Visualizing the Large Language Model Industry

Process Book & Project Proposal

Ashkan Aleshams - Lorena Buciu

ASIP SQUAD
CSC316
Caronlina Nobre
April 1, 2025

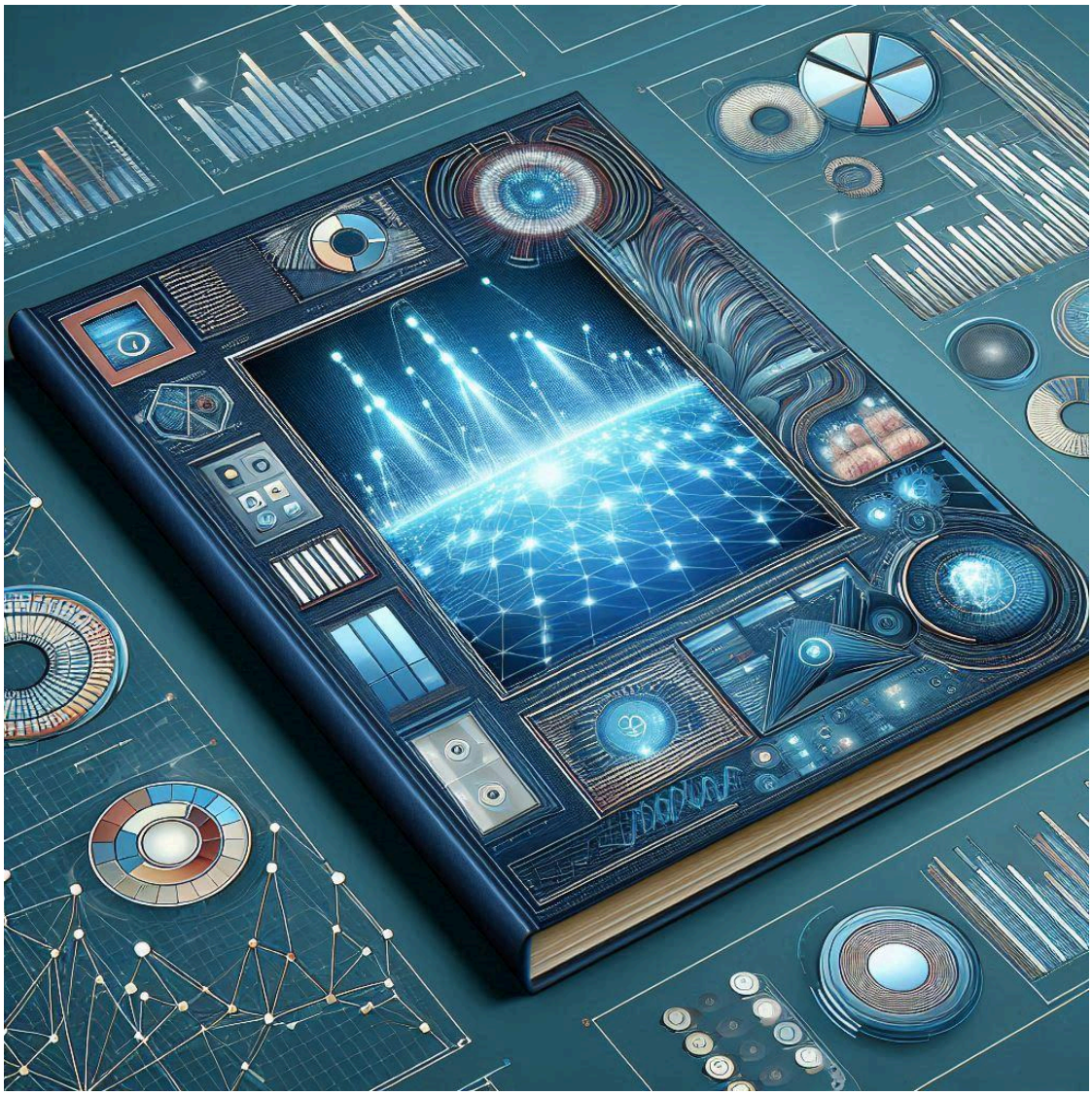


Table of Contents

Table of Contents **1**
Project Proposal **2**
Basic Information 2
Abstract 2
Justification of Topic - Background and Motivation 2
Related Work 3
Data 5
Audience and Questions 5
Data Cleanup 5
Team Agreement 5
Project Mapping **8**
Discussion 8
Focused Target Audience 9
Initial Questions 9
Datasets **10**
Initial Visualizations 11
Planning **18**
Sketch Step 18
Decision step 23
Storyboarding 26
Prototyping **28**
V1 Prototype 28
V2 Prototype 30
Testing **33**
Think Aloud Study 33
Post-testing Reflection 34
Final **36**
Final Reflection 36
Demo and Link 36
Bibliography **37**

Project Proposal

Basic Information

Project Title: Visualizing the Large Language Model Industry.

Team Name: ASIP SQUAD

Team Info:

Role	Name	Utorid	Student #	email
Lead Author	Ashkan Aleshams	Aleshams	1007263621	Ashkan.aleshams@mail.utoronto.ca
Author	Lorena Buciu	buciulor	1006759456	lorena.buciu@mail.utoronto.ca

Abstract

Some argue that the current AI craze is equivalent to the .com boom of the early 21st century. Similar to that era, economists and experts such as Brent Thill suggest that AI's potential and capabilities are overestimated in the short term but underestimated in the long term (CNBC, 2024). This project examines the LLM (Large Language Model) industry, analyzing key players, user adoption, and industry statistics from an entrepreneurial software development perspective. By leveraging data visualization, we aim to provide insights into the evolving landscape of AI-driven development, highlighting trends, opportunities, and challenges for entrepreneurs and developers navigating this rapidly growing sector.

Justification of Topic - Background and Motivation

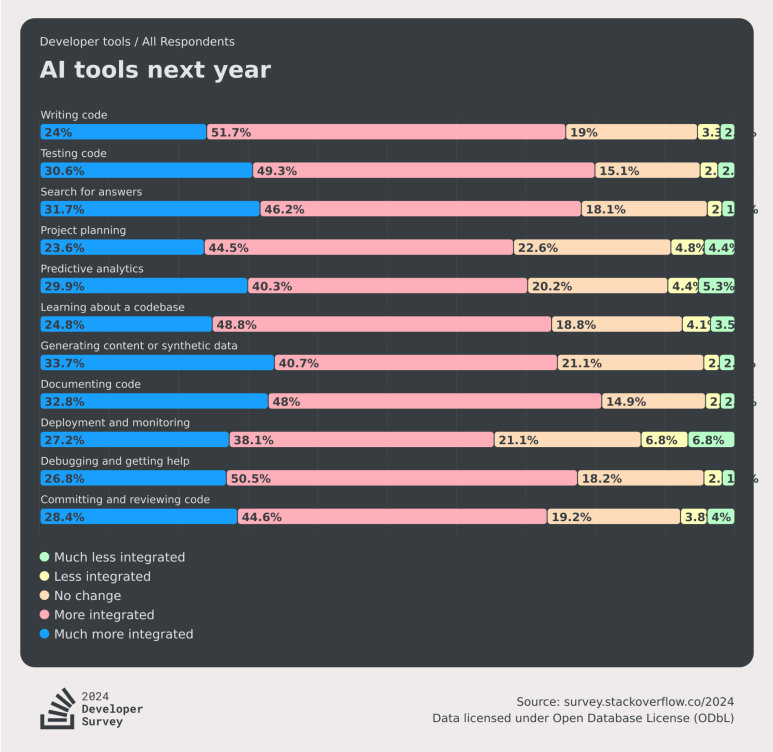
As computer science students and software developers, we interact with generative AI and Large Language Models (LLMs) daily, yet we recognize gaps in our understanding of their capabilities, market dynamics, and cost structures. Given the rapid adoption of LLMs in the

software industry—where 76% of developers are already using or planning to use AI tools (Stack Overflow, 2024)—it is crucial to explore how these technologies are evolving. Additionally, as aspiring entrepreneurs, we seek insights into LLM pricing models and business applications to inform future projects and startups.

Related Work

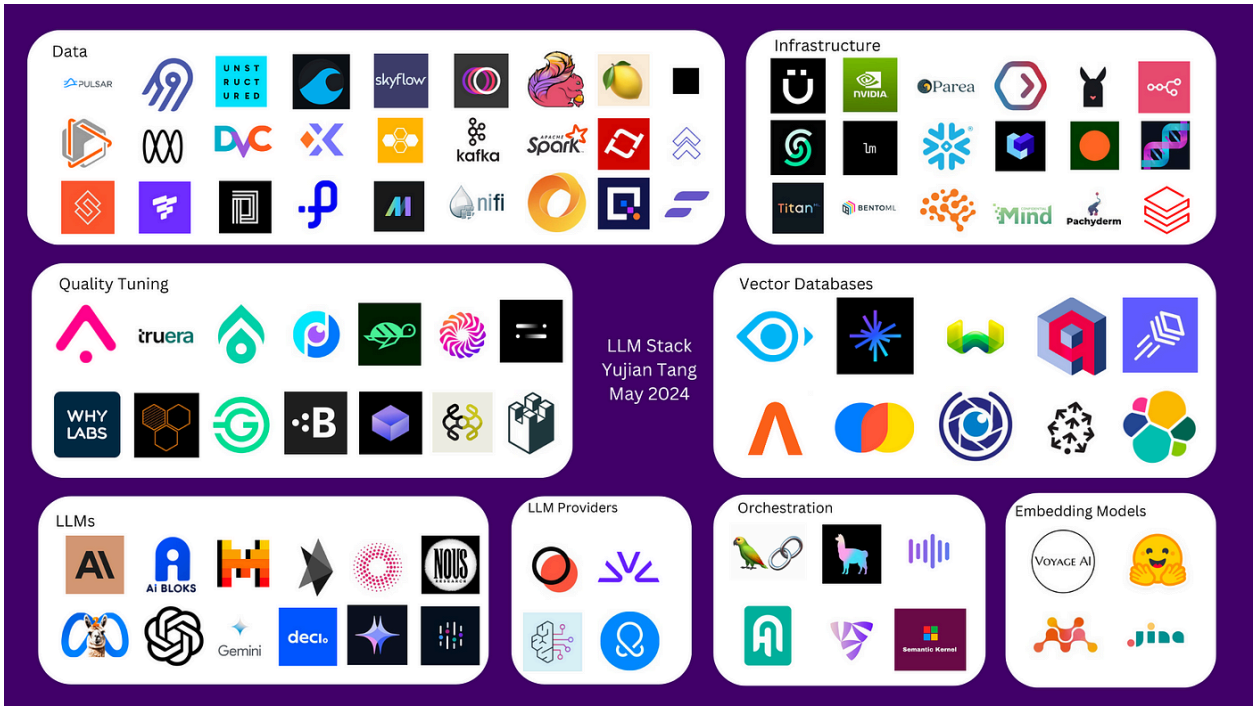
The following visualizations inspired us to create our own project surrounding AI. We found it difficult to find truly unique data visualizations in this space, so we feel that we can add valuable insights with our visualizations.

Interactive Visualization of a survey of 65,000 Software Developer by Stackoverflow:



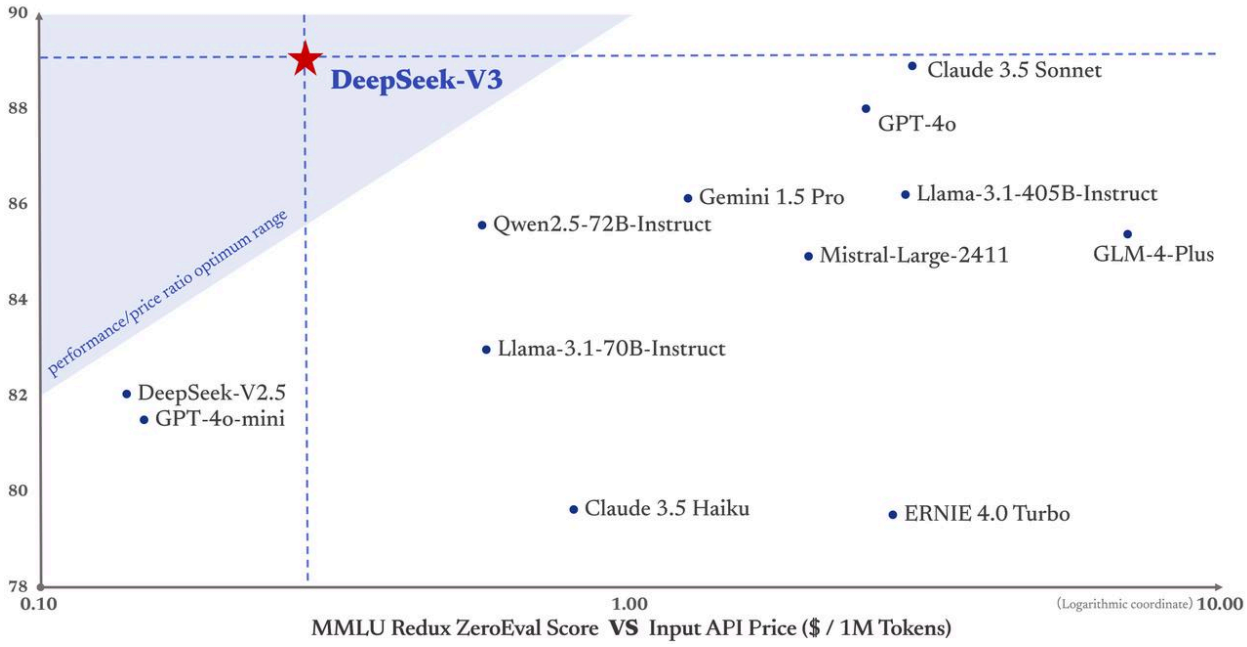
Source: (Stackoverflow)

This graphic was a motivation to look at the different players in the LLM market, with a lot of questions and doubts surrounding the LLM/AI industry, visualizations will help us address these.



Source: (Tang, 2024)

The release doc of deepseek v3 features an effective and jaw dropping visualization on performance of LLMs vs their cost per token.



Source: (*Introducing DeepSeek-V3*)

Data

We will primarily rely on datasets produced by companies in the LLM industry, including financial reports, benchmarks, and API usage statistics published by major AI providers such as OpenAI, Google DeepMind, and Meta. Additionally, we will consult credible third-party organizations such as Statista, McKinsey, and academic research papers to cross-check details and validate industry claims. Public datasets from surveys, GitHub AI adoption trends, and market research reports will also provide valuable insights into LLM adoption, developer usage patterns, and pricing models.

Audience and Questions

Our target audience is software developers, entrepreneurs and computer science students who are interested in the LLM space. Our goal is to answer several key questions about the LLM industry by examining the data described above. These include how software developers are leveraging LLMs currently and identifying any trends involved in their usage. We will also aim to explain the profitability of the LLM industry by visualizing financial reports, including how much LLM tools make, the costs of research and development and the cost of tokens.

Our overall goal of this project is to help software developers and entrepreneurs decide which LLM is best suited for their needs and how profitable a potential company based on that LLM could be.

Data Cleanup

Some data cleanup is expected upon deriving the data from the sources mentioned previously. Since we plan on using data from various third party organizations, we may need to standardize some data such as currencies or compute financial performance metrics ourselves if there is missing data. We plan on deriving quantities such as performance-to-cost ratios, cost per token, market share, and usage statistics over time to visualize the evolution of the industry. We will perform any data processing in JavaScript using a D3 script to clean and modify the dataset for our purposes.

Team Agreement

Communication

Use Discord as the primary team communication channel. All team members must respond within 24 hours maximum, preferably within 4 hours during business hours (9am-5pm EST). For urgent matters, ping individuals.

Respond to comments and resolve them both in Google Drive and Github in a timely manner before submission.

Weekly team meetings held every Friday at 11am-12pm EST via Zoom. Each member presents a 2-minute status update including progress, blockers and estimated completion time. Meeting notes stored in shared Google Doc. Attendance mandatory - notify other members 24hrs ahead for absences.

Occasionally, we will additionally meet outside the regular weekly meeting and the time for that meeting will be determined by a When2Meet link sent by the team leader (lead author). Attendance mandatory - notify other members 24hrs ahead for absences.

If you are unable to complete the tasks assigned to you, it is your responsibility to let everyone else know and find a team member to cover your tasks.

Code Guidelines and Version Control

We will work as an agile team using github (already has been created and members have been added) to manage our codebase, features (issues) and branches.

Main branch protected - no direct commits.

Merging a branch with Main requires approval from at least one other team member.

All code must be done in its appropriate branch.

All complex code requires inline comments explaining the logic. Comment on any workarounds or technical debt.

Approved pull requests should be merged using squash and merge.

Branch naming convention: **<feature or bug>/<issue #>-<good description>**

- e.g., feature/123-login-page

Commit convention: **<type>[optional scope]: <description> [optional body]**

- e.g., https://www.conventionalcommits.org/en/v1.0.0-beta.2/

Tasks

Together as a team, we will create specific and actionable tasks (Issues on Github) on a rolling basis with the direction of the team leader (lead author).

Tasks between members will be distributed equally and assigned by the team leader. However, team members are welcome to switch tasks amongst each other or volunteer for certain tasks.

Tasks (Issue on Github) will have descriptive names and descriptions with proper labeling. Update task status and progress frequently. Include time estimates, dependencies, and relevant screenshots/documentation

We will use Github projects to manage a kanban board showing the status of our tasks (in progress, in review, blocked, done)

The team leader is responsible for submitting assignments and making Github releases.

Quality Standards

All code must follow team style guide

No failing tests in main branch

Documentation updated with code changes

Successful deployment to staging required before merge

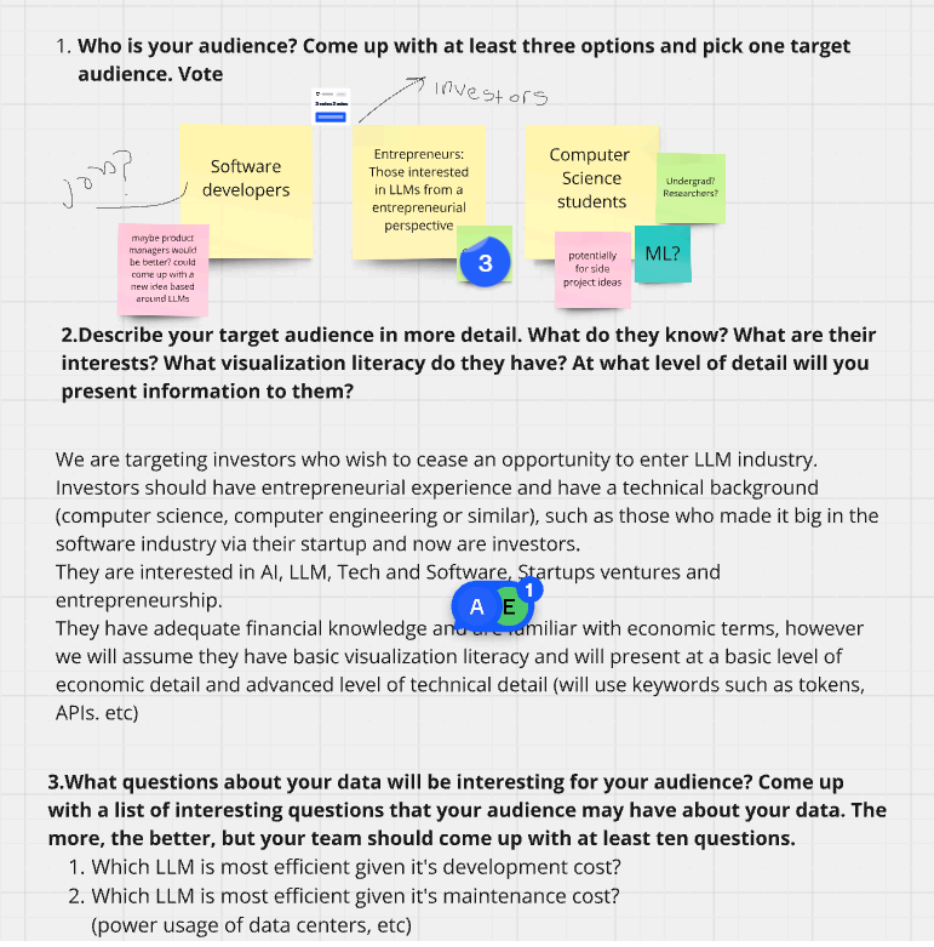
Team Signature and Acknowledgement

Role	Name	Date	Full name as Signature
Lead Author	Ashkan Aleshams	Feb 5, 2025	Ashkan Aleshams
Author	Lorena Buciu	Feb 5, 2025	Lorena Buciu

Project Mapping

Discussion

We used the discussion prompts provided to ideate and diverge about our audience and the direction of the project on this virtual whiteboard on Miro. With respect to our previous audience and the questions we aimed to answer as shown in the Audience and Questions section, we picked our top three possible target audiences which were (1) Software Developers, (2) Entrepreneurs and (3) Computer Science Students and brainstormed more about them. Then, we took a vote and unanimously selected Entrepreneurs, however this was still too vague so then we pivoted toward Investors with an entrepreneurial passion and technical background seeking to enter the LLM industry. We then continued to further brainstorm on the other prompt questions.



Screenshot of Miro virtual board

Focused Target Audience

Our focused target audience consists of investors with an entrepreneurial and technical background, such as in computer science or computer engineering, who are interested in opportunities within the LLM industry. Their interests span AI, LLMs, technology, software, startup ventures, and entrepreneurship. They have adequate financial knowledge from their entrepreneurship background and would be familiar with most economic terms, such as metrics from industry financial reports and investor relations insights. We will assume that they have basic visualization literacy, and we will provide a basic level of economic detail in our visualizations. A big focus of our visualizations is around the technical aspects of the LLM industry, so we will be presenting a more advanced level of technical detail with specific keywords.

Initial Questions

1. Which LLM is most efficient given its development cost?
2. Which LLM is most efficient given it's maintenance cost?(power usage of data centers, etc)
3. Which LLM is most economical for the user given its subscription or token cost?
4. What does the performances of LLMs look like given their input API price ($ /1M Tokens)
5. Which AI/LLM company has the most users?
6. Which AI/LLM company has been most profitable?
7. Which segment of the LLM industry (Data, Infra, LLMs, Quality Tuning, etc) is undervalued at the moment?
8. Which segment of the LLM industry (Data, Infra, LLMs, Quality Tuning, etc) has been most profitable for investors?
9. What might the acquisition cost of LLM providers look like? Which company in the LLM space is most expensive (Valuation, Stock price, etc)
10. What has been the customer acquisition cost for LLMs companies such as OpenAI, Deepseek, etc?
11. Which AI/LLM tools are currently used in the software development industry and who are their parent companies?
12. Which LLM is most widely used in sub industries (health tech, fintech, education etc)?
13. How scalable are LLMs for high volume use cases?
14. What are the barriers to entry in creating a company around an LLM?
15. What is the environmental impact of LLMs?

Datasets

Our first dataset is called LLMStats (Chavez Tamales) and it pertains to the pricing and performance of various LLMs and the providers that own them. The attribute “name” refers to the name of the AI provider and is categorical. It then contains an attribute called “providermodels” which contains a list of all the models offered by the provider. Within each model there is, “model_id” which is a unique human readable name for the model. The following attributes are all quantitative. “price_per_input_token” which is the price per input token in USD. “price_per_output_token”, which is the price per output token in USD. “throughput”, which is the processing speed in tokens per second. “latency”, which is the average response time for the model in milliseconds.

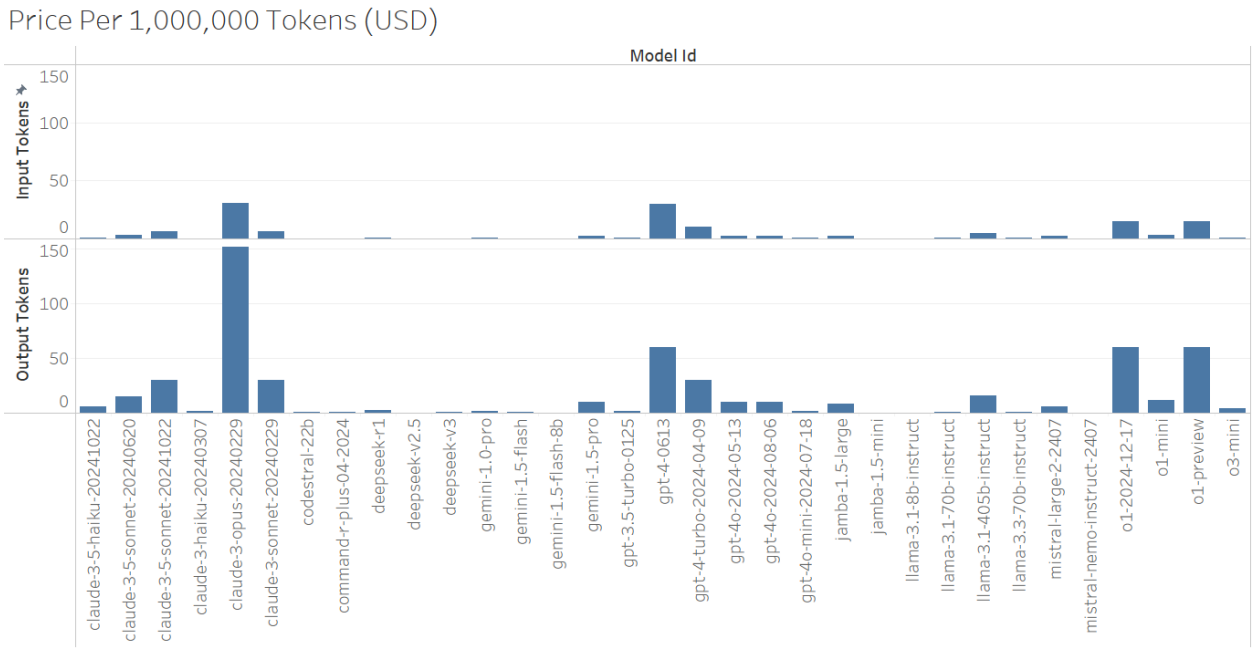
The Open LLM Leaderboard contents dataset by Hugging Face provides aggregated results for the Open LLM leaderboard which aims to compare LLMs in a reproducible way. The data consists of the following attributes: “eval name” - the name of the LLM, “average” - a weighted average of normalized scores from all benchmarks, “architecture” - the structure and design of the model, “CO2 cost” - co2 emissions of the model (kg), and “# params (B)” - the number of trainable parameters in the LLM, in billions

The use of artificial intelligence by businesses and organizations in producing goods or delivering services over the last 12 months, second quarter of 2024 (Statistics Canada, 2024) dataset is created by Statistics Canada and contains AI use percentage by various canadian business characteristics. The attribute “Business characteristics” refers to various business characteristics such as industry, company size, etc. The “Value” attribute refers to the percentage of AI used in that business characteristic.

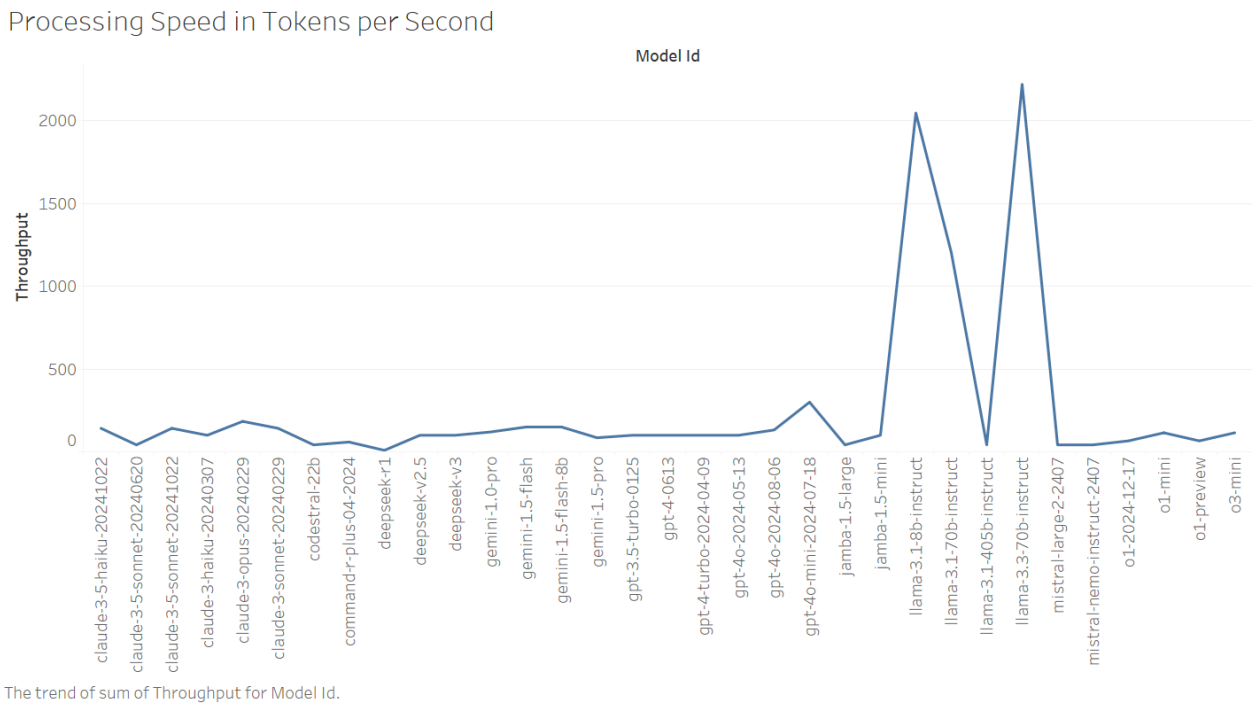
The Processing the artificial intelligence dataset (Innovation, Science and Economic Development Canada) is a dataset created by the Government of Canada to report the worldwide collection of approximately 85,000 AI patented inventions over the 1998-2017 period. We scalped this dataset and created a json file called “Ai_patents_wroldwide.json”. Its attributes are “Grouping” which refers to the type of AI patent, “Patent Inventions” which refers to number of patents in that grouping and “Percentage” which refers to the percentage of the grouping with respect to the 85,000 total AI patents.

Initial Visualizations

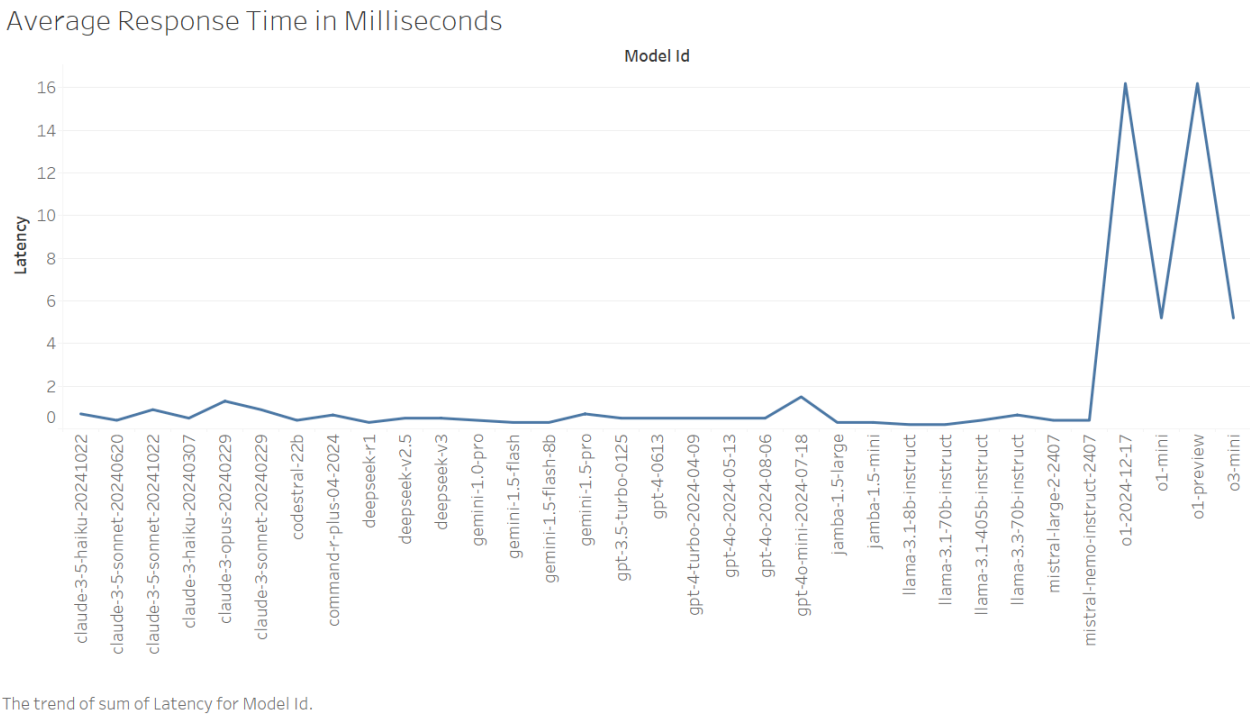
Initial visualizations (Collectively)



Source: (Chavez Tamales)



Source: (Chavez Tamales)

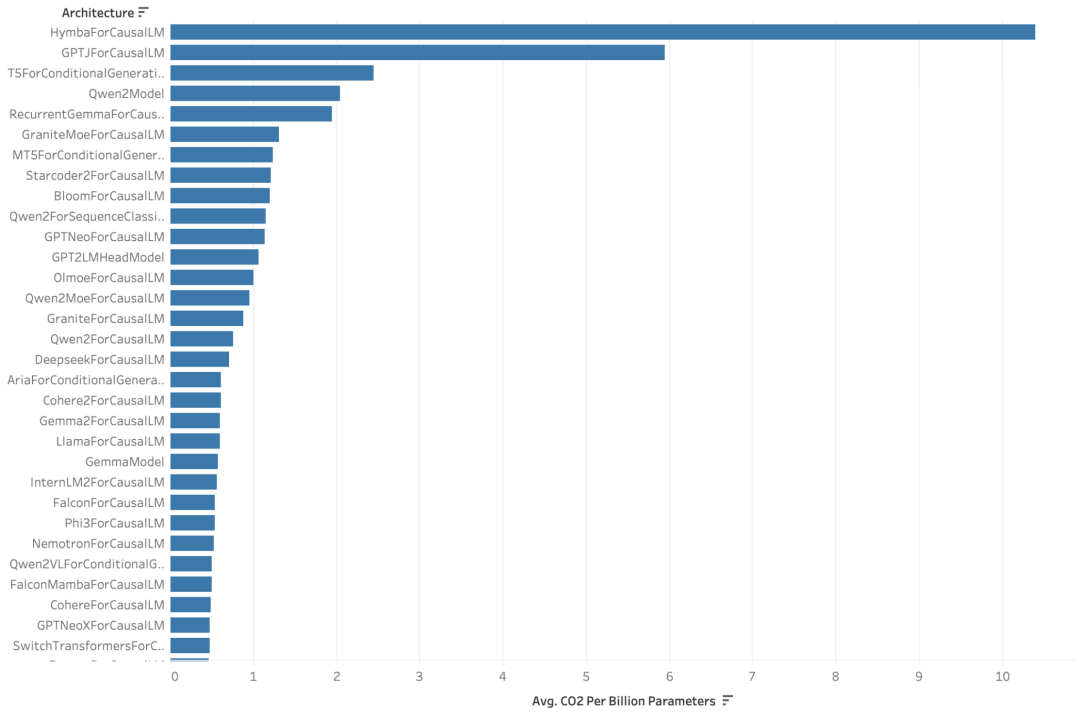


Source: (Chavez Tamales)

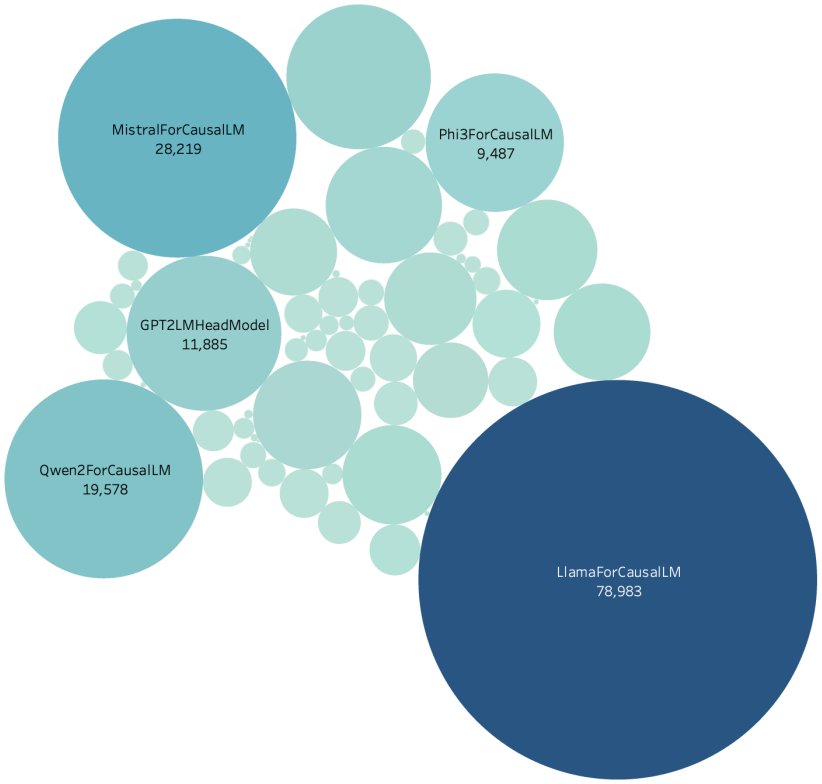
With these 3 visualizations, we answered initial questions 3 and 4. The last two visualizations show the performance of the various models in slightly different ways. One is latency and the other is the processing speed in tokens. The first visualization outlines the cost per million tokens for each model. We stuck with the original questions we came up with as pricing and performance would be very important to know for entrepreneurs looking to enter the LLM space.

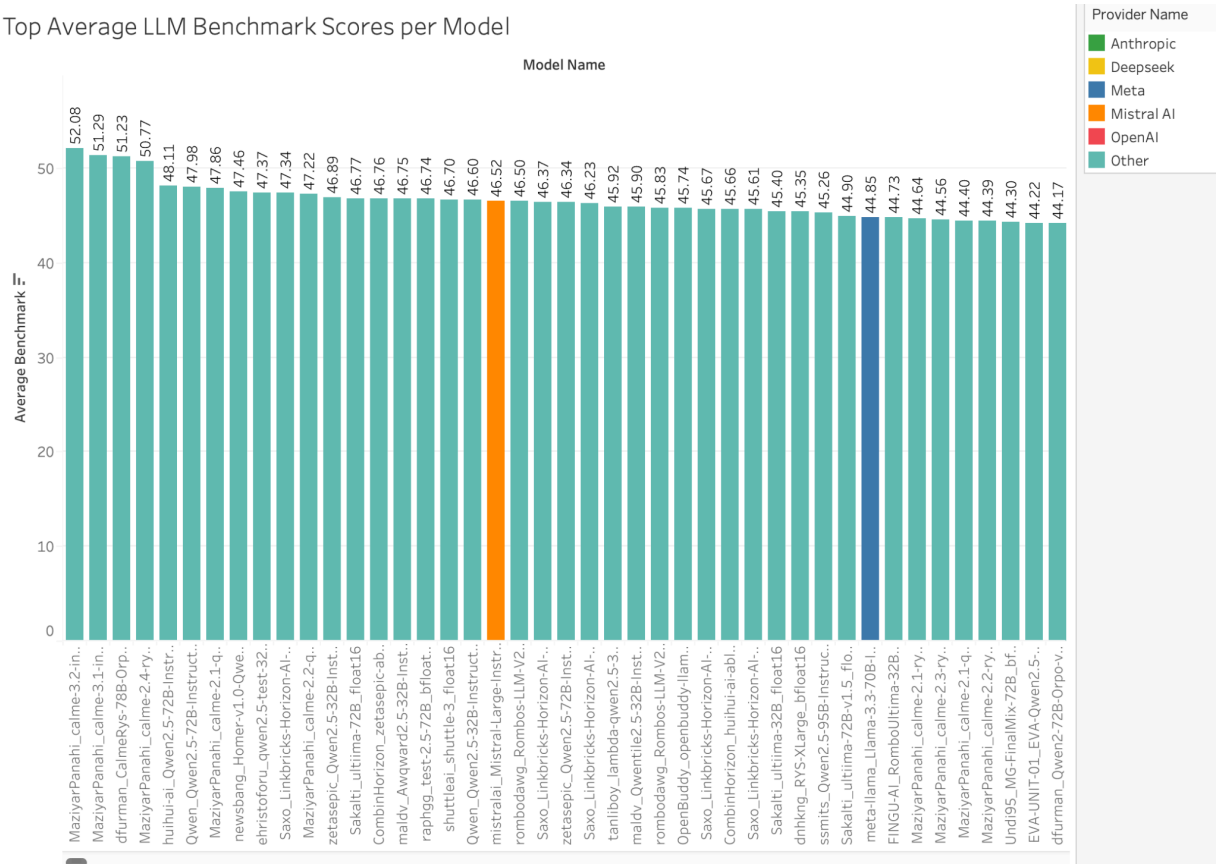
Lorena's initial visualizations

Average CO_2 emissions per billion parameters for LLM architectures



LLM Architecture Popularity Based on Hugging Face Hub Likes

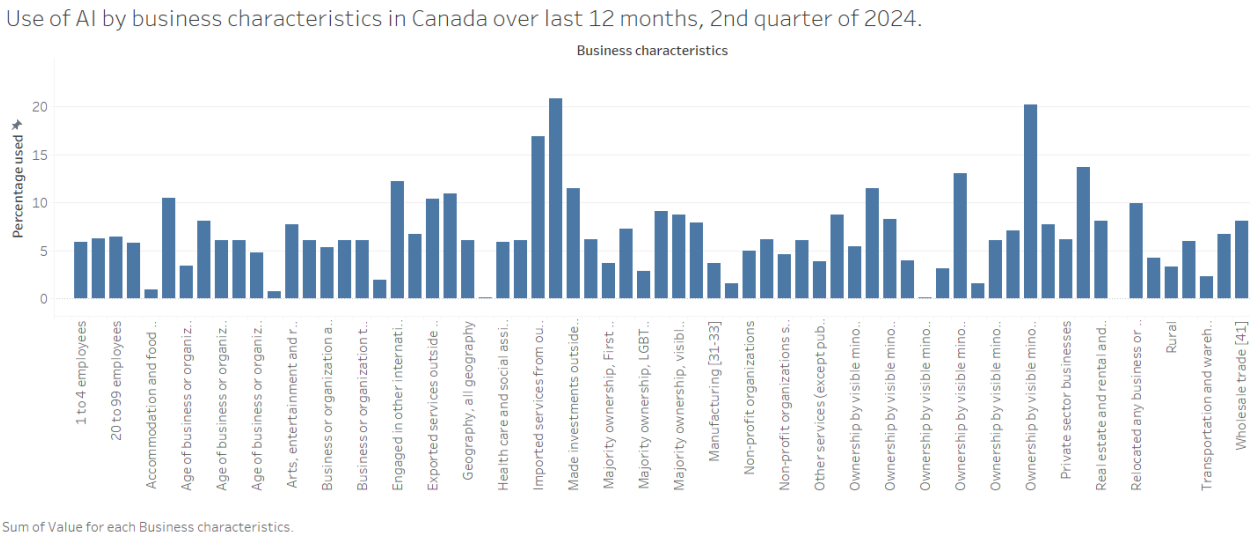




Data Source for all visualizations:
https://huggingface.co/datasets/open-llm-leaderboard/contents

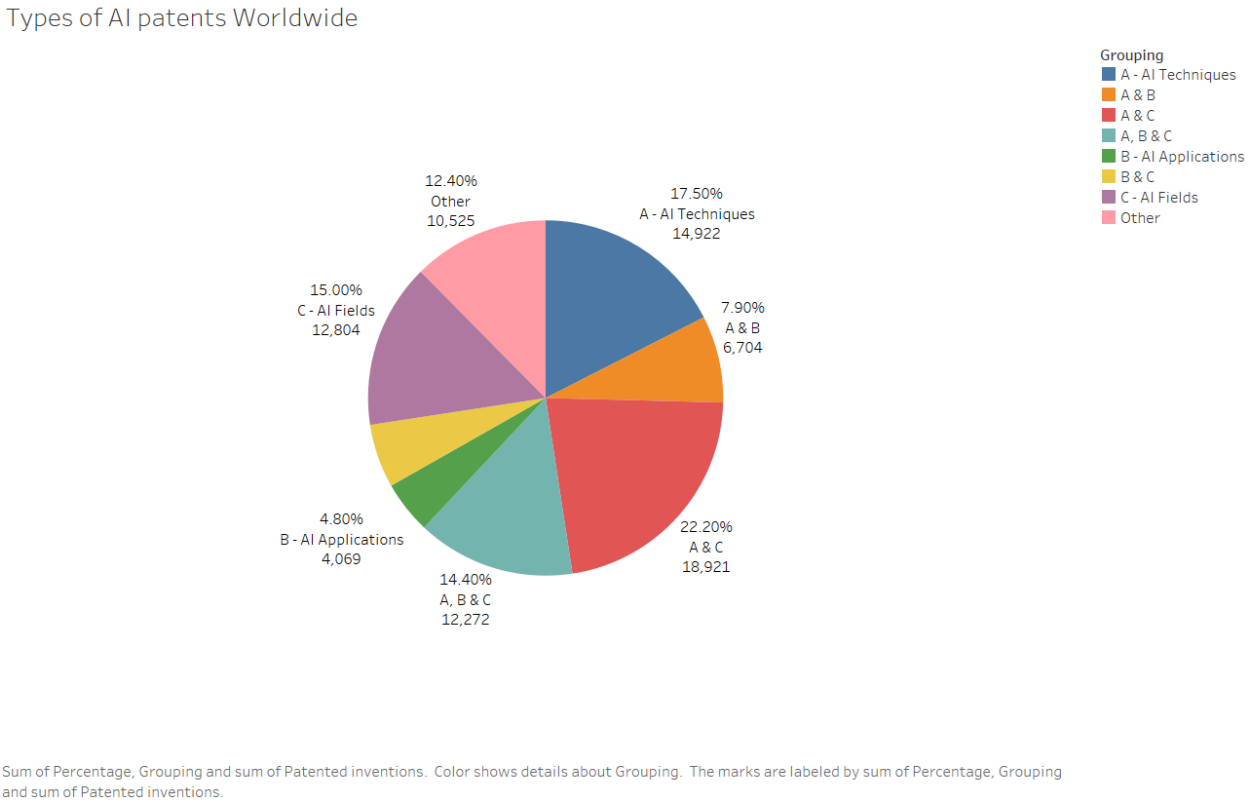
For visualization #1, the visualization slightly differs from the original question “What is the environmental impact of LLMs?” as it is more specific analyzing CO2 emissions per billion parameters for different model architectures. I think this is better to show the specific environmental impact from the data that is available, the question could be better framed as “What are the CO2 contributions of different LLM models?”. Visualization #2 shows the popularity of the model by likes on Hugging Face, which is not necessarily in our question list but is related to “Which AI/LLM company has the most users?, since the dataset used doesn’t have any information on users. Visualization #3 aims to answer “How scalable are LLMs for high volume use cases?” using the data available from the dataset, comparing average benchmarks for each LLM to get an idea of which LLMs have the best performance. We should aim to find specific metrics related to scalability.

Ashkan's initial visualizations



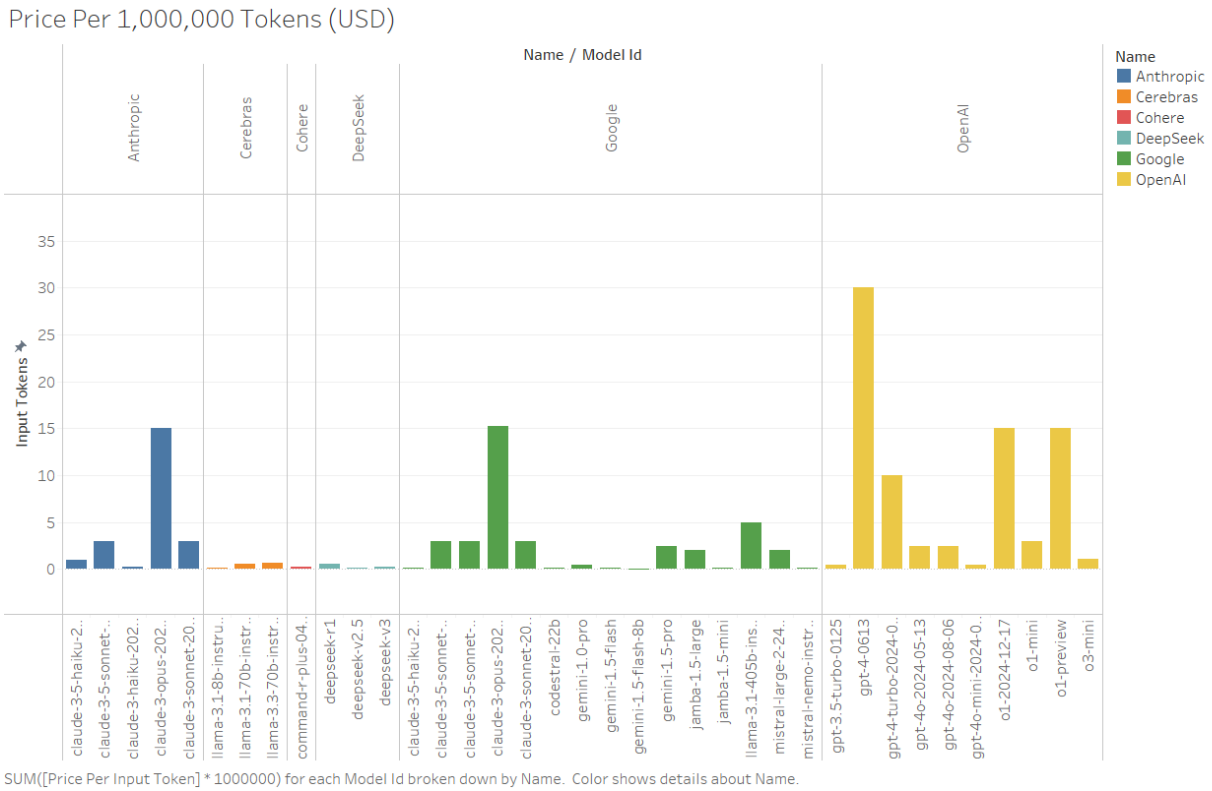
Source: (Statistics Canada, 2024)

The Tableau visualization, which answers the question: "What is the prevalence of AI usage across different business characteristics in Canada within the last 12 months, specifically for the second quarter of 2024?", differs significantly from our team's initial research questions. Our original questions aimed to explore the competitive landscape of the LLM industry, focusing on aspects like cost-efficiency, market share, and profitability. However, the LLM market is so new that comprehensive datasets for these metrics simply don't exist. Financial information is proprietary, and standardized performance benchmarks are still under development. Consequently, we used an available dataset on general AI adoption in Canada, which, while valuable, doesn't isolate LLMs. This data availability dictated the questions we could realistically answer. While our initial questions remain relevant, the lack of data forced a shift to a broader perspective on AI adoption across various business sizes, sectors, demographics, and activities. This broader view, while not directly addressing our LLM-specific questions, still provides valuable context about the overall AI market and potential demand for LLM services as the industry matures and more data becomes available.



Source: (Innovation, Science and Economic Development Canada)

The questions we answered in Tableau were largely shaped by the available dataset and the visualization methods that best represented the information. While our team initially explored broader questions about AI and LLM trends in relation to finances, we pivoted to a more data-driven approach, leading to the question: "How many AI patents have been filed, and what trends do they depict?" The key difference is that our original questions aimed for open-ended insights, whereas the Tableau analysis required more specific, quantifiable questions. As a result, the visualizations focused on breaking down AI patents by category and their distribution across different AI fields. Some of these visualized questions proved more effective because they provided clearer, data-backed insights, while our original questions encouraged broader discussions. Ultimately, we adjusted our analysis to align with the available data, ensuring meaningful and interpretable results.



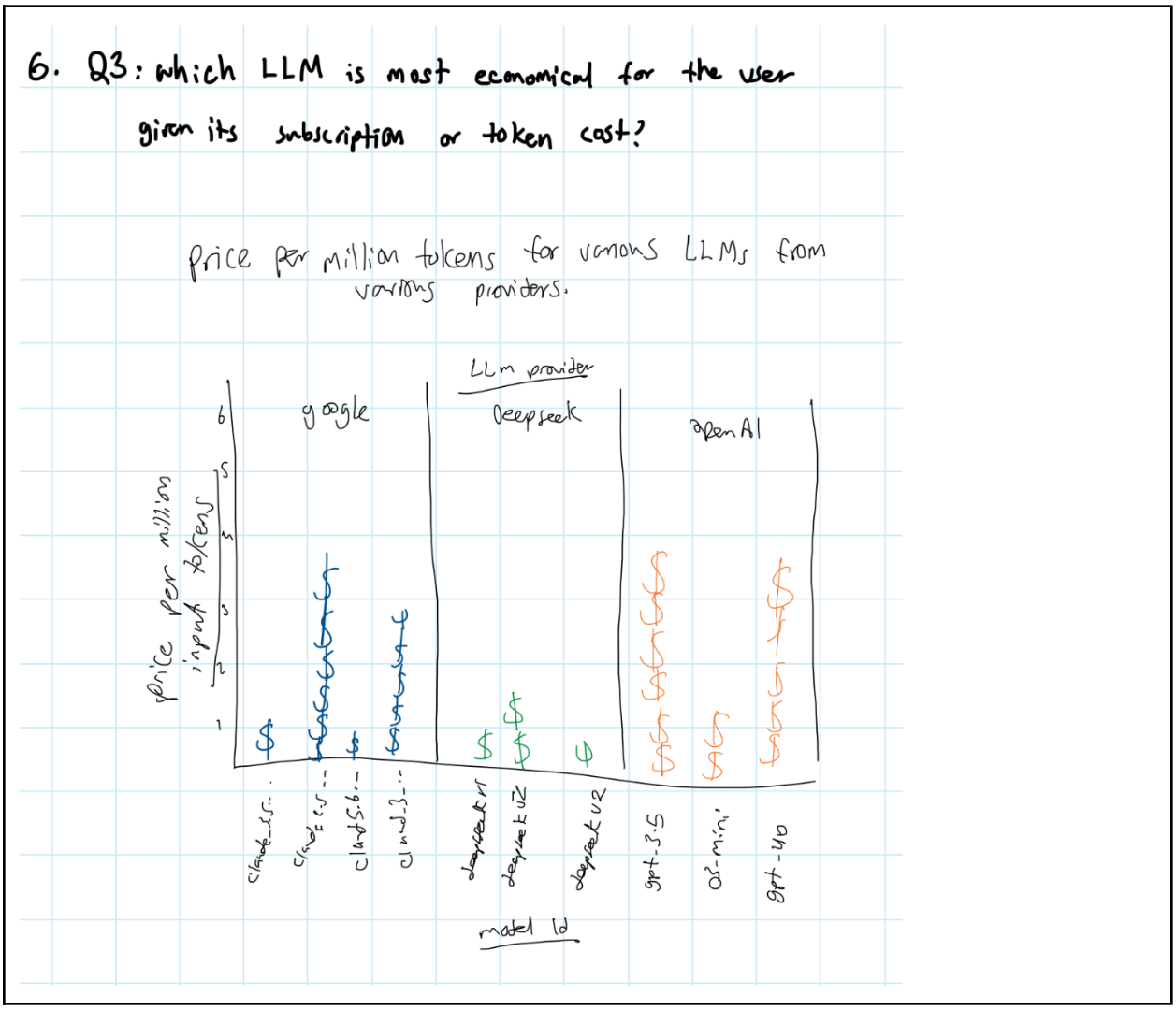
Source: (Chavez Tamales)

This bar chart visualizes the input price per million tokens for a range of Large Language Models (LLMs) from various providers, directly addressing the question of user cost efficiency (Question 3). The chart clearly displays the price differences between models, allowing for a quick comparison of relative cost. Color-coding groups models by their respective providers (Anthropic, Cerebras, Cohere, DeepSeek, Google, and OpenAI), making it easy to see the pricing strategies of different companies. However, while this visualization effectively shows the raw price per million tokens, it doesn't yet account for the nuances of token usage in real-world tasks. To more directly address user cost efficiency, future iterations could incorporate estimated token consumption for representative tasks, allowing users to compare the actual cost per task (e.g., cost to summarize a document or generate a blog post). Furthermore, the chart currently lacks any performance metrics, limiting its ability to address Question 4, which explores the relationship between price and performance. Adding a secondary axis or creating small multiples to display performance data alongside price would significantly enhance the visualization, enabling users to make more informed decisions based on the price-performance trade-offs of different LLMs. Finally, tooltips providing additional context, such as specific model details and pricing tier information, would further improve the chart's usability.

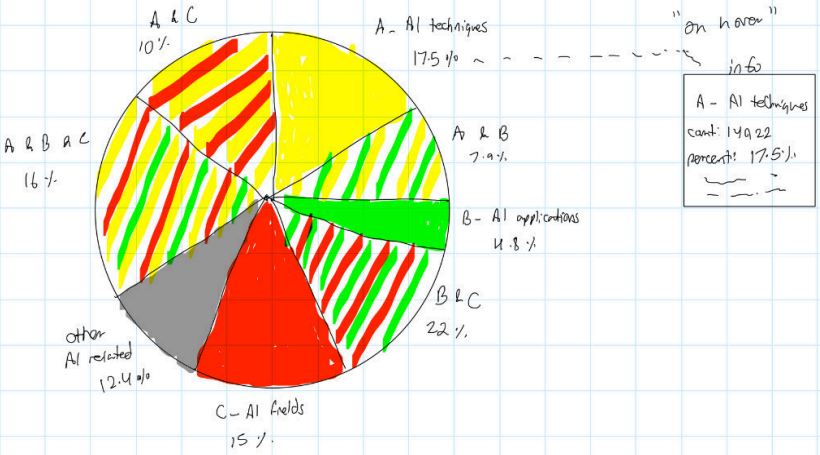
Planning

Sketch Step

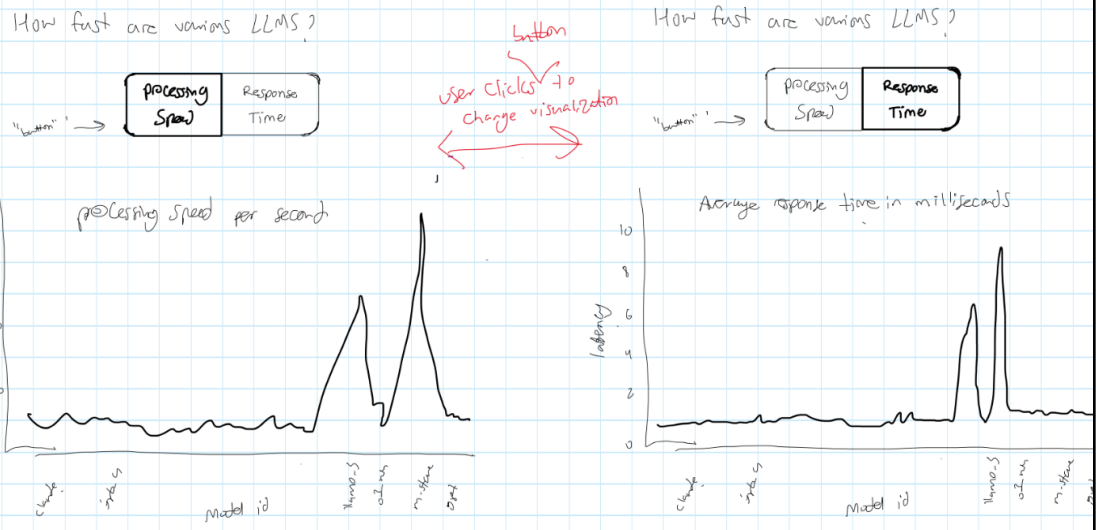
Ashkan's Sketches:



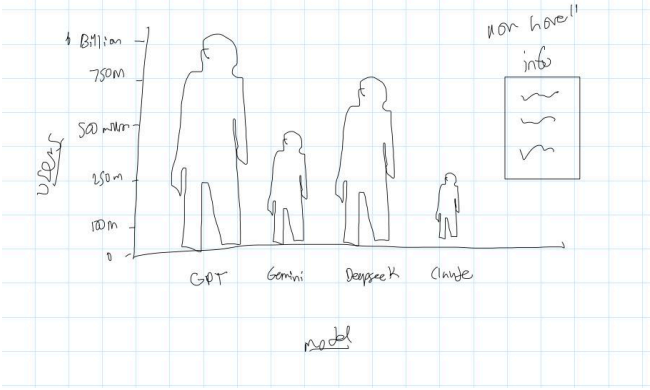
7. New question: How many AI patents have been filed and what trends do they depict?



8. Q4: What does the performance of LLMs look like given their input API price (C$ / 1M token)

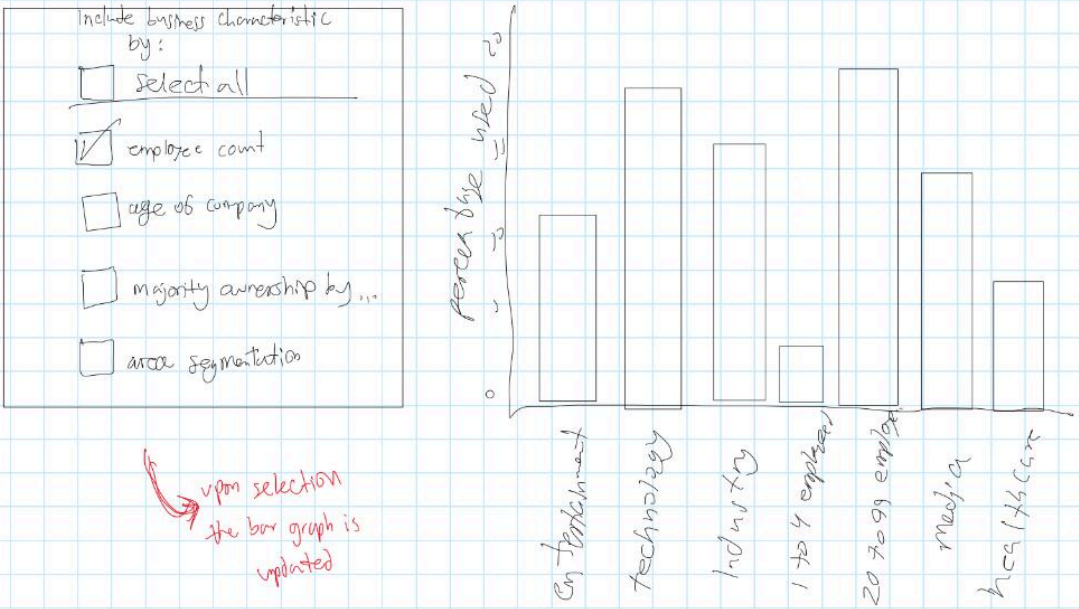


Q: Q5: which AI/LLM has the most users?

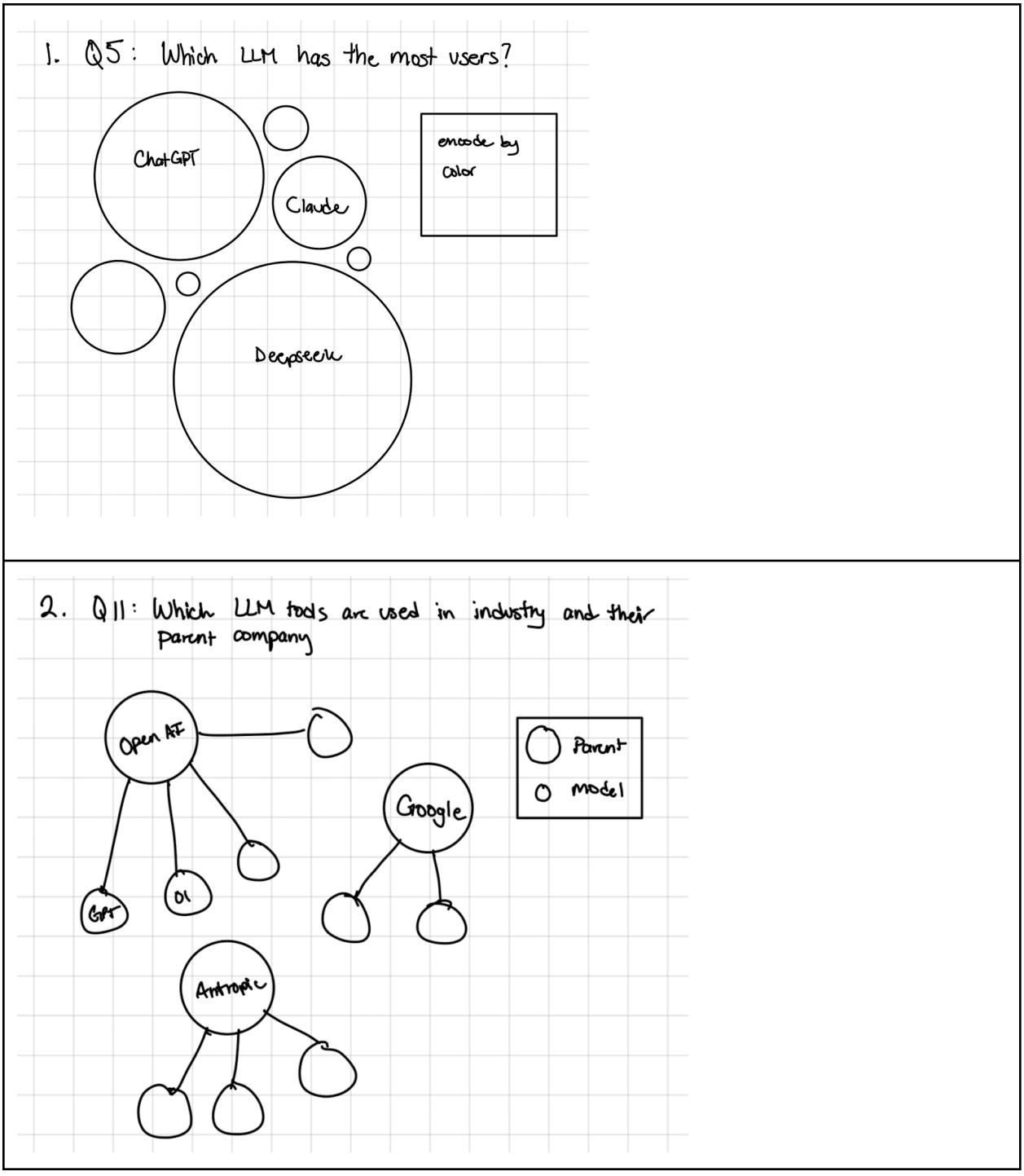


10. New question: use of AI by business characteristics in Canada

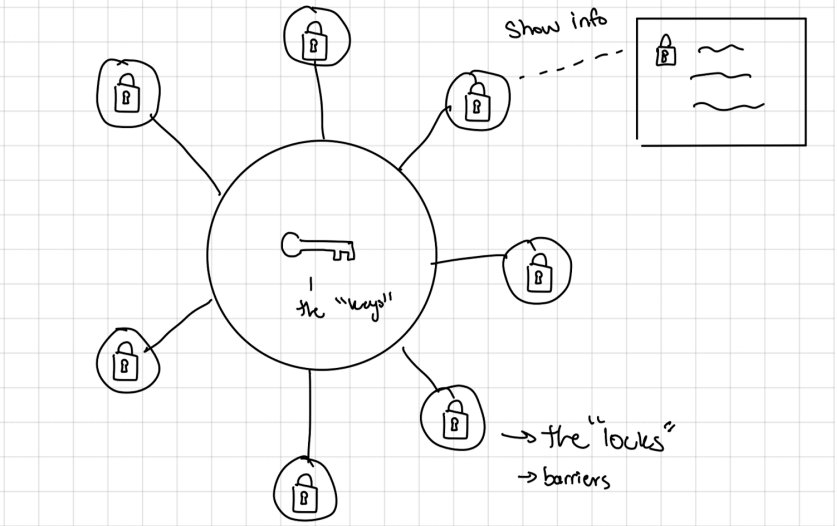
- use of AI by business characteristics in Canada



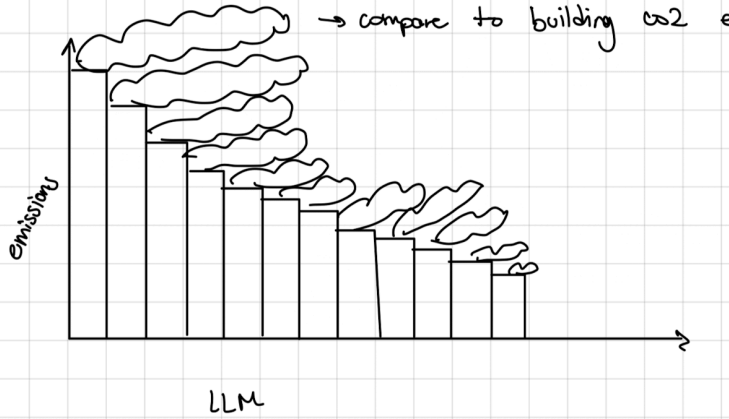
Lorena's Sketches:



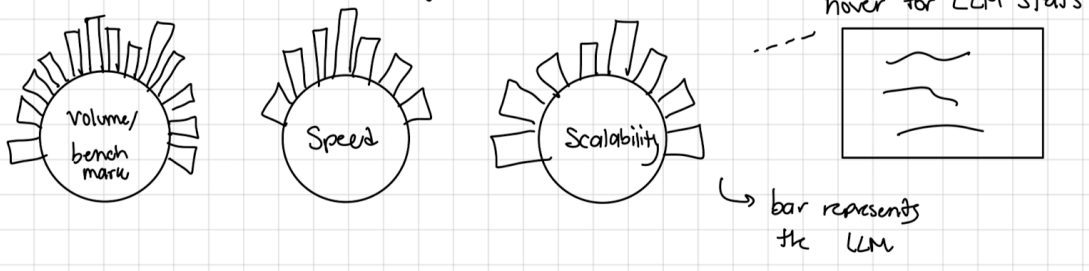
3. Q14: What are the barriers to entry in the LLM industry?



4. Q15: What is the environmental impact of LLMs



5. Q13: How scalable are LLMs for high volume use cases?



Decision step

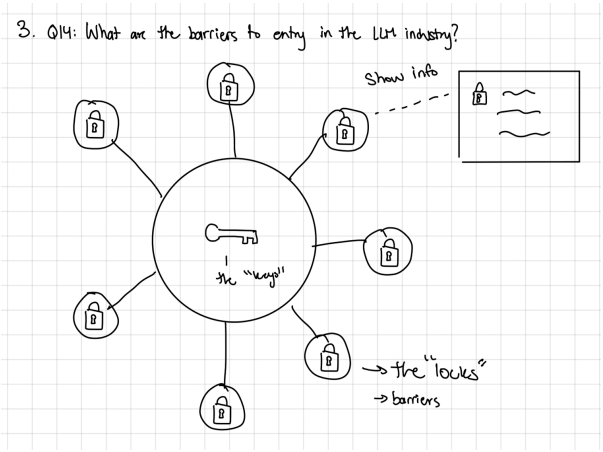
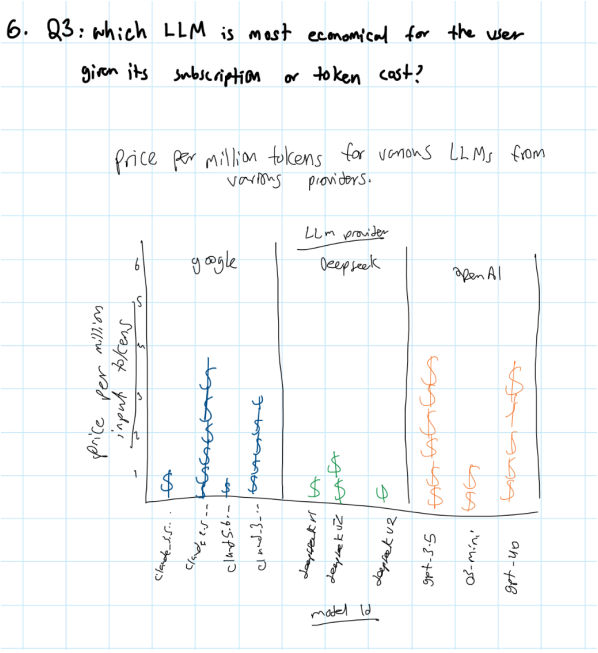
Cleaned Dataset
A Google Drive Link to our cleaned dataset called *LLMStats_ChavezTamales_Data*
https://drive.google.com/file/d/1uQMIgwJOwXinuTB0qG4tsC4_qvmLFBJz/view?usp=sharing

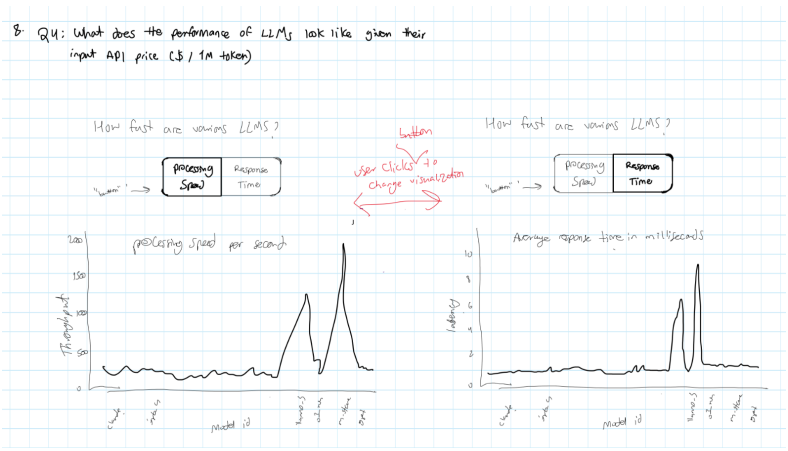
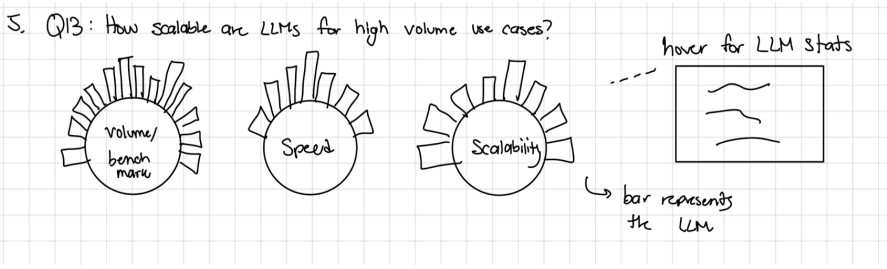
Affinity Diagram

Sketch ID	Question ID	Author	Votes
6	Q3	Ashkan	+1 +1=2
7	Q16: How many AI patents have been filled and what trends do they depict? **(New Question)**	Ashkan	
8	Q4	Ashkan	+1 +1=2
9	Q5	Ashkan	
10	Q17. What is the prevalence of AI usage across different business characteristics in Canada within the last 12 months **(New Question)**	Ashkan	
1	Q5	Lorena	+1 = 1
2	Q11	Lorena	+1=1
3	Q14	Lorena	+1 +1=2
4	Q15	Lorena	+1=1
5	Q13	Lorena	+1=1

Selected Sketches

We selected the following 4 sketches in order of most relevant to least relevant:

Relevancy Ranking	Sketch ID	Sketch
1st	3	
2nd	6	

3rd	8	
4th	5	

Decision Summary and Rational

We initially created 10 sketches and conducted a voting process using an affinity diagram. Seven out of the 10 sketches received at least one vote, with sketches 6, 8, and 3 receiving two votes each. Our goal was to select four sketches that would provide a solid range of initial questions while keeping the scope manageable, especially given that we are a two-person team. To begin, we selected sketches 6, 8, and 3, as they received the highest number of votes. Next, we considered the remaining four sketches that had at least one vote. We noticed that our selected sketches focused on LLMs, and we wanted to maintain this theme rather than expanding into the broader AI market. Additionally, we aimed to explore LLMs from an entrepreneurial and financial perspective, making our insights more relevant to individuals looking to enter the space. Therefore, we eliminated sketch ID 4, which explored the environmental impact of LLMs and did not align with our chosen direction. Additionally, we lacked proprietary data for this sketch, further supporting our decision to exclude it. We then chose sketch ID 5 because it offered greater depth and value compared to sketches 1 and 2, and it enabled us to narrow our datasets down to a single source.

Storyboarding

Insights:

Ashkan’s Insights:

1. Two specific models; “jambe-1.5-large" and "llama-3-70b-instruct," demonstrate exceptionally high throughput, peaking well above 1500 tokens per second. High-performing models often have “large” or “70b” (70 billion parameters) in their name meaning larger models, despite their complexity can achieve higher throughput which means higher processing speed.
2. Despite the wide price difference between models, most models have similar performances, with most showing throughput below 500 tokens per second and average response time close to zero seconds.
3. Deepseek models are best for high-volume and cost-sensitive applications due to their low price and acceptable throughput and latency.
4. The return on investment for Claude models is questionable for high-volume applications, as their performance gains do not align with their significant cost.

Lorena’s Insights:

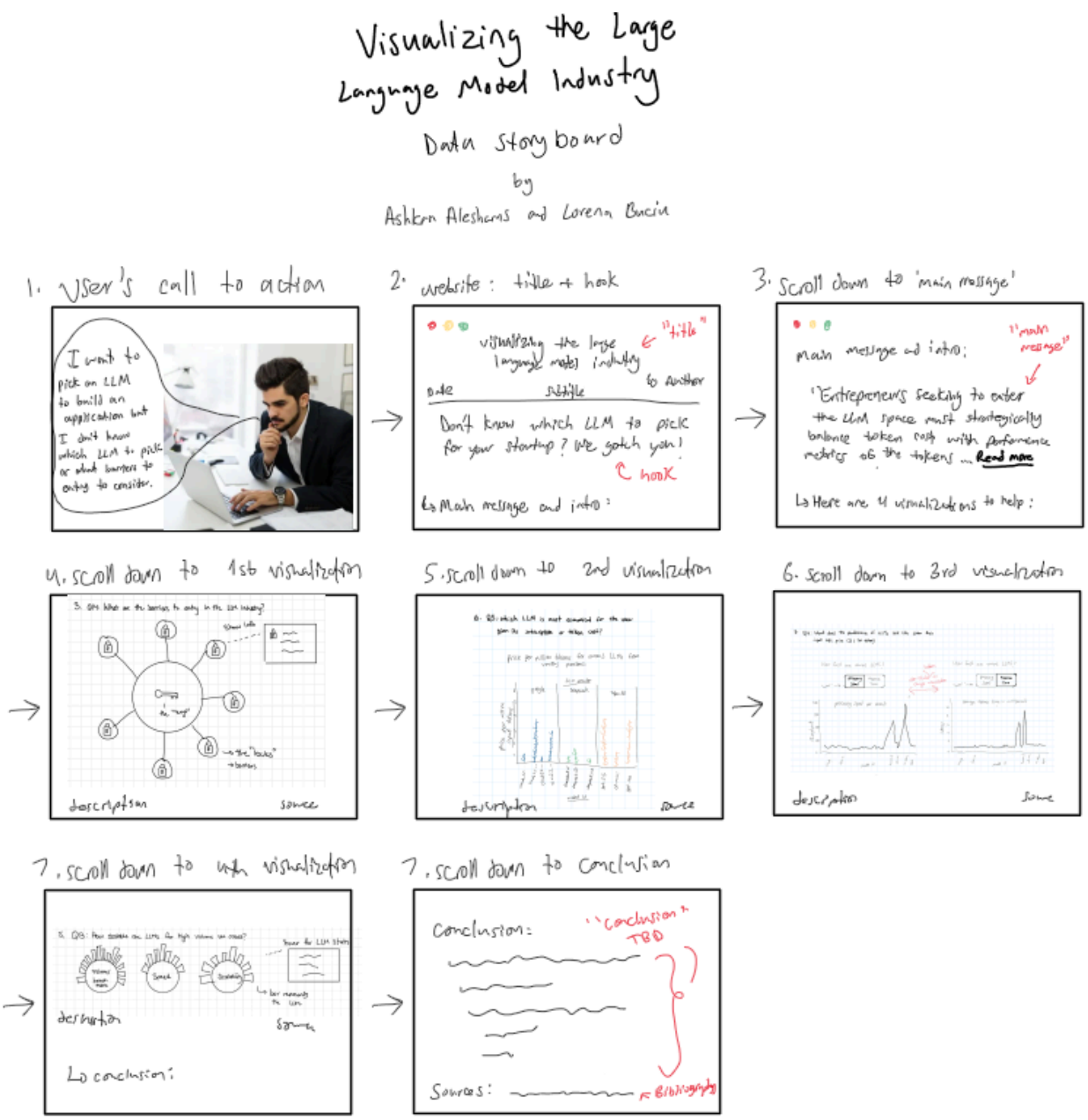
1. Google has the most number of models at 15 LLMs, with OpenAI at second with 10 models and Anthropic at third place with 5 models.
2. While “Mini” models are generally expected to be faster (have a lower latency) due to their smaller size, we can see that "01-2024-12-17" and "03-mini" show high latency, suggesting size is not the only factor determining response time. Other factors at play might be the model architecture, hardware/software and the network the LLM is hosted on.
3. It is very important to pair the cost of the tokens, with the speed of the tokens. A model that is cheap, but slow, may end up costing more, than a model that is fast, and a little more expensive, due to the amount of time it takes to complete the task.

Main Message:

“Entrepreneurs seeking to enter the LLM space must strategically balance token cost with performance metrics of the tokens like latency and throughput based on the application's specific needs. Large, complex models like those with '70b' parameters offer high throughput but come at a higher cost, making models like Deepseek more economically viable for high-volume, cost-sensitive applications, and highlighting the questionable ROI of premium models like Claude in such scenarios, while also being mindful that size alone doesn't guarantee speed, as demonstrated by the unexpected latency in certain 'mini' models.**”**

This is highly relevant to our target audience of entrepreneurs with a tech background looking to enter the LLM space. We've combined several of our key insights to provide them with valuable, actionable takeaways to help them strategically balance token costs with critical performance metrics like latency and throughput, ensuring alignment with their application's specific needs.

Data Storyboard:



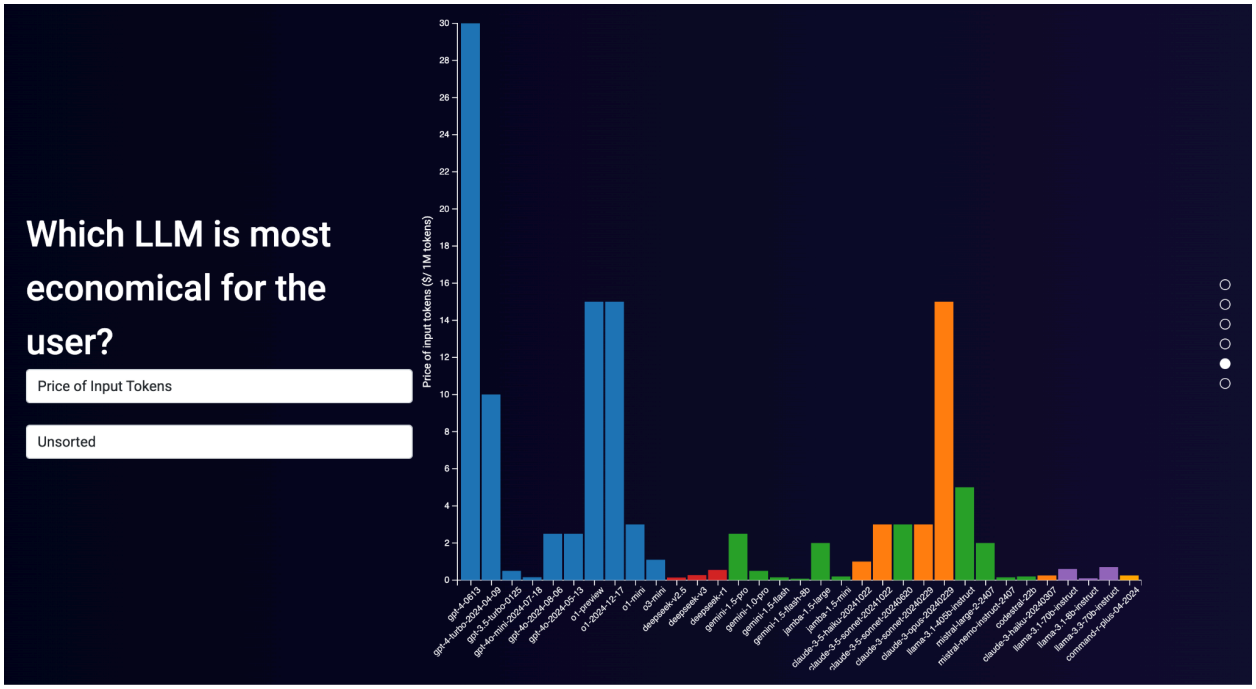
Prototyping

V1 Prototype

Progress Updates

1. Setup the Github repository. Setup Github Actions for deployment and deployed website to https://ashkanaleshams.github.io/ASIP-SQUAD/
2. Created the project layout including rough webpage design and structure.
3. First Implementation of the '**innovative view'** visualization completed which is the first visualization (barriers to entry) on the webpage.
4. Implemented visualization ID 6 which is the second visualization (economics) on the webpage.
5. LLMStats data cleaned up by flattening all the Provider/LLM into a simple json list: https://drive.google.com/file/d/1XOLKZBzyxe5q0QU1y7vROFd1Xw31ehne/view?usp=sharing
6. Partially Implemented visualization ID 5 - may change to a single full page circular bar plot and have a slider and dropdown to select the benchmark to view and which models to view
7. Decided to also use the hugging face open LLM data so we can also gain insights into open source LLMs rather than just proprietary ones and the dataset also contains more benchmarks
8. Cleaned up hugging face open LLM data, converted to a json file and cleaned up unicode characters from the fields and values
    a. https://github.com/AshkanAleshams/ASIP-SQUAD/blob/main/data/open_data.json

Prototype Screenshots
Website Link: https://ashkanaleshams.github.io/ASIP-SQUAD/



Frameworks & Sources Used

Datasets:

- Barriers data source (Deeper Insights, n.d)
    - Used chatgpt to generate a json dataset with the information from the article
- Open LLM data (Leaderboard, O.L, n.d)
- LLM Stats data (Chavez Tamales, n.d)

Libraries and assets:

- Svg lock source: https://www.svgrepo.com/svg/136946/lock
- Svg key source: https://www.svgrepo.com/svg/535465/key-skeleton

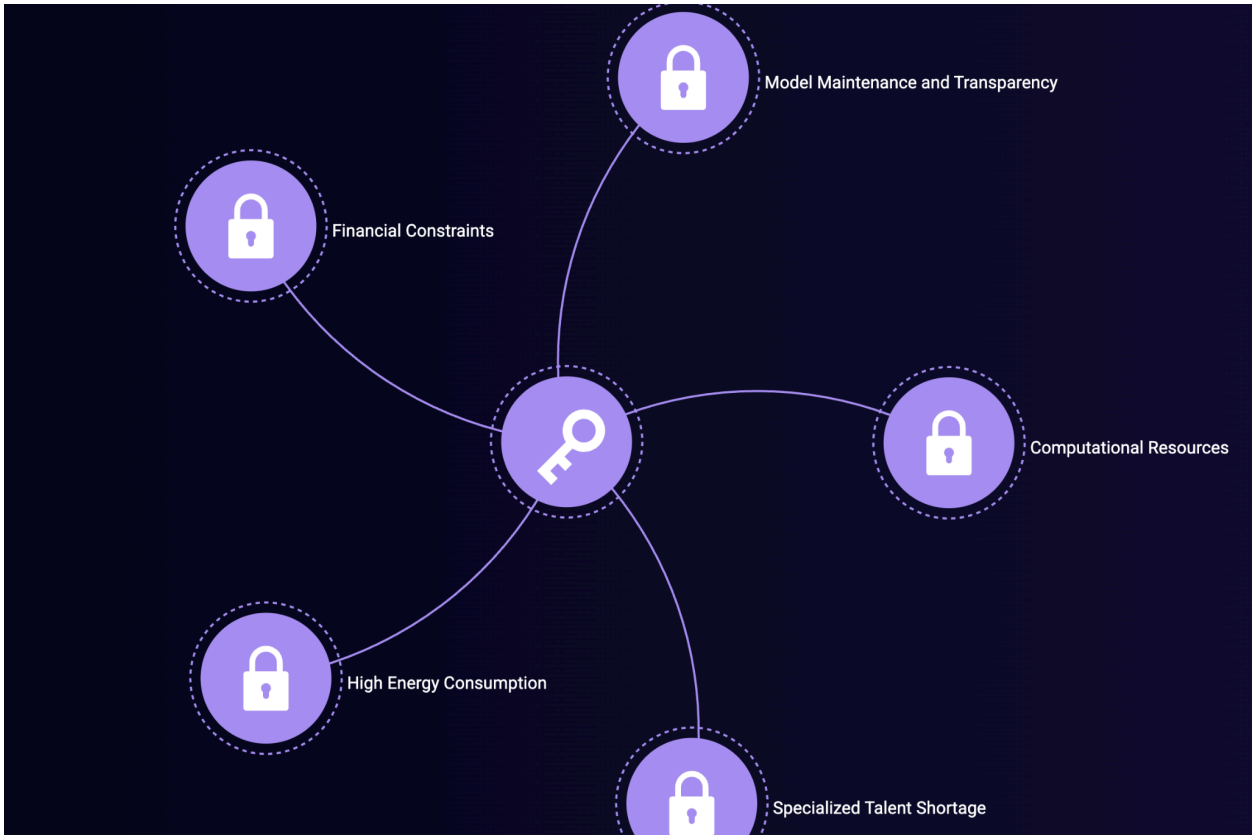
- Nav bar implementation: https://www.cssscript.com/one-page-scroll-dot-nav/

V2 Prototype

Progress Update:

- Added scrolling effects and text to present the data story
- Enhanced and completed all visualizations
- Added an additional comparison scatter plot visualization ‘secret vis’.

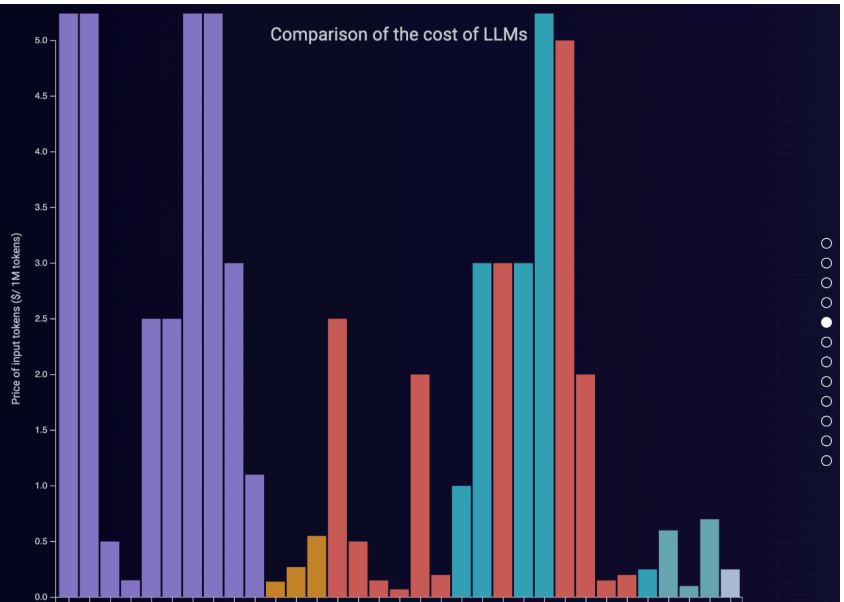
Screenshots:



How does the cost of LLMs compare?

Price of Input Tokens

Unsorted

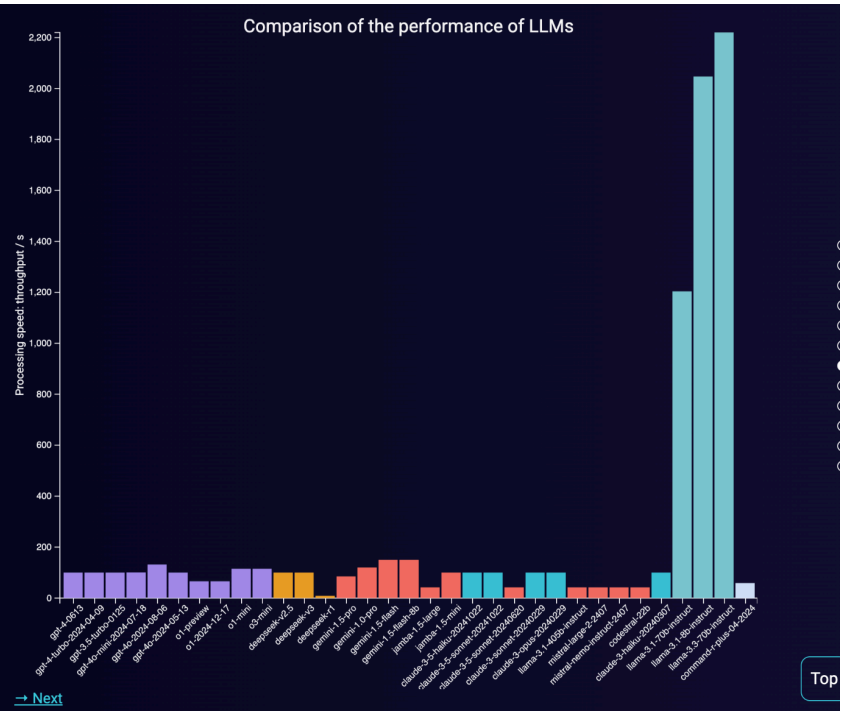


→ Next

Top

What does the performance of LLMs look like?

Processing speed



→ Next

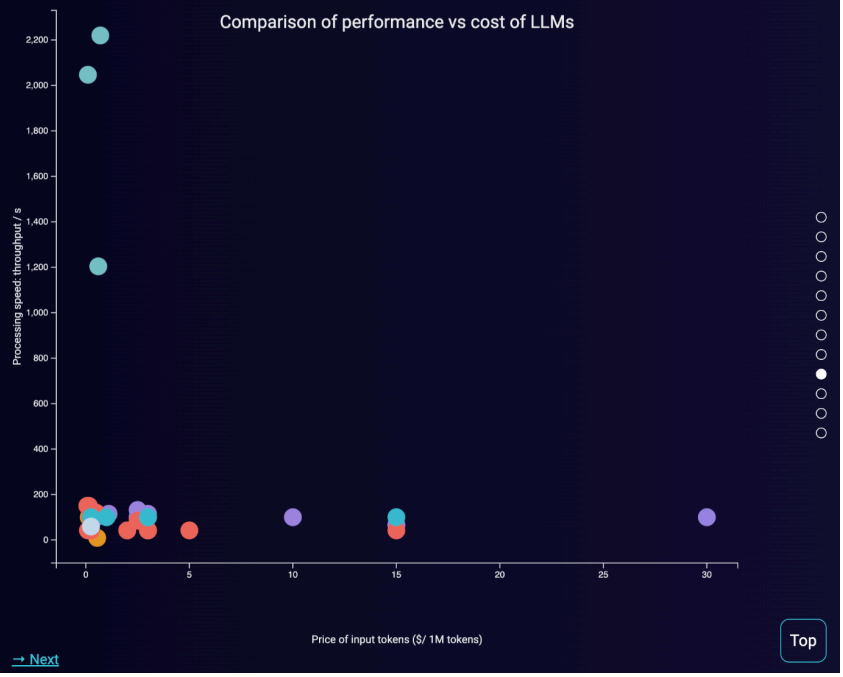
Top

How does the performance of LLMs compare to their cost?

Processing speed

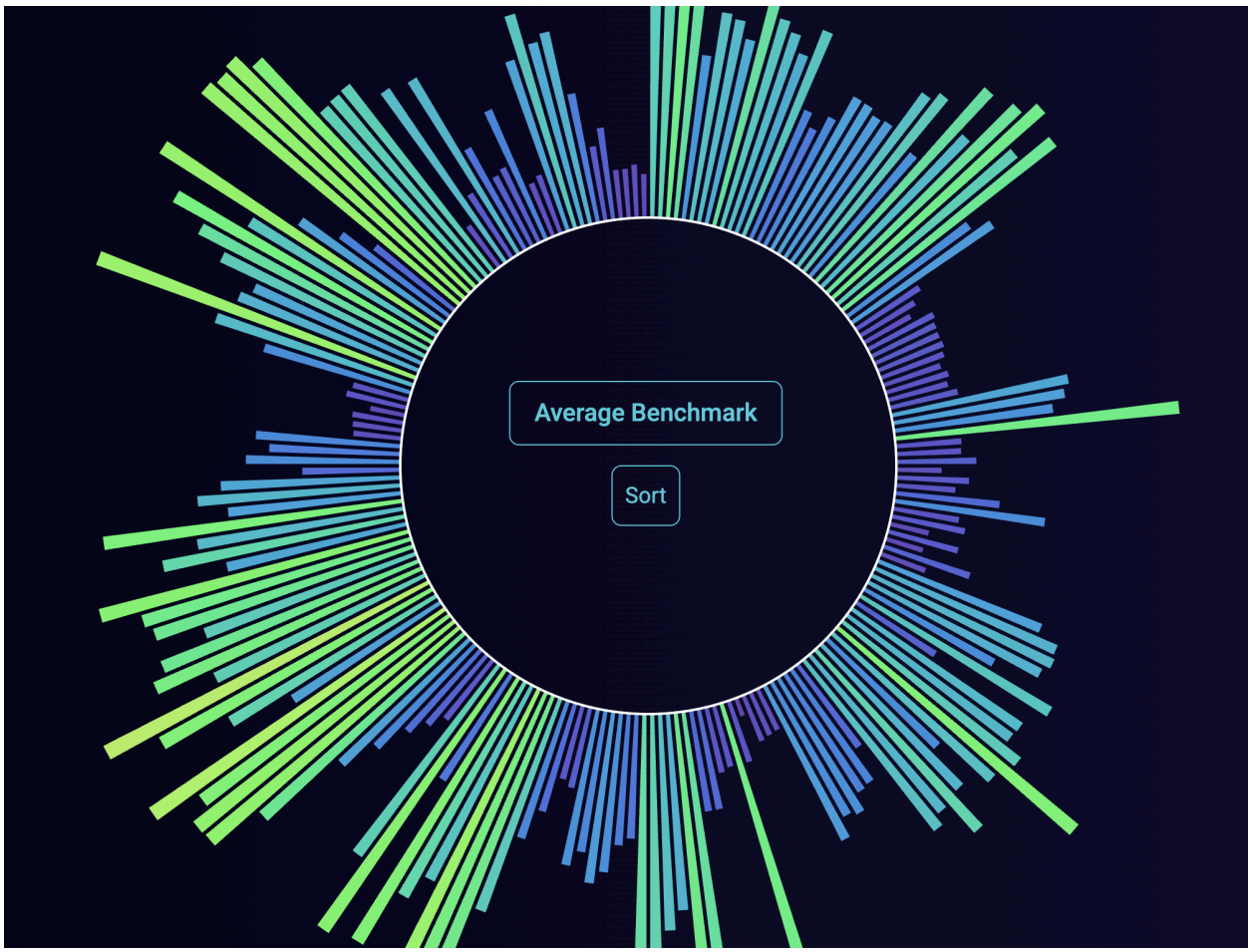
vs

Price of Input Tokens



→ Next

Top



Added libraries:
GSAP animations: https://gsap.com/
Lenis: https://github.com/darkroomengineering/lenis for smooth scroll

Testing

Think Aloud Study

Tester Name	Harsh Shah - harsh.shah@mail.utoronto.ca Group: Viz-ards
Describe any usability issues or confusion the tester encountered while using the prototype.	The scroll down experience was buggy and not smooth which led to usability issues, many times the section scroll feature would cause the screen to only show one half of the visualization and scrolling to see the full visualization was difficult (on a laptop screen) as it would either scroll too far or not scroll at all.
Was the tester able to understand the main message of the data story? (e.g., Yes/No + why/why not?)	Yes, the tester was able to understand that this storyboard was for tech entrepreneurs seeking to enter the LLM space, they understood this via the first visualization and did not read the header introducing the storyboard. This was because of the section subheader 'What are the barriers to entry in the LLM industry?'. This shows the effectiveness of the first visualization and it's section subheader/title acting as the hook to grab the user's attention and motivate the main message of the datastory. Then throughout the other visualization, this was further solidified. The user also credited the 'How does the performance of LLMs compare to their cost?' scatterplot visualization which was an additional new visualization added in V2 to being the most important visualization and the one acting as the conclusion to the main message.
What parts of the interface or visualization did the tester find most engaging or effective?	The user found the following engaging and/or effective: - 1st visualization: liked it's design and message - Sort feature and its animation for visualizations, especially for the last visualization. - The 4th visualization: found the ability to compare performance type vs type of token useful and effective to depict the storyboard's message.

	- The tooltips and how they include all dataset details for a given LLM regardless of the visualization.
What parts did the tester find confusing or less effective?	The user was confused about terms such as ‘CO2 cost’, ‘latency’, ‘throughput’ and ‘benchmark’ and did not know what they meant. Overall, we should explain what these mean.
Did the tester encounter any inconsistencies in design, data, or narrative?	The sort button for last visualization is different from the one in 2nd visualization. Also a sort feature is missing for the 3rd visualization which the tester believed could be useful.
Were there any unexpected interactions or insights that emerged during the session?	The fade-in animation added for the visualization is buggy and causes the visualization to not be displayed if the user does not scroll to a specific point, however the user was able to hover over the bars and display the tooltip even though the visualization itself was hidden.
What specific improvements or changes did the tester suggest for the prototype?	The user suggested we add a colour legend for 2nd, 3rd and 4th visualizations to quickly communicate to the user which colours represent which companies (with their logo) even though the tooltip mentions the company name.
Did the tester suggest any additional insights or visualizations to include?	No, however the user suggested describing the terms we used via tooltips or a small footnote. The tester suggested fixing the sizing and the position of the ‘next’ button which is too small and inconsistent. The tester suggested increasing the size of axes labels.
General observations or comments from the tester.	The user enjoyed the entire storyboard experience including website layout, design and visualizations and found it useful.

Post-testing Reflection

Based on the results of your ‘think aloud’ study, what would you improve in your data story?

We should improve the scrolling experience and visualization animations. We should also ensure that the design is responsive for different screen sizes. We should also add some more information in the last visualization to explain what the different benchmarks mean. Adding a legend for some of the visualizations may be useful. There are also some inconsistencies with

the ui components such as the sort button, and the next button. There are also some visual bugs with the tooltips that could be improved.

Are there any additional insights and visualizations you would use? Would you amplify or change your message? Did your narrative work? Did the tester get your takeaways?

The narrative of our data story was clear to the tester. Overall, the phrasing leading into the visualizations would make the goal and insights of them clear. For some visualizations the takeaways could be enhanced by providing information about what some of the LLM language means. Such as the different benchmark scores.

Decide as a team which of these improvements you will implement and write down your decisions and why you made them in your process book as a numbered list.

- ☑ ~~1. Add definitions for the keyword used, including in the LLM benchmarks in the last visualization~~
- ☑ ~~2. Make the sort button consistent across the visualizations~~
- ☑ ~~3. Add sorting for the third visualization~~
- ☑ ~~4. Improve the scrolling experience~~
- ☑ ~~5. Add legends for the visualizations that use a color scheme~~
- ☑ ~~6. Fix bugs with tooltips showing when they shouldn't be~~
- ☐ 7. Add zoom for fourth visualization

We decided to implement these improvements because they will enhance the user experience of the data story and also allow the user to get more insights out of the visualization.

Final

Final Reflection

We are very proud of our data story and believe it effectively uses visualizations to guide the target audience in understanding the LLM industry and key considerations based on specific application needs. Our design is sleek and minimalistic, ensuring a visually appealing experience, while interactive features further enhance learning and engagement. By allowing users to explore data dynamically, we make complex concepts more accessible and intuitive. Since our previous think-aloud study, we have significantly improved the user experience by refining our interface, optimizing navigation, and implementing six out of seven planned enhancements. These improvements include better data organization, more responsive interactions, and clearer visual hierarchy, all aimed at making the story more compelling and informative. The only exception was adding zoom functionality to the fourth visualization. After careful evaluation, we decided to scope this feature out, as it would be too time-consuming to implement and could introduce navigation challenges, potentially making the application feel unresponsive. Instead, we focused on optimizing the existing visuals to ensure clarity and user-friendly design.
Overall, we are pleased with the outcome and confident that our data story provides an engaging and informative experience for users, effectively bridging the gap between raw data about LLM models and actionable insights for the LLM industry.

Demo and Link

Final Project Web-based Data Story

https://ashkanaleshams.github.io/ASIP-SQUAD/

Final Project Video Demo

https://drive.google.com/file/d/1sD6C7_CUIIZFH9CJBej8aCvpp7kAzSyW/view?usp=sharing

Bibliography

AI Tools Next Year. AI | 2024 Stack Overflow Developer Survey. (n.d.). https://survey.stackoverflow.co/2024/ai#developer-tools-ai-next

Chavez Tamales, J. (n.d.). *LLMStats: A comprehensive set of LLM Benchmark scores and provider prices.* GitHub. https://github.com/JonathanChavezTamales/LLMStats

CNBC. (2024, October 3). *Ai is overestimated near term but underestimated long term, says Jefferies' Brent Thill*. CNBC. https://www.cnbc.com/video/2024/10/03/ai-is-overestimated-near-term-but-underestimated-long-term-says-jefferies-brent-thill.html

Deeper Insights. (n.d.). *The unspoken challenges of large language models - deeper insights*. Deeper Insights Cisco. https://deeperinsights.com/ai-blog/the-unspoken-challenges-of-large-language-models

Innovation, Science and Economic Development Canada. (n.d.).*Processing the artificial intelligence dataset*. Canadian Intellectual Property Office. https://ised-isde.canada.ca/site/canadian-intellectual-property-office/en/processing-artificial-intelligence-dataset

Introducing DeepSeek-V3. DeepSeek API Docs. (n.d.). https://api-docs.deepseek.com/news/news1226

Stackoverflow (2024). *Developer profile* . Developer Profile | 2024 Stack Overflow Developer Survey. (n.d.). https://survey.stackoverflow.co/2024/developer-profile#demographics

Statistics Canada. (2024, May, 27). Table 33-10-0825-01 Use of artificial intelligence by businesses and organizations in producing goods or delivering services over the last 12 months, second quarter of 2024. https://www150.statcan.gc.ca/t1/tbl1/en/cv.action?pid=3310082601

Tang, Y. (2024, February 12). *The LLM App Stack*. Medium. https://medium.com/plain-simple-software/the-llm-app-stack-2024-eac28b9dc1e7

Leaderboard, O. L. (n.d.). Open-LLM-leaderboard/contents · datasets at hugging face. open-llm-leaderboard/contents · Datasets at Hugging Face. https://huggingface.co/datasets/open-llm-leaderboard/contents