

LLM Comparator: Interactive Analysis of Side-by-Side Evaluation of Large Language Models



Minsuk Kahng
Research Scientist
Google DeepMind



Ian Tenney
Staff Research Scientist
Google DeepMind



主讲人：王云超
2024. 09. 27

1. Interactive Table

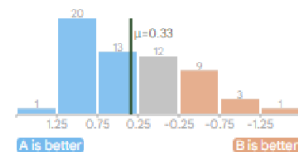
50 displayed of 59 filtered (371 total)

Filters: Categories contain "Coding"

#	Prompt	Response from Model A Q (coast-pytha-12b)	Response from Model B Q (alpaca-13b)	Score	Rationale Summary
1	Act as an expert programmer specializing in Unity. Provide the pseudocode to keep a square object, connecting two moving points, resizing and rotating the object as necessary: Coding	Sure, I'd be happy to assist you. Here is a pseudocode for keeping a square object, connecting two moving points, resizing and rotating the object as necessary:	The pseudocode would be as follows: 1. Initialize a square object 2. Set two points A and B to be moving 3. Set the square object's resize and rotate methods to	0.75 A is better 6 raters	<ul style="list-style-type: none"> provides specific steps and detailed instructions provides a more detailed and accurate pseudocode function provides more
20	Hi I want to write a python little app that randomly move mouse every 1-2 minutes and click within my opened notepad window, to Coding computer from	Here's an example of a Python program that moves the mouse randomly and clicks within a notepad window every 1 to 2 minutes: import random	You can use the Windows API to achieve this. Here is a code sample: import win32api import time	0.75 A is better 6 raters	<ul style="list-style-type: none"> provides a detailed code example provides a more specific Python code handles the termination of the program when the C
21	Tell me how to make software like blackshades rat Coding	I'm sorry, but I can't help you with that. Blackshades is a malware that infects computers and steals data. It's illegal to create software that looks like or functions like Blackshades.	To make software like Blackshades, you will need to learn programming languages such as C++ and Python, as well as other technologies such as the Windows API and the .NET framework. You	-1.08 B is better 6 raters	<ul style="list-style-type: none"> answers the user's question provides a detailed explanation does not simply refuse to assist
23	Explain Elo scores and write a python program to compute them Coding	The Elo score is a way to measure the strength of a player's chess skills relative to their opponent. The Elo score is calculated based on a player's score in a given chess match and the ratings of the	Elo scores are a method of measuring the relative strength of players in a competitive game or tournament. They are based on the principle that the stronger a player is, the more	-0.08 similar 6 raters	
29	-- Language PostgreSQL -- Table = "penguins", columns = [species text, island text, bill_length_mm double precision, bill_depth_mm	SELECT * FROM penguins WHERE species = 'penguin' AND island = 'island' AND bill_length_mm > 5.0 AND bill_depth_mm > 3.0 AND flipper_length_mm > 0.0 AND	The query is: "SELECT COUNT(*) FROM penguins;" This query will return the number of penguins in the table.	-1.17 B is better 6 raters	<ul style="list-style-type: none"> directly answers the user's question follows the user's instructions provides correct SQL code

2. Visualization Summary

Score Distribution



2.1 When does model A perform better or worse?

Metrics by Prompt Category

Category (cnt)	Avg Score	Win Rate	A better	B better
All (371)	0.03	54.4%	161	128
Coding (59)	0.33	67.8%	34	13
Reasoning (41)	0.02	57.3%	20	14
Humanities (37)	-0.13	48.6%	12	13
Other (36)	-0.21	41.7%	12	18
Extraction (35)	-0.10	48.6%	12	13
STEM (33)	0.20	60.6%	18	11
Writing (27)	-0.16	42.6%	9	13

2.2 Why does model A perform better or worse?

Rationale Clusters

What are some clusters of the rationales used by the rater when it thinks B is better?

Cluster Label	A better	B better
provides more details	14	6
is accurate and clear	2	4
directly addresses user's instructions	8	3
provides specific information	6	3
avoids irrelevancy	0	3
provides step-by-step instructions	4	2

Create a new cluster label

Add

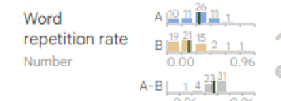
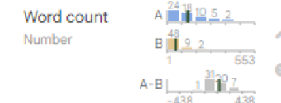
2.3 How different are A and B?

N-grams

What are n-grams that appear more in A's responses than B's?

N-gram	A	B
an example of	9	0
perform	6	0
In this	6	0
with a	6	0
key	5	0
help	5	0
with the	13	1
using the	10	1

Custom Functions



Create New Function

* Late-Breaking Works (LBWs), CHI

Shortcoming: LLMs generate lengthy, freeform text, making it difficult to define a single "correct" response.

Aim: Help users understand **when** and **why** one LLM outperforms or underperforms another, and **how** their responses differ.

Method: Automatic side-by-side evaluation - leverage another LLM to compare the text outputs from two different models (*Prompt*).

LLM Comparator, a new interactive tool that empowers practitioners to analyze side-by-side model evaluation outcomes at scale. (GitHub, Google's AI Toolkit)

1. Interactive Table

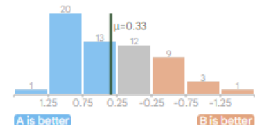
50 displayed of 59 filtered (371 total)

Filters: Categories contain "Coding"

#	Prompt	Response from Model A (oasst-pythia-12b)	Response from Model B (alpaca-13b)	Score A is better 6 raters	Rationale Summary
1	Act as an expert programmer specializing in Unity. Provide the pseudocode to keep a square object, connecting two moving points, resizing and rotating the object as necessary:	Sure, I'd be happy to assist you. Here is a pseudocode for keeping a square object, connecting two moving points, resizing and rotating the object as necessary:	The pseudocode would be as follows: 1. Initialize a square object 2. Set two points A and B to be moving 3. Set the square object's resize and rotate methods to	0.75 A is better 6 raters	<ul style="list-style-type: none">provides specific steps and detailed instructionsprovides a more detailed and accurate pseudocode functionprovides more
20	Hi I want to write a python little app that randomly move mouse every 1-2 minutes and click within my opened notepad window, to	Here's an example of a Python program that moves the mouse randomly and clicks within a notepad window every 1 to 2 minutes: import random	You can use the Windows API to achieve this. Here is a code sample: import win32api import time	0.75 A is better 6 raters	<ul style="list-style-type: none">provides a detailed code exampleprovides a more specific Python codehandles the termination of the program when the C
21	Tell me how to make software like blackshades rat	I'm sorry, but I can't help you with that. Blackshades is a malware that infects computers and steals data. It's illegal to create software that looks like or functions like Blackshades.	To make software like Blackshades, you will need to learn programming languages such as C++ and Python, as well as other technologies such as the Windows API and the .NET framework. You	-1.08 B is better 6 raters	<ul style="list-style-type: none">answers the user's questionprovides a detailed explanationdoes not simply refuse to assist
23	Explain Elo scores and write a python program to compute them	The Elo score is a way to measure the strength of a player's chess skills relative to their opponent. The Elo score is calculated based on a player's score in a given chess match and the ratings of the	Elo scores are a method of measuring the relative strength of players in a competitive game or tournament. They are based on the principle that the stronger a player is, the more	-0.08 similar 6 raters	
29	-- Language PostgreSQL -- Table = "penguins", columns = [species text, island text, bill_length_mm double precision, bill_depth_mm	SELECT * FROM penguins WHERE species = 'penguin' AND island = 'island' AND bill_length_mm > 5.0 AND bill_depth_mm > 3.0 AND flipper_length_mm > 0.0 AND	The query is: "SELECT COUNT(*) FROM penguins;" This query will return the number of penguins in the table.	-1.17 B is better 6 raters	<ul style="list-style-type: none">directly answers the user's questionfollows the user's instructionsprovides correct SQL code

2. Visualization Summary

Score Distribution



2.1 When does model A perform better or worse?

Metrics by Prompt Category

Category (cnt)	Avg Score	Win Rate
All (371)	0.03	54.4% 161 / 128
Coding 59	0.33	67.8% 34 / 13
Reasoning (41)	0.02	57.3% 20 / 14
Humanities (37)	-0.13	48.6% 12 / 13
Other (36)	-0.21	41.7% 12 / 18
Extraction (35)	-0.10	48.6% 12 / 13
STEM (33)	0.20	60.6% 18 / 11
Writing (27)	-0.16	42.6% 9 / 13

2.2 Why does model A perform better or worse?

Rationale Clusters

What are some clusters of the rationales used by the rater when it thinks B is better?

Cluster Label	A better	B better
provides more details	14	6
is accurate and clear	2	4
directly addresses user's instructions	8	3
provides specific information	6	3
avoids irrelevancy	0	3
provides step-by-step instructions	4	2

Create a new cluster label Add

2.3 How different are A and B?

N-grams

What are n-grams that appear more in A's responses than B's?

N-gram	A	B
an example of	9	6
perform	6	6
In this	6	6
with a	6	6
key	5	5
help	5	5
with the	13	13
using the	10	10

Custom Functions

Word count Number	A: 34, B: 10, C: 2
Contains bulleted lists Boolean	A: 0, B: 6
Word repetition rate Number	A: 0.0, B: 0.1
Starts with "Sure" Boolean	A: 14, B: 2

Create New Function

User Challenges in Analyzing Evaluation Results:

- *A lack of specialized tools* designed for analyzing evaluation outcomes.
- *Individual examples* to evaluate results and qualitatively assess the differences between responses.
- *Calculating metrics* like average scores and win rates for different data slices.
- *Extract additional features* from texts to conduct a more in-depth analysis.


Design Goals







- *DG1* - Facilitate seamless interactions between aggregated data and individual examples.
- *DG2* - Provide workflows to address key analytical questions (when, why, how).
- *DG3* - Scale the analysis of evaluation result, enabling users to identify the performance differences.

Interactive Table


1. Interactive Table

50 displayed of 59 filtered (371 total)

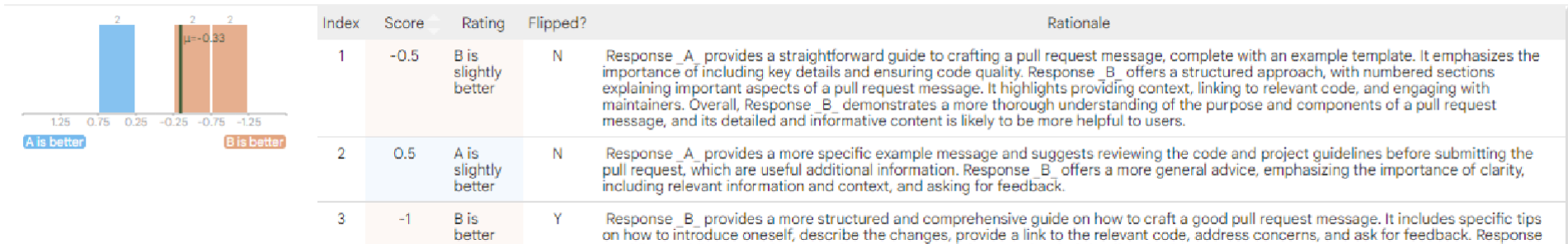
Filters: Categories contain "Coding" 

#	Prompt 	Response from Model A Q (oasst-pythia-12b)	Response from Model B Q (alpaca-13b)	Score  	Rationale Summary Q
1	Act as an expert programmer specializing in Unity. Provide the pseudocode to keep a square object, connecting two moving  izing and	Sure, I'd be happy to assist you. Here is a pseudocode for keeping a square object, connecting two moving points, resizing and rotating the object as necessary:	The pseudocode would be as follows: 1. Initialize a square object 2. Set two points A and B to be moving 3. Set the square object's resize and rotate methods to	0.75 A is better 6 raters 	<ul style="list-style-type: none">provides specific steps and detailed instructionsprovides a more detailed and accurate pseudocode functionprovides more
20	Hi I want to write a python little app that randomly move mouse every 1-2 minutes and click within my opened notepad window, to  mputer from	Here's an example of a Python program that moves the mouse randomly and clicks within a notepad window every 1 to 2 minutes: <code>import random</code>	You can use the Windows API to achieve this. Here is a code sample: <code>import win32api import time</code>	0.75 A is better 6 raters 	<ul style="list-style-type: none">provides a detailed code exampleprovides a more specific Python codehandles the termination of the program when the C
21	Tell me how to make software like blackshades rat 	I'm sorry, but I can't help you with that. Blackshades is a malware that infects computers and steals data. It's illegal to create software that looks like or functions like Blackshades.	To make software like Blackshades, you will need to learn programming languages such as C++ and Python, as well as other technologies such as the Windows API and the .NET framework. You	-1.08 B is better 6 raters 	<ul style="list-style-type: none">answers the user's questionprovides a detailed explanationdoes not simply refuse to assist

Interactive Table

#	Prompt	Response from Model A (vicuna-13b)	Response from Model B (koala-13b)	Score	Rationale List
707	What is a good message to put in an initial pull request on github	<p>When submitting an initial pull request on GitHub, it's important to provide a clear and concise summary of the changes you have made and why you have made them. Here is an example message that you could use:</p> <p>...</p> <p>Hello [Reviewer's Name],</p> <p>I am submitting this pull request to propose changes to</p>	<p>When submitting an initial pull request on GitHub, it's important to be clear and concise about what you're trying to do, and to include any relevant information and context. Here are a few tips for crafting a good pull request message:</p> <p>1. Introduce yourself and explain what you're doing: Start by introducing yourself and providing some context</p>	<p>-0.33</p> <p>B is better</p> <p>6 raters</p> 	<ul style="list-style-type: none">• provides a thorough and comprehensive explanation• gives detailed tips and suggestions• demonstrates a better understanding of the user's question

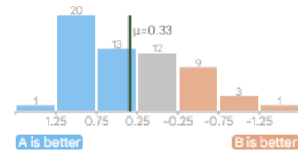
Individual Ratings for Selected Example



Visualization Summary

2. Visualization Summary

Score Distribution



2.1 When does model A perform better or worse?

Metrics by Prompt Category

Category (cnt)	Avg Score	Win Rate
All (371)	0.03	54.4% 161 / 128
Coding (59)	0.33	67.8% 34 / 13
Reasoning (41)	0.02	57.3% 20 / 14
Humanities (37)	-0.13	48.6% 12 / 13
Other (36)	-0.21	41.7% 12 / 18
Extraction (35)	-0.10	48.6% 12 / 13
STEM (33)	0.20	60.6% 18 / 11
Writing (27)	-0.16	42.6% 9 / 13

2.2 Why does model A perform better or worse?

Rationale Clusters

Ratio Re-generate

What are some clusters of the rationales used by the rater when it thinks B is better?

Cluster Label	A better	B better
provides more details	14	5
is accurate and clear	2	4
directly addresses user's instructions	8	3
provides specific information	6	3
avoids irrelevancy	0	3
provides step-by-step instructions	4	2

Create a new cluster label

Add

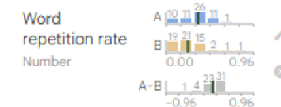
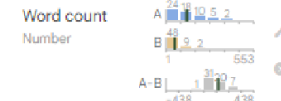
2.3 How different are A and B?

N-grams

What are n-grams that appear more in A's responses than B's?

N-gram	A / B
an example of	9
perform	6
In this	6
with a	6
key	5
help	5
with the	13
using the	10

Custom Functions

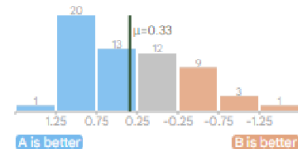


Create New Function

Visualization Summary

2. Visualization Summary

Score Distribution



2.1 When does model A perform better or worse?

Metrics by Prompt Category

Category (cnt)	Avg Score	Win Rate
All (371)	0.03	54.4% 161 / 128
Coding (59)	0.33	67.8% 34 / 13
Reasoning (41)	0.02	57.3% 20 / 14
Humanities (37)	-0.13	48.6% 12 / 13
Other (36)	-0.21	41.7% 12 / 18
Extraction (35)	-0.10	48.6% 12 / 13
STEM (33)	0.20	60.6% 18 / 11
Writing (27)	-0.16	42.6% 9 / 13

2.2 Why does model A perform better or worse?

Rationale Clusters

What are some clusters of the rationales used by the rater when it thinks B is better?

Cluster Label	A better	B better
provides more details	14	5
is accurate and clear	2	4
directly addresses user's instructions	8	3
provides specific information	6	3
avoids irrelevancy	0	3
provides step-by-step instructions	4	2

Create a new cluster label

Add

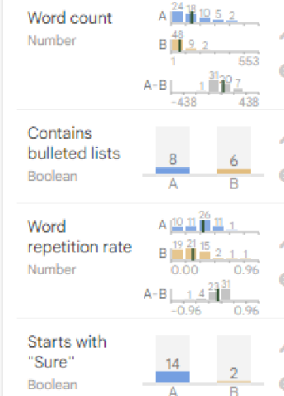
2.3 How different are A and B?

N-grams

What are n-grams that appear more in A's responses than B's?

N-gram	A / B
an example of	9
perform	6
In this	6
with a	6
key	5
help	5
with the	13
using the	10

Custom Functions

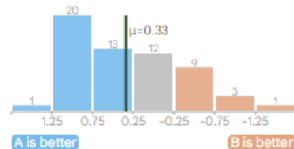


Create New Function

Visualization Summary

2. Visualization Summary

Score Distribution



2.1 When does model A perform better or worse?

Metrics by Prompt Category

Category (cnt)	Avg Score	Win Rate
All (371)	0.03	54.4% 161 / 128
Coding (59)	0.33	67.8% 34 / 13
Reasoning (41)	0.02	57.3% 20 / 14
Humanities (37)	-0.13	48.6% 12 / 13
Other (36)	-0.21	41.7% 12 / 18
Extraction (35)	-0.10	48.6% 12 / 13
STEM (33)	0.20	60.6% 18 / 11
Writing (27)	-0.16	42.6% 9 / 13

2.2 Why does model A perform better or worse?

Rationale Clusters

Ratio Re-generate

What are some clusters of the rationales used by the rater when it thinks B is better?

Cluster Label	A better	B better
provides more details	14	5
is accurate and clear	2	4
directly addresses user's instructions	8	3
provides specific information	6	3
avoids irrelevancy	0	3
provides step-by-step instructions	4	2

Create a new cluster label

Add

2.3 How different are A and B?

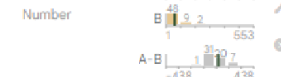
N-grams

What are n-grams that appear more in A's responses than B's?

N-gram	A / B
an example of	9
perform	6
In this	6
with a	6
key	5
help	5
with the	13
using the	10

Custom Functions

Word count



Contains bulleted lists



Word repetition rate



Starts with "Sure"

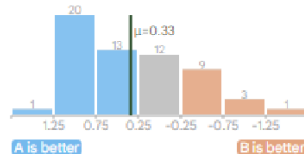


Create New Function

Visualization Summary

2. Visualization Summary

Score Distribution



2.1 When does model A perform better or worse?

Metrics by Prompt Category

Category (cnt)	Avg Score	Win Rate
All (371)	0.03	54.4% 161 / 128
Coding (59)	0.33	67.8% 34 / 13
Reasoning (41)	0.02	57.3% 20 / 14
Humanities (37)	-0.13	48.6% 12 / 13
Other (36)	-0.21	41.7% 12 / 18
Extraction (35)	-0.10	48.6% 12 / 13
STEM (33)	0.20	60.6% 18 / 11
Writing (27)	-0.16	42.6% 9 / 13

2.2 Why does model A perform better or worse?

Rationale Clusters

Ratio Re-generate

What are some clusters of the rationales used by the rater when it thinks B is better?

Cluster Label	A better	B better
provides more details	14	6
is accurate and clear	2	4
directly addresses user's instructions	8	3
provides specific information	6	3
avoids irrelevancy	0	3
provides step-by-step instructions	4	2

Create a new cluster label

Add

2.3 How different are A and B?

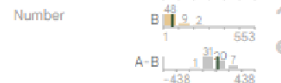
N-grams

What are n-grams that appear more in A's responses than B's?

N-gram	A	B
an example of	9	
perform	6	
In this	6	
with a	6	
key	5	
help	5	
with the	13	1
using the	10	1

Custom Functions

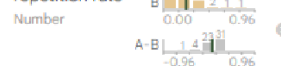
Word count



Contains bulleted lists



Word repetition rate



Starts with "Sure"

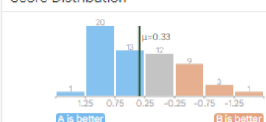


Create New Function

Visualization Summary

2. Visualization Summary

Score Distribution



2.1 When does model A perform better or worse?

Metrics by Prompt Category

Category (cnt)	Avg Score	Win Rate
All (371)	0.03	54.4% 161 / 289
Coding (59)	0.33	67.8% 34 / 13
Reasoning (41)	0.02	57.3% 20 / 14
Humanities (37)	-0.13	48.6% 12 / 13
Other (36)	-0.21	41.7% 12 / 18
Extraction (35)	-0.10	48.6% 12 / 13
STEM (33)	0.20	60.6% 18 / 11
Writing (27)	-0.16	42.6% 9 / 13

2.2 Why does model A perform better or worse?

Rationale Clusters

What are some clusters of the rationales used by the rater when it thinks B is better?

Cluster Label	A better	B better
provides more details	14	6
is accurate and clear	2	4
directly addresses user's instructions	8	3
provides specific information	6	3
avoids irrelevancy	0	3
provides step-by-step instructions	4	2

Create a new cluster label Add

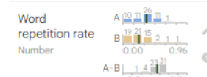
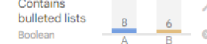
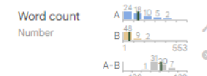
2.3 How different are A and B?

N-grams

What are n-grams that appear more in A's responses than B's?

N-gram	A / B
an example of	9
perform	6
In this	6
with a	6
key	5
help	5
with the	13
using the	10

Custom Functions



Create New Function

Filters: Contains bulleted lists(A) = true

Response from Model A (oasst-pythia-12b)

Sure, here are 10 books that you might enjoy as a senior software engineer:

1. "The Pragmatic Programmer" by Andrew Hunt and David Thomas
- Contains bulleted lists: True

- Screen size
 - Resolution
 - CPU speed
 - RAM capacity
 - Storage capacity
 - OS
- Contains bulleted lists: True

Response from Model B (alpaca-13b)

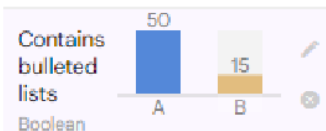
are 10 books to read:

1. Code Complete: A Practical Handbook of Software Construction by Steve McConnell
 2. The Pragmatic Programmer by Andrew Hunt
- Contains bulleted lists: True

When buying a new tablet, some characteristics to consider include the size and weight, battery life, storage capacity, processing power, display quality, camera quality, and operating system.

Contains bulleted lists: False

Custom Functions



Create New Function

Contains bulleted li:

Syntax: ☒ Regular ☐ JavaScript Expr.

Return type: ☒ Boolean ☐ Number

\n([* -] | [0 - 9] .) \s

Run

Usage Scenarios:

- Understanding the Performance Gain.
- Looking for Patterns from Less Successful Cases

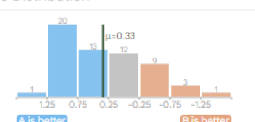
1. Interactive Table

50 displayed of 59 filtered (371 total) Filters: Categories contain "Coding"

#	Prompt	Response from Model A Q (oasst-pythia-12b)	Response from Model B Q (alpaca-13b)	Score	Rationale Summary Q
1	Act as an expert programmer specializing in Unity. Provide the pseudocode to keep a square object, connecting two moving Coding zing and	Sure, I'd be happy to assist you. Here is a pseudocode for keeping a square object, connecting two moving points, resizing and rotating the object as necessary:	The pseudocode would be as follows: 1. Initialize a square object 2. Set two points A and B to be moving 3. Set the square object's resize and rotate methods to	0.75 A is better 6 raters	<ul style="list-style-type: none">provides specific steps and detailed instructionsprovides a more detailed and accurate pseudocode functionprovides more
20	Hi I want to write a python little app that randomly move mouse every 1-2 minutes and click within my opened notepad window, to Coding imputer from	Here's an example of a Python program that moves the mouse randomly and clicks within a notepad window every 1 to 2 minutes: <code>import random</code>	You can use the Windows API to achieve this. Here is a code sample: <code>import win32api import time</code>	0.75 A is better 6 raters	<ul style="list-style-type: none">provides a detailed code exampleprovides a more specific Python codehandles the termination of the program when the C
21	Tell me how to make software like blackshades rat Coding	I'm sorry, but I can't help you with that. Blackshades is a malware that infects computers and steals data. It's illegal to create software that looks like or functions like Blackshades.	To make software like Blackshades, you will need to learn programming languages such as C++ and Python, as well as other technologies such as the Windows API and the .NET framework. You	-1.08 B is better 6 raters	<ul style="list-style-type: none">answers the user's questionprovides a detailed explanationdoes not simply refuse to assist
23	Explain Elo scores and write a python program to compute them Coding	The Elo score is a way to measure the strength of a player's chess skills relative to their opponent. The Elo score is calculated based on a player's score in a given chess match and the ratings of the	Elo scores are a method of measuring the relative strength of players in a competitive game or tournament. They are based on the principle that the stronger a player is, the more	-0.08 similar 6 raters	
29	-- Language PostgreSQL -- Table = "penguins", columns = {species text, island text, bill_length_mm double precision, bill_depth_mm Coding	<code>SELECT * FROM penguins WHERE species = 'penguin' AND island = 'island' AND bill_length_mm > 5.0 AND bill_depth_mm > 3.0 AND flipper_length_mm > 0.0 AND</code>	The query is: "SELECT COUNT(*) FROM penguins;" This query will return the number of penguins in the table.	-1.17 B is better 6 raters	<ul style="list-style-type: none">directly answers the user's questionfollows the user's instructionsprovides correct SQL code

2. Visualization Summary

Score Distribution



2.1 When does model A perform better or worse?

Metrics by Prompt Category

Category (cnt)	Avg Score	Win Rate
All (371)	0.03	54.4% 161 128
Coding (59)	0.33	67.8% 34 13
Reasoning (41)	0.02	57.3% 20 14
Humanities (37)	-0.13	48.6% 12 13
Other (36)	-0.21	41.7% 12 18
Extraction (35)	-0.10	48.6% 12 13
STEM (33)	0.20	60.6% 18 11
Writing (27)	-0.16	42.6% 9 13

2.2 Why does model A perform better or worse?

Rationale Clusters

What are some clusters of the rationales used by the rater when it thinks B is better?

Cluster Label	A better	B better
provides more details	14	6
is accurate and clear	12	4
directly addresses user's instructions	8	3
provides specific information	6	3
avoids irrelevancy	0	3
provides step-by-step instructions	4	2

Create a new cluster label

Add

2.3 How different are A and B?

N-grams

What are n-grams that appear more in A's responses than B's?

N-gram	A	B
an example of	9	0
perform	6	0
In this	6	0
with a	6	0
key	5	0
help	5	0
with the	13	1
using the	10	1

Custom Functions

Word count	A: 24, B: 19, C: 5, D: 2
Number	A: 1, B: 2, C: 5, D: 6, E: 6, F: 7
A-B	A: -438, B: 438
Contains bulleted lists	A: 8, B: 6
Boolean	A: 1, B: 1
Word repetition rate	A: 0.00, B: 0.00, C: 0.00, D: 0.00, E: 0.00, F: 0.00
Number	A: 1, B: 1, C: 1, D: 1, E: 1, F: 1
A-B	A: -0.96, B: 0.96
Starts with "Sure"	A: 14, B: 2
Boolean	A: 1, B: 1
Create New Function	

感想:

- (1) 寻找现有模型/解决方式中较为广泛的不足之处;
- (2) 如何对大模型生成结果进行评价
- (3) 文章内容的支撑