Brian Buras Manager -
Applied Research Team

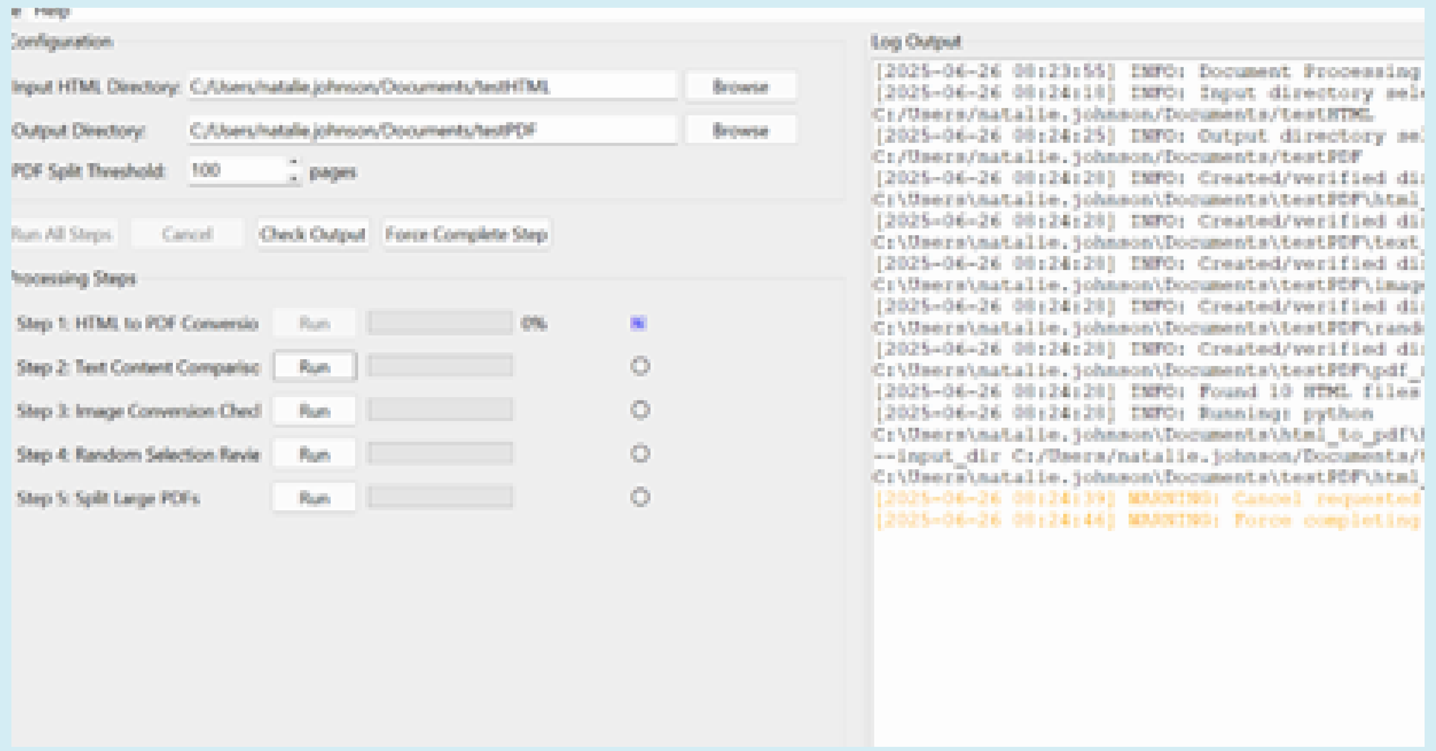
natalie lam johnson .com

Summer 2025

What is Retrieval Augmented Generation?

Combines LLM with an external knowledge base (ex. TDC) to provide more accurate answers

RAG HTML → PDF
gui app



Input HTML Directory: C:/Users/natalie.johnson/Documents/testHTML

Output Directory: C:/Users/natalie.johnson/Documents/testPDF

PDF Split Threshold: 100 pages

Step 1: HTML to PDF Conversio 0% ●

Step 2: Text Content Comparisc ○

Step 3: Image Conversion Chec ○

Step 4: Random Selection Revie ○

Step 5: Split Large PDFs ○

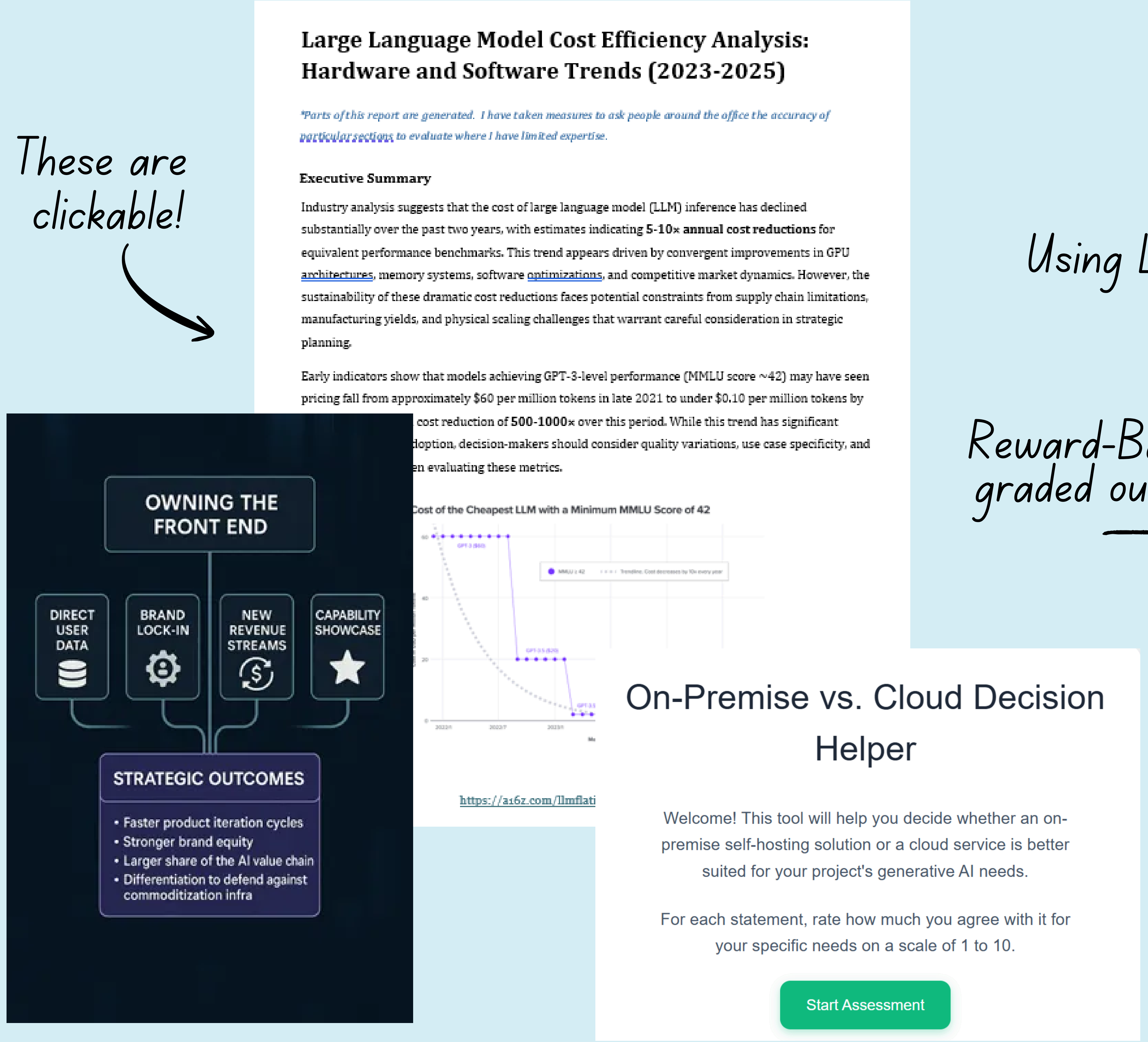
```
[2025-06-26 08:23:55] INFO: Document Processing
[2025-06-26 08:24:18] INFO: Input directory sel
C:/Users/natalie.johnson/Documents/testHTML
[2025-06-26 08:24:25] INFO: Output directory se
C:/Users/natalie.johnson/Documents/testPDF
[2025-06-26 08:24:28] INFO: Created/verified di
C:\Users\natalie.johnson\Documents\testPDF\html_
[2025-06-26 08:24:28] INFO: Created/verified di
C:\Users\natalie.johnson\Documents\testPDF\text_
[2025-06-26 08:24:28] INFO: Created/verified di
C:\Users\natalie.johnson\Documents\testPDF\image
[2025-06-26 08:24:28] INFO: Created/verified di
C:\Users\natalie.johnson\Documents\testPDF\rand
[2025-06-26 08:24:28] INFO: Created/verified di
C:\Users\natalie.johnson\Documents\testPDF\pdf_
[2025-06-26 08:24:28] INFO: Found 10 HTML files
[2025-06-26 08:24:28] INFO: Running: python
C:\Users\natalie.johnson\Documents\html_to_pdf\
--input_dir C:/Users/natalie.johnson/Documents/
C:\Users\natalie.johnson\Documents\testPDF\html_
[2025-06-26 08:24:39] WARNING: Cancel requested
[2025-06-26 08:24:44] WARNING: Force completing
```

WeShare Pages

- Costs of Generative AI
- On-Premise (Open Weight) vs Cloud Service (Closed Weight)
- 5 Year CapEx for on-Prem vs. Cloud Server Hosting
- Case for Newer (more Expensive) Models

Economics of Gen AI

These are clickable!



Large Language Model Cost Efficiency Analysis: Hardware and Software Trends (2023-2025)

**Parts of this report are generated. I have taken measures to ask people around the office the accuracy of particular sections to evaluate where I have limited expertise.*

Executive Summary

Industry analysis suggests that the cost of large language model (LLM) inference has declined substantially over the past two years, with estimates indicating **5-10× annual cost reductions** for equivalent performance benchmarks. This trend appears driven by convergent improvements in GPU architectures, memory systems, software optimizations, and competitive market dynamics. However, the sustainability of these dramatic cost reductions faces potential constraints from supply chain limitations, manufacturing yields, and physical scaling challenges that warrant careful consideration in strategic planning.

Early indicators show that models achieving GPT-3-level performance (MMLU score ~42) may have seen pricing fall from approximately $60 per million tokens in late 2021 to under $0.10 per million tokens by … cost reduction of **500-1000×** over this period. While this trend has significant … doption, decision-makers should consider quality variations, use case specificity, and … en evaluating these metrics.

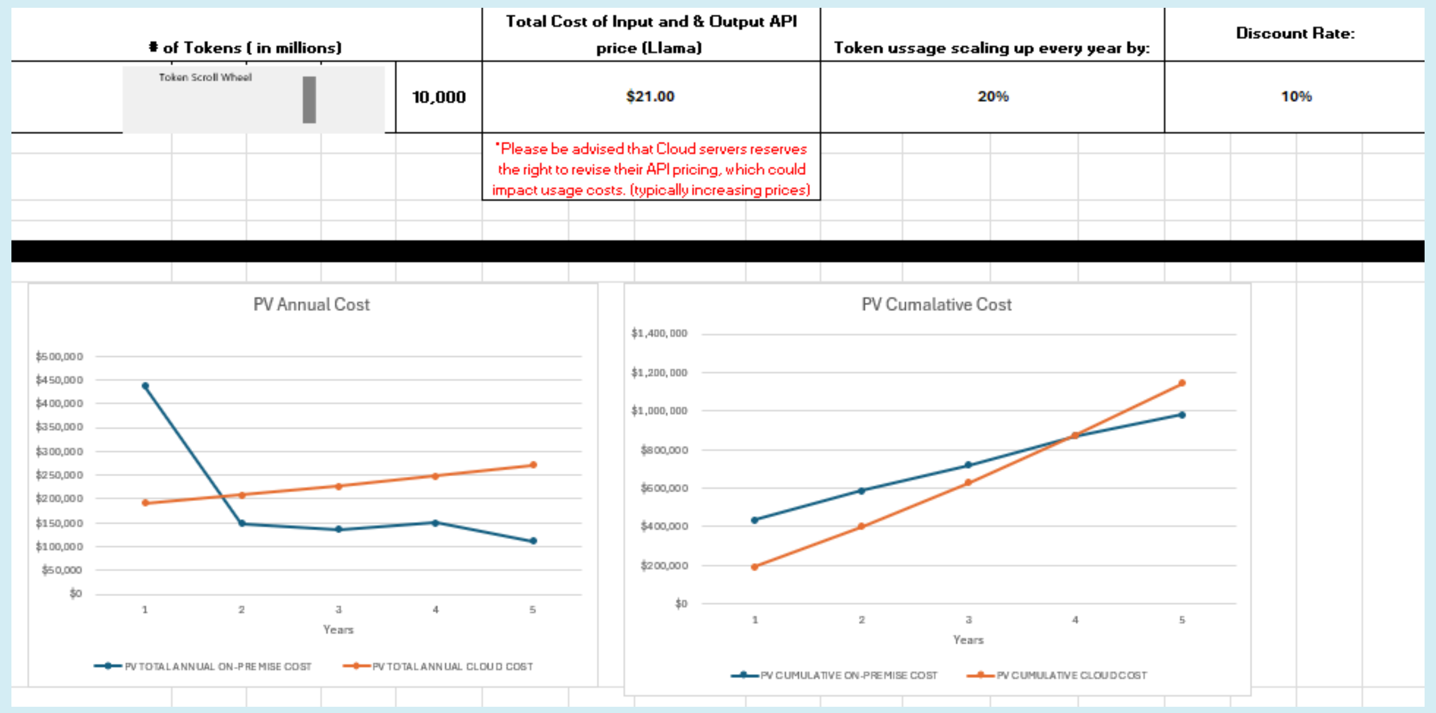
https://a16z.com/llmflati

- Faster product iteration cycles
- Stronger brand equity
- Larger share of the AI value chain
- Differentiation to defend against commoditization infra

On-Premise vs. Cloud Decision Helper

Welcome! This tool will help you decide whether an on-premise self-hosting solution or a cloud service is better suited for your project's generative AI needs.

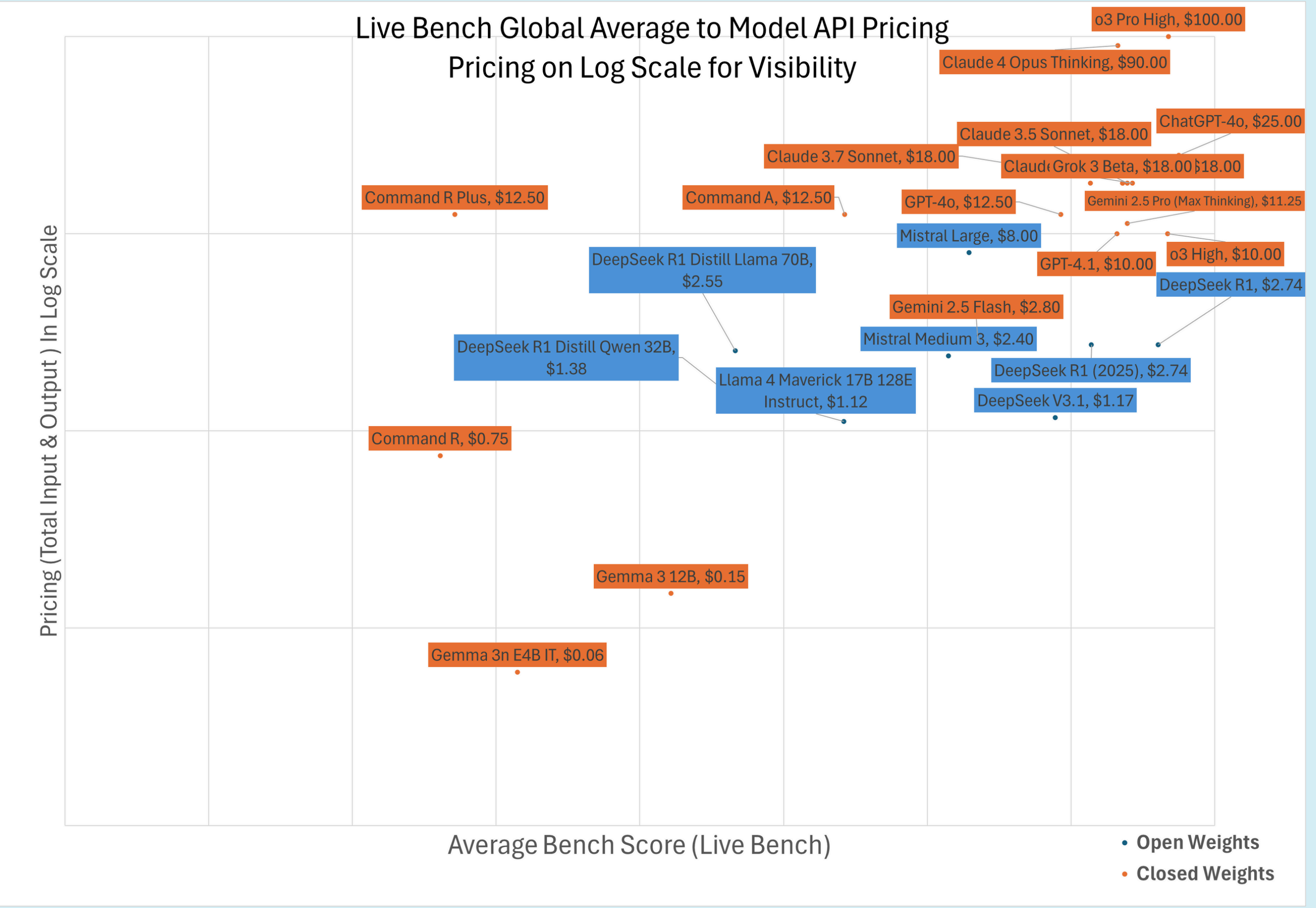
For each statement, rate how much you agree with it for your specific needs on a scale of 1 to 10.



# of Tokens (in millions)		Total Cost of Input and & Output API price (Llama)	Token ussage scaling up every year by:	Discount Rate:
Token Scroll Wheel	10,000	$21.00	20%	10%
		*Please be advised that Cloud servers reserves the right to revise their API pricing, which could impact usage costs. (typically increasing prices)		

These are clickable!

*Filtered Out Compressed & Chinese Models



Using Labled Data

Vocab I've Learned

Reward-Based, graded output

- SFT: Supervised Fine Tuning
- RFT: Reinforcement Fine Tuning
- Mixed Precision: Combining number precisions for faster computation
- Hill Climbing: read here
- NRE Assessment: non-recurring engineering cost
- POC: Proof of Concept