Universidad del Valle de Guatemala
Ciencia de Datos
Sección 10

Batalla cara a cara con datos del ChatBot Arena
Análisis Exploratorio

María José Gil -20337
Fabián Juárez - 21440
Joshua Chicoj -20566
Sofi Lam Méndez - 21548

30 de septiembre 2024

Investigación del tema:

El proyecto se enfoca en la predicción de preferencias de usuarios entre respuestas generadas por dos modelos de lenguaje distintos. Se podrían llevar a cabo distintas técnicas para poder resolver problemas relacionados al procesamiento del lenguaje natural como se da en este caso, veamos algunas de ellas:

Modelos / Técnicas de detección de patrones en el Procesamiento del Lenguaje Natural:

- Modelos de lenguajes grandes (LLMs):
  Es un tipo de modelo de aprendizaje automático el cual es pre-entrenados con grandes cantidades de datos textuales, capaces de generar texto coherente y relevante. Este se utiliza para tareas de procesamiento del lenguaje natural. Estos modelos son de propósito general que se destaca en un amplia gama de tareas *(IBM, s.f)* .
- Reinforcement Learning form Human Feedback (RLHF):
  Técnica clave en la que los modelos aprenden a partir de las preferencias expresadas por los usuarios a base de ejemplos. Esto implica ajustar el comportamiento del modelo de acuerdo con retroalimentación específica de los humanos (*What Is RLHF? - Reinforcement Learning From Human Feedback Explained - AWS, s. f.*) .
- Clasificación con SVM:
  Es un algoritmo de aprendizaje supervisado que se utiliza para la clasificación de datos. Este puede combinarse con otros métodos como los histogramas de gradientes orientados para el reconocimiento y detección de objetos. *(Pattern Recognition, s.f)*

Situación Problemática:

Veamos como en los últimos años ha habido un crecimiento y uso exponencial de los chatbots impulsados por inteligencia artificial. Cada vez más empresas los utilizan para cosas como servicio al cliente, resolución de problemas, ayudar con clases, etc. Esto ha hecho que la demanda para sistemas mejores, más inteligentes, y a la vez más eficientes ha incrementado. Esto entonces nos da paso al problema en cuestión que es, que debemos mejorar las respuestas y el rendimiento de los chatbots, ya que muchas veces sigue siendo necesaria la interacción humana, y los chatbots no son capaces de manejar problemas o situaciones complejas.

Problema científico:

El problema en este caso es de qué manera podemos evaluar y mejorar a los chatbots para que estos elaboren siempre contenido relevante, coherente, que tenga que ver con la situación y el contexto que se le están proveyendo, que suene de forma natural. De modo que de alguna

manera debemos ser capaces de comparar y evaluar a estos Chatbots, como en la Chatbot Arena para lograr hacer esto de manera eficiente.

Objetivos:

General:

Lograr identificar patrones a nivel general que nos indiquen qué tipo de respuestas son las que los usuarios suelen preferir

Específicos:

1. Identificar, si dentro de los chatbots datos existe uno que sea mucho más dominante que los otros
2. Diseñar métricas efectivas para poder evaluar las respuestas de los diferentes chatbots y poder así determinar cuál es la preferente por el usuario.
3. Utilizar distintas técnicas de Machine Learning para identificar cuál es la más efectiva para realizar las evaluaciones y las predicciones

Análisis exploratorio:

Descripción de los datos:

El conjunto de datos se divide en tres categorías, el conjunto train, el conjunto test y el conjunto sample. Cada uno cuenta con las siguientes variables:

Train:

1. Id: identificador de la fila
2. model_[a/b]: la identificación del modelo [a / b]
3. prompt: solicitud inicial dada por el usuario
4. response_[a/b]: la respuesta del modelo a la solicitud anterior
5. winner_model_[a/b/tie]: columna en binario que determina la elección del juez.

Test

1. Id: identificador único de la fila
2. promtp: solicitud brindada por el usuario
3. response_[a/b]: respuestas del modelo a la solicitud

Sample

1. id: identificador único de la fila
2. winner_model_[a/b/tie]: lo que se predice del conjunto test.

El conjunto de datos train, cuenta con 57,477 observaciones y 9 variables distintas. Las cuáles 3 son categóricas, 1 es numérica y 5 variables nominales.

Limpieza de datos:

Para el conjunto de datos las operaciones de limpieza que se realizaron son las siguientes:

1. Primero, se revisaron si existen valores nulos o faltantes, de los cuáles no se encontraron. Además, se revisó si existen observaciones duplicadas, las cuáles tampoco hubo, por lo que no fue necesario eliminar datos.

2. Posteriormente, vemos que tanto los prompts del usuario como las respuestas de los modelos son una lista, de prompts y respuestas respectivamente. Entonces, convertimos esto a lista, para poder acceder a cada elemento individualmente.
3. Luego, observamos que las respuestas de los modelos de la IA's están dadas en formato LaTeX, por lo que, cambiamos el formato de cada elemento de las listas de respuestas a formato texto. Para esto quitamos todos los comandos existentes en el formato LaTeX tales como: "\begin\{itemize\}", "\textbf{}", \textit{}", entre otros.
4. Finalmente, pasamos tanto las listas de los prompts como las de las respuestas ya en texto de las IA's por una función pre-definida de limpieza de texto, que se encarga de eliminar caracteres innecesarios y stopwords, convertir cada palabra del texto a minúsculas, y realiza la tokenización.

Análisis exploratorio

Se buscó en cada uno de los modelos las distintas victorias en total para cada uno de ellos. Se muestran algunos de los modelos con las mayores victorias en el conjunto de datos

gpt-4-1106-preview	7387
gpt-3.5-turbo-0613	7083
gpt-4-0613	6165
claude-2.1	5583
claude-instant-1	4136

Figura 1. Victorias por modelo.

Se encontró que unos de estos tenían más prompts y respuestas que otros por los que para ver el win rate se tuvo que llevar a cabo una comparativa basándose en las victorias con el total de usos en el modelo e igualmente con los que perdieron y empataron.

	model	win	loss	tie	count
0	RWKV-4-Raven-14B	0.228843	0.461140	0.310017	1158
1	alpaca-13b	0.253742	0.446187	0.300071	1403
2	chatglm-6b	0.172086	0.482950	0.344964	1261
3	chatglm2-6b	0.129433	0.567376	0.303191	564
4	chatglm3-6b	0.158746	0.543984	0.297270	989

Figura 2. Victorias por modelo estandarizadas.

Consiguientemente, para poder ver esto de manera más clara, se realizaron gráficas tanto para los más preferidos, los menos preferidos, y los que tenían una mayor cantidad de empates

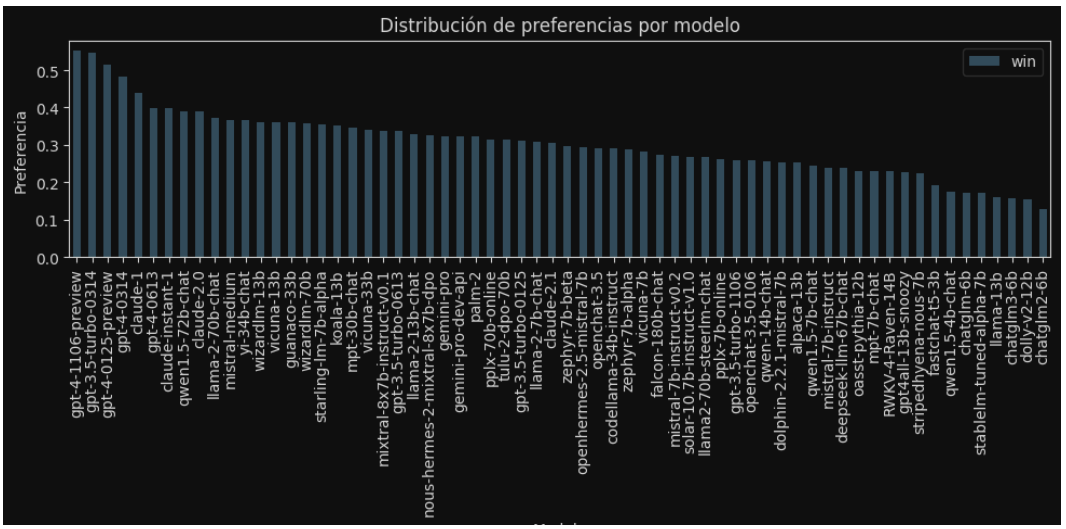

Figura 3. Gráficas de la frecuencia de Victorias estandarizadas.

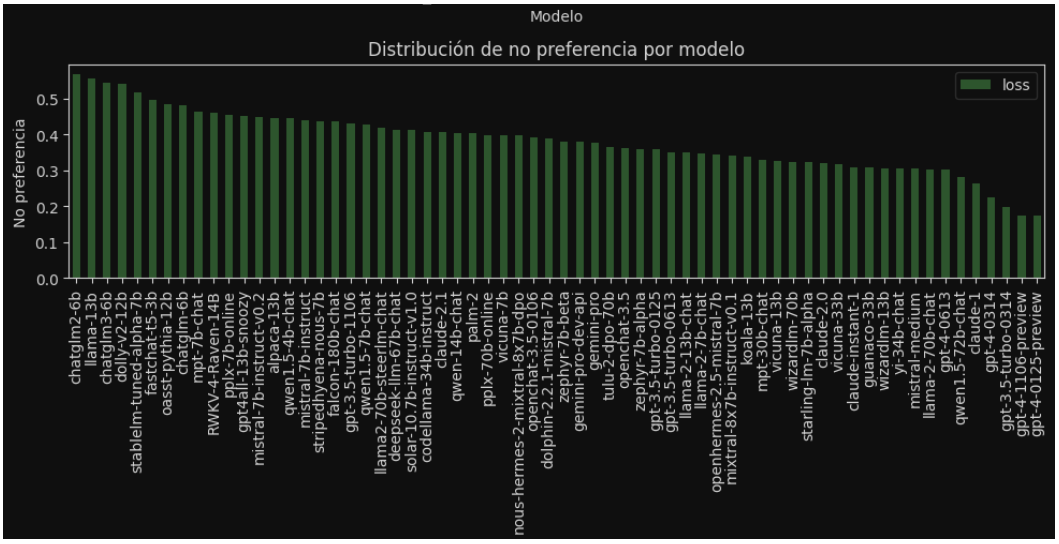

Figura 4. Gráficas de la frecuencia de Derrotas estandarizadas.

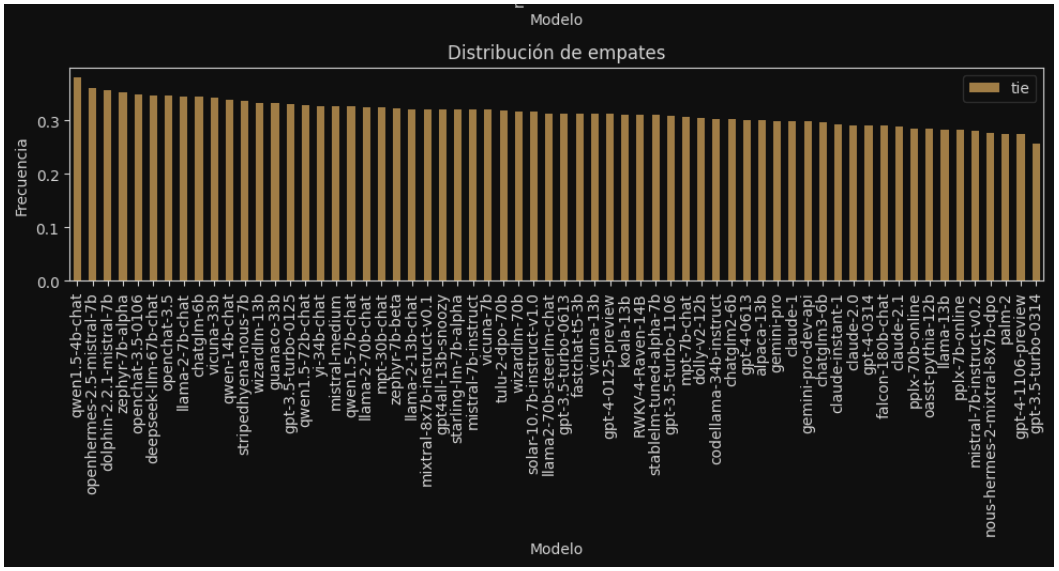

Figura 5. Gráficas de la frecuencia de Empates estandarizadas.

Los resultados muestran que el modelo gpt-4-1106-preview es consistentemente uno de los más preferidos por los usuarios, con 3678 preferencias en el rol de Model A y 3709 en Model B. Este modelo destaca significativamente en las interacciones, lo que sugiere que sus respuestas son frecuentemente valoradas como las mejores. Le sigue el modelo gpt-3.5-turbo-0613, con un rendimiento igualmente fuerte en ambas posiciones: 3553 preferencias en Model A y 3530 en Model B, lo que lo convierte en un competidor muy robusto.

En cuanto a los empates, gpt-3.5-turbo-0613 lidera con 2217 empates, seguido de cerca por gpt-4-1106-preview con 2027, lo que indica que los usuarios encuentran difícil elegir entre las respuestas de estos dos modelos en muchas ocasiones. Otros modelos relevantes incluyen gpt-4-0613 y claude-2.1, que también reciben una cantidad considerable de preferencias en las tres categorías (Model A, Model B y empates), aunque en menor medida. Por otro lado, modelos como gpt-4-0314 y claude-instant-1 presentan una participación más modesta. Estos datos sugieren que los modelos GPT, en sus diversas versiones, tienden a dominar en términos de preferencias del usuario, con frecuentes empates entre ellos, lo que resalta la alta calidad y competitividad de sus respuestas.

Por otro lado, veamos que a la vez, modelos como claude-instant-1 y gpt-4-0314 están también en el rango de poca preferencia como en el de mucha preferencia. Como están en ambos rangos, podemos suponer que estos modelos son buenos en dar respuestas para ciertos temas, pero considerablemente malos para otros, es decir, están más especializados y no son tan generales como otros modelos más dominantes. Respecto a los empates, como se estableció anteriormente, tenemos chatbots como el GPT-3.5-turbo-0613 y GPT-4-1106-preview que tienen una cantidad considerable de empates. ¿Qué implica esto? Que la efectividad de los modelos es comparable, por lo que abre a la pregunta de si esto sucede en queries específicas o en las generales.

Finalmente, es importante darnos cuenta de que, inclusive los modelos que suelen no ser preferidos, no tienen por qué ser necesariamente malos. Es decir, es posible que estos modelos están enfocados a cosas muy específicas, como servicio médico, de autos, etc. y consiguientemente fallen más cuando se les hacen queries generales. Además de esto, hay que considerar que las preferencias no sean necesariamente por el contenido de la respuesta, sino por el estilo en el que estas respuestas están redactadas. Siempre existe la posibilidad de que simplemente el modelo no sea tan bueno, pero en este momento es importante que consideremos todas las posibilidades hasta profundizar más con la investigación de los datos.

A partir de esto, consideramos que varias de nuestras dudas podían ser aclaradas a través de contar las repeticiones de palabras en los prompts, tanto para los mejores, los peores y los más empatados. Consiguientemente, para hacerlo de manera visual, decidimos hacer Word Clouds del top 5 de cada una de las categorías, tanto para prompts como para las respuestas dadas por los modelos. Antes de hacer esto, decidimos también eliminar más palabras, el “equivalente” a stop words en preguntas, estas palabras fueron: 'use', 'used', 'data', 'help',

'need', 'time', 'may', 'one', 'would', 'could', 'like', 'also', 'using' y 'make'. Estas Word Clouds son visibles en nuestro repositorio de GitHub.

A partir de las word cloud, podemos confirmar las suposiciones hechas previamente. Veamos que los top model todos tienen word clouds muy similares. Esto nos indica que la gente suele preferir el uso de las mismas palabras, indicándonos que tal vez, más allá del contenido, que definitivamente es importante, también es significativo la manera en la que las respuestas son redactadas. Por otro lado, los non prefered también tienen las mismas palabras entre sí, particularmente las palabra example resalta entre todos. Esto nos indica que probablemente estos modelos no sean buenos generalizando, o profundizando en las respuestas más allá de lo prácticamente fáctico.

Finalmente, el análisis más interesante es el de los top tied models. Veamos qué son las word clouds más diferentes entre sí, y parecen ser de temas muy específicos, ya sea programar, poemas, videos, entre otros. Esto nos responde a la pregunta que los ties suelen ocurrir en queries más especializadas, y menos en preguntas generales, indicándonos que, algunos bots, aunque no son tan buenos generalizando, siguen siendo bastante buenos en temas específicos.

Hallazgos y Conclusiones

A partir de todo esto podemos concluir varias cosas, tanto de las respuestas de los chatbots como de las preferencias de las personas. Veamos que por lo observado, un factor significativo en la elección de una respuesta o no, es la manera en la que la respuesta está redactada. Además de esto, luego de investigar un poco, nos dimos cuenta de que todos los top bots son “State of the Art”. Los modelos en el top son modelos state-of-the-art, esta es una medida global de qué tan bueno es un modelo luego de haberlo puesto a prueba realizando varias tareas. Dichos modelos cumplen más de una funcionalidad, es decir, son capaces de procesar imágenes, audio, generar texto, etc. Los modelos en el bottom, además de no ser SOTA, también realizan tareas más específicas o no tienen soporte para múltiples lenguajes. Lo cual confirmaba nuestra teoría de que los Chatbots más generalizables eran los mejores

Por otro lado, es interesante notar que los bots en general mejoran, y esto lo podemos ver en los empates, cuando se les piden cosas específicas. Es decir, cuando, por ejemplo, se le pide al chatbot que programe algo, explique un contenido preciso y otros tipos de prompts que no sean tan abiertos y puedan obtener la información fácil, y devolverla de manera clara sin que la redacción y el estilo sean tan importantes. Podemos ver que también esta es la razón para explicar a los Chatbots menos preferidos. Veamos que la palabra que más se repite es “example”, es decir, les cuesta más dar un ejemplo de una situación, cosas que no son puramente fácticas o no necesariamente están explicitadas en algún lado. Les cuesta más ser “creativos”, por lo que su preferencia fue menor.

Link a la presentación

La presentación fue trabajada en el siguiente enlace.

Link al repositorio

El código fue trabajado en el siguiente enlace.

Referencias

Reconocimiento de patrones (Pattern Recognition). (s. f.). MATLAB & Simulink. https://la.mathworks.com/discovery/pattern-recognition.html

What are Large Language Models (LLMs)? | IBM. (s. f.). https://www.ibm.com/topics/large-language-models

What is RLHF? - Reinforcement Learning from Human Feedback Explained - AWS. (s. f.). Amazon Web Services, Inc. https://aws.amazon.com/what-is/reinforcement-learning-from-human-feedback/