Anthropic Compute Capacity Allocation Framework

Problem: Compute allocation requires tradeoffs. It can be difficult to quantify and compare the benefits of compute allocations across business verticals.

Goal: Allocate compute resources in a way that ensures Anthropic upholds its commitment to safety, remains at the frontier, and supports revenue generation.

This paper provides a framework to allocate Anthropic's compute capacity between research teams and customer facing products. Specifically, compute allocations are analyzed between research, training, and inference teams at the total company level. This categorization is used because these verticals represent the majority of raw compute allocation and require fundamentally different evaluation mechanisms. Further, the framework provided can be used to allocate between these teams with any level of granularity. Finally, it is assumed that team leaders are given the autonomy to make suballocations and can use the framework to empower their decision making.

In order to perform this analysis, the following information is needed: (a) available compute supply, (b) workload demand normalized in H100 equivalent units, (c) accelerator throughput (FLOPs), (d) accelerator pricing by cloud provider and total cost of ownership for any owned or "on-premise" infrastructure, (e) API and subscription based revenue as well as revenue coming from other Anthropic lines of business, (f) anticipated duration of compute requests, (g) revenue attribution per model, (h) estimated (future) and historical (past) training hours required per model, (i) expected accelerator hours required to serve inference demand.

Compute allocations will follow a tiered priority framework. The first tier, P0, is a non-negotiable bucket of capacity. This is capacity for safety evaluations, red teaming, and ensuring the reliability and security of Anthropic's products and systems. P0 will also contain capacity to satisfy any contracted SLAs. The second tier, P1, is for company-wide strategic initiatives. These are initiatives that leadership determines are imperative for the company's success. An example of a strategic initiative could be researching and implementing a new training or scaling architecture that has demonstrated efficiencies or improvements in model capabilities. The third tier, P2, is flexible capacity. This bucket of capacity is left over after P0, P1, and a buffer for unforeseen circumstances, incidents, or unexpected spikes in demand. P2 is where tradeoffs occur and where decision makers must analyze the costs and benefits of capacity allocations. For the purposes of this analysis, the available categories that will receive capacity in both P1 and P2 are inference, training, and research. In practice, the framework works as follows: If Anthropic identifies a strategic initiative, the corresponding category will be elevated from P2 to P1. The strategic initiative may also inform how capacity is split between the two remaining categories, likely resulting in a "lean" towards one bucket relative to the other.

Global variables that are used in conjunction with the framework are stated below. Global variables include total accelerator supply, total H100 hours used to train industry models and future Anthropic models, and revenue per accelerator hour during inference. All sources and assumptions used to derive these global variables can be found on the assumptions page. These assumptions are used directionally and are not intended to be exactly accurate with respect to Anthropic's operations.

Estimated Anthropic Accelerator Supply:

Supply[1]					
Provider	**Chip**	**Quantity**	**H100e Ratio (N Chip = 1 H100)**	**H100e**	**Annual Instance Hours**
AWS	Trainium 2	1,000,000	1.50 : 1	666,667	5,840,000,000
GCP	TPU	1,000,000	4.00 : 1	250,000	2,190,000,000
Azure/Nvidia	B200	156,556	0.33 : 1	469,667	4,114,285,714
Total H100 Equivalent Supply					**12,144,285,714**

Estimated H100 Training Hours for Industry Models & Future Annual Training Cost:

Estimated Total H100 Hours to Train Industry Models[2]					
Model	**Total FLOP**	**H100 FLOP/hr**	**MFU**	**Training FLOP/H100hr**	**H100 hours**
GPT 5	5E+25	3.564E+18	50%	1.782E+18	28,058,361
GPT 4	2.1E+25	3.564E+18	50%	1.782E+18	11,784,512
GPT 4.5	6.4E+25	3.564E+18	50%	1.782E+18	35,914,703
Sonnet 4	5E+25	3.564E+18	50%	1.782E+18	28,058,361
Opus 4	1.5E+26	3.564E+18	50%	1.782E+18	84,175,084
Sonnet 3.7	3.4E+25	3.564E+18	50%	1.782E+18	19,079,686
Grok 2	3E+25	3.564E+18	50%	1.782E+18	16,835,017
Grok 3	4.6E+26	3.564E+18	50%	1.782E+18	258,136,925
Llama 3.1	3.8E+25	3.564E+18	50%	1.782E+18	21,324,355

Estimated Hours Needed to Train Anthropic's Models[2]	
Assumed Increase in Model Total FLOP YOY	400%
Assumed H100 hrs for Training (based on Opus)	336,700,337
Number of Training Runs per Year (Pre-Training)	2
Assumed % for New Variants in Same Family	20%
New Variants per Training Run (Post-Training)	3
Assumed H100 hours Per Post-Training	67,340,067
Total H100 Hours per Year	**1,077,441,077**

Estimated Anthropic Revenue per Accelerator Hour During Inference:

Pricing Matrix and Weighted Average Price per Million Input and Output Tokens[3]						
Model	**Release**	**Assumed % of Total Tok Usage**	**Input**	**Output**	**Wtd Avg Price per M Input**	**Wtd Avg Price per M Output**
Claude Opus 4.6	2/5/2026	10%	$5.00 / MTok	$25.00 / MTok	$5.57	$27.84
Claude Opus 4.5	11/24/2025	10%	$5.00 / MTok	$25.00 / MTok		
Claude Opus 4.1	8/5/2025	10%	$15.00 / MTok	$75.00 / MTok		
Claude Opus 4	5/22/2025	10%	$15.00 / MTok	$75.00 / MTok		
Claude Sonnet 4.5	9/29/2025	25%	$3.00 / MTok	$15.00 / MTok		
Claude Sonnet 4	5/22/2025	25%	$3.00 / MTok	$15.00 / MTok		
Claude Haiku 4.5	10/15/2025	3%	$1.00 / MTok	$5.00 / MTok		
Claude Haiku 3.5	11/4/2024	3%	$0.80 / MTok	$4.00 / MTok		
Claude Haiku 3	3/13/2024	3%	$0.25 / MTok	$1.25 / MTok		

H100 Tok Throughput (Nvidia), H100 Tok Throughput (likely), Tok Attribution per Model Size, and Input/Output Tok per H100 Hr at 1/2 Ratio[4]							
ISL / OSL	**Model Size**	**H100 Output Tok/s (Nvidia)**	**20% Resource Constraints**	**Weight**	**Wtd Avg Output Tok/s**	**Output Tok/H100 hr**	**Input Tok/H100 hr**
1000, 2000	8B	13,505	2,701	10%	484	1,743,833	871,916
1000, 2000	70B	1,854	371	50%			
1000, 2000	405B	361	72	40%			

Utilization Sensitivity[5]			
Utilization	**Revenue per H100 Hr**	**Discount**	**Revenue per H100 Hr**
20%	$10.68	30%	$7.48
30%	$16.02	30%	$11.22
40%	$21.36	30%	$14.95
50%	$26.70	30%	$18.69
60%	$32.04	30%	$22.43

With global variables and assumptions stated, scenario analysis can be done. The charts below use the allocation framework to produce recommended allocation percentages as well as allocations under two different strategic initiatives. As mentioned above, the framework assumes individual team leaders have the autonomy to allocate compute within their respective teams and could use the framework for any suballocations.

Scenario #1: Recommendation

Supply	12,144,285,714
Revenue	$18,000,000,000

	Total Per Tier		
	Priority Tier	Percentage of Total Capacity	H100 Hours
Safety/SLA	P0	10%	1,214,428,571
Strategic Initiative	P1	0%	0
Flex	P2	75%	9,108,214,286
	Buffer	15%	1,821,642,857

	P1 and P2 Breakout		
	Inference	Training	Research
Strategic Initiative	-	-	-
Percentage Within P2	26%	24%	50%
H100 Hours P2	2,407,419,160	2,154,882,155	4,545,912,971
Percentage of Total	20%	18%	37%

Scenario #1 does not claim a strategic initiative but rather states a recommended allocation based on Anthropic's identity as a research company, and on estimated training and revenue data. The first assumption is that for Anthropic to remain at the frontier, it must train and produce two model families with three model variants per family every year (*1.077B H100 hours, chart 2*). It is imperative that Anthropic continue producing models with state-of-the-art capabilities. This increases adoption and allows Anthropic to maintain its market share. The required number of H100 hours to train Anthropic's models is multiplied by 2 to account for inefficiencies (restarts and dead-end training runs) and the overstatement of supply (not all accelerators are on 100% of the time and capacity additions occur throughout the year, not all at once). A deeper analysis is required to estimate the true multiple on estimated training hours. As a result, Anthropic must dedicate 18% of total capacity towards training to remain competitive and produce industry leading models.

Second, it is important to allocate sufficient inference capacity to sustain current run rate projections and meet basic inference reliability standards. Public information like Anthropic's $14 billion revenue run rate and third party FY revenue projections is used to estimate demand at $18 billion. Inference serving H100 hours are calculated by dividing revenue by the imputed revenue per inference hour (*$7.48, chart 5*). As a result, 20% of total capacity must be allocated towards inference to ensure demand is served reliably.

The remaining capacity, 37%, is allocated towards research as Anthropic is fundamentally a research lab and should invest heavily in exploration. Training alone will not ensure that Anthropic remains at the frontier and must be accompanied by strong research. This is because model capability, safety, and reliability will allow Anthropic to have outsized impact and, in turn, steer the industry towards safety. Without frontier level capabilities, Anthropic models would have less demand and Anthropic's ability to impact the world would diminish, leaving room for another lab to take its place. Research is the key to model improvement, staying at the frontier, and increasing impact.

Further, the downstream effects of AI systems have yet to be fully realized. As such, there are many unknowns with the creation and diffusion of this technology. Prioritizing safety research helps to anticipate potential risks of LLMs and can help mitigate future consequences. Without research focused specifically on safety and interpretability, Anthropic might build systems that are

inherently misaligned with societal objectives. This can hinder humanity's progress rather than accelerate it. Continued research to understand, guide, and improve Anthropic's artificial intelligence systems is essential.

Importantly, the percentage estimates are meant to demonstrate how the framework can be used. All percentages can be toggled to arrive at different allocations and analyze tradeoffs between revenue and other objectives. A key feature of this framework is the ability to quantify decision making. Here is an extreme hypothetical example: Leadership recommends spending 70% of compute on research. This framework gives the user an estimate of the revenue that might be forgone and the amount of training Anthropic can execute relative to the recommendation. While such directives are unlikely, the framework quantifies decisions so they can be better understood. For reference, tradeoffs in different scenarios can be seen below.

Sensitivity Analysis						
Scenario	Inference	Training	Research	Serviceable Revenue	% of Recommended Training Hrs	Research Hours
Recommendation	20%	18%	37%	$18,000,000,000	100%	4,545,912,971
Revenue Grows to $25B (pricing constant)	28%	18%	30%	$25,000,000,000	100%	3,609,694,409
Additional Training Run Needed	20%	22%	33%	$18,000,000,000	125%	4,007,192,433
Revenue Per Inference Hour Drops to ($5.50), Revenue Target Unchanged	27%	18%	30%	$18,000,000,000	100%	3,680,604,858
Revenue Per Inference Hour Drops to ($5.50), Revenue Target Changes	20%	18%	37%	$13,240,805,378	100%	4,545,912,971
Training Hours Multiple Declines to 3x	20%	13%	42%	$18,000,000,000	75%	5,084,633,510
Leadership Recommends 70% on Research	3%	3%	70%	$2,270,036,171	14%	8,501,000,000

Scenario #2: Anthropic Prioritizes Serviceable Revenue

Supply	**12,144,285,714**
Revenue	**$31,780,506,396**

	Total Per Tier		
	Priority Tier	**Percentage of Total Capacity**	**H100 Hours**
Safety/SLA	P0	10%	1,214,428,571
Strategic Initiative	P1	35%	4,250,500,000
Flex	P2	40%	4,857,714,286
	Buffer	15%	1,821,642,857

	P1 and P2 Breakout		
	Inference	**Training**	**Research**
Strategic Initiative	4,250,500,000	-	-
Percentage Within P2	0%	60%	40%
H100 Hours P2	0	2,914,628,571	1,943,085,714
Percentage of Total	35%	24%	16%

Scenario #2 assumes that Anthropic prioritizes serviceable revenue. Reasons for this might be mass model adoption, proving the viability of the business, implications for a public offering, dramatic increases in existing customer demand, or potentially an opportunity to increase market share relative to competitors. In this scenario, 35% of all capacity would serve inference and support $31.8 billion in revenue. The remaining capacity after this strategic initiative would "lean" towards training as producing additional models quickly could further increase revenue.

Scenario #3: Anthropic Prioritizes Training and Shipping New Models

Supply	12,144,285,714
Revenue	$14,528,231,495

	Total Per Tier		
	Priority Tier	Percentage of Total Capacity	H100 Hours
Safety/SLA	P0	10%	1,214,428,571
Strategic Initiative	P1	35%	4,250,500,000
Flex	P2	40%	4,857,714,286
	Buffer	15%	1,821,642,857

	P1 and P2 Breakout		
	Inference	Training	Research
Strategic Initiative	-	4,250,500,000	-
Percentage Within P2	40%	0%	60%
H100 Hours P2	1,943,085,714	0	2,914,628,571
Percentage of Total	16%	35%	24%

In Scenario #3, Anthropic's strategic initiative is to prioritize training and shipping new models. This might be motivated by a new training technique or scaling architecture that has proven successful. In this scenario, the remaining capacity would "lean" towards research rather than inference as additional research may lead to breakthroughs that improve training efficiency. Further, training and research can have multiplicative effects when explored in tandem. In isolation, research and training can lead to breakthroughs, but when combined, benefits compound. The capacity allocated for inference in Scenario #3 would support $14.5 billion in revenue.

Communication:

To communicate allocations to each team, it is critical that methodologies and any value functions are clearly described and available to all teams. This is true at the company level and at the team/workload level. Documentation explaining company strategic initiatives as well as how value is attributed to each bucket of capacity will be provided. While allocation decisions can be difficult to deliver to teams that receive less than requested capacity, clearly stated company goals and frameworks create a standard that each team can expect.

From a logistics perspective, biweekly allocation meetings will be held with representatives from each organization requesting capacity (inference, training, research). A potential allocation plan will be presented to the allocation committee prior to this meeting. During the meeting, director level advocates will get the chance to present their rationale for additional compute requests beyond the original allotment and will use the tiered priority framework to discuss how they plan to use the additional capacity within their organization. If additional capacity is requested, an ROI calculation must be presented and tradeoffs of the capacity must be addressed by the requester. The goal is not to make additional capacity requests difficult, but to empower leaders to make a compelling case for why their organization should receive more compute.

After these meetings, the Compute Strategy and Operations team will make final capacity allocation decisions and then communicate those decisions to leaders. This structure creates a dialogue between leaders and decision makers so that allocations can be made transparently. Leadership is provided the opportunity to advocate on their organization's behalf while using a shared framework, or set of rules, that each organization adheres to. This also allows decision makers to strategically incentivize the different lines of business by stating the value functions the company uses to analyze capacity decisions.

Once capacity has been allocated to an organization, it will be segmented into shared resource pools. Teams within that organization will be grouped and assigned to a resource pool based on workload type (specific resource pooling and team grouping should be managed by the individual organizations). Teams will then submit a request form for capacity from their shared pool. The form will include fields for accelerator type requested, cores needed, the H100 equivalent of this capacity, the duration for the capacity, the workload that this capacity will serve, the nature of the request, and ROI (optional). The nature of the request field will contain three potential options: critical, alignment with company goal or objective, and growth bet. Priority first goes to critical workloads, then to workloads that align with company goals, and then to growth bets, which are evaluated based on ROI. ROI will be decomposed further with a timeline when this ROI can be expected, expected efficiency or revenue increase, and confidence interval. These inputs allow team leads to evaluate the likelihood of the ROI, which helps with allocation decisions. Operationally, each category within "nature of this request" will have its own priority ranking per workload, and workloads will be executed accordingly.

Compute Capacity Management:

To manage compute capacity, visibility into how capacity is used is essential. First, a live utilization dashboard must be created and monitored so that decision makers, organizational leaders, and individual teams have insight into the amount of compute used and how efficiently it is used. If a team or organization has requested capacity and is only using a fraction of it, this comes at a cost to the business and to the other groups that requested that capacity. Often, capacity is requested to satisfy peak demand. Most of the time, capacity does not serve peak demand, and for this reason can be left underutilized if allocated to one workload specifically. This underutilization is why a shared resource pool is critical. Teams can flex their additional capacity to serve other workloads when not needed, balancing the load across accelerators and increasing both performance and utilization. A dashboard with visibility into utilization, as well as the available shared pool of compute, creates transparency and leads to flexibility. Ultimately, this results in better performance and more productive conversations around capacity requests.

The next critical component to managing compute capacity is determining if compute allocations have led to successful outcomes. While not perfect, outcomes can be analyzed by tracking capability improvements per training hour and by tracking margin per inference request. The trend over time for both of these metrics will tell a story about the effectiveness of research.

Capabilities per training hour will be measured by a carefully selected list of benchmarks. If capabilities per training hour are increasing, each training hour is more effective, which would likely indicate that research improved model capabilities or efficiency. If capabilities per training hour decrease for consecutive quarters, allocations towards research or between research teams might need to be reevaluated. One caveat to this metric is that benchmarks become saturated. There may be little headroom to demonstrate increased model performance even if models might be improving on un-benchmarked dimensions. Saturation is crucial to look out for when using benchmarks as a proxy for model capability and training efficiency.

The second metric worth tracking is margin per inference request. If margin per inference request is increasing over time, then a few things may be true. First, users might be trusting the model for more complex requests and using more tokens. Second, the cost to serve that model might be

decreasing due to efficiencies realized through research. It is important to segment both revenue and costs by model and by customer tier to gain granular insights as to why one model might perform differently than another. Here is a rule to govern compute allocations with this metric: If margin per inference request declines year over year, research will be prioritized to find more efficient ways to train models, cost effective ways to serve models, or ways to further increase capabilities. Notably, margin per inference request can be misleading as it does not provide information on total volume. A high margin per inference request might contribute very little to total revenue due to low request volumes. Conversely, a low margin per inference request may contribute significantly to revenue due to high volume. The scale of inference is meaningful, as small margin improvements can result in dramatic revenue increases. It is important to consider total volume when using the metric.

Conclusion:

Anthropic is faced with the difficult challenge of allocating compute capacity between research teams and customer facing products. This decision comes with tradeoffs. Using a tiered priority framework that emphasizes transparency, managing compute capacity with tools that provide visibility, and analyzing metrics that measure the impact of compute allocations can make these decisions easier. Compute is the backbone of any research organization, and if allocated properly, it can be a tactical asset rather than a costly bottleneck. The allocations and operational recommendations in this document demonstrate how compute can be used strategically to ensure Anthropic stays at the frontier, prioritizes its mission, and supports recurring revenue. Further, the best allocation frameworks change over time and reflect the goals of the companies they serve.

Assumptions

Supply Chart Assumptions[1]:
Capacity is on 100% of the time
Fleet size is an estimate based on publicly available information
Conversion factors are high level estimates, in practice this is different for every workload
Assumes 5-year Azure/Nvidia deal, B200 chips run continuously, $3.50/hr committed pricing
Assumes GPU compute is 80% of the $30B Azure/Nvidia commitment
Nvidia/Azure capacity derived from ($30B * 80%) / (5 * 365 *24) / (3.5)
Assumes TPUs are V5p
Assumes no "On-Premise" capacity, all compute capacity is through cloud providers

Training Hours Assumptions[2]:
H100 throughput assumes TF32 Tensor Cores
H100 assumed throughput is 990 TFLOP/s and then converted to FLOP/hr
Total Training FLOP to H100 hours derived from Total FLOP / (H100 FLOP/hr * MFU)
MFU used to convert Nvidia reported FLOP/s to FLOP/s achieved in practice
Total FLOP needed for pre-training increases 4X YOY
Post-Training Total FLOP is 20% of Pre-Training
Two Pre-Training Runs per Year with 3 Post-Training Runs for Every Pre-Training Run

Price per Million Input/Output Tokens Assumptions[3]:
Pricing Direct from Anthropic Website
Pricing Only Considers Input/Output Tokens, Excludes Other Pricing Components
Percentage of Token Usage is Derived from Weighting in Model Size and Throughput Table

Token Throughput and Weighted Average Input/Output Tokens per H100 Hour Assumptions[4]:
Output Tok/s, Input Seq. Length to Output Seq. Length ratio, and model size from Nvidia
Assumes 20% of Nvidia output Tok/s guidance due to practical resource utilization constraints with non-synthetic workloads
Model weighting represents assumed percentage of tokens that will come from each model
Model weighting assumes mid-size model (Sonnet) is most active and largest model (Opus) is second most active
Uses output tokens per model size from Nvidia as a proxy for output tokens when serving Anthropic models

Utilization and Revenue per Inference Serving H100 Hour Assumptions[5]:
Pre-Discount is likely overstated: Only analyzes API revenue, does not consider the impact of subscription revenue or 3p revenue, does not assume any enterprise discounts on APIs
30% revenue per hour discount is a rough estimate
Utilization percentage assumed for accelerator hours as accelerator is not inference serving 100% of every hour
Parameter estimates may be low which results in higher throughput values per accelerator hour than what might happen in reality

Links[1]:
https://newsletter.semianalysis.com/p/amazons-ai-resurgence-aws-anthropics-multi-gigawatt-trainium-expansion
https://www.aboutamazon.com/news/aws/aws-project-rainier-ai-trainium-chips-compute-cluster
https://www.googlecloudpresscorner.com/2025-10-23-Anthropic-to-Expand-Use-of-Google-Cloud-TPUs-and-Services
https://blogs.nvidia.com/blog/microsoft-nvidia-anthropic-announce-partnership/
https://newsletter.semianalysis.com/p/microsofts-ai-strategy-deconstructed
https://epoch.ai/data/ai-chip-sales-documentation#amazon
https://skymod.tech/inside-googles-tpu-and-gpu-comparisons/#:~:text=TPU%20v5e%2D8%20offers%201%2C576,(3%2C342%20TFLOPS%20%C3%97%202)
https://www.clarifai.com/blog/nvidia-b200-vs-h100#:~:text=Performance%20efficiency:%20The%20NVIDIA%20B200,power%20consumption%2C%20and%20infrastructure%20complexity

Links[2]:
https://galileo.ai/blog/llm-model-training-cost
https://ecosystem.aethir.com/blog-posts/the-llm-infrastructure-revolution-how-gpu-requirements-are-reshaping-the-ai-industry#:~:text=Strategic%20approaches%20vary%20across%20the,exaFLOPS%20of%20AI%20training%20compute.
https://www.cudocompute.com/blog/what-is-the-cost-of-training-large-language-models
https://epoch.ai/blog/what-will-ai-look-like-in-2030#:~:text=We%20argue%20that%20AI%20scaling,capabilities%20across%20science%20and%20beyond.
https://epoch.ai/data-insights/open-models-threshold
https://epochai.substack.com/p/notes-on-gpt-5-training-compute
https://epoch.ai/data-insights/models-over-1e25-flop
https://www.golodiuk.com/news/stanford-course-language-modeling-how-long-to-train-70b-llm/
https://www.coreweave.com/blog/nvidia-h100-gpu-benchmark-results-what-we-learned-from-large-scale-gpu-testing#:~:text=51%2D52%25%20MFU%20on%20NVIDIA,Llama%203's%20reported%20numbers
https://www.nvidia.com/en-us/data-center/h100/
https://www.itpro.com/technology/artificial-intelligence/dollar100-billion-to-build-an-ai-model-anthropic-ceo-dario-amodei-predicts-soaring-ai-training-costs-but-models-will-become-far-more-powerful

Links[3,4,5]:
https://www.nvidia.com/en-us/data-center/h100/
https://platform.claude.com/docs/en/about-claude/pricing
https://nvidia.github.io/TensorRT-LLM/blogs/H100vsA100.html
https://nvidia.github.io/TensorRT-LLM/0.21.0/performance/perf-overview.html
https://aiven.io/tools/llm-leaderboard
https://siliconangle.com/2026/02/12/anthropic-closes-30b-round-annualized-revenue-tops-14b/

https://medium.com/@onurbingul/how-to-keep-your-gpu-busy-part-1-maximizing-throughput-in-llm-inference-8a9f63c44974
https://www.reuters.com/technology/anthropic-hikes-2026-revenue-forecast-20-information-reports-2026-01-28/#:~:text=Reuters%20Plus-,Anthropic%20hikes%202026%20revenue%20forecast%2020%25%2C%20The%20Information%20reports,Sign%20up%20here.
https://www.clarifai.com/blog/mi300x-vs-h100#:~:text=Theoretical%20vs%20Real%E2%80%91World%20Throughput,sharding%20and%20has%20shorter%20TTFT.
https://openinfer.io/news/2025-03-21-unlocking-the-full-potential-of-gp-us-for-ai-inference/#:~:text=One%20of%20the%20most%20important,SM)%2C%20reducing%20overall%20parallelism.
https://www.latimes.com/business/story/2026-01-23/from-4-billion-to-9-billion-anthropics-revenue-doubles-in-six-months
https://www.anthropic.com/news/anthropic-raises-30-billion-series-g-funding-380-billion-post-money-valuation