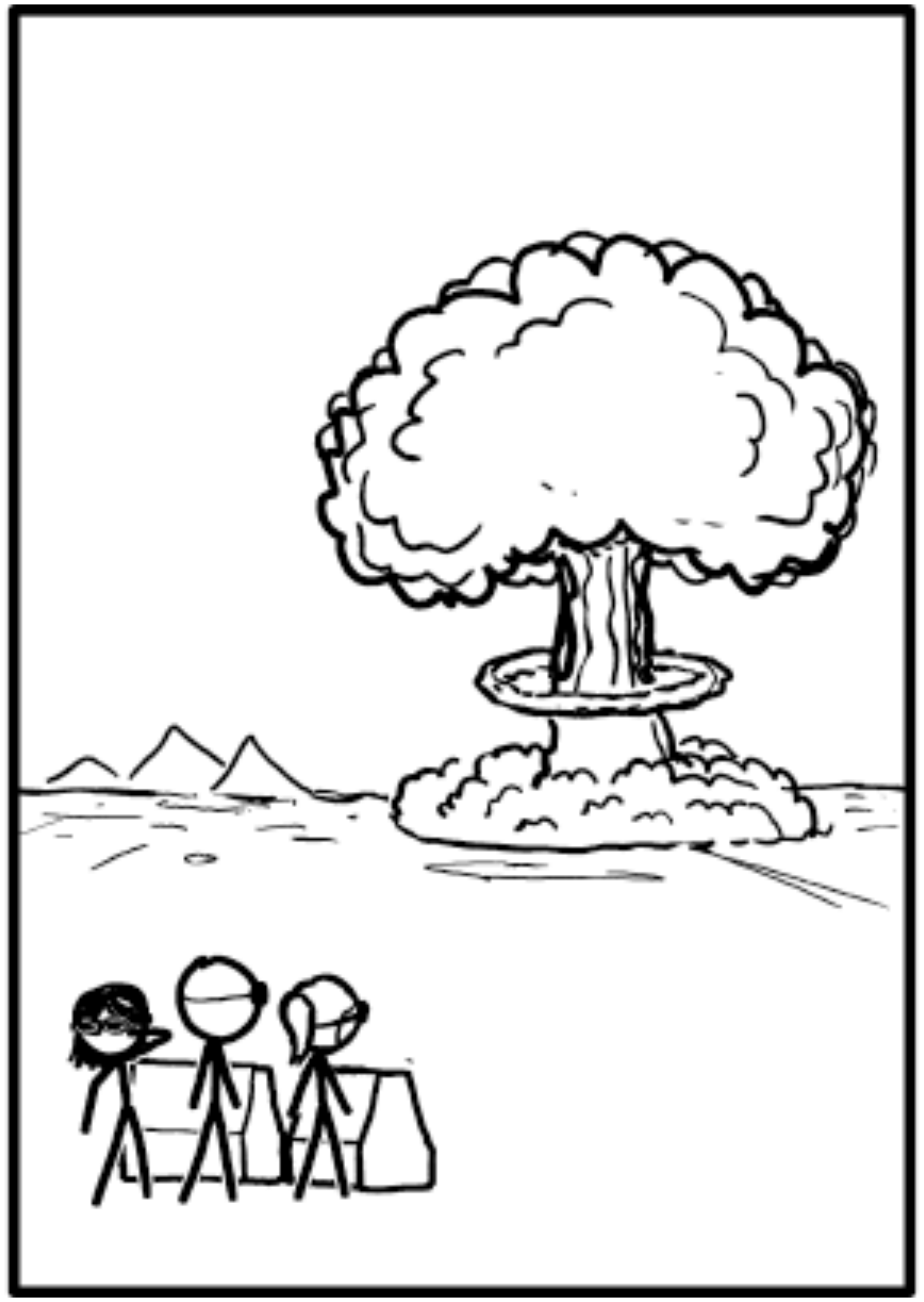


Image source: xkcd

Perfect app experience is correlated, but independent from models.

https://elmo.chat/

Elmo is your AI companion to create summaries and insights

Featured on Chrome Web Store

Developed by Established Publisher

Install from Chrome Web Store

Free / No Account Required / Supports Multiple Languages

Work seamlessly across HTML, Youtube, PDF, and Google Docs.

Guillermo Rauch

This looks really handy

Bing Xu

Elmo is the co-pilot for Chrome. Super helpful when reading new

Diogo Santos

Elmo is your AI extension 🧠 for Chrome to create summaries, insights, and extended knowledge. What does Elmo offer? 😳 ✓ Summaries and highlights; ✓ Keep asking questions; ✓ Dive deep into keywords; ✓ Chat with PDFs; ✓

Alvin-GenAI

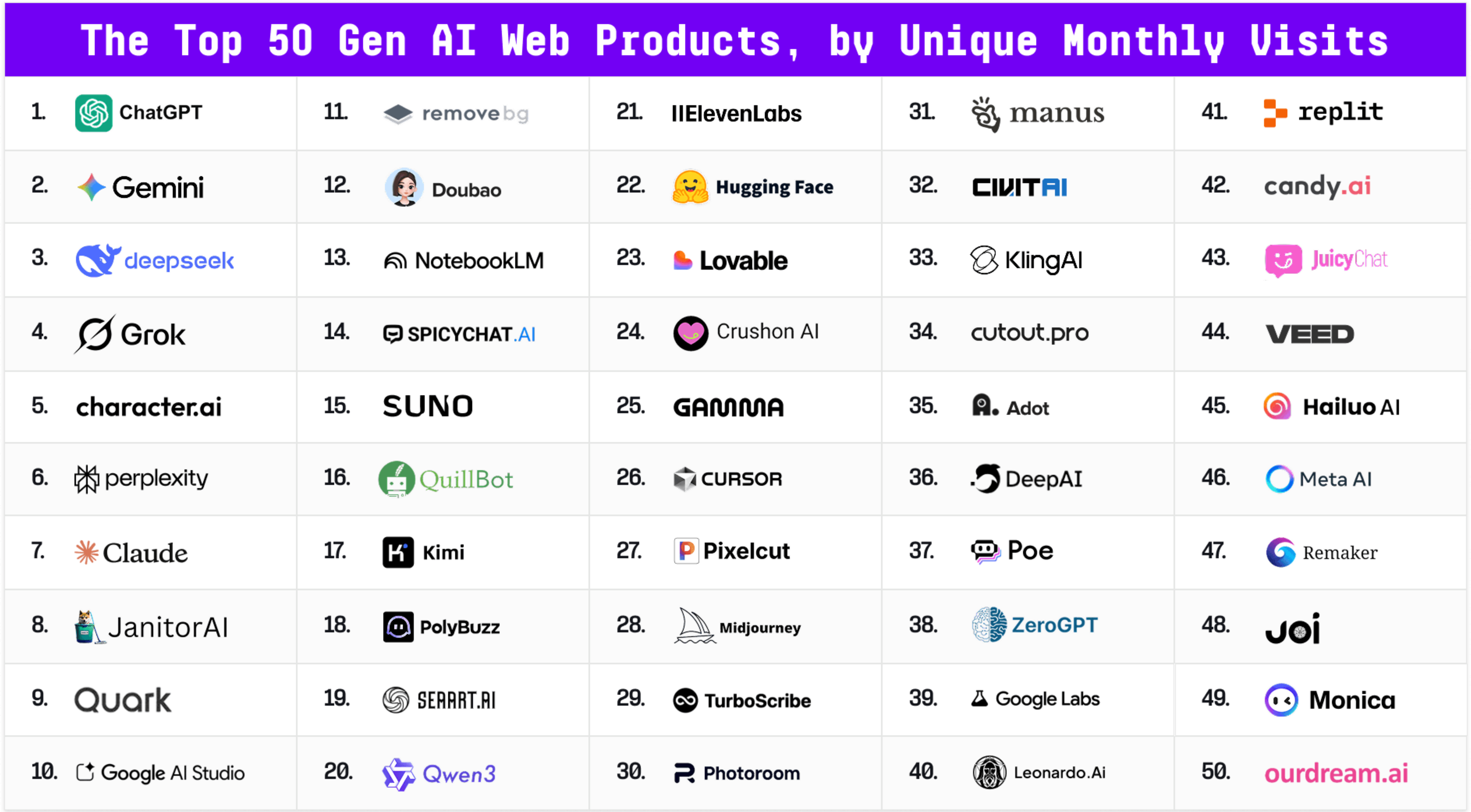
I've downloaded https://elmo.chat an AI Chrome extension to create summaries, insights and extended knowledge. It does - summarizes your websites right next to the page, - summarizes PDF files too, - summarizes

Tulsi Prasad

I tried out this AI chrome extension yesterday! 1. No product hunt launches 2. Featured on Chrome Web Store like an OG 3. No signup needed, just plug and play! Works pretty accurate on blogs and textual content, but videos are

Consumer App Landscape is Highly Fluid

Due to the continued improvement of foundational models



The Top 50 Gen AI Web Products, by Unique Monthly Visits

1. ChatGPT	11. removebg	21. IIElevenLabs	31. manus	41. replit
2. Gemini	12. Doubao	22. Hugging Face	32. CIVITAI	42. candy.ai
3. deepseek	13. NotebookLM	23. Lovable	33. KlingAI	43. JuicyChat
4. Grok	14. SPICYCHAT.AI	24. Crushon AI	34. cutout.pro	44. VEED
5. character.ai	15. SUNO	25. GAMMA	35. Adot	45. Hailuo AI
6. perplexity	16. QuillBot	26. CURSOR	36. DeepAI	46. Meta AI
7. Claude	17. Kimi	27. Pixelcut	37. Poe	47. Remaker
8. JanitorAI	18. PolyBuzz	28. Midjourney	38. ZeroGPT	48. Joi
9. Quark	19. SEAART.AI	29. TurboScribe	39. Google Labs	49. Monica
10. Google AI Studio	20. Qwen3	30. Photoroom	40. Leonardo.Ai	50. ourdream.ai

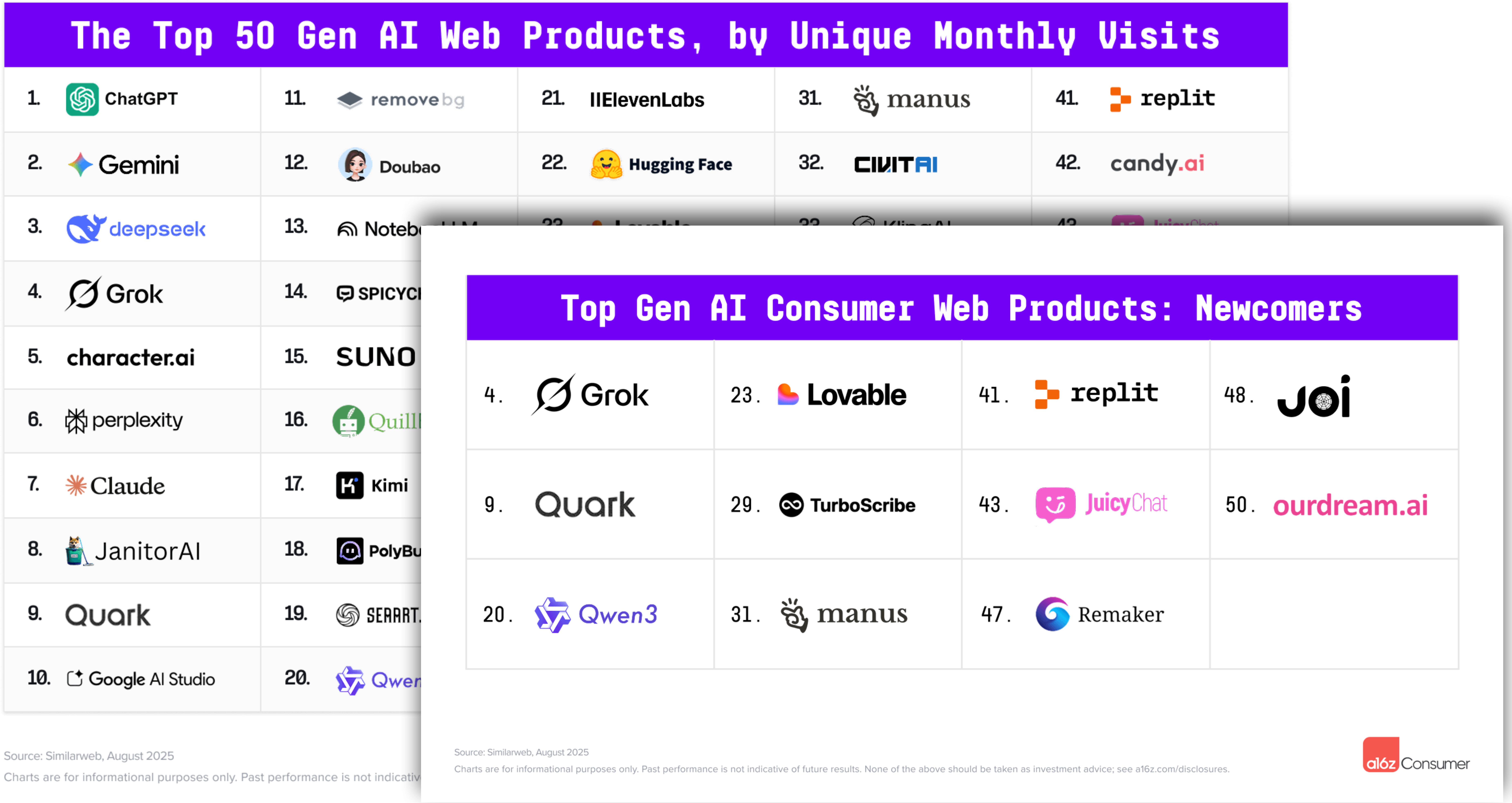
Source: Similarweb, August 2025

Charts are for informational purposes only. Past performance is not indicative of future results. None of the above should be taken as investment advice; see a16z.com/disclosures.

a16z Consumer

Consumer App Landscape is Highly Fluid

Due to the continued improvement of foundational models



The Top 50 Gen AI Web Products, by Unique Monthly Visits

1. ChatGPT	11. removebg	21. IIElevenLabs	31. manus	41. replit
2. Gemini	12. Doubao	22. Hugging Face	32. CIVITAI	42. candy.ai
3. deepseek	13. Noteb			
4. Grok	14. SPICYC			
5. character.ai	15. SUNO			
6. perplexity	16. Quill			
7. Claude	17. Kimi			
8. JanitorAI	18. PolyBu			
9. Quark	19. SEAART.			
10. Google AI Studio	20. Qwen			

Source: Similarweb, August 2025

Charts are for informational purposes only. Past performance is not indicativ

Top Gen AI Consumer Web Products: Newcomers

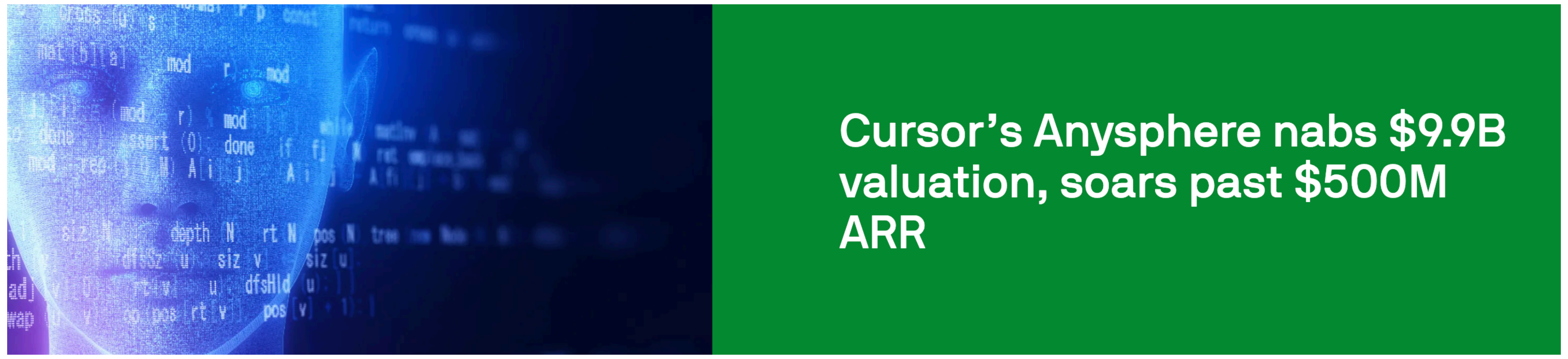
4. Grok	23. Lovable	41. replit	48. Joi
9. Quark	29. TurboScribe	43. JuicyChat	50. ourdream.ai
20. Qwen3	31. manus	47. Remaker	

Source: Similarweb, August 2025

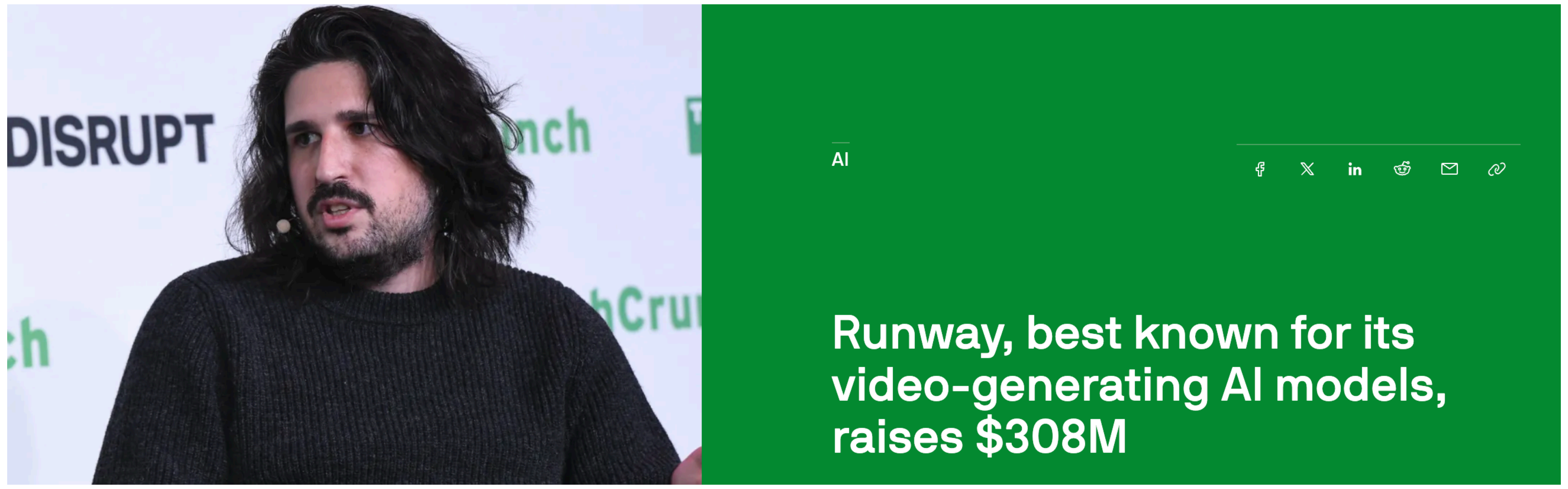
Charts are for informational purposes only. Past performance is not indicative of future results. None of the above should be taken as investment advice; see a16z.com/disclosures.

a16z Consumer

Prosumers: willingness to pay drives revenue

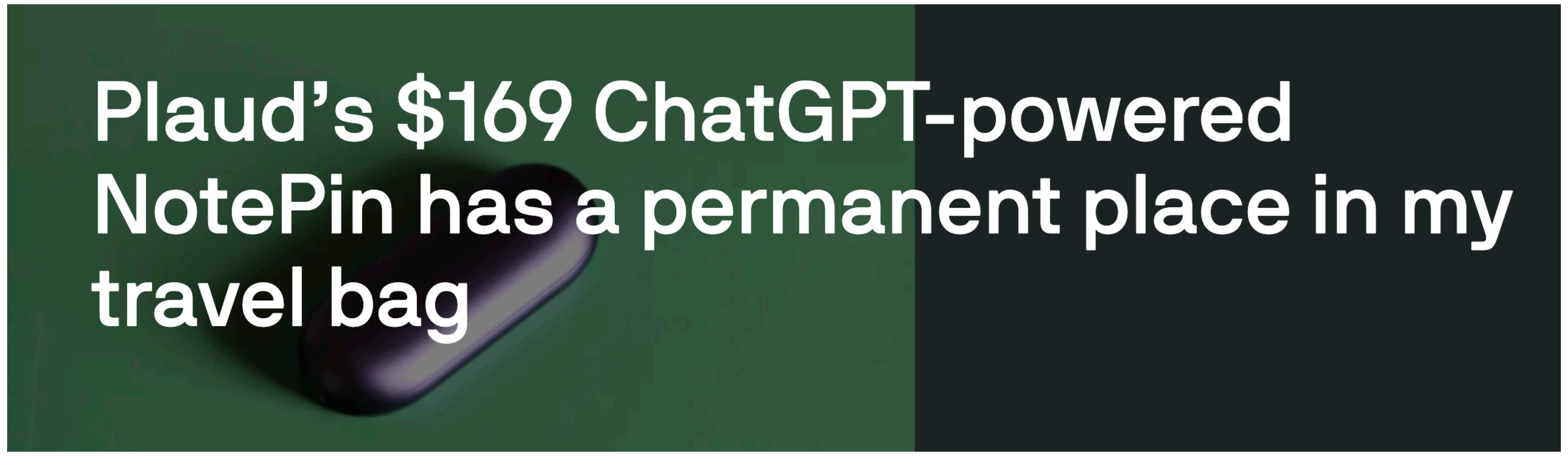


Cursor’s Anysphere nabs $9.9B valuation, soars past $500M ARR

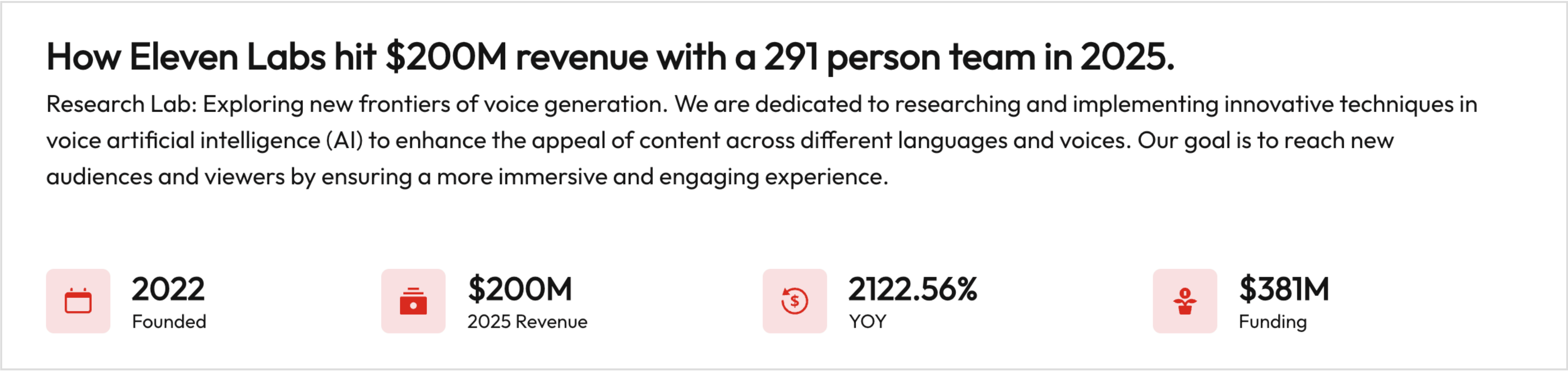


AI

Runway, best known for its video-generating AI models, raises $308M



Plaud’s $169 ChatGPT-powered NotePin has a permanent place in my travel bag



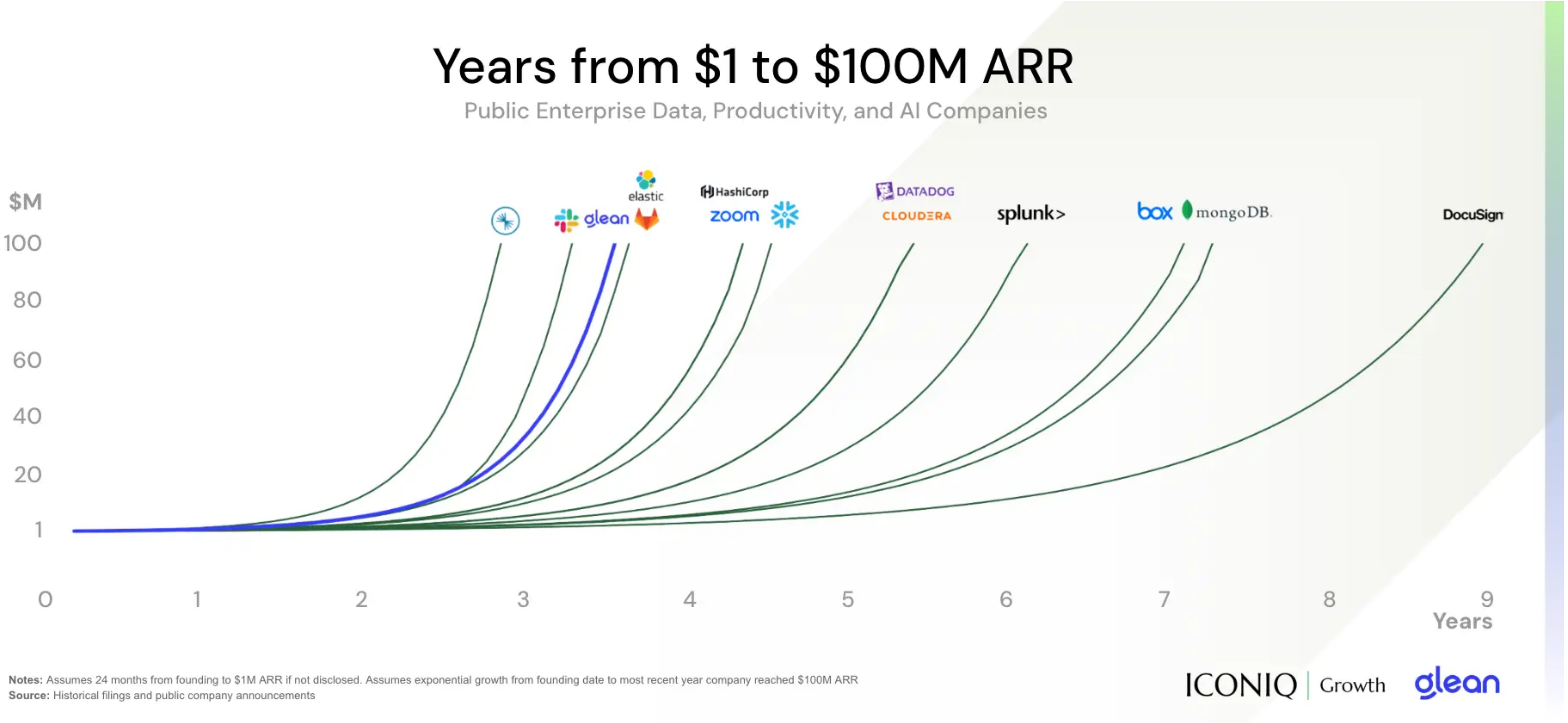
How Eleven Labs hit $200M revenue with a 291 person team in 2025.

Research Lab: Exploring new frontiers of voice generation. We are dedicated to researching and implementing innovative techniques in voice artificial intelligence (AI) to enhance the appeal of content across different languages and voices. Our goal is to reach new audiences and viewers by ensuring a more immersive and engaging experience.

2022	$200M	2122.56%	$381M
Founded	2025 Revenue	YOY	Funding

Still, there may be a big enterprise market

Enterprises used to be slow. Now faster.

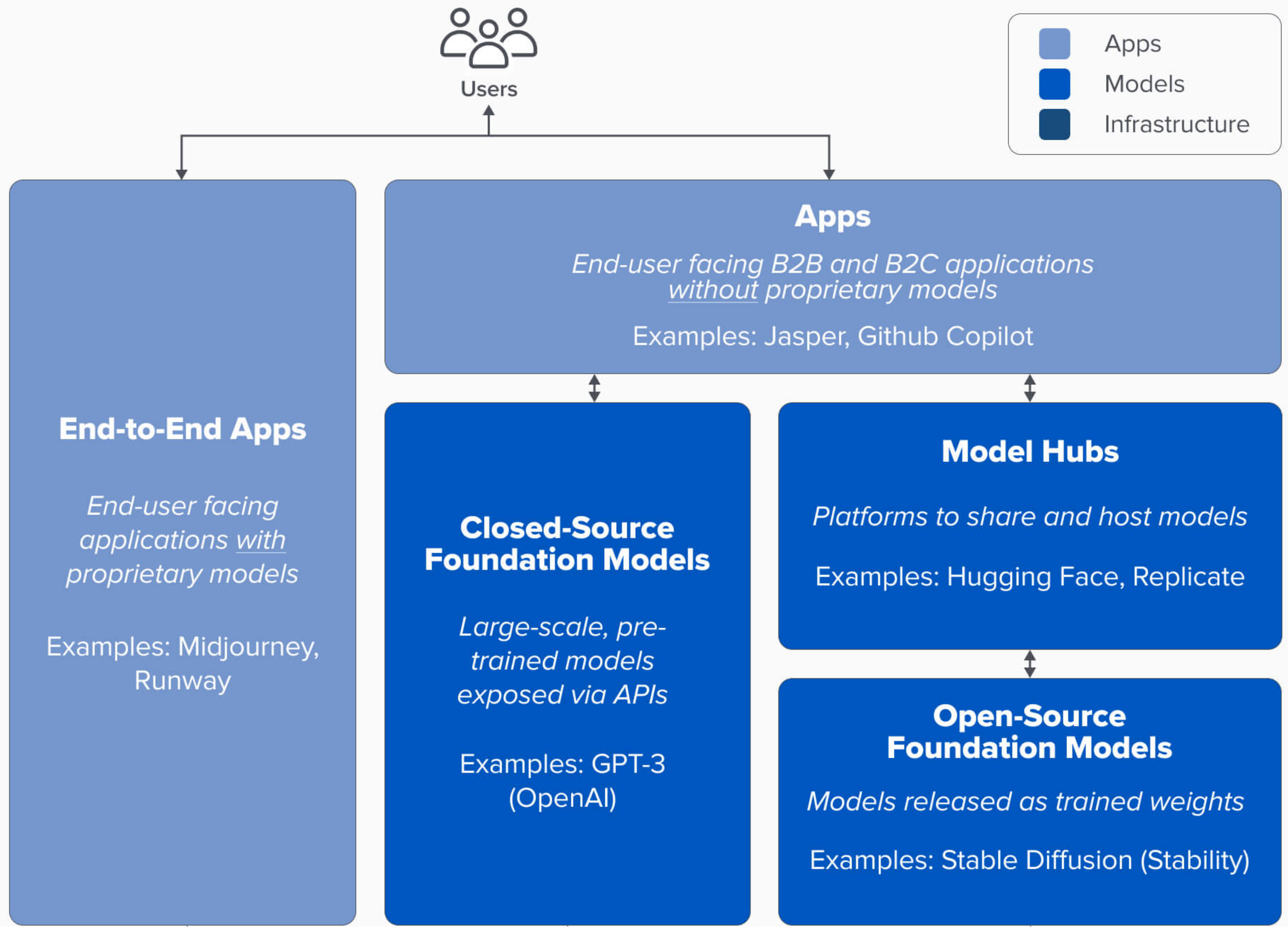

Years from $1 to $100M ARR

Public Enterprise Data, Productivity, and AI Companies

Notes: Assumes 24 months from founding to $1M ARR if not disclosed. Assumes exponential growth from founding date to most recent year company reached $100M ARR

Source: Historical filings and public company announcements

Preliminary generative AI tech stack



Users

🟦 Apps
🟦 Models
🟦 Infrastructure

Apps

End-user facing B2B and B2C applications without proprietary models

Examples: Jasper, Github Copilot

End-to-End Apps

End-user facing applications with proprietary models

Examples: Midjourney, Runway

Closed-Source Foundation Models

Large-scale, pre-trained models exposed via APIs

Examples: GPT-3 (OpenAI)

Model Hubs

Platforms to share and host models

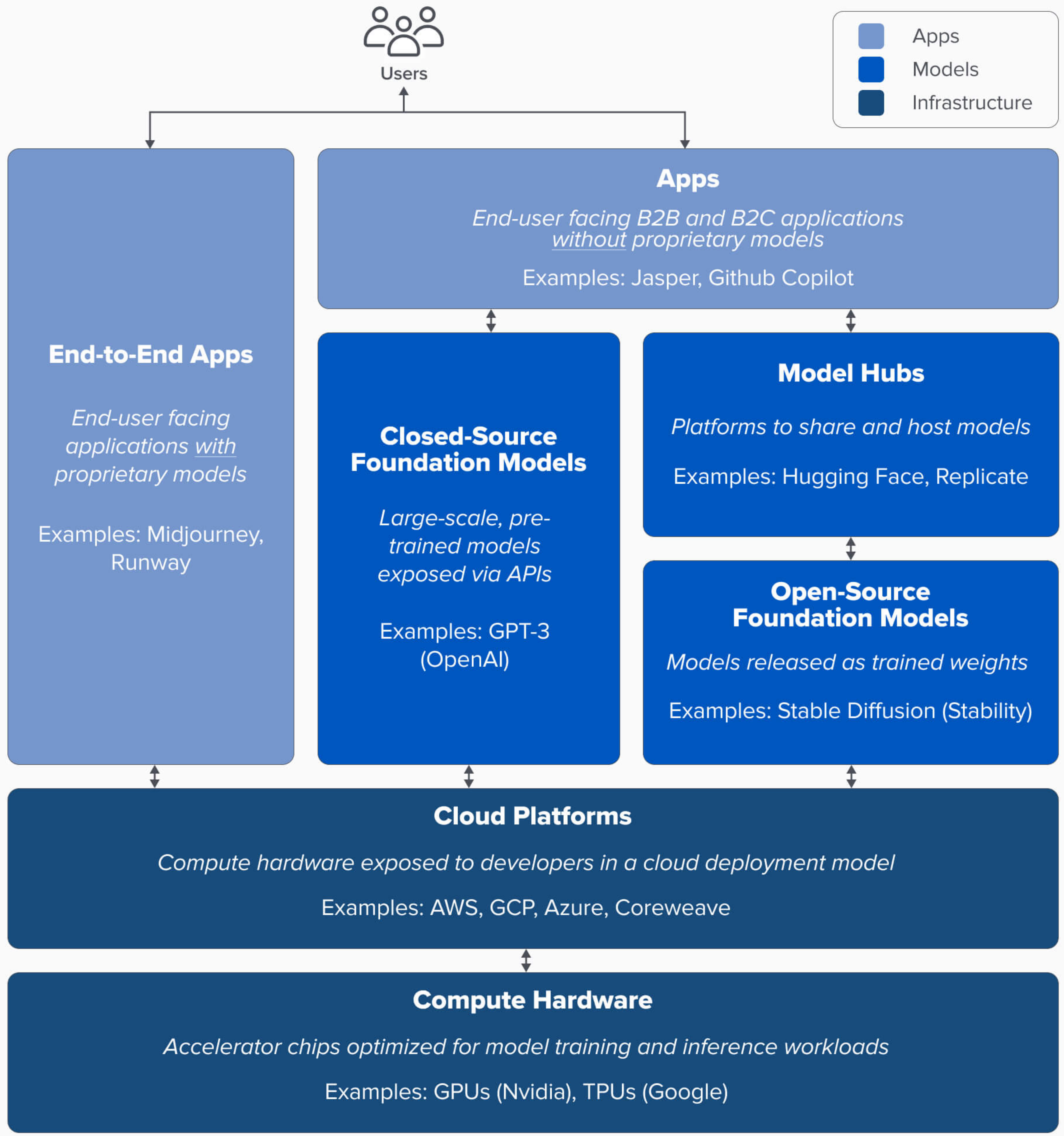
Examples: Hugging Face, Replicate

Open-Source Foundation Models

Models released as trained weights

Examples: Stable Diffusion (Stability)

Preliminary generative AI tech stack



Users

- 🟦 Apps
- 🟦 Models
- 🟦 Infrastructure

Apps

End-user facing B2B and B2C applications without proprietary models

Examples: Jasper, Github Copilot

End-to-End Apps

End-user facing applications with proprietary models

Examples: Midjourney, Runway

Closed-Source Foundation Models

Large-scale, pre-trained models exposed via APIs

Examples: GPT-3 (OpenAI)

Model Hubs

Platforms to share and host models

Examples: Hugging Face, Replicate

Open-Source Foundation Models

Models released as trained weights

Examples: Stable Diffusion (Stability)

Cloud Platforms

Compute hardware exposed to developers in a cloud deployment model

Examples: AWS, GCP, Azure, Coreweave

Compute Hardware

Accelerator chips optimized for model training and inference workloads

Examples: GPUs (Nvidia), TPUs (Google)

3
AI infra is the 3rd pillar in IT strategy.

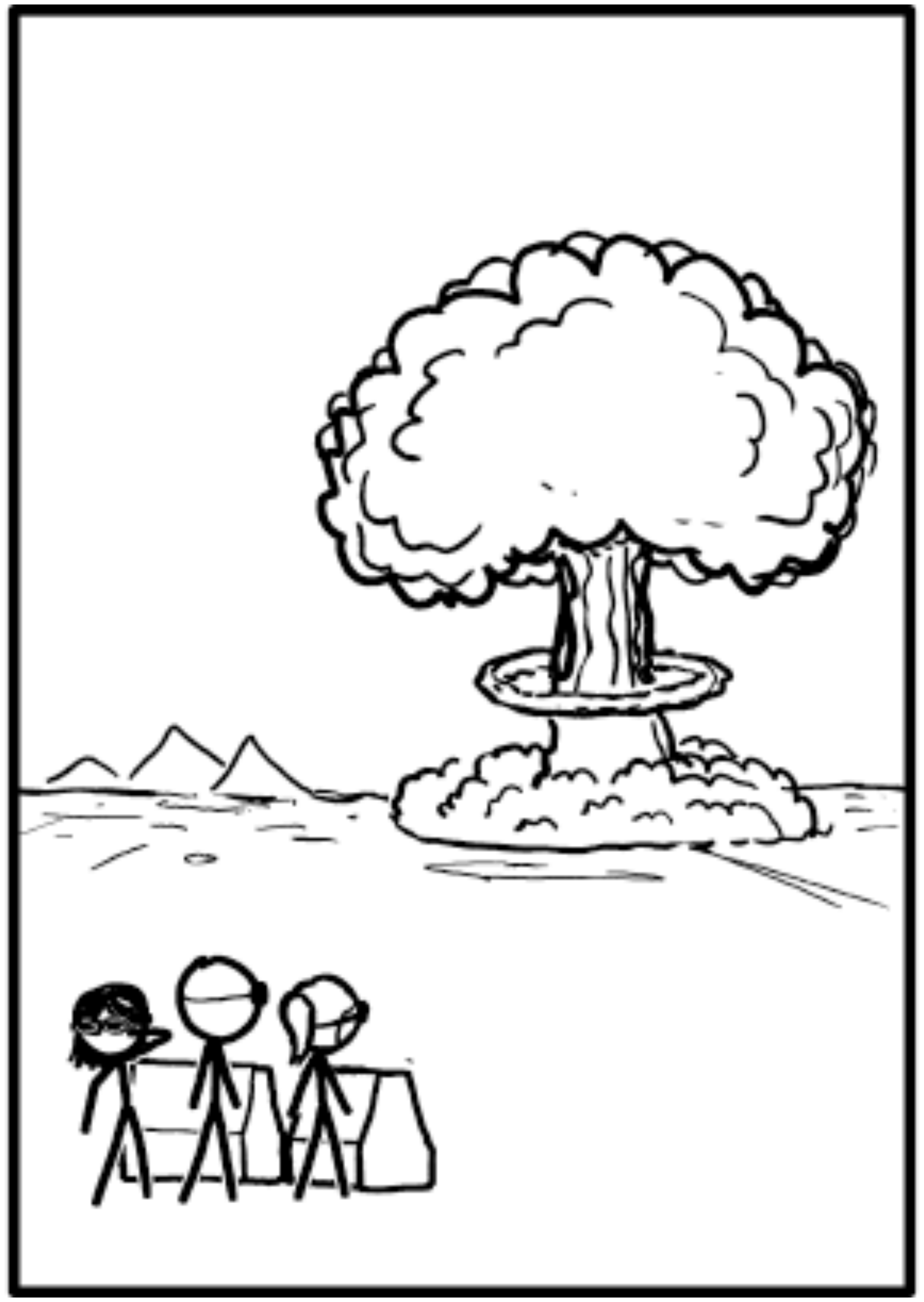
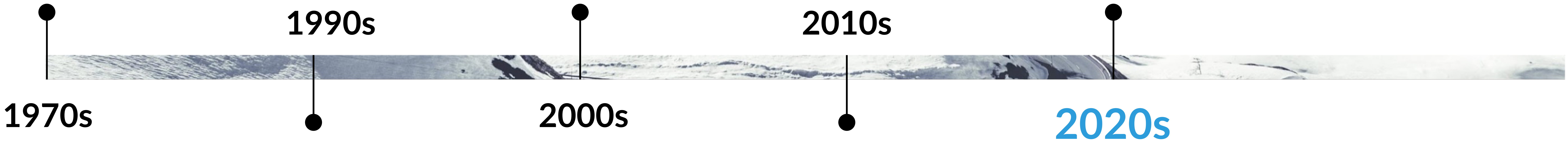


Image source: xkcd

“The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin.”

- Professor Richard Sutton, “The Bitter Lesson”

The Third Pillar



1970s

Scientific Computing

Computers are used for large-scale physics / weather simulations. Large clusters of scientific computing machines.

1990s

Virtual Private Servers

First application of managed machines - but still limited offering of software and applications on top of the machines.

2000s

“Web Service Cloud”

Spearheaded by Amazon Web Services, the public cloud does a great job in web serving: moving data around, like webpages / images / videos.

2010s

“Data Cloud”

E-commerce has called for the processing power of exabyte scale data. Snowflake and Databricks emerged and grew into unicorns eventually.

2020s

Need for AI Cloud

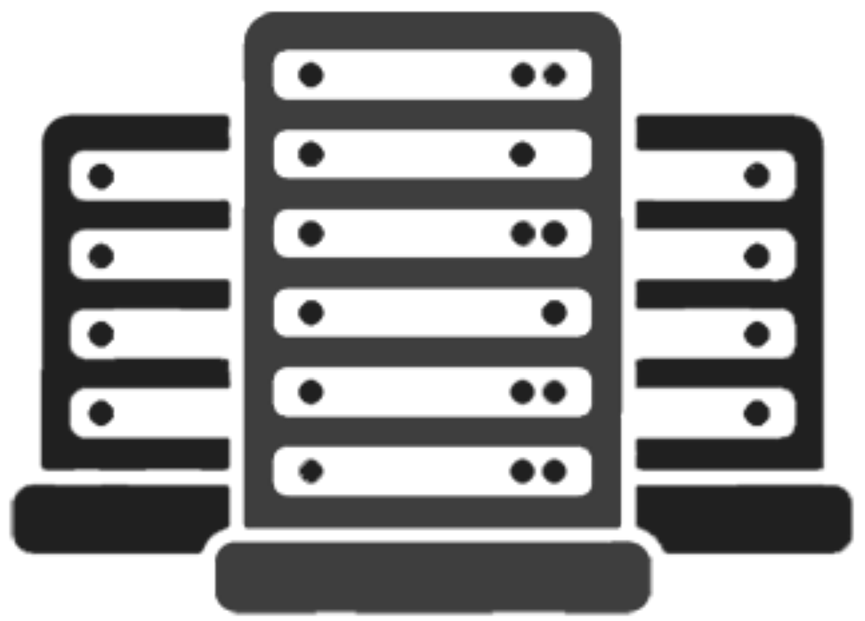
Modern AI applications call for exaflops computation power, over high performance, heterogeneous and cloud native infra.
This has never been seen in the history of cloud computing.

Why? AI is different from conventional compute

Data Compute

- IO >> compute
- Simple abstraction
- Very Distributed Systems

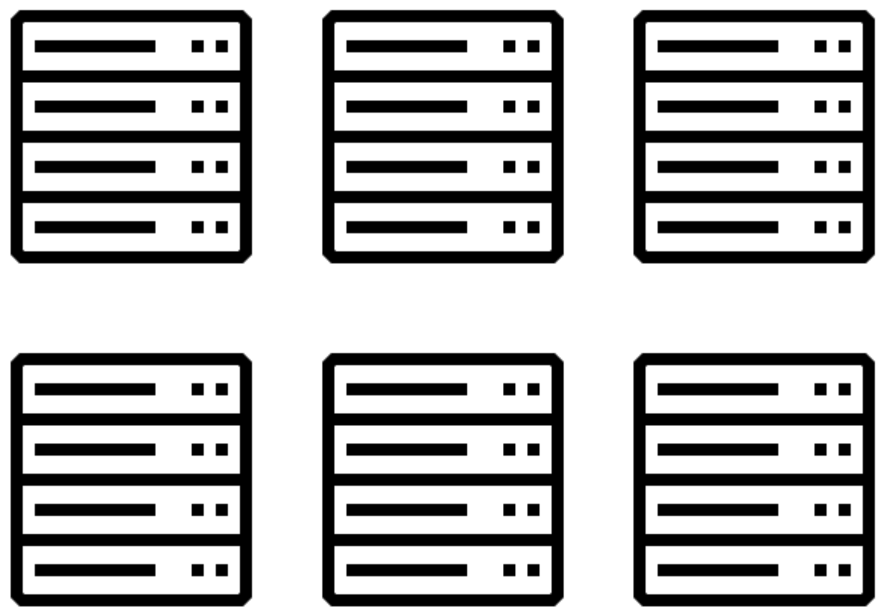
Easy to use
Hard for infra



Web Services

- IO > compute
- Arbitrary code
- Embarrassingly Parallel system

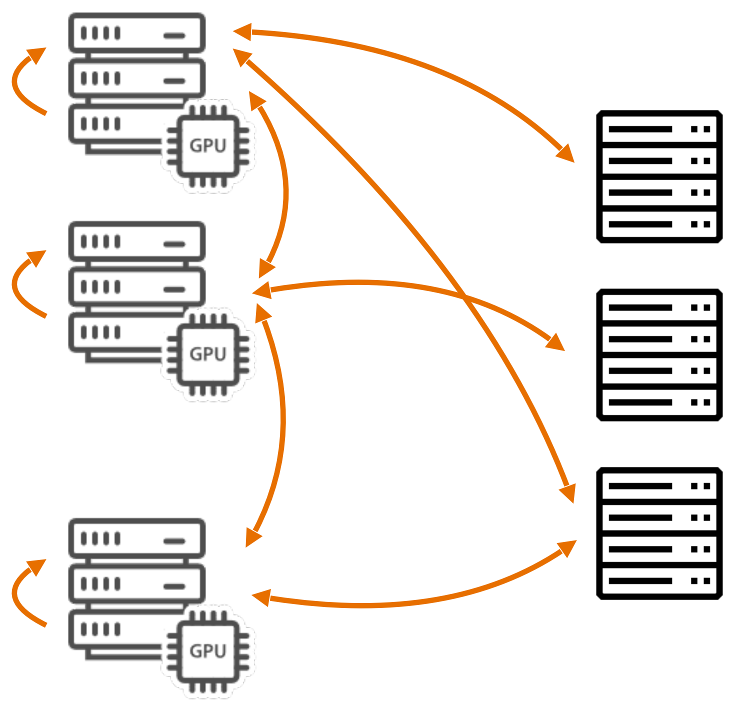
(Kinda) easy to use
(Kinda) easy for infra



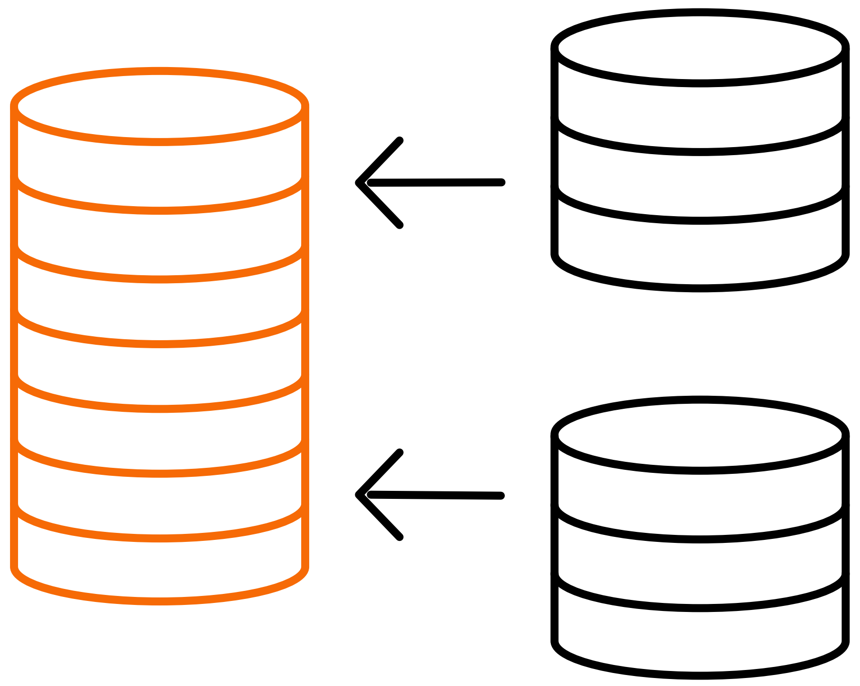
AI Compute

- Compute >> IO
- Arbitrary code
- Very Distributed Systems

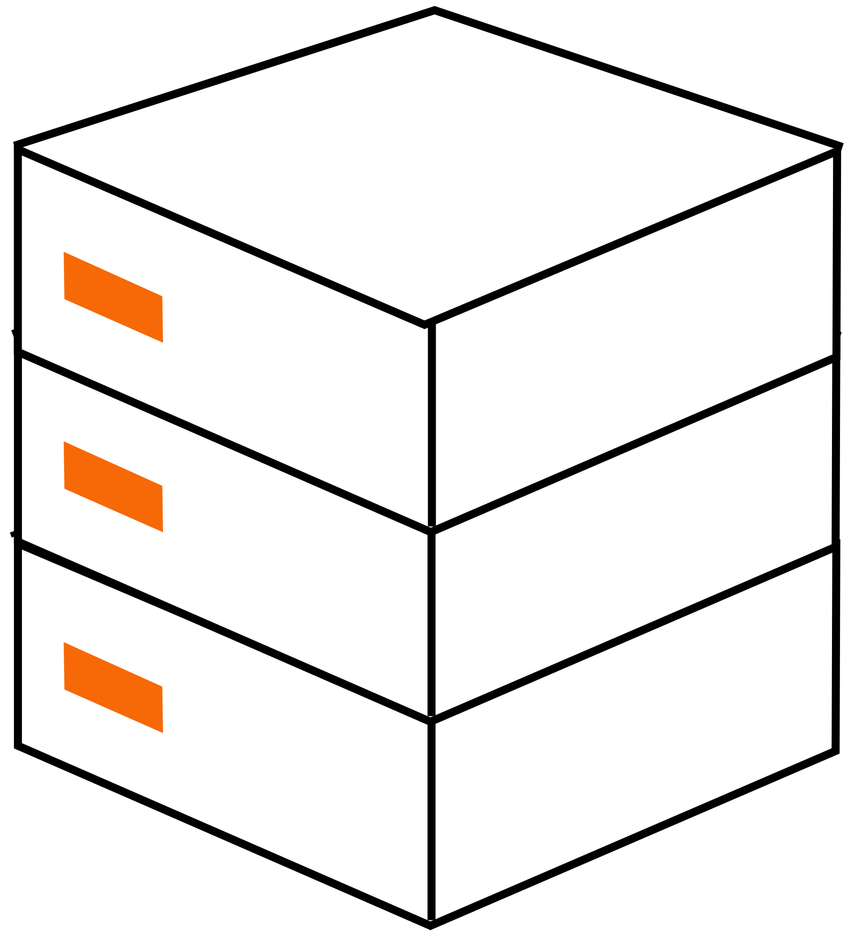
(Pretty) hard to use
(Pretty) hard for infra



Conventional cloud value proposition no longer holds…



SOFTWARE

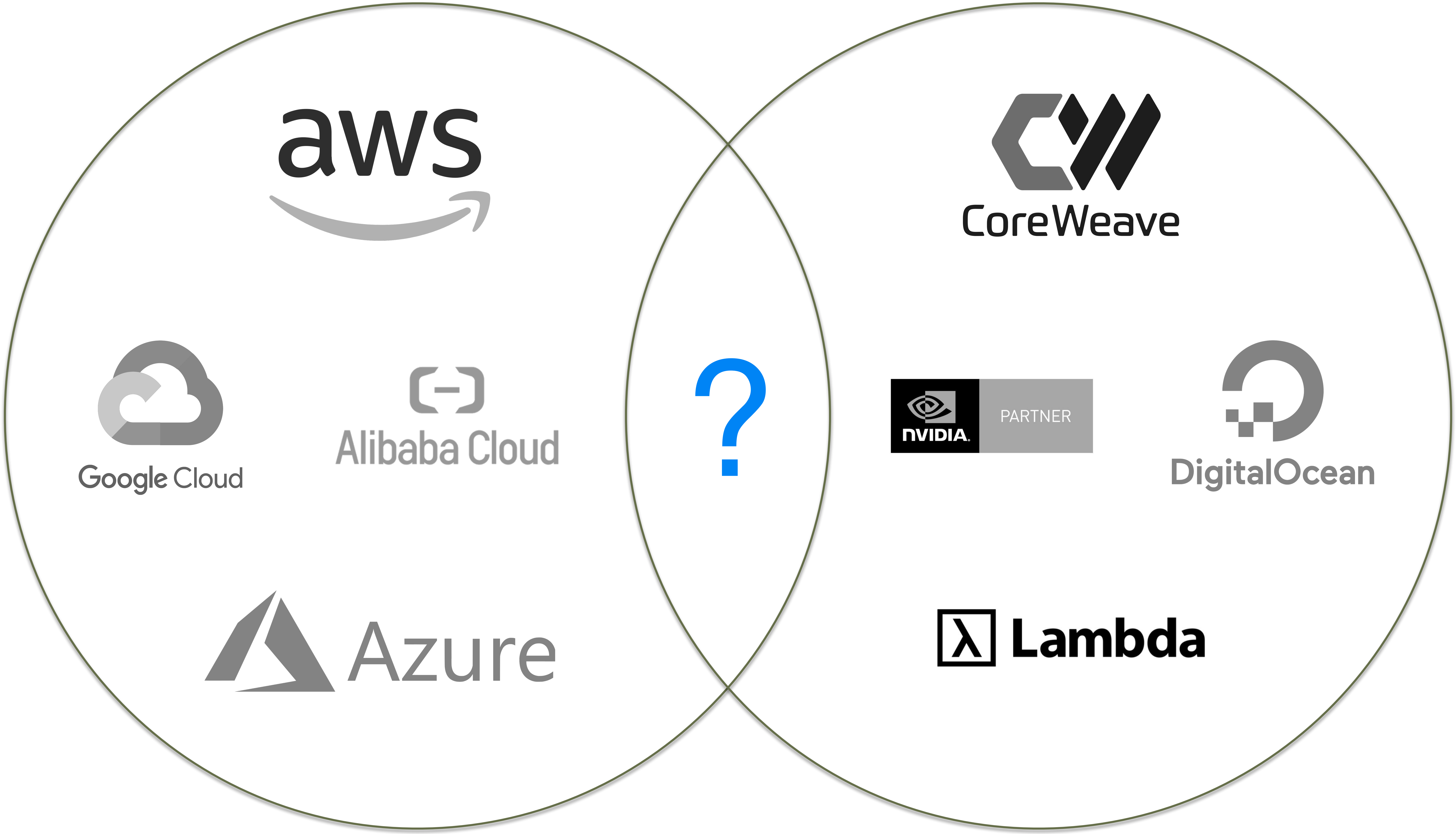


SUPPLY CHAIN

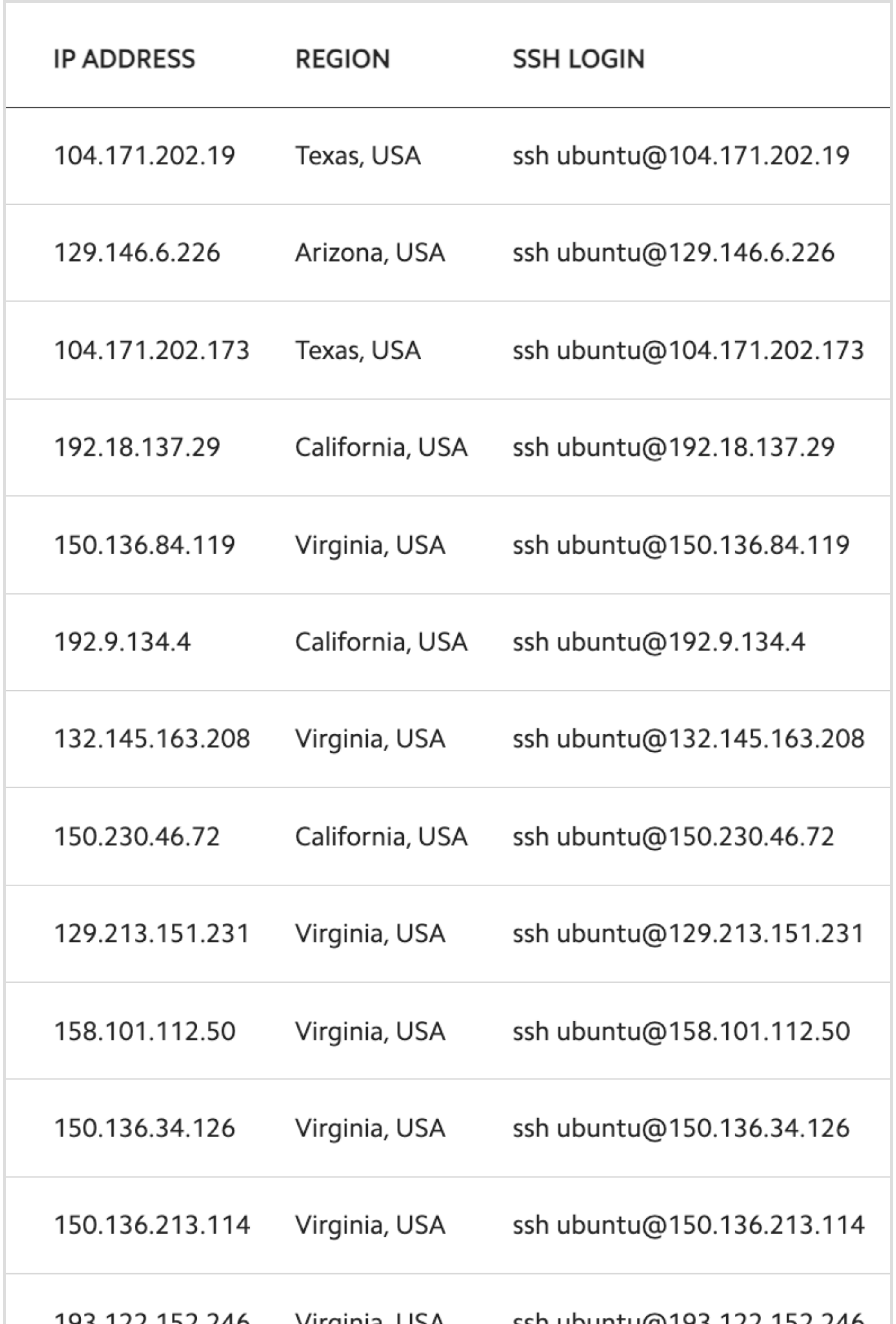
RETURN OF THE MPI

Conventional cloud value proposition no longer holds…

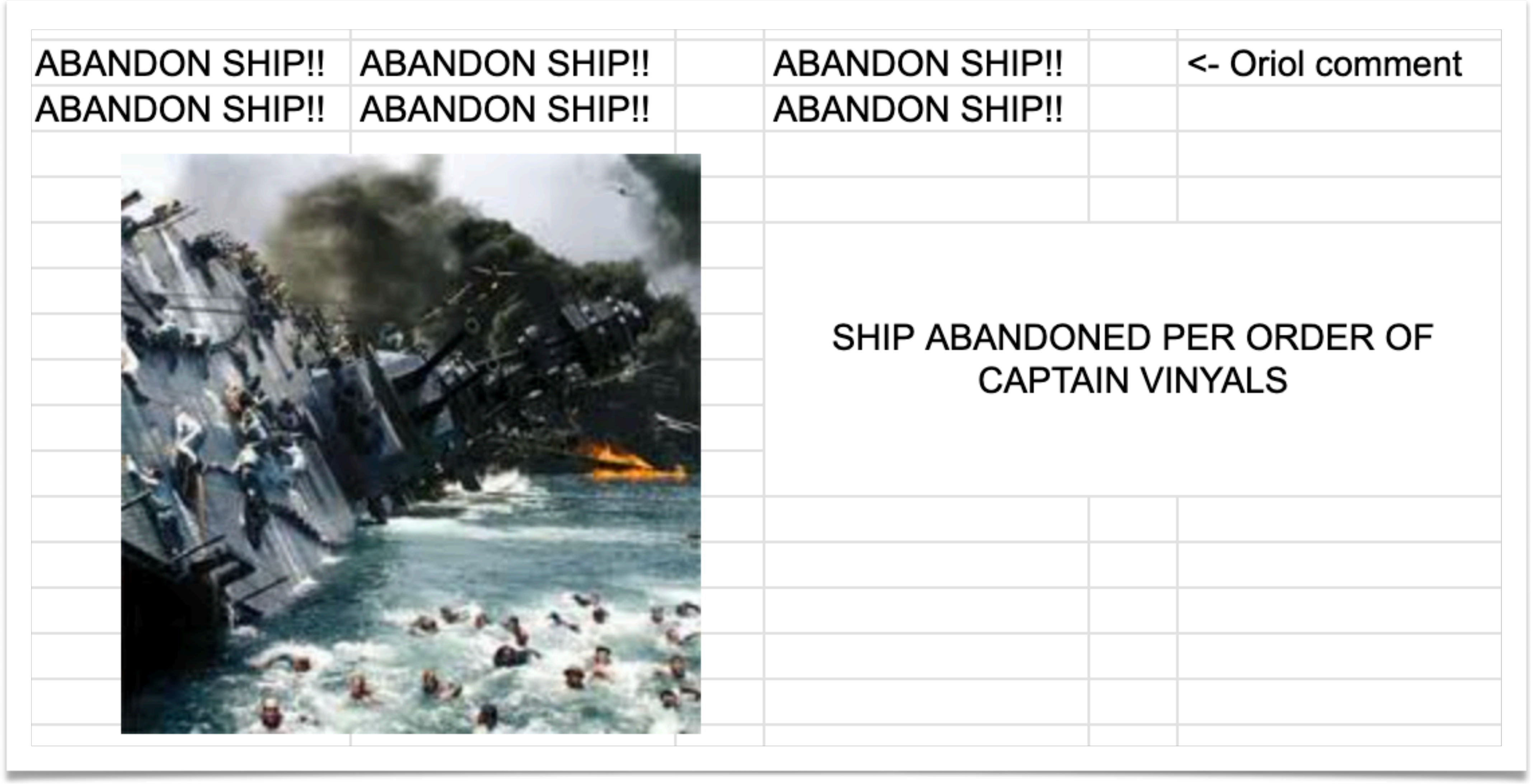
	Conventional Cloud	AI Cloud
Software: Variety	Complicated: Many applications Middleware	Simple: “AI frameworks”
Software: Workload	Varied: Compute, storage, network, big data, database, etc.	Unified: Numerical Computation
Supply Chain: Flexibility	High: CPU based Virtualization/Migration	Low: Large training Hard to live migrate
Supply Chain: Interchangeability	High: VMs can do many different jobs	Low: Really just doing AI compute


aws
Google Cloud
Alibaba Cloud
Azure
?
CoreWeave
NVIDIA PARTNER
DigitalOcean
λ Lambda

You shouldn’t run on baremetal. Also, K8s is wrong.

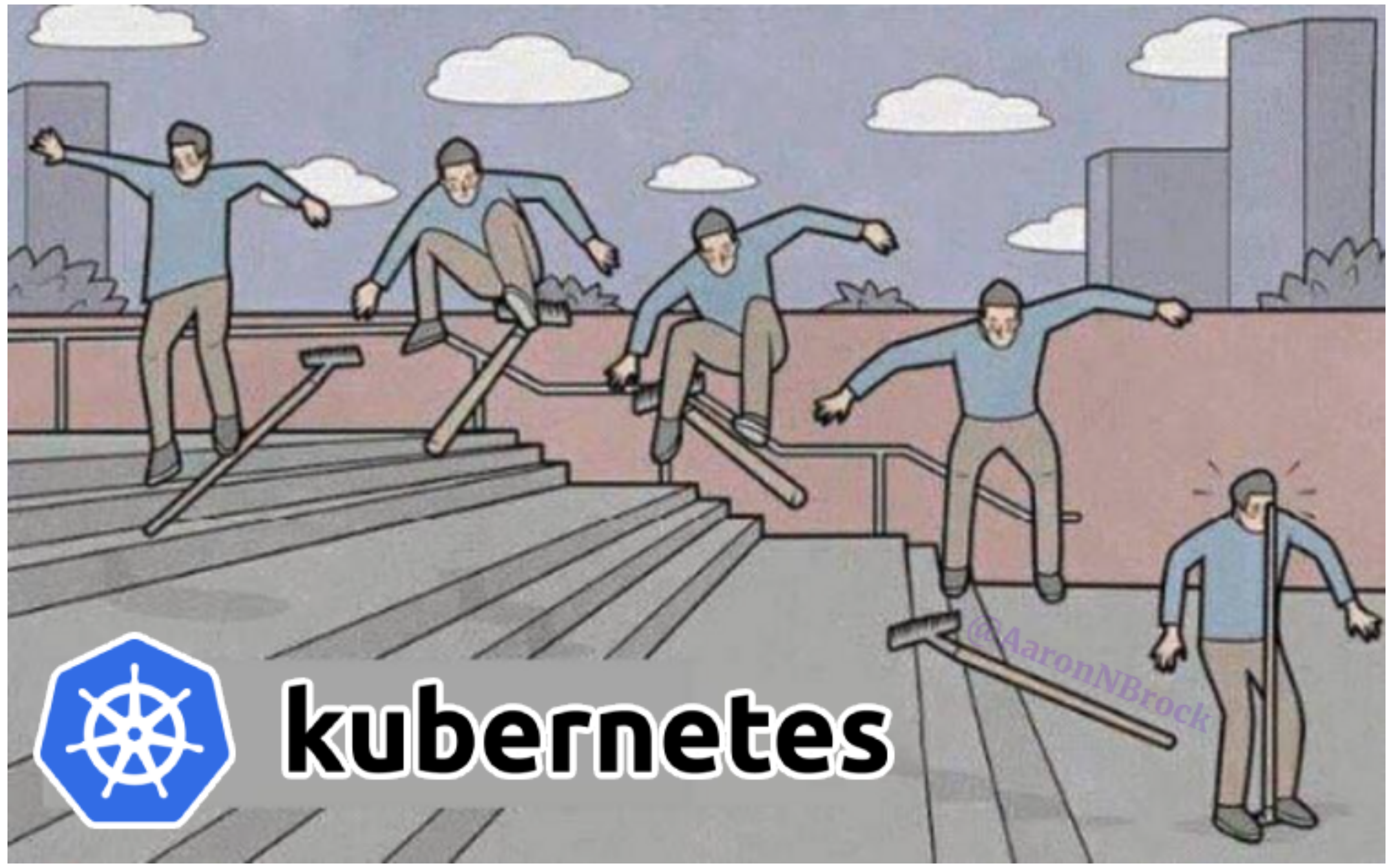


IP ADDRESS	REGION	SSH LOGIN
104.171.202.19	Texas, USA	ssh ubuntu@104.171.202.19
129.146.6.226	Arizona, USA	ssh ubuntu@129.146.6.226
104.171.202.173	Texas, USA	ssh ubuntu@104.171.202.173
192.18.137.29	California, USA	ssh ubuntu@192.18.137.29
150.136.84.119	Virginia, USA	ssh ubuntu@150.136.84.119
192.9.134.4	California, USA	ssh ubuntu@192.9.134.4
132.145.163.208	Virginia, USA	ssh ubuntu@132.145.163.208
150.230.46.72	California, USA	ssh ubuntu@150.230.46.72
129.213.151.231	Virginia, USA	ssh ubuntu@129.213.151.231
158.101.112.50	Virginia, USA	ssh ubuntu@158.101.112.50
150.136.34.126	Virginia, USA	ssh ubuntu@150.136.34.126
150.136.213.114	Virginia, USA	ssh ubuntu@150.136.213.114
193.122.152.246	Virginia, USA	ssh ubuntu@193.122.152.246



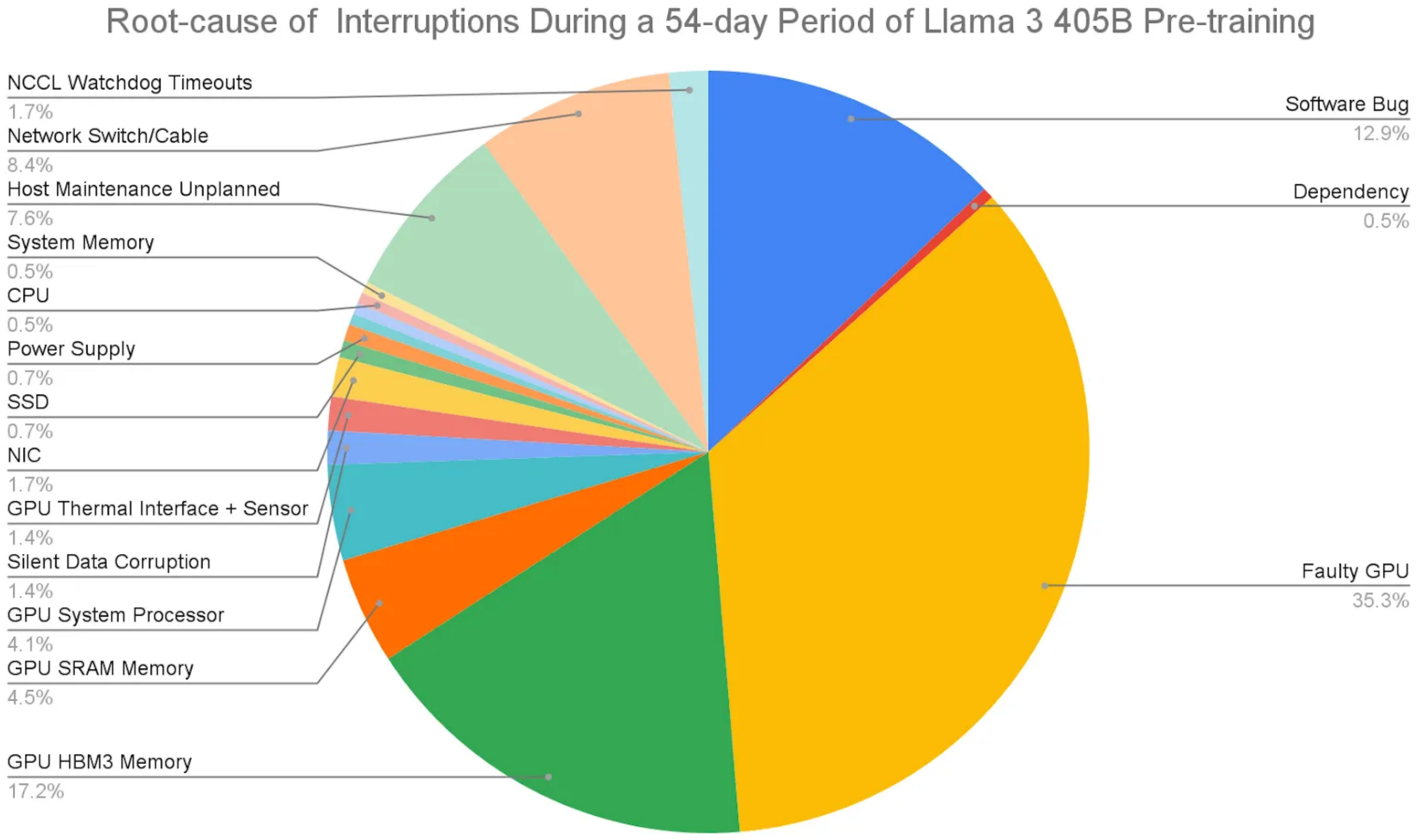
ABANDON SHIP!!	ABANDON SHIP!!		ABANDON SHIP!!		<- Oriol comment
ABANDON SHIP!!	ABANDON SHIP!!		ABANDON SHIP!!		

SHIP ABANDONED PER ORDER OF CAPTAIN VINYALS



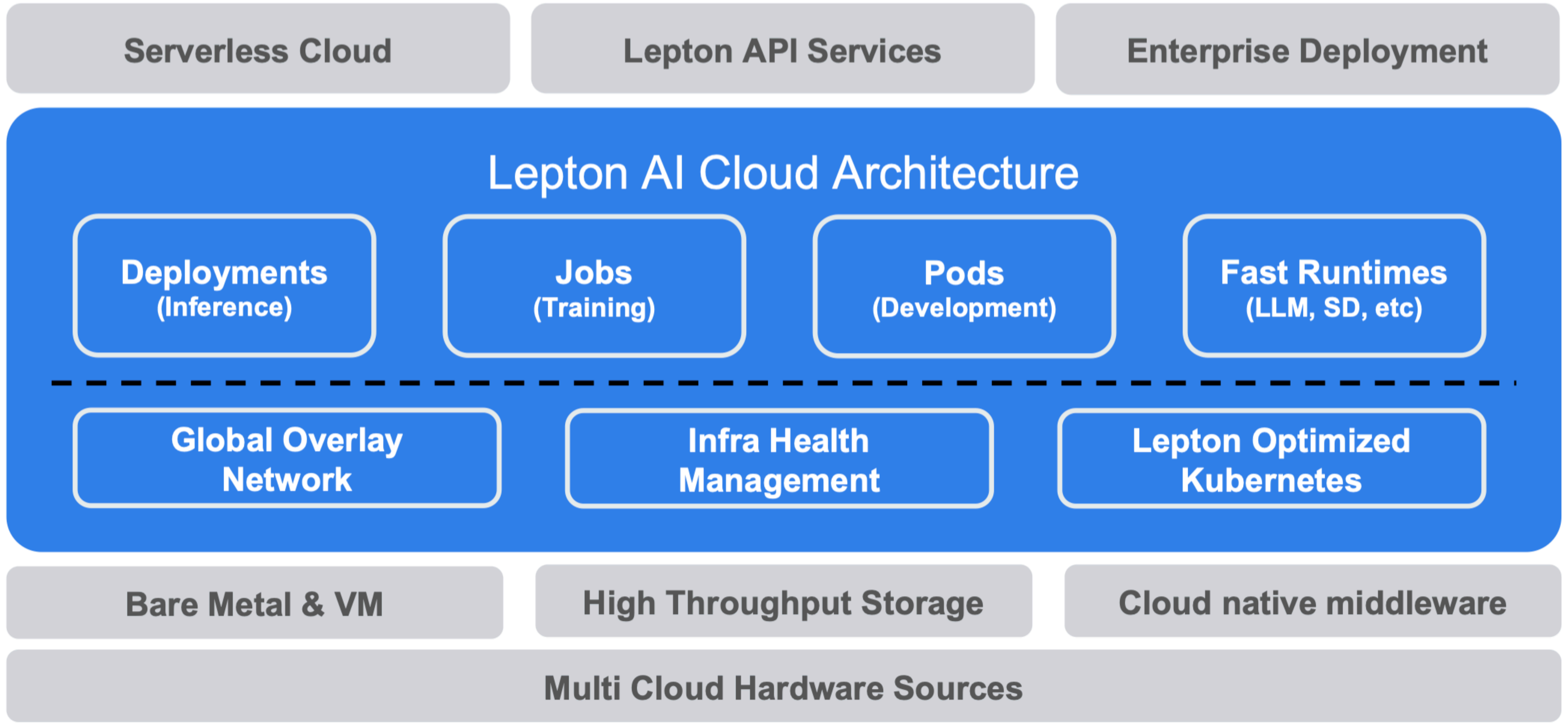
What you want to care about…

- Developer efficiency
  - Developer's time is precious
- Infra efficiency
  - GPUs do die
  - (More frequently than you think)



Best Practices?

- Multi-cloud supply chain management
- Elasticity and utilization management
- AI native platform to unify dev, training, and inference
- Build your own team around model and applications



Serverless Cloud | Lepton API Services | Enterprise Deployment

Lepton AI Cloud Architecture

Deployments (Inference) | Jobs (Training) | Pods (Development) | Fast Runtimes (LLM, SD, etc)

Global Overlay Network | Infra Health Management | Lepton Optimized Kubernetes

Bare Metal & VM | High Throughput Storage | Cloud native middleware

Multi Cloud Hardware Sources

AI infra is different, but also the same.

4
HW and SW design: back to the future?

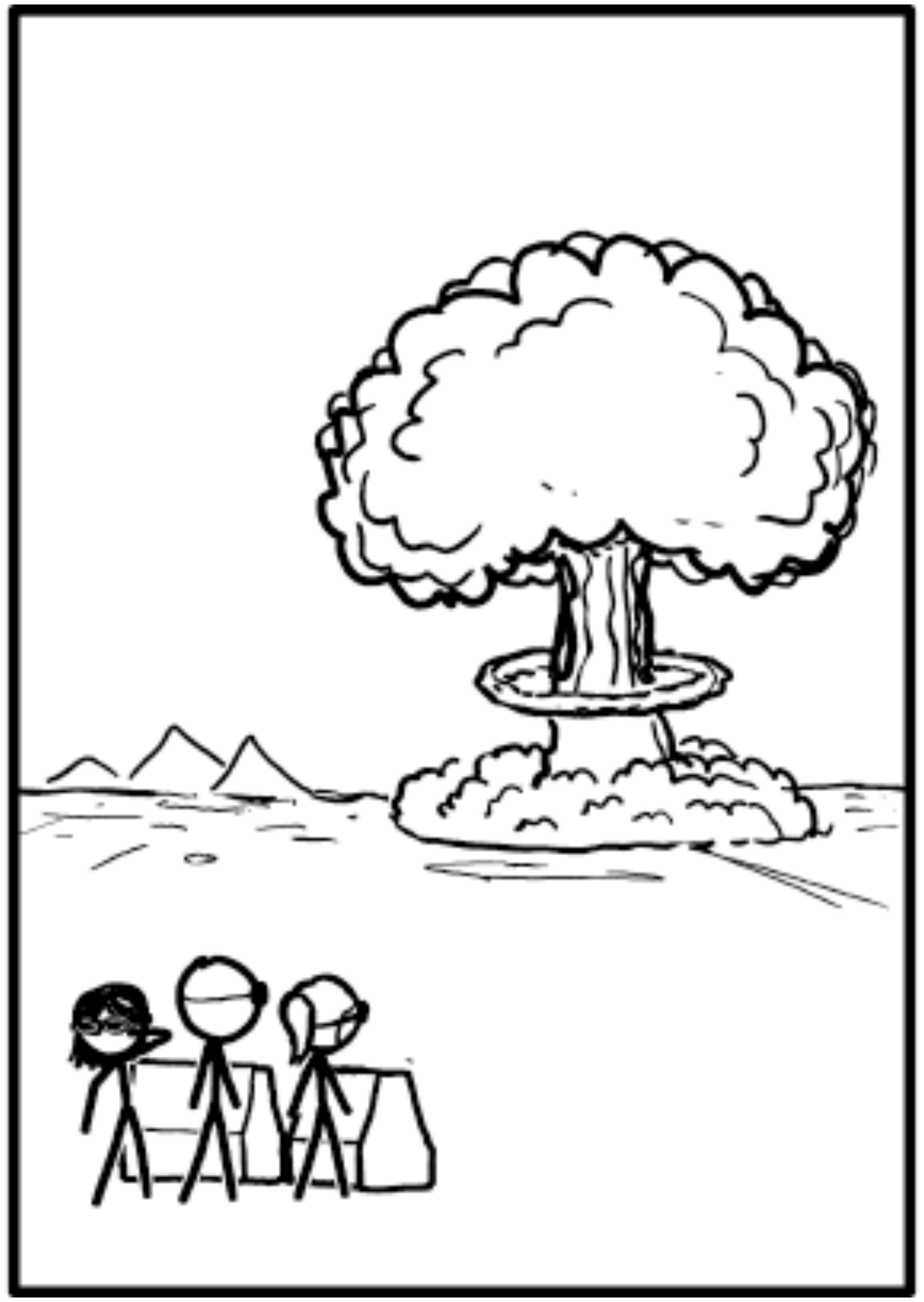
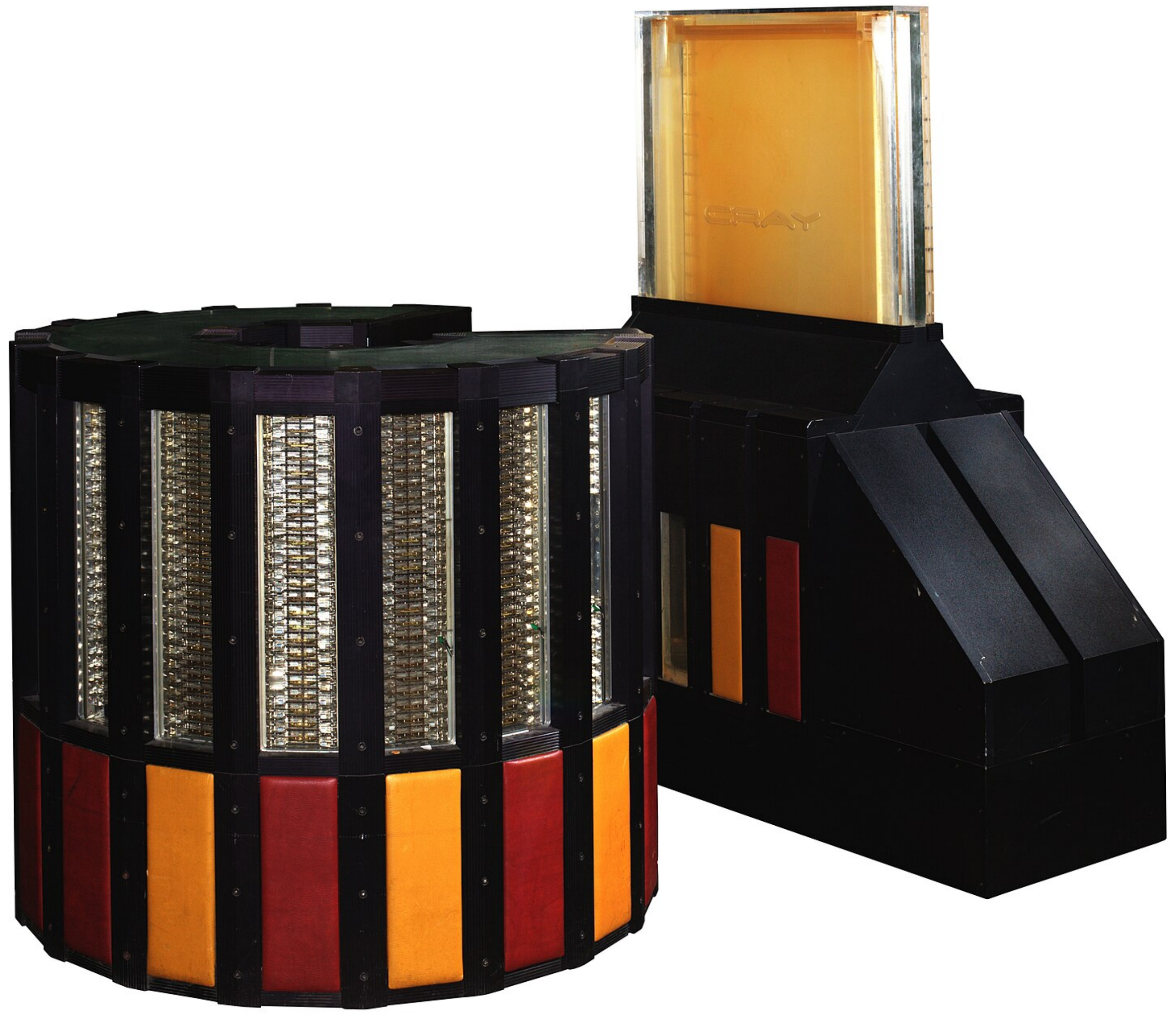
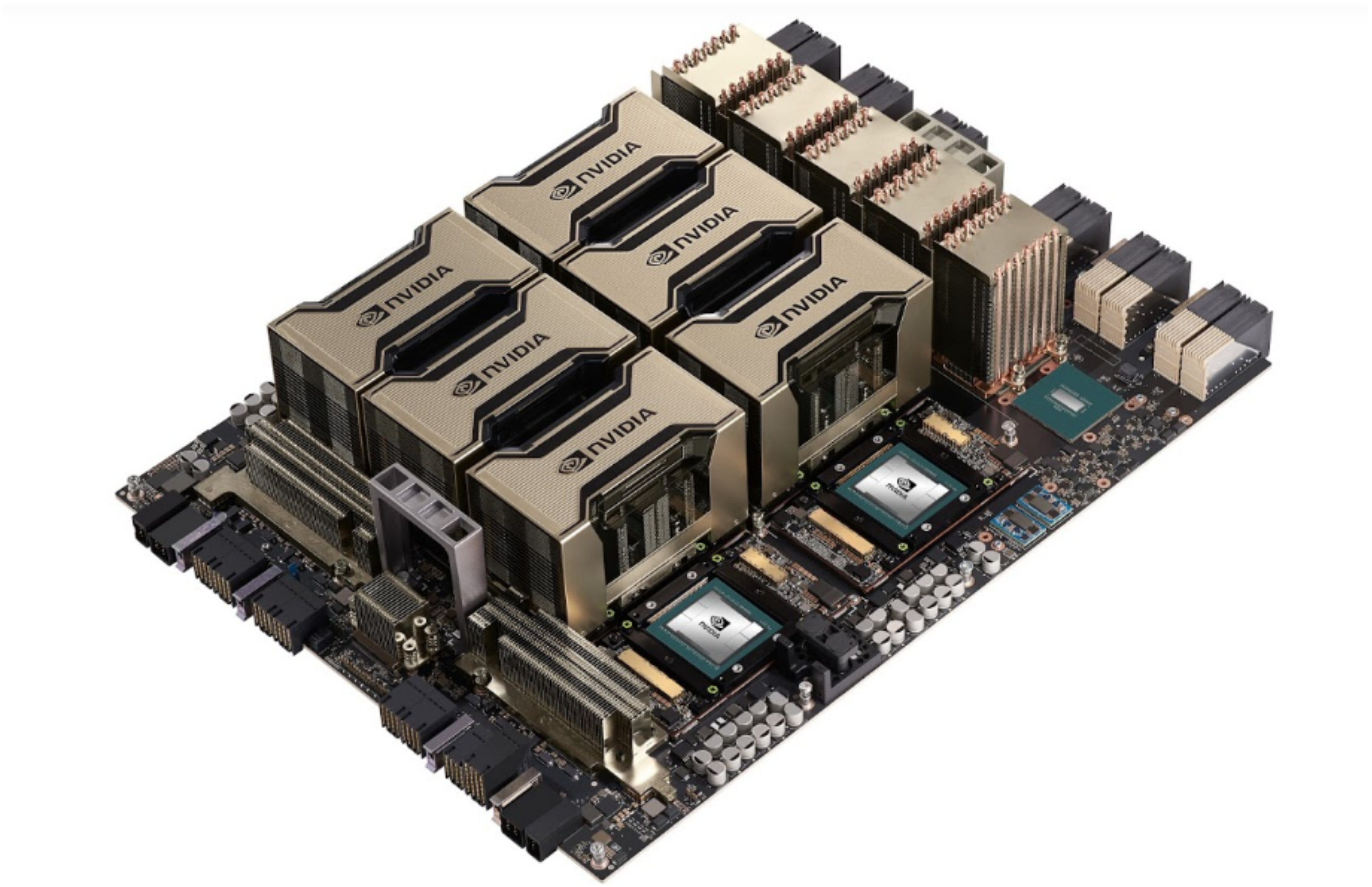
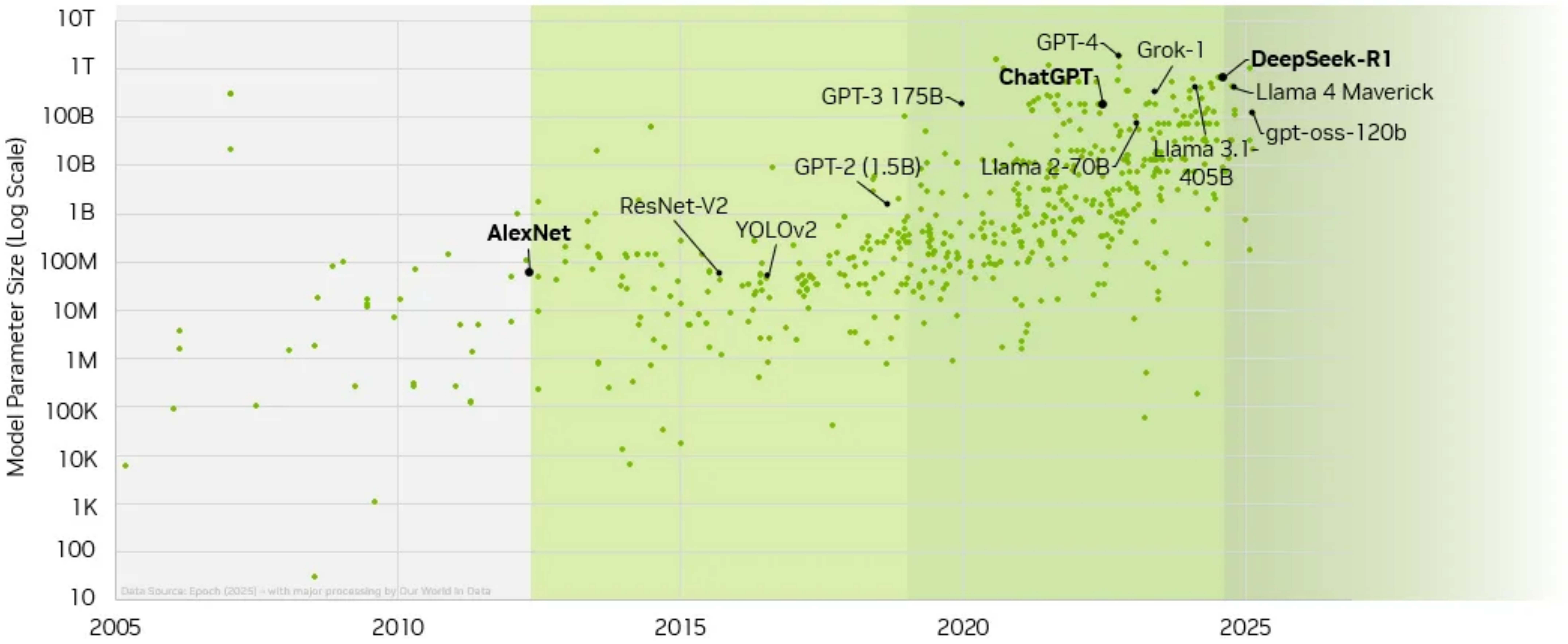


Image source: xkcd


CRAY

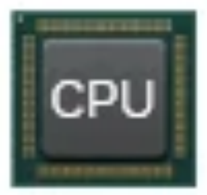

NVIDIA



Data Source: Epoch (2025) – with major processing by Our World in Data

Early AI Era

≤100M parameters.
Follows Moore's Law – Model size doubling every 20 months
Inference runs on CPUs



GPU AI Era

~100M to ~1B parameters
Doubling every 6 months
Inference runs on 1 GPU

Multi-GPU AI Era

~1B to multi-trillion parameters
Commercial AI Driven
Doubling every 10 months
Inference running up to 8 GPUs

Age of AI Reasoning at Scale

Drastic Increase in Compute for Reasoning
Expansion of Distributed Parallelism Techniques
Large Scale Mixture of Experts
Inference running up to 72 GPUs

UNITED

Evolution of System Designs from an AI Engineer Perspective

Yangqing Jia

Lepton AI (now part of NVIDIA)