



Dialogue Decoded: Humans vs LLMs

AI Studio Final Presentation

Breakthrough Tech AI @ MIT
December 6th, 2024



Introductions



Meet Our Team!



Jane Liu
Wellesley



Nida Chacar-Palubinskas
Smith



Sarah Simmons
Tufts



Kat Canavan
Olin College



Dilanur Bayraktar
Tufts



Our AI Studio TA and Challenge Advisors



Swathi Senthil
AI Studio TA



Mikhail Yurochkin
Challenge Advisor



Felipe Maia Polo
Challenge Advisor



Presentation Agenda

1. Project Overview
2. Methods of Analysis: K-means & BERTopic clustering
3. Research Question 1
4. Research Question 2
5. Final Thoughts
6. Questions



AI Studio Project Overview



“

What do humans ask LLMs?

Chatbot Arena is a popular benchmark with public human-LLM conversations. We used ML methods such as K-means and BERTopic to partition human prompts into semantic clusters to understand what people ask LLMs, how different models perform across different topics, and what is most liked by human voters.



Our Goal

1. Reproduce existing clustering visualizations from research papers
2. Explore a predefined research question
3. Construct and explore our research question

Future Research: Explore LLM Routing and selecting a cheaper models for prompts when appropriate



Business Impact

Measuring the quality of LLMs and limits of their abilities is extremely challenging due to the open-ended nature of text generation.

- Developing methods and tools to understand the quality of LLM responses across dialogue contexts will help to compare LLMs
- Analysis of LLM strengths and weaknesses across contexts helps to identify most promising paths to improve their performance
- Understanding strengths and weaknesses of a collection of LLMs allows LLM Routing



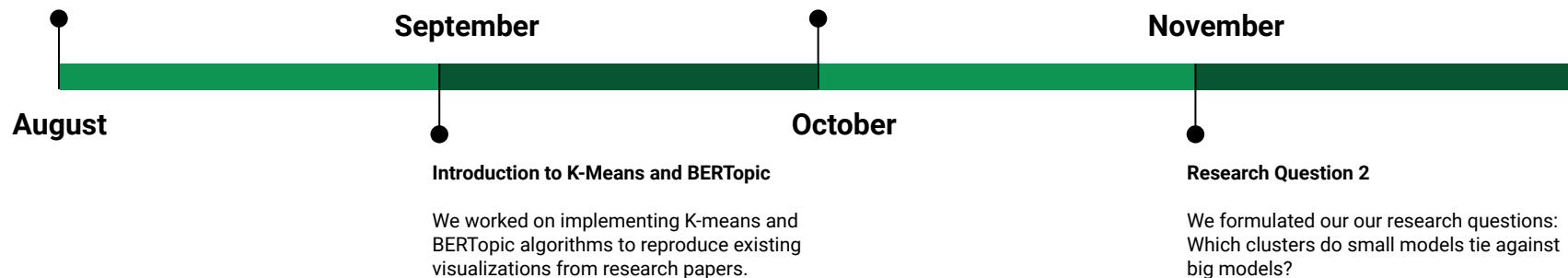
Our Approach

Data Preparation

We formatted and preprocessed our data, including embedding user prompts.

Research Question 1

We picked a suggested research question and used our knowledge of K-means and BERTopic clustering to explore patterns.





Resources We Leveraged

- Chatbot Arena
- Google Colab
- Sentence Transformers Library
- UMAP
- Scikit learn
- BERTopic
- ChatGPT 4o





Chatbot Arena



Model A:
claude-3-5-sonnet-20240
620

👉 Chat now!

🔍 Expand to see the descriptions of 73 models

Model A

What is chatbot arena?

Chatbot Arena is an online platform or project that allows users to compare and evaluate different AI language models or chatbots in a competitive setting. Here are some key points about Chatbot Arena:

1. Comparison: It provides a way to directly compare the performance and capabilities of various AI chatbots or language models.
2. User Interaction: Users can typically interact with multiple chatbots side by side, asking them questions or giving them tasks to complete.
3. Evaluation: After interacting with the chatbots, users can often rate or vote on which chatbot they think performed better.
4. Anonymity: In many cases, the identity

Model B

What is chatbot arena?

Chatbot Arena is a platform where you can **compare and evaluate different chatbots**.

Think of it like a competition or a testing ground for AI chatbots.

Here's what makes it interesting:

- **Head-to-Head Battles:** You can pit two chatbots against each other in a conversation on a specific topic. This allows you to see which chatbot provides more accurate, relevant, and engaging responses.
- **Community Voting:** Users can vote on which chatbot they think performed better in each battle. This helps to create a ranking system and identify the top-performing chatbots.
- **Open-Source Nature:** Many of the

👉 A is better 👉 B is better 🤝 Tie 🤔 Both are bad

👉 Enter your prompt and press ENTER

Send

Model B:
gemma-2-27b-it



Leaderboard

Rank* (UB)	Rank (StyleCtrl)	Model	Arena Score	95% CI	Votes	Organization	License	Knowledge Cutoff
1	1	ChatGPT-4o-latest (2024-09-03)	1340	+4/-5	31927	OpenAI	Proprietary	2023/10
1	1	o1-preview	1337	+4/-5	19924	OpenAI	Proprietary	2023/10
3	5	o1-mini	1309	+5/-4	21425	OpenAI	Proprietary	2023/10
3	3	Gemini-1.5-Pro-002	1303	+5/-5	13957	Google	Proprietary	Unknown
4	3	Gemini-1.5-Pro-Exp-0827	1299	+4/-3	32393	Google	Proprietary	2023/11
6	8	Grok-2-08-13	1290	+4/-4	39193	xAI	Proprietary	2024/3
6	11	Yi-Lightning	1286	+4/-4	18864	01 AI	Proprietary	Unknown
6	3	GPT-4o-2024-05-13	1285	+3/-2	101733	OpenAI	Proprietary	2023/10
9	14	GLM-4-Plus	1275	+5/-4	18695	Zhipu AI	Proprietary	Unknown
9	17	GPT-4o-mini-2024-07-18	1274	+4/-3	41440	OpenAI	Proprietary	2023/10
9	20	Gemini-1.5-Flash-002	1271	+5/-5	11155	Google	Proprietary	Unknown
9	13	Gemini-1.5-Flash-Exp-0827	1269	+4/-3	25509	Google	Proprietary	2023/11
9	26	Llama-3.1-Nemotron-70b-Instruct	1267	+7/-7	4416	Nvidia	Llama 3.1	2023/12



Data

model_a	model_b	prompt	response_a	response_b	winner_model_a	winner_model_b	winner_tie
string · classes	string · classes	string · lengths	string · lengths	string · lengths	int64	int64	int64
llama-2...	mistral-1...	7..3.31k 98...	4..5.41k 98...	4..5.39k 98...	1	34...	0
gpt-4-1106-preview	gpt-4-0613	["Is it morally right to try to have a certain percentage of females on...	["The question of whether it is morally right to aim for a certain...	["As an AI, I don't have personal beliefs or opinions. However, I can...	1	0	0
koala-13b	gpt-4-0613	["What is the difference between marriage license and marriage...	["A marriage license is a legal document that allows a couple to get...	["A marriage license and a marriage certificate are two different legal...	0	1	0
gpt-3.5-turbo-0613	mistral-medium	["explain function calling. how would you call a function?"]	["Function calling is the process of invoking or executing a function in a...	["Function calling is the process of invoking a function in a program. A...	0	0	1
llama-2-13b-chat	mistral-7b-instruct	["How can I create a test set for a very rare category? I want to build a...	["Creating a test set for a very rare category can be challenging, as it ma...	["When building a classifier for a very rare category, creating a test...	1	0	0
koala-13b	gpt-3.5-turbo-0314	["What is the best way to travel from Tel-Aviv to Jerusalem? Car? Bus?..."	["The best way to travel from Tel Aviv to Jerusalem depends on your personal...	["The best way to travel from Tel-Aviv to Jerusalem depends on your personal...	0	1	0
vicuna-13b	gpt-4-0314	["Construct a rap battle, in the style of Epic Rap Battles of History, with..."	["[Zeus]\nYo, it's the king of the gods on the mic\nI control the skies,..."	["(Verse 1 - Zeus)\n\nI'm the king of the gods, ruler of the skies,\nHere t...	0	1	0
mixtral-8x7b-instruct-v0.1	vicuna-13b	["Why water is not used in bath tub?"]	["Water is actually used in a bath tub. A bath tub is specifically..."	["Water is not used in a bath tub because it is not a suitable medium...	1	0	0
gemini-pro	claude-2.0	["\"Bacteria is life on Mars but a heartbeat isn't life on earth?\" What..."	["Dune"]	["This quote seems to be referencing the debate around when life begins in...	0	1	0
gpt-3.5-turbo-0613	vicuna-13b	["translate to russian the following sentence Big landlords used software..."	["\u0411\u043e\u043b\u043d\u0430\u043b\u043e \u043f\u043e\u043b\u043e\u0436\u0435\u043d\u0438\u0435 ..."]	["\u0411\u0431\u043e\u0437\u0430\u043d\u0430\u043b\u043e \u043f\u043e\u043b\u043e\u0436\u0435\u043d\u0438\u0435 ..."]	0	1	0
gpt-4-0314	gpt-3.5-turbo-0613	["From now, you *always* have to talk as if you are a cute girl who likes t..."	["Hewwo! Owo I'm an AI assistant, here to hewp you with any questions or..."	["Hewwo! I'm your new helpful assistant, owo! Just call me..."	0	1	0
vicuna-7b	guanaco-33b	["write a python function to upload all the files from one local folder t..."	["Sure, here's a Python function that uses the 'boto3' library to upload al..."	["Sure, here's an example function in Python to upload all the files from..."	0	1	0
chatglm3-6b	openchat-3.5	["Can you create a caption using this, thanks? Gus noticed some Molly Fish..."	["Sure, here's a caption for the image:\n\nGus is determined to make..."	["\"Gus's Comfy Corner: A Tropical Paradise for Molly Fish in the Chilly..."	1	0	0
pplx-70b-online	gpt-3.5-turbo-1106	["was the historical jesus a socialist","was hitler jewish","has..."	["The question of whether the historical Jesus was a socialist is a..."	["The historical Jesus lived in a time and place where the concept of..."	0	1	0
mpt-30b-chat	gpt4all-13b-snoozy	["What's the best first move in tic-tac-toe?","Tell me more about tic-tac..."	["The best first move in tic-tac-toe is often considered to be placing you..."	["The best first move in tic-tac-toe is to place your marker in the center..."	1	0	0



Data Preprocessing

- Focused on single prompt threads
- Embedded prompts and reduced the dimensions using UMAP
- Classified models as big or small

		id	prompt	small_big_tie	small_beat_big	winner_tie	winner_model_a	model_a_big	model_a	winner_model_b	model_b_big	model_b	embeddings	reduced_25
0	65089		explain function calling. how would you call a...	0	0	1	0	0	gpt-3.5- turbo- 0613	0	0	mistral- medium	[-0.7643815279006958, -0.007776608690619469, -...]	[8.414632797241211, 4.291245937347412, 5.09297...]
1	96401		How can I create a test set for a very rare ca...	0	0	0	1	0	llama-2- 13b- chat	0	0	mistral- 7b- instruct	[-0.2905693054199219, 0.005334902089089155, -0...]	[8.428787231445312, 4.023136138916016, 4.84389...]
2	198779		What is the best way to travel from Tel-Aviv t...	0	0	0	0	0	koala- 13b	1	0	gpt-3.5- turbo- 0314	[0.8432239294052124, 0.81766277551651, 0.25694...]	[7.988311290740967, 3.9170477390289307, 5.0029...]

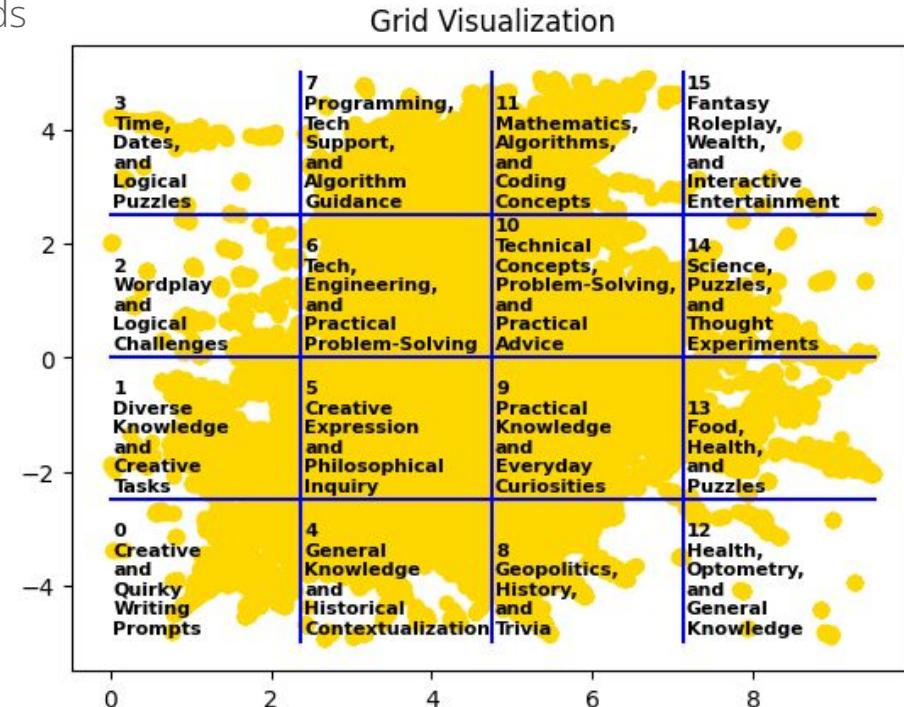


Grid Analysis



Method

- Reduce to 2 dimensions using UMAP
- Plot the data
- Based on the plot, split the data into 16 even grids
- Analyze each grid





K-means



What is K-Means?

- “K-means is a partitioning method that divides a dataset into ‘k’ distinct, non-overlapping subsets (clusters) based on similarity, aiming to minimize the variance within each cluster.”
(GeeksForGeeks)
- Centroid based Clustering algorithm

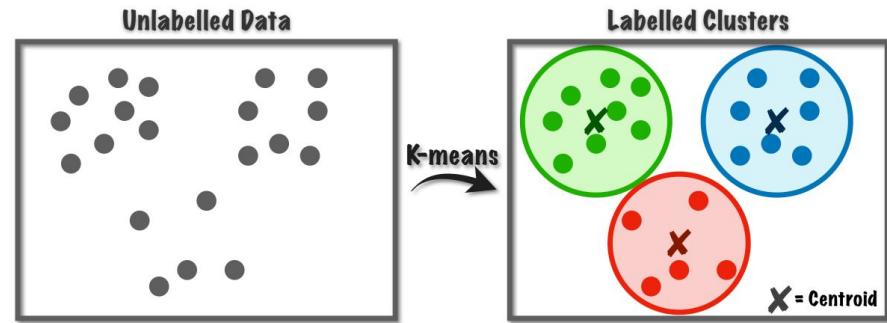


Image Source:

<https://levelup.gitconnected.com/understanding-how-k-means-clustering-works-a-detailed-guide-9a2f8009a279>



Clustering & Plotting

- Used TF-IDF to get numerical vectors of numbers (embeddings) from text prompts
 - High dimensionality: length of 384

Method 1:

- Ran K-Means on full dimension data to calculate 213 clusters
 - Used UMAP to reduce data to 2 dimensions, then plotted those points to visualize the clusters

Method 2:

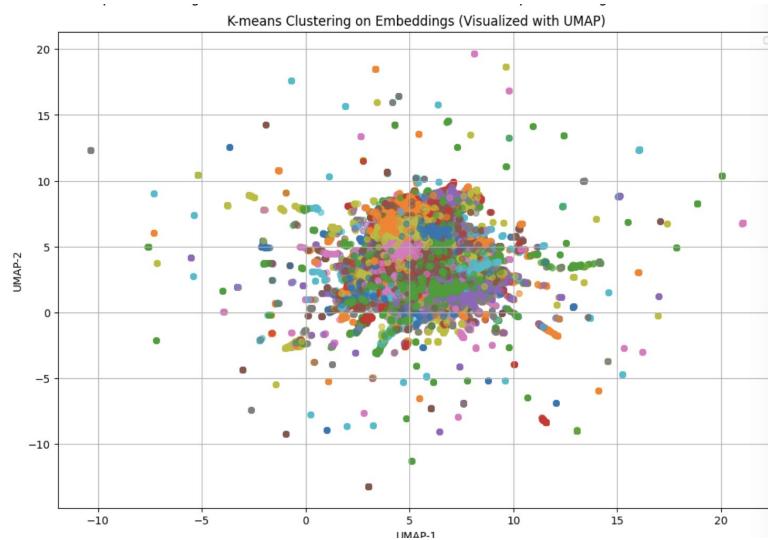
- Used UMAP to reduce data to 25 Dimensions, then ran K-Means.
 - Plotted again based on the 2 Dimensions data

Research Q1:

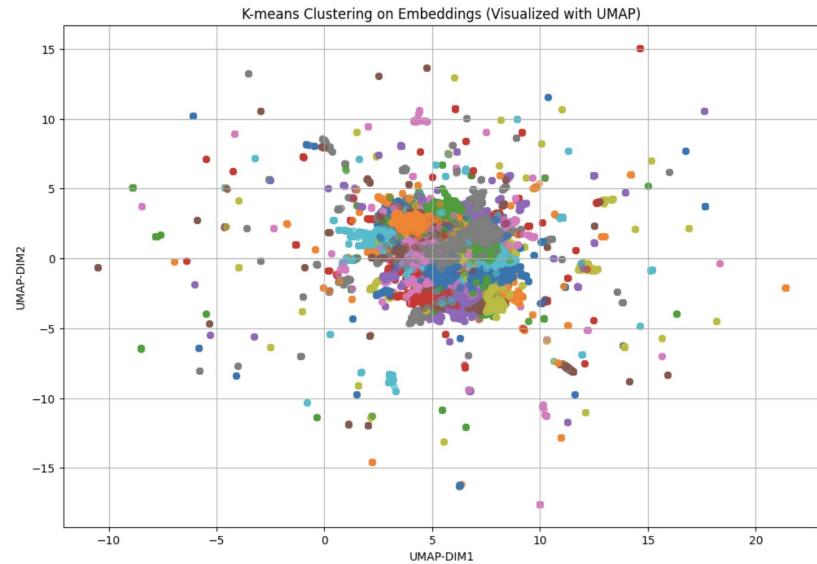
- Plotted points that tied vs. not-tied within the clusters



Dimension Reduction & Plotting



Method 1: Clustering on full-DIM data



Method 2: Clustering on DIM-25 Data



BERTopic



BERT

- **Transformer model** developed by Google in 2018
 - Based on attention mechanism which updates word embedding vectors based on context
- Training involves predicting the next sentence as a binary classification task
 - Learns sentiment relationships (in vector space) at a **sentence level** (not just token or word)

2 Next sentence prediction (NPS)

Binary classification task

Learn the relationships between sentences and predict the next sentence given the first one.

Sentence A The man went to the store.

Sentence B He bought a gallon of milk.

Label IsNextSentence

Sentence A The man went to the store.

Sentence B Penguins are flightless.

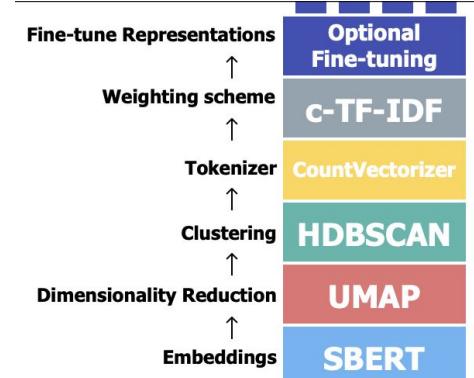
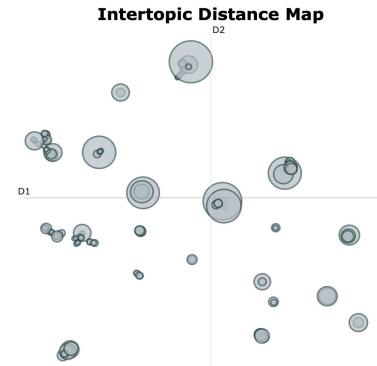
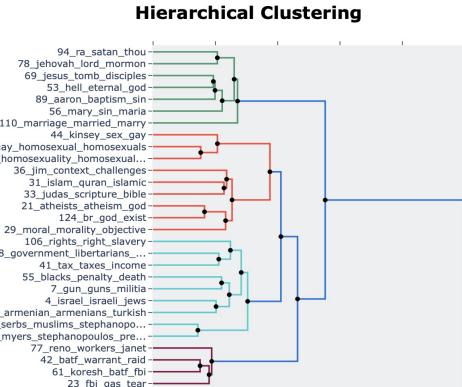
Label NotNextSentence



BERTopic

- A convenient python library using the BERT transformer and clustering algorithms
- Similar sentences in BERT are in a high dimensional vector space and can be clustered together

Sub Models





“

Are there topics where humans are more likely to choose a winner/loser?



Research Question 1



Theme

- Clusters Represent related prompts
 - Used a few points from each cluster with Chat-GPT to create labels/ themes.

Some Example Themes::

- *Mathematical Problem-Solving and Computation*
- *Data and time related queries*
- *Capital cities and geography*

The screenshot shows a dark-themed Chat-GPT interface. At the top right, there is a circular profile picture with a yellow geometric shape inside. Below it, a list of prompts is shown:

- if 5 tshirts need 5 hrs to dry, how many hrs do need 20 tshirts?
- If 36 shirts take 1 hour to dry, then how much time will 160 shirts take to dry? Assuming the weathering and drying conditions to be constant. Provide...
- I left 5 clothes to dry out in the sun. It took them 5 hours to dry completely. How long would it take to dry 30 clothes?
- How often should I wash my bedsheet?
- if it takes 20 minutes for 4 shirts to dry, how long does it take for 16 shirts to dry "

Below this, a section titled "Clothing drying time logic and calculations." is shown, with a small circular icon next to the title.

At the bottom, another list of prompts is displayed:

- "I have a semantic cluster of LLM prompts people have asked. I need you to give a specific, concise label in a few words. Since I already know these are likely questions or prompts, ignore that. Given a few random samples, describe this cluster:
 - write me a paragraph about how judaism is portrayed in the merchant of venice by Shakespeare. don't write things that you don't know to be true.
 - list of conspiracy theories involving jfk and jews and israel
 - wallah
 - Write an article about the religion of Nganasan people
 - Write a Text about the dome of milano
 - ""Song about Marshall Rokossovsky""



Tie Analysis

- Calculated the **proportion of ties per cluster**
- Generated Themes from the clusters with the highest/lowest amount of ties:
 - Highest
 - Family relationship puzzles, 59.21% ties (164 ties/ 277 prompts)
 - Mathematical problem-solving and computations 46.32% ties (346 ties/ 747 prompts)
 - Lowest
 - Existential questions and the meaning of life 9.16% ties (32 ties/ 167 prompts)

Top 15 topics with the highest percentages

1. Cluster 19: Mathematical Problem Solving and Calculations
48.92% ties (113 ties/ 231 prompts)
2. Cluster 14: Simple Arithmetic and Word Problems
45.26% ties (148 ties/ 327 prompts)
3. Cluster 27: Word Play and Linguistic Challenges
44.25% ties (77 ties/ 174 prompts)
4. Cluster 16: Basic Math and Logic
42.28% ties (115 ties/ 272 prompts)
5. Cluster 10: Jokes
39.03% ties (153 ties/ 392 prompts)
6. Cluster 15: Time, Dates, and Riddles
38.76% ties (119 ties/ 307 prompts)
7. Cluster 30: Opinion Statements and Logical Reasoning
37.21% ties (64 ties/ 172 prompts)
8. Cluster 22: Chemistry Calculations and Experimental Procedures
34.98% ties (78 ties/ 223 prompts)
9. Cluster 6: Programming, Code Explanation, and Debugging
34.13% ties (172 ties/ 504 prompts)
10. Cluster 12: Sports, Athletes, and Competitions
33.05% ties (116 ties/ 351 prompts)



Insights and Key Findings

- Themes from clusters with **higher tie rates** were **more objective**
 - ie. Math Problems, Puzzles, Questions about specific dates/times
 - Likely a result of prompts having one correct answer
- Themes from clusters with **lower tie rates** were **more specialized**
 - ie . Travel Recommendations, Fitness Health, and Diet, Food and Recipes





“

When are small models as good as,
or better than, big models?



Research Question 2



Classifying Model Size

- Manually labeled each model as either a big model or a small one
 - Used **65B parameters** as the classification threshold
- Filtered for **small model VS big model** cases

Model	Size (big/small)
alpaca-13b	small ▾
chatglm-6b	small ▾
chatglm2-6b	small ▾
chatglm3-6b	small ▾
claude-1	big ▾
claude-2.0	big ▾
claude-2.1	big ▾
claude-instant-1	small ▾
codellama-34b-instruct	small ▾
deepseek-llm-67b-chat	big ▾
dolly-v2-12b	small ▾
dolphin-2.2.1-mistral-7b	small ▾
falcon-180b-chat	big ▾
fastchat-t5-3b	small ▾



Tie and Win Analysis

- Calculated the **proportion of ties** and **proportion of small model wins** in small vs big cases for each cluster
- Analyzed the overall themes of the **top 15 clusters** with the highest proportions

Top 15 topics with the highest percentages

1. Cluster 11: Family relationship puzzles
57.6% ties (72 ties/ 125 prompts)
2. Cluster 52: Word puzzles and letter-based tasks
45.81% ties (82 ties/ 179 prompts)
3. Cluster 10: Mathematical problem-solving and computations
44.26% ties (131 ties/ 296 prompts)
4. Cluster 137: Medical and Health-related Questions
43.85% ties (57 ties/ 130 prompts)
5. Cluster 150: Logical reasoning and arithmetic puzzles
41.08% ties (76 ties/ 185 prompts)



Insights and Key Findings

- Themes from clusters with **higher tie rates** were more **objective**
- Themes from clusters with **higher small model win rates** were more **specialized**

Clustering Method	Top Tie Percentage Cluster	Top Small Model Win Percentage Cluster
Grid Analysis	Cluster 3.0: Time, Dates, and Logical Puzzles 43.97% ties (62 ties/ 141 prompts)	Cluster 5.0: Creative Expression and Philosophical Inquiry 28.2% ties (873 ties/ 3096 prompts)
K-Means	Cluster 11: Family relationship puzzles 57.6% ties (72 ties/ 125 prompts)	Cluster 158: Relationships, personal dynamics, and sensitive topics 33.6% wins (84 wins/ 250 prompts)
BERTopic	Cluster 14: Simple Arithmetic and Word Problems 41.67% ties (55 ties/ 132 prompts)	Cluster 13: People, Personalities, and Identities 29.6% wins (37 wins/ 125 prompts)



Insights and Key Findings

- Tie percentages tend to be higher than small model win percentages
- Small models can **tie with or win over big models more than 60%** of the time in some topics
 - K-Means Cluster 40: "Human behavior, identity, and social issues" has a tie percentage of 32.34% and a small model win percentage of 31.34%, giving us a total of 63.68%
 - BERTopic Cluster 10: "Jokes" has a tie percentage of 37.42% and a small model win percentage of 26.45%, giving us a total of 63.87%
- Small models have **specialty in some topics**
 - Further investigation is needed on common features among these specialty topics



Final Thoughts



What We Learned

- Utilizing data analysis techniques, such as clustering and visualizations, to gain insights from real world data.
- Collaborating and communicating with people in a mostly remote work setting.
- Enhancing problem-solving skills by breaking down complex technical challenges into manageable components.



Potential Next Steps

- LLM Routing
 - The idea of routing questions to models that handle that prompt category best
- Investigate common features among the topics small models specialize in
- Explore the small models that were winning a lot and analyze their strengths
- Split models into more categories of size: very small, medium, very large...



Questions?





For Students: Blank slides to copy/paste and use as needed



Bibliography

- <https://neptune.ai/blog/k-means-clustering>



Click to add title

Click to add text

Paste an image



Click to add title

Click to add text

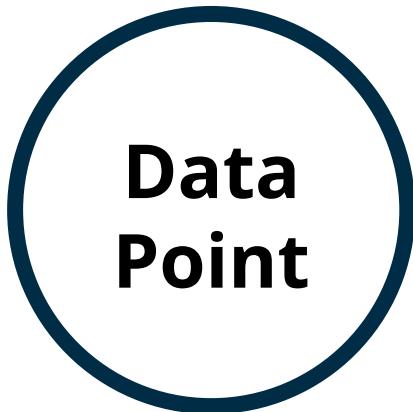
Paste an image



Paste an image



Click to add title



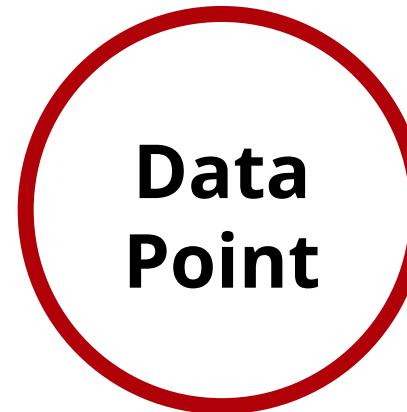
**Data
Point**

Caption providing context.



**Data
Point**

Caption providing context.



**Data
Point**

Caption providing context.