Unravelling of OpenAI

TEAM MARSH — IIT ROORKEE

Moulik Gupta
Agam Pandey
Ridhi Mahajan
Satyam Sinha
Hemant Bidasaria
Keshav Goyal

ICC
Indian
Case
Challenge

BUSINESS CLUB

KSHITIJ2024
THE TECHNO-MANAGEMENT FEST

Executive Summary

- Open AI' release of GPT Builder and GPT Store raises oversaturation, low quality GPTs and huge computational needs for the company.
- A **change in Pricing model** for ChatGPT is proposed, GPT 3.5 (Free), GPT 4 + GPTStore ($15/month) and GPT 4+ GPTBuilder+GPTStore ($20/month)
- GPT Builder will have 2 models for creating custom GPTs , **monetizable GPT/ Non-monetizable GPT**.
- Computational expenses would be tackled by **limitation on Data storage/ Token generation for a free credit/month**, money add on following exhausted credits.
- To ensure quality GPTs and tackle oversaturation, **85% uniqueness, Data Privacy and Non-objectionable guidelines** for GPT Store suggested.
- AI chips act as the powerhouse for large language models, enabling their ability to comprehend and produce intricate text through efficient handling of vast amounts of information simultaneously.
- Bridging the supply-demand gap in AI chips is crucial for sustaining technological progress and meeting the growing demands of an advancing industry
- **Nvidia and Graphic Core are the key leaders** in terms of favourable parameters required for bridging the Supply demand gap found out after a detailed analysis of startups and firms
- Comparing different firms and startups based on different economic metrics

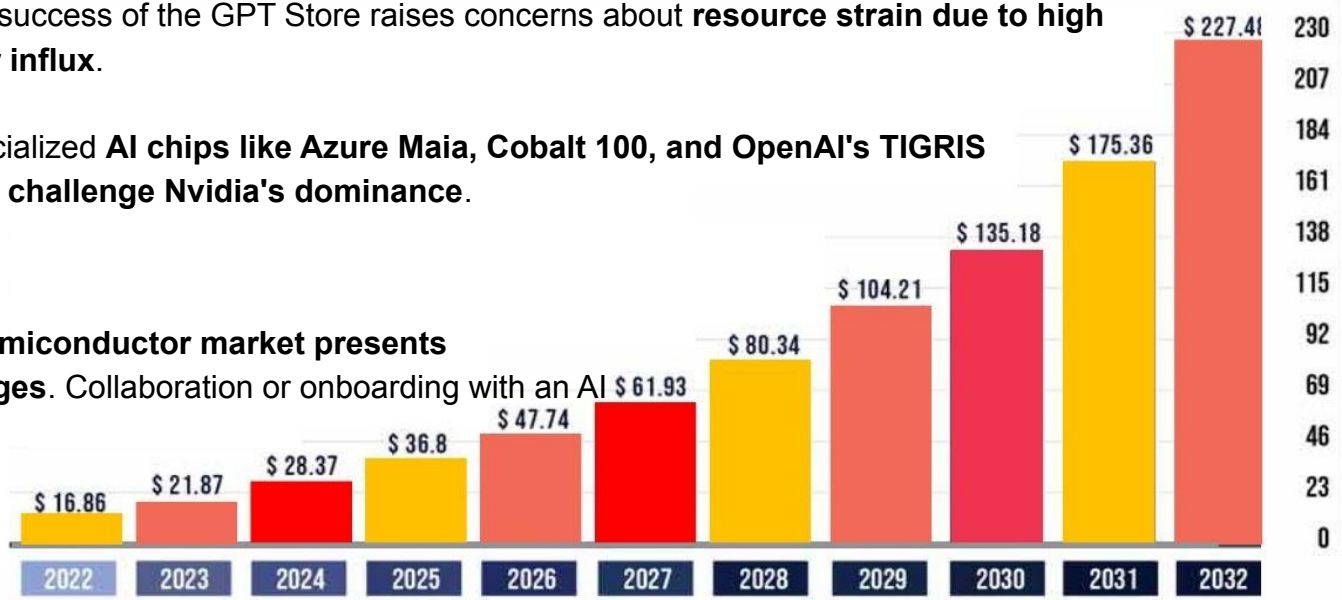
OpenAI's GPT Store Challenges: Resource Strain, Revenue Models, and Semiconductor Ventures

- OpenAI's GPT4 Turbo and GPTs, enhances API integration While the GPT Store and builder promote accessibility, restricting GPT Builder to premium users poses a **tradeoff between inclusivity and incentives**, impacting OpenAI's AI development commitment. **Balancing revenue models is crucial for growth and engagement.**

- The success of the GPT Store raises concerns about **resource strain due to high user influx**.

- Specialized **AI chips like Azure Maia, Cobalt 100, and OpenAI's TIGRIS chip challenge Nvidia's dominance**.

- **OpenAI's entry into the semiconductor market presents opportunities and challenges**. Collaboration or onboarding with an AI chip design firm/startup



Timeline

GPT Store Developer Guidelines

Proposed guidelines that a Developer must follow to monetize and publish custom GPT

Types of Chips

Different type of chips in use and technical evaluation.

Startups & Firms Analysis

Comparing Chip startups and diving deep into

GPT Builder & GPT Store Features

ChatGPT model with GPT Builder allows users to create personalized GPTs for non-commercial use, enabling customization and innovation within limits for personal purposes.

Cost Analysis | GPTs Pricing Model

Revenue Splitting and Revenue / day analysis through ChatGPT Builder and GPT Store

AI Chips Demand & Supply

Current trend of AI chip shortage and the need in AI industry

Economic Analysis of Acquisition

Comparing acquisition costs and revenue prospects of various startups to select one aligning with your company's goals and offering strategic value, innovative technology, and sustainable growth potential.

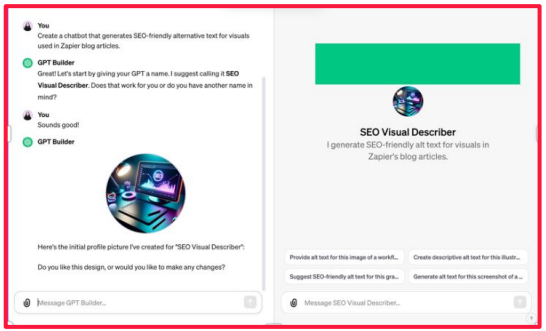
GPT Builder will be split into Monetizable and Non-monetizable versions to address the issue of excessive low-quality GPTs on the GPT Store

Features and overview of ChatGPT model for creating customized GPTs using GPT Builder for personal and non- monetizable use

System Integration and Plug-ins

Allowing **GPT-4 Plus users** to create custom GPT and use it with registered mail as plug-in with websites for personal use.
Measures for implementing one usability/subscriber

- **Ensure Plug-in functionality** with registered subscriber
- Smooth working with Google/Microsoft services (other major Social medias)



Flagging similar personal GPTs & recommendation to use available models

10GB data storage limit for Custom GPTs

Assuming a total number of 90,000 - 100,000 users/month of GPT-4 Plus users, to manage & reduce data storage cost/GPT, a **10GB cap for users**
100 GB cap for enterprise suggested.

Tackling Oversaturation & Optimizing GPTs

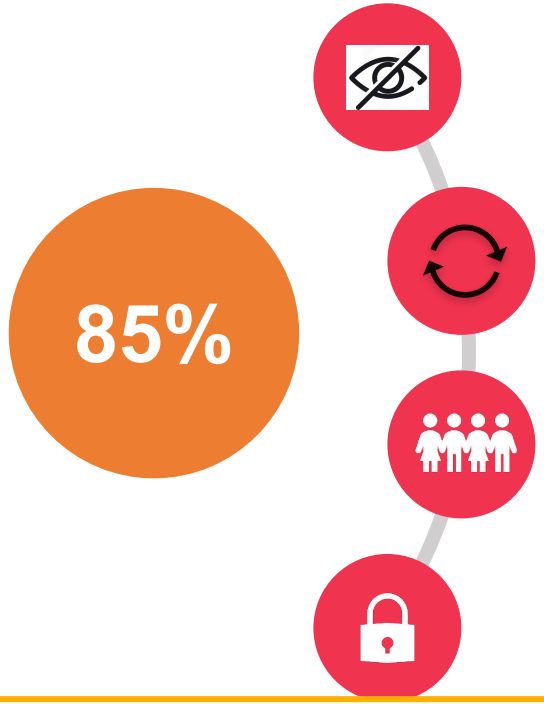
Pushing users to use monetized GPTs more to prevent oversaturation (similar personal GPT with available model to be flagged and recommendations given)

- **Algorithms to flag similar GPTs**
- Suggest alternative GPTs available while training model

***Considering the volatile customer growth of ChatGPT,cost analysis and figure used are as of 29-12-2023**

Enforcing 85% Uniqueness, Data Privacy, Non-Objectionable content Requirement for Quality GPTs and Addressing Oversaturation.

Each custom GPT will be evaluated by setting 85% (subject to change) uniqueness through Dataset,Context in lines with WorldQuant Brain's alpha published by Developers



Objectionable Content

Gpts should not include content : offensive, insensitive ,upsetting intended to disgust, in exceptionally poor taste.Defamatory, discriminatory, or mean-spirited content, including references or commentary about religion, race, sexual orientation, gender, national/ethnic origin

Redundancy

There shouldn't be redundancy in the gpts by the developers.The GPTs satisfying all other policy guidelines would be selected and will be considered.

Children Safety

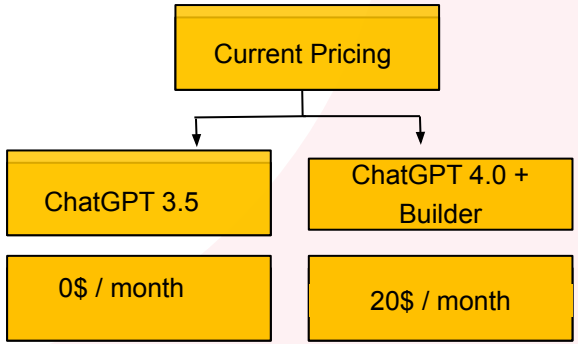
Before submitting an app that targets children to the you are responsible for ensuring your app is appropriate for children and compliant with all relevant laws.

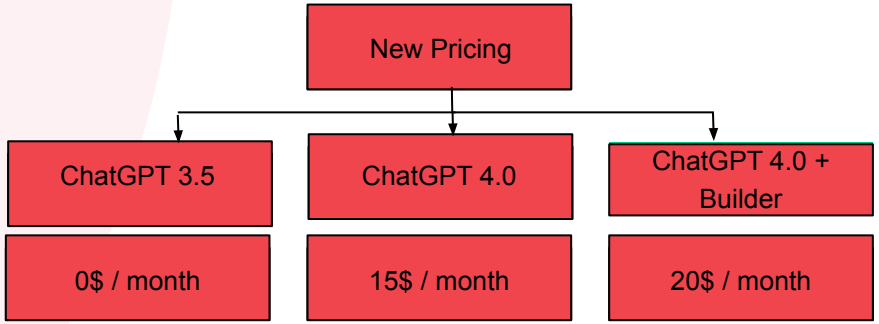
Privacy

Explain its data retention/deletion policies.Gpts should only request access to data relevant to the core functionality of the app and should only collect and use data that is required to accomplish the relevant task.

The guiding principle of the GPT Store is simple— to provide a safe experience for users to get GPTS and a great opportunity for all developers to be successful. Every GPT is reviewed by experts and an editorial team helps users discover new apps every day.

Three-Tier Pricing Model: Free ChatGPT, Paid ChatGPT-4, and Paid ChatGPT 4+GPTBuilder, Offering GPT Store Access





Free

0$ per person/month

Features:

- GPT 3.5
- Regular model updates

Plus

15$ per person/month

Features:

- GPT 4.0
- Access to GPT Store
- Advanced-Data analytics
- Early access to beta version
- Regular model updates

Plus + Builder

20$ per person/month

Features:

- GPT 4.0
- Build 10 custom GPT models
- Monetize Custom GPT
- Access to GPT Store
- Advanced-Data analytics
- Early access to beta version
- Regular model updates

Cost Analysis Of the Pricing Models

OPERATIONAL COST	Daily users	Per tokens charge	Tokens per day	Total cost per day
ChatGPT 3.5	20 M	0.000225$	40	180000$
ChatGPT Plus	300K	0.000225$	500	33500 $
ChatGPT Plus + Builder	200K	0.000225$	700	15700 $
				Total = 230K $

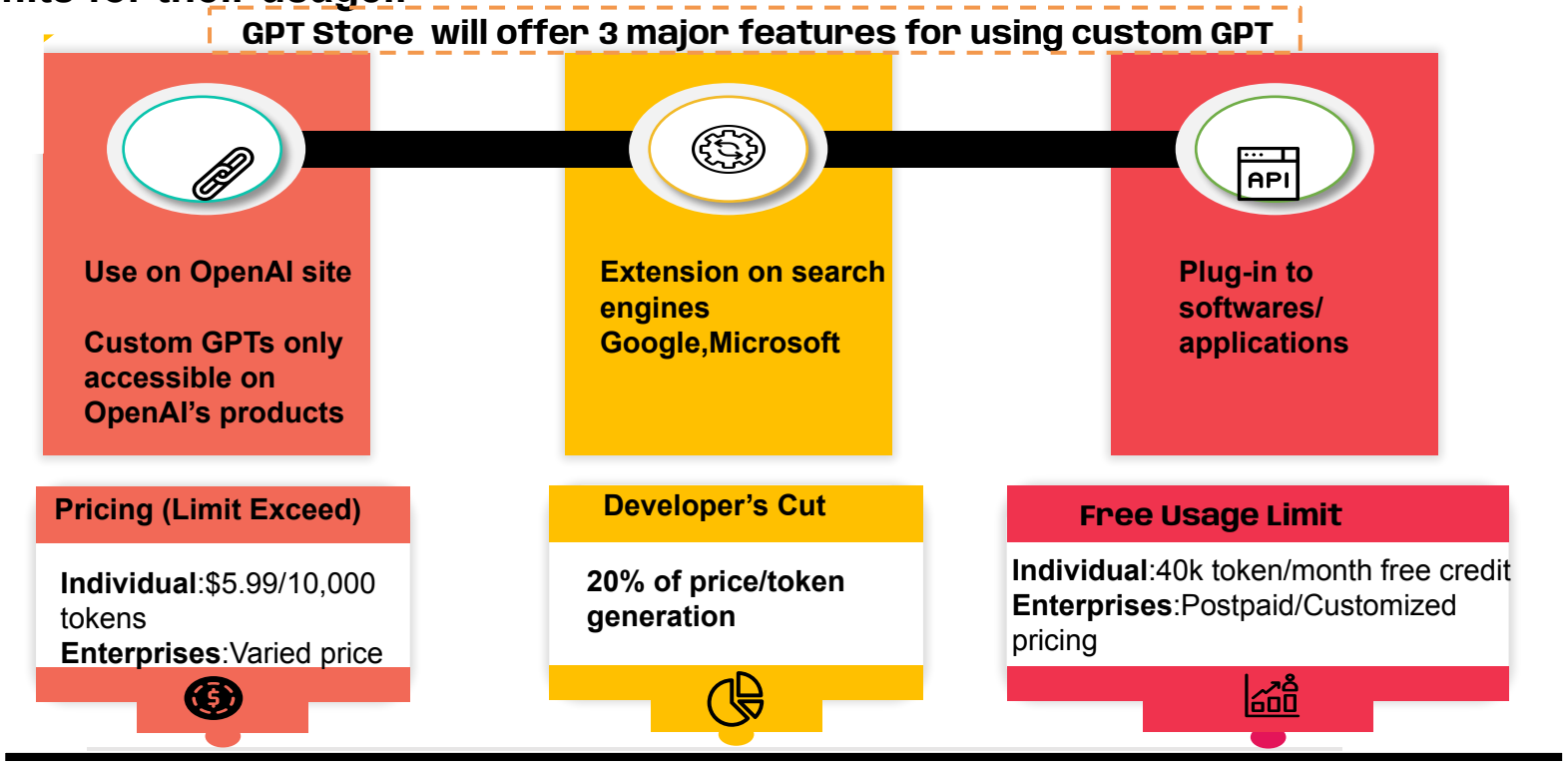
REVENUE	Daily users	Per tokens charge	Total Cost per day	
ChatGPT 3.5	20M	0	0	
ChatGPT Plus	300K	230K $ / day	150K $ / day	Total = 280K $
ChatGPT Plus + Builder	200K	230K $ / day	130K $ / day	

Even Though OpenAI just tries to breakeven cost through these models, we are able to generate a profit of $50K

Note- High profit margins considered here act as a leverage to the R&D cost

- ***Assumptions are added in the appendix.***

The GPT Store will offer custom GPT models usable on-site as plug-ins/extensions, with set token limits for their usage..



**Usage requirements varies for Enterprises,setting a hard figure isn't possible*

Developers will receive 20% of the pricing per token for their contributions to these custom GPTs, incentivizing their involvement and encouraging innovation within the platform

Revenue-Sharing between Developer and GPT Store for ChatGPT, shows ChatGPT can generate $50,000/day from custom GPTs while incurring cost of $500,000/day

Cost Analysis of custom GPTs by subscribers/day

Assumptions

- Token Generation Cost: $0.0003/word (Azure A100 Single GPU)
- Max. Token Length / Message: 4096 tokens = 3072 words
- Avg. Message: 30% text = 125 tokens/100 words
- User Conversation Assumptions: 8-10 messages, each containing 800 words (1067 tokens)

Calculations

- Usage Metrics: 25 times/day
- Cost Calculation: $0.32 - $0.5 / custom GPT
- Subscriber Statistics: 0.2% - 0.3% active users using GPT (241,500 - 300,000)
- GPT Builder Usage: 30% of subscribed users = 90,000/month
- Average GPT Usage: 10 GPTs, 7-8 at a time per user

Revenue analysis/day of Custom GPT's

GPT Builder

- Price of GPT Builder model (Tier 3)= $20/month = $0.67/day
- Number of GPTs created = 7-8 (assumption)
- Tokens / conversation = 1067 = 1100 tokens / conversation
- Total usage times= 25 times /day
- Number of users = 200,000
- Revenue/day = 200,000*0.67~$13500/day

GPT Store

For every token generated by a custom GPT on GPT Store, a 20% percent commision will be given to the developer.

Cost of running = 90,000 X 7 X 25 times/day X $(0.32-0.5)= $500,000/day

AI chips: powerhouse for large language models, enabling their ability to comprehend and produce intricate text through efficient handling of vast amounts of information simultaneously.

Importance

Improved Performance

AI chips significantly enhance performance by efficiently processing complex tasks, leading to faster and more powerful execution of various applications.

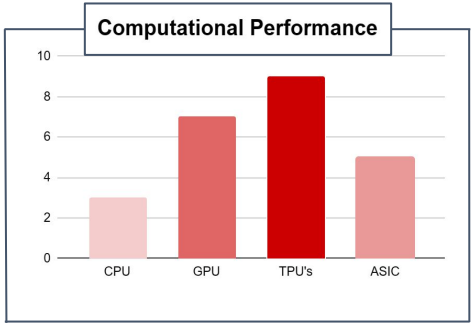
Reduced Training Time

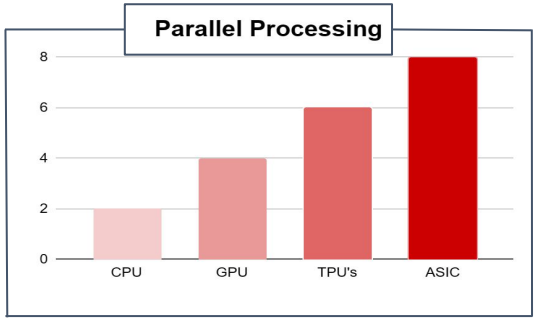
They expedite model training by leveraging parallel processing, significantly reducing the time required for large language models to learn and optimize their parameters.

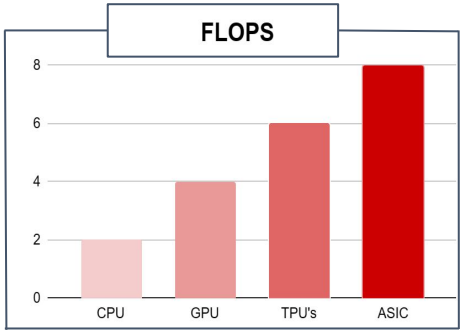
Lower costs

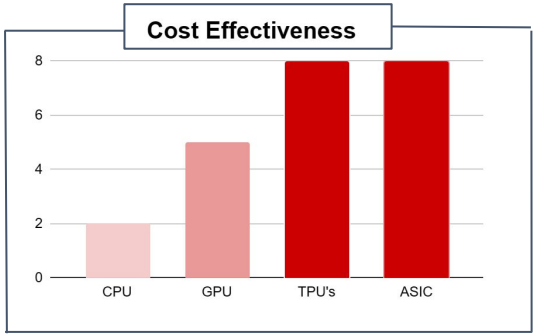
AI chips cut costs by efficiently managing energy and computation, leading to more affordable deployment and operation of large language models.

Analysis









Addressing AI chip Supply-Demand Gap is vital for technological advancement

Demand Drivers

Soaring AI Adoption:
High demand for powerful chips due to AI's widespread use to handle complex algorithms and train datasets.

Global AI Momentum:
Government initiatives worldwide have boosted demand for AI chips, as well as the surge in massive data generation has intensified the need for efficient AI chips across diverse applications.

Diversified Applications:
Edge computing, from smart devices to autonomous drones, necessitates a different breed of energy-efficient AI chips for on-device processing.

Supply Challenges

Manufacturing Challenges:
Complex semiconductor production process poses challenges in meeting rising chip demand as expansion takes time and huge investment.

Geopolitical Impact:
Trade conflicts and political instability disrupts supply chains, affecting access to crucial materials and facilities.

Talent Shortage and innovation:
Shortage of skilled experts delays the design and production of advanced AI chips. Intense competition for superior AI chips sometimes hinders immediate supply.

Impact on OpenAI

- Open-Source Initiatives
- In-House Expertise
- Diverse Collaborations for supply

Future Landscape

- Growing AI Chip Market
- Bridging the Gap between demand and supply
- Using Adaptable Approach to innovate

Startup Analysis

Parameter	Market Size	Market Share	USP	Team Expertise	Financial Stability	Customer Base	Compatibility	Cultural Fit	Reach
Rain Neuromorphics	Niche ($18.6bn - Neuromorphic)	Very small	Pioneering Neuromorphic Processors						
Graphcore	Established ($63.9bn - AI Accelerator)	Mid Range	High-Performance & Good Accelerators,intelligent processing unit" (IPU) architecture						
Cerebras	Established ($73.9bn - AI Accelerator)	Small	Unsurpassed Processing Power, wafer scale chip architecture						
Blaize	Growing ($12.5bn)	Very Small	Efficient & Low-Power Edge Processors						
Mythic	Niche ($18.6bn - Neuromorphic)	Negligible	Analog AI for Performance & Efficiency						

For complete analysis: Click here

Overall Leader
Graphcore

Inferences:-

- **Graphcore emerges as a well-rounded leader,** excelling in financial stability, customer base, infrastructure, and reach. Their compatibility with OpenAI might need work, but overall, they offer a stable and proven solution.
- **Blaize shines in compatibility and infrastructure,** making them ideal for real-world deployments close to data sources. Their cultural fit aligns well with OpenAI, making collaboration promising.
- **Rain Neuromorphics and Mythic are high-potential options for research-focused collaborations**. Their cutting-edge technologies align with OpenAI's values, but early-stage challenges require long-term commitment.
- The average of all scores based on colour palettes was taken to identify the overall score

Firm Analysis

Parameter	Market Size	Market Share	USP	Team Expertise	Financial Stability	Customer Base	Compatibility	Cultural Fit	Reach
NVIDIA	>$25 billion	78%	A100 GPU: 54 billion transistors, 400W						
Google Cloud TPUs	>$10 billion	7%	TPUv4 Pod: 4,096 chips, 900W, 1.1 exaflops						
Intel Ponte Vecchio	>$15 billion	Emerging	Xe Link interconnect: 256 GB/s, multi-die architecture						
Marvell	>$8 billion	10%	ThunderX3: 96 Arm cores, 150W, 100 Gbps Ethernet						
Tencent Cloud XuanTie	>$5 billion	25%	XuanTie 910: 256 cores, 140W, high memory bandwidth						

For complete analysis refer : Click here

Overall Leader
NVIDIA

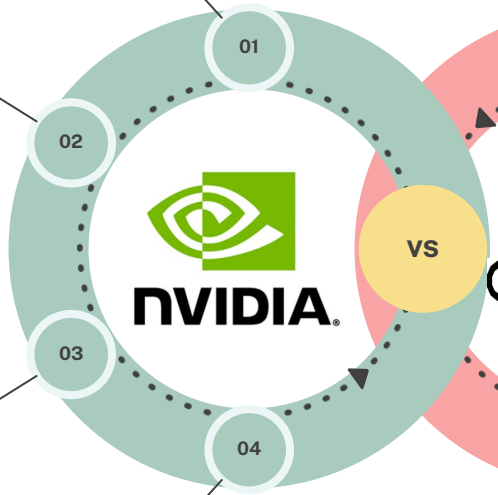
Inferences:-

- Nvidia has achieved the highest financial stability, customer base and Reach which can help OpenAI cater to maximum audience
- Google Cloud has a high Team Expertise and is culturally fit for OpenAI
- Intel is highly compatible with OpenAI but lacks in Reach
- The average of all scores based on colour palettes was taken to identify the overall score

Recent Venture Analysis of OpenAI

	ANTHROP\C Anthropic (Silicon Design Startup)	RAIN Investing in Brain-Inspired Chips	OpenAI Project Tigris (Internal Initiative)
IMPACT	10x faster LLM training 50-70% cost reduction for LLMs Increased control over AI hardware	Explore alternative chip architectures Improve energy efficiency and performance	Challenge NVIDIA's dominance Increase diversity and innovation in AI chip landscape
OPPORTUNITIES	Dominate LLM chip market segment Foster open-source collaboration Democratize access to powerful AI	Pioneer next-gen AI chips with human-like intelligence Uncover new applications for neuromorphic chips beyond AI	Offer differentiated AI chips for specific needs Cater to beyond LLMs and specific AI tasks
FIGURES	$8-10 billion funding sought GPT-3: 60-70% LLM market share AI chip market: $372.7 billion by 2028	$51 million invested in Rain AI's neuromorphic chips Research on brain function and neurotechnology advancement	Potential billions in funding sought NVIDIA GPU dominance in AI (61% market share)

FIRM VS STARTUP



01. Market leadership

NVIDIA dominates the AI chip market with a vast ecosystem of developers and established relationships with cloud providers. OpenAI could leverage this reach to accelerate its impact.

02. Proven Technologies

NVIDIA GPUs are proven and powerful, offering immediate access to high-performance hardware for OpenAI's research.

03. Financial stability

NVIDIA's massive resources could ensure long-term stability and potentially fund OpenAI's research ambitions.

04. Reduced Competition

Merging with a competitor could eliminate market rivalry and create a combined force driving AI innovation.

01. Complementary Technology

Graphcore's IPUs are specifically designed for AI workloads, unlike NVIDIA's GPUs repurposed for AI. This could offer better performance and efficiency for OpenAI's research.

02. Open Source Alignment

Graphcore's commitment to open-sourcing its software and hardware aligns well with OpenAI's values of democratizing AI. This could foster closer collaboration and transparency.

03. Lower risk of cultural clash

Both companies have a strong focus on research and innovation, potentially creating a smoother integration.

04. Acquisition cost

An acquisition would likely be cheaper than a merger with a giant like NVIDIA.

Conclusion

Integration complexity

Both options involve integration challenges, but merging with a larger company like NVIDIA could be significantly more complex and risky.

Loss of autonomy

Merging with NVIDIA could dilute OpenAI's independence and influence on its technology roadmap

Intellectual property

NVIDIA's closed-source approach could clash with OpenAI's commitment to open-sourcing its work.

Recommendation:

The ideal decision depends on OpenAI's priorities and risk tolerance.

- If prioritizing cutting-edge AI technology, open-source principles, and cultural fit, acquiring **Graphcore** might be preferable.
- If market reach, immediate access to powerful hardware, and financial stability are top priorities, merging with **NVIDIA**

Additional considerations:

OpenAI could also explore a partnership or joint venture with either company, gaining benefits without full commitment. It could continue developing its own chips alongside an acquisition or merger, diversifying its access to technology.

APPENDIX

Assumptions in calculation of Cost Analysis

- Currently, there are 20 million daily active users on the free version of ChatGPT, and an additional 500,000 daily active users on the Plus version. OpenAI incurs a cost of $0.000225 per token processed.
- The majority of OpenAI's revenue and profit stems from the enterprise model and API integrations, both of which feature negotiable pricing. Consequently, for the standard model accessible to the general public, OpenAI aims to achieve a breakeven between its operational costs and revenue.

Analysis of different available chips

Chips / Parameter	TPU v3	v100	a100	Cerebras WSE	GraphCore IPU
Efficiency	80-100%	70-93%	70-93%	33%	61%
Energy Efficiency	9/10	8/10	10/10	5/10	9/10
Memory/Model Size	9/10	8/10	9/10	6/10	7/10
Memory Efficiency	9/10	8/10	9.5/10	6/10	7/10
Area Efficiency	10/10	7/10	8/10	4/10	6/10

For complete data refer :[TPU vs GPU vs Cerebras vs Graphcore: A Fair Comparison between ML Hardware | by Mahmoud Khairy | Medium]

Estimating OpenAI's Annual Nvidia GPU Costs: A Guesstimate

Assumptions:

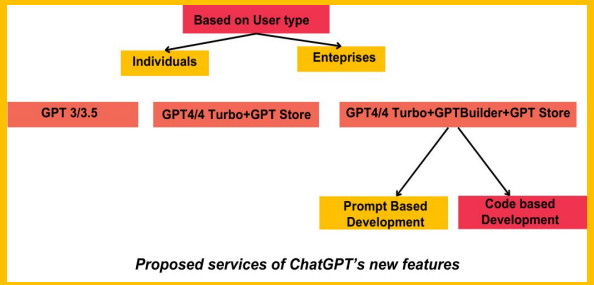
- Upgrade Frequency: We assume OpenAI upgrades their entire system annually, though partial upgrades are possible.
- System Scale: We base estimates on the 5,760 A100 GPUs previously used in their supercomputer, potentially increasing for H100 usage.
- Chip Mix: We consider a scenario with both A100s and H100s, potentially shifting towards more H100s in the future.

Cost per Chip:

- A100: $50,000 (based on market data and estimations)
- H100: $70,000-$80,000 (estimated range based on current market trends and A100 pricing)

Scenarios:

1. Pure A100 System (Baseline):
    - Annual Cost: 5,760 chips * $50,000/chip = $288 million
2. 50% A100, 50% H100 Mix:
    - Annual Cost: (2,880 A100 chips * $50,000/chip) + (2,880 H100 chips * $75,000/chip) = $381 million (assuming an average H100 cost of $75,000)
3. Full H100 System (Maximum Estimate):
    - Annual Cost: 5,760 chips * $75,000/chip = $432 million



Proposed services of ChatGPT's new features

Reasons for Gap in Supply and Demand of AI chips

Reason	Details	Numbers and Statistics	References
Limited Production Capacity	- Concentration of production in few major foundries (TSMC, Samsung) - Long lead times for building new fabs (2-3 years)	- Global fab utilization rate over 90% in 2023 - Lead time for advanced AI chips up to 52 weeks	- Gartner report: "Market Trends: Semiconductor Manufacturing Capacity" - Semiconductor Engineering article: "Fab Utilization at 92%, Lead Times Remain Long"
Geopolitical Tensions	- Export controls and trade restrictions (e.g., US-China trade war) - Disruptions to supply chains due to geopolitical events (e.g., Ukraine war)	- 25% tariff on some Chinese-made AI chips due to trade war - Disruption in neon gas supply from Ukraine for chip production	- The Information article: "Trade War Takes Bite Out of AI Chip Supply" - Reuters article: "Ukraine War Threatens Neon Supply for Chipmaking"
Rapidly Evolving Technology	- Constant advancements in AI chip architectures and technologies - Shifting demand for specific AI chips as applications evolve	- Neuromorphic computing advancements requiring new production processes - Edge AI chip market expected to grow at 35.7% CAGR from 2023 to 2028	- IDC report: "Emerging Technologies & Trends in the AI Chip Market" - VentureBeat article: "Edge AI Boom Drives Demand for Specialized Chips"

Project Tigris

Investment Guesstimate: $1.5 billion initial with a 5-year target of $7.5 billion valuation.

Competition: Nvidia holds 80% of the AI chip market share (AI Hardware Market Report 2023 by McKinsey & Company). Toppling them requires substantial war chest.
Technology Risk: Developing a competitive AI chip takes 2-4 years (Forbes article: "The AI Chip Race Heats Up"). Assuming similar timelines, factoring in R&D and manufacturing complexities justifies a higher valuation ceiling.
Growth Potential: The AI market is expected to reach $1.6 trillion by 2025 (Statista AI Market Forecast). A 5% share for Tigris in 5 years could translate to a significant valuation.
Data Points:

Similar AI chip venture Cerebras raised $425 million (Crunchbase), indicating the initial investment ballpark.
AI chip design cost estimates range from $200 million to $1 billion (MIT Technology Review article: "The Cost of Building an AI Chip").
Rain Neuromorphic Chips:

Deal Guesstimate: $75 million over 3 years, with potential for extension.

OpenAI's Investment: $51 million letter of intent reported (Wired article: "OpenAI Agreed to Buy $51 Million of AI Chips"), suggesting a larger final deal is possible.
Rain's Stage: Early-stage startups typically secure funding in rounds of $10-25 million (CB Insights report: "Early-Stage Startup Funding Trends"). A multi-year, phased investment aligns with this pattern.
Shared Benefits: Collaboration allows OpenAI access to advanced technology and Rain secures a reliable customer and development partner.

Similar neuromorphic chip startup Sentient Technologies raised $80 million (Crunchbase), providing a valuation reference.
AI chip development partnerships often involve multi-year technology licensing agreements, suggesting a longer engagement potential.
Jony Ive Collaboration:

Project Status Guesstimate: On hold, with earlier funding discussions exceeding $5 billion.

Altman's Departure: His exit creates leadership uncertainty, potentially stalling the project (Reuters article: "OpenAI CEO Altman Sought Billions for AI Chip Venture").
Complexity and Cost: Consumer AI devices involve high design and manufacturing costs (IEEE Spectrum article: "The High Cost of Building Artificial Intelligence"). A $10 billion+ valuation reflects this risk.
Market Uncertainties: Similar consumer AI devices haven't achieved widespread success, making investors cautious about large upfront investments.

Apple's HomePod, a voice-activated AI device, reportedly cost over $1 billion to develop (The Information article: "Inside Apple's HomePod Flop"). This highlights the potential cost scale.
Similar AI-powered smart displays like Amazon Echo Show have seen limited user adoption (Statista report: "Smart Display Market Share"). This signifies market uncertainties.

REFERENCES

- Gartner report: "Market Trends: Semiconductor Manufacturing Capacity"
- https://www.semi.org/sites/semi.org/files/2022-12/glo-csi-dhl-resilience-of-the-semiconductor-supply-chain.pdf
- The Information article:"Trade War Takes Bite Out of AI Chip Supply"
- https://www.reuters.com/technology/exclusive-ukraine-halts-half-worlds-neon-output-chips-clouding-outlook-2022-03-11/
- IDC report: "Emerging Technologies & Trends in the AI Chip Market" VentureBeat article: "Edge AI Boom Drives Demand for Specialized Chips"
- https://www.engineering.com/story/openai-may-design-their-own-chips
- https://techcrunch.com/2023/10/06/openai-said-to-be-considering-developing-its-own-ai-chips/
- https://www.datacenterdynamics.com/en/news/openai-considers-making-its-own-ai-chips-acquiring-hardware-business/
- https://arstechnica.com/information-technology/2023/10/openai-may-jump-into-ai-hardware-amid-high-costs-supply-constraints/
- https://techcrunch.com/2023/10/06/openai-said-to-be-considering-developing-its-own-ai-chips/
- https://www.forbesindia.com/article/cryptocurrency/elon-musk-sets-sights-on-openais-chatgpt-with-new-ai-startup/84451/1
- https://openai.com/blog