



IIC3670 Procesamiento de Lenguaje Natural

<https://github.com/marcelomendoza/IIC3670>

- EVALUACIÓN -

GLUE: General Language Understanding Evaluation

CoLA (Corpus of Linguistic Acceptability): Esta es una tarea de clasificación binaria en la que los sistemas deben predecir si una oración en inglés es gramaticalmente correcta o no.

SST-2 (Stanford Sentiment Treebank): Esta es otra tarea de clasificación binaria en la que los sistemas deben predecir el sentimiento de una oración, ya sea positivo o negativo.

MRPC (Microsoft Research Paraphrase Corpus): En esta tarea, los sistemas deben determinar si dos oraciones son paráfrasis la una de la otra.

STS-B (Semantic Textual Similarity Benchmark): Aquí, los sistemas deben predecir qué tan similar es el significado de dos oraciones en una escala de 0 a 5.

QQP (Quora Question Pairs): En esta tarea, los sistemas deben determinar si dos preguntas son semánticamente equivalentes.

MNLI (Multi-Genre Natural Language Inference): Esta tarea requiere que los sistemas determinen la relación entre una oración de premisa y una oración de hipótesis: si la hipótesis es un entailment, una contradicción, o neutra.

QNLI (Question Natural Language Inference): Esta es una versión de la tarea de Natural Language Inference, en la que se presenta a los sistemas una pregunta y un párrafo, y deben determinar si la respuesta a la pregunta se puede inferir del párrafo.

RTE (Recognizing Textual Entailment): Similar a MNLI, pero las parejas de oraciones provienen de varios sets de datos anteriores.

WNLI (Winograd NLI): Esta tarea está diseñada para evaluar la capacidad de un sistema para la comprensión de la referencia de pronombres, es decir, a qué se refiere un pronombre en un texto.

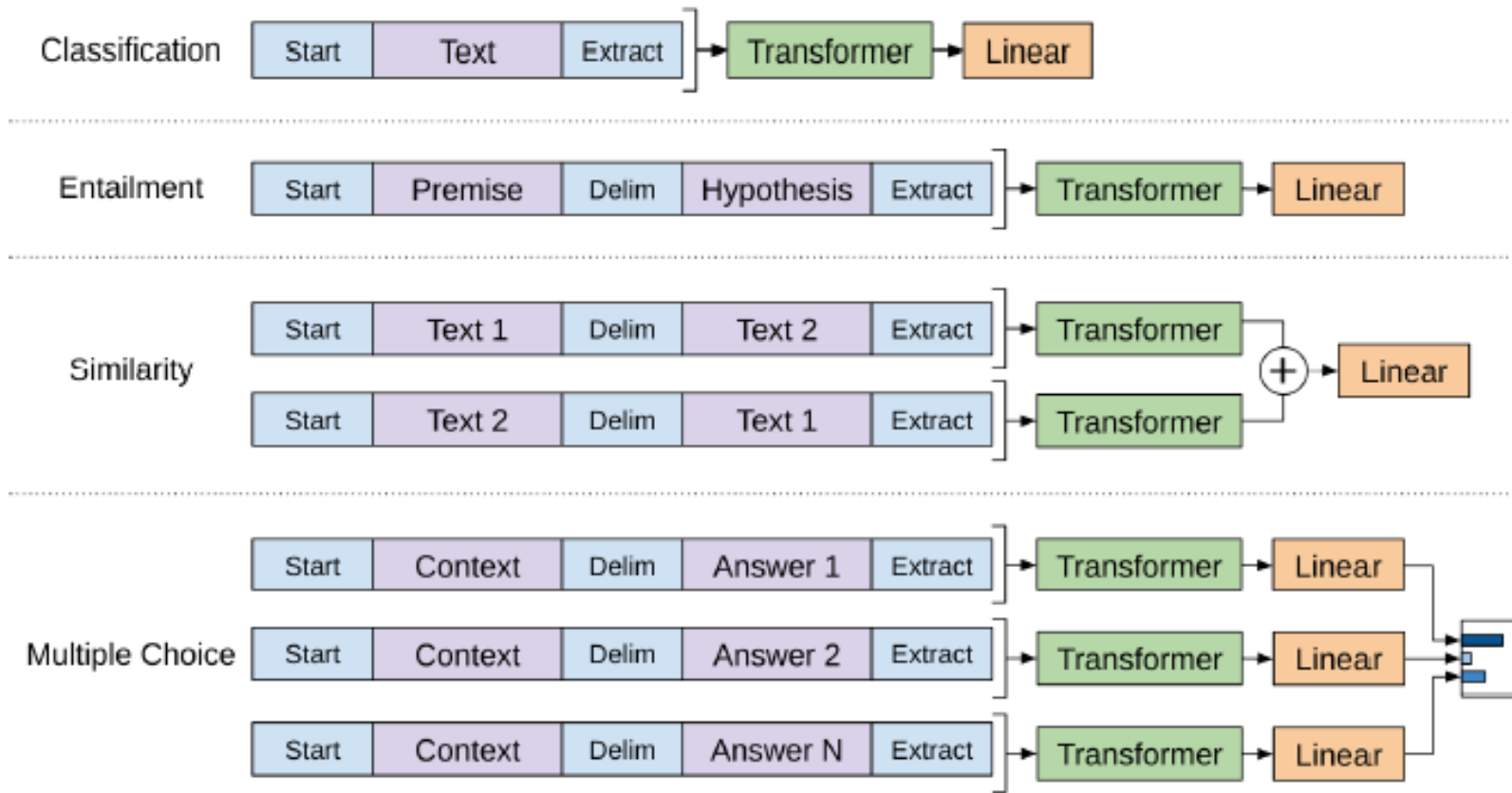
 <https://gluebenchmark.com/>

GLUE: General Language Understanding Evaluation

Corpus	Train	Test	Task	Metrics	Domain
Single-Sentence Tasks					
CoLA	8.5k	1k	acceptability	Matthews corr.	misc.
SST-2	67k	1.8k	sentiment	acc.	movie reviews
Similarity and Paraphrase Tasks					
MRPC	3.7k	1.7k	paraphrase	acc./F1	news
STS-B	7k	1.4k	sentence similarity	Pearson/Spearman corr.	misc.
QQP	364k	391k	paraphrase	acc./F1	social QA questions
Inference Tasks					
MNLI	393k	20k	NLI	matched acc./mismatched acc.	misc.
QNLI	105k	5.4k	QA/NLI	acc.	Wikipedia
RTE	2.5k	3k	NLI	acc.	news, Wikipedia
WNLI	634	146	coreference/NLI	acc.	fiction books

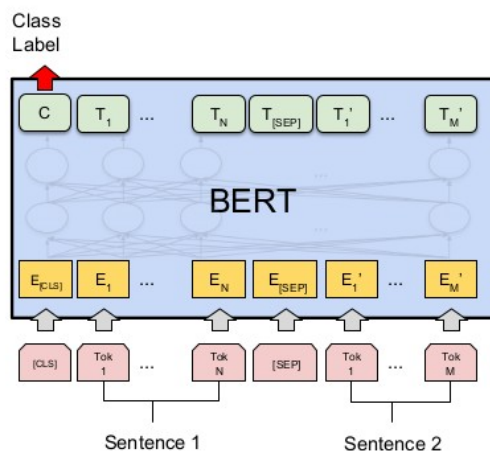
GLUE: General Language Understanding Evaluation

Fine-tuning:

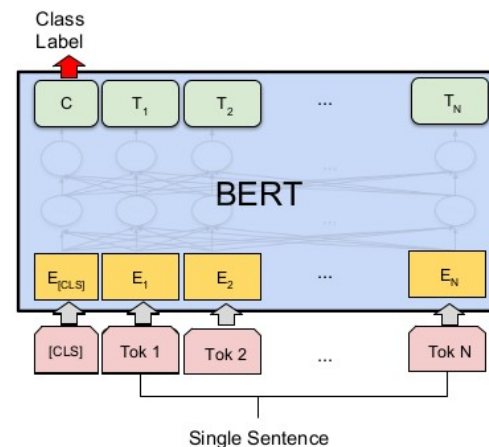


GLUE: General Language Understanding Evaluation

Codificación para GLUE (fine tuning), también se evalúa en SquAD y CoNLL NER:

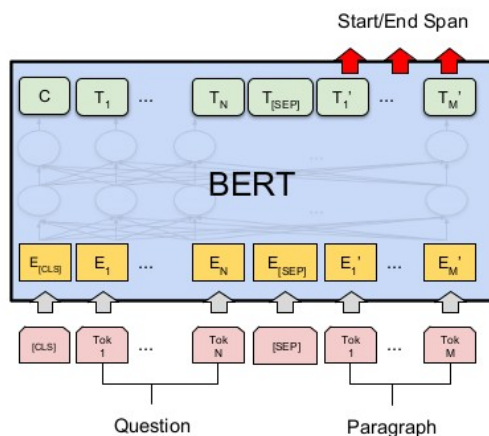


(a) Sentence Pair Classification Tasks:
MNLI, QQP, QNLI, STS-B, MRPC,
RTE, SWAG

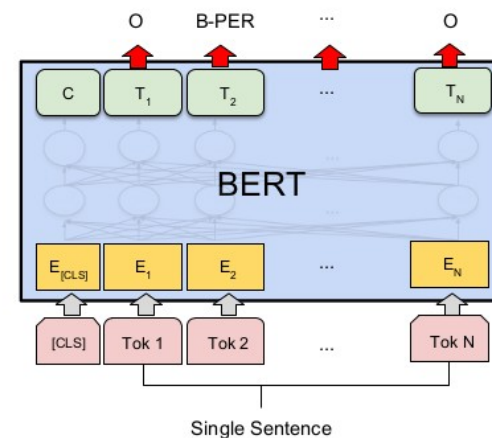


(b) Single Sentence Classification Tasks:
SST-2, CoLA

Predecir el fragmento del párrafo que responde a la pregunta



(c) Question Answering Tasks:
SQuAD v1.1



(d) Single Sentence Tagging Tasks:
CoNLL-2003 NER

GLUE: General Language Understanding Evaluation

Resultados (GLUE):

System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	Average -
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
BERT _{BASE}	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
BERT _{LARGE}	86.7/85.9	72.1	92.7	94.9	60.5	86.5	89.3	70.1	82.1



Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, Samuel R. Bowman: GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding.
BlackboxNLP@EMNLP2018: 353-355

SuperGLUE

BoolQ (BoolQ boolean questions): Los participantes deben responder preguntas de sí o no basadas en un párrafo corto de un artículo de Wikipedia.

CB (CommitmentBank): Se evalúa la capacidad de un modelo para predecir la postura de un hablante hacia una afirmación que ha hecho, basándose en la evidencia proporcionada.

COPA (Choice of Plausible Alternatives): Los modelos tienen que seleccionar la causa o el efecto más plausible de una situación dada.

MultiRC (Multi-Sentence Reading Comprehension): Los modelos deben responder preguntas sobre un párrafo, donde cada pregunta tiene múltiples respuestas correctas.

WiC (Word-in-Context): Esta tarea evalúa si un modelo puede entender el significado de una palabra en dos oraciones diferentes.

WSC (Winograd Schema Challenge): Es una tarea de resolución de la correferencia de pronombres que requiere una comprensión profunda del sentido común y del contexto.

RTE (Recognizing Textual Entailment): Los modelos deben determinar si una oración es verdadera (entails), es falsa (contradicts) o es neutral (neither) en base a una oración previa.

SQuAD v1.1 (Stanford Question Answering Dataset): Es una tarea de respuesta a preguntas basada en la comprensión de un párrafo.



<https://super.gluebenchmark.com/>

SuperGLUE

Corpus	Train	Dev	Test	Task	Metrics	Text Sources
BoolQ	9427	3270	3245	QA	acc.	Google queries, Wikipedia
CB	250	57	250	NLI	acc./F1	various
COPA	400	100	500	QA	acc.	blogs, photography encyclopedia
MultiRC	5100	953	1800	QA	F1 _a /EM	various
ReCoRD	101k	10k	10k	QA	F1/EM	news (CNN, Daily Mail)
RTE	2500	278	300	NLI	acc.	news, Wikipedia
WiC	6000	638	1400	WSD	acc.	WordNet, VerbNet, Wiktionary
WSC	554	104	146	coref.	acc.	fiction books



Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, Samuel R. Bowman: SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. [NeurIPS2019](#): 3261-3275

SuperGLUE

Model	Avg	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC	AX _b	AX _g
Metrics		Acc.	F1/Acc.	Acc.	F1 _a /EM	F1/EM	Acc.	Acc.	Acc.	MCC	GPS Acc.
Most Frequent	47.1	62.3	21.7/48.4	50.0	61.1 / 0.3	33.4/32.5	50.3	50.0	65.1	0.0	100.0/ 50.0
CBoW	44.3	62.1	49.0/71.2	51.6	0.0 / 0.4	14.0/13.6	49.7	53.0	65.1	-0.4	100.0/ 50.0
BERT	69.0	77.4	75.7/83.6	70.6	70.0 / 24.0	72.0/71.3	71.6	69.5	64.3	23.0	97.8 / 51.7
BERT++	71.5	79.0	84.7/90.4	73.8	70.0 / 24.1	72.0/71.3	79.0	69.5	64.3	38.0	99.4 / 51.4
Outside Best	-	80.4	- / -	84.4	70.4*/24.5*	74.8/73.0	82.7	-	-	-	- / -
Human (est.)	89.8	89.0	95.8/98.9	100.0	81.8*/51.9*	91.7/91.3	93.6	80.0	100.0	77.0	99.3 / 99.7

ROUGE (Recall-oriented understudy for Gisting Evaluation)

Métrica para evaluación de resúmenes automáticos.

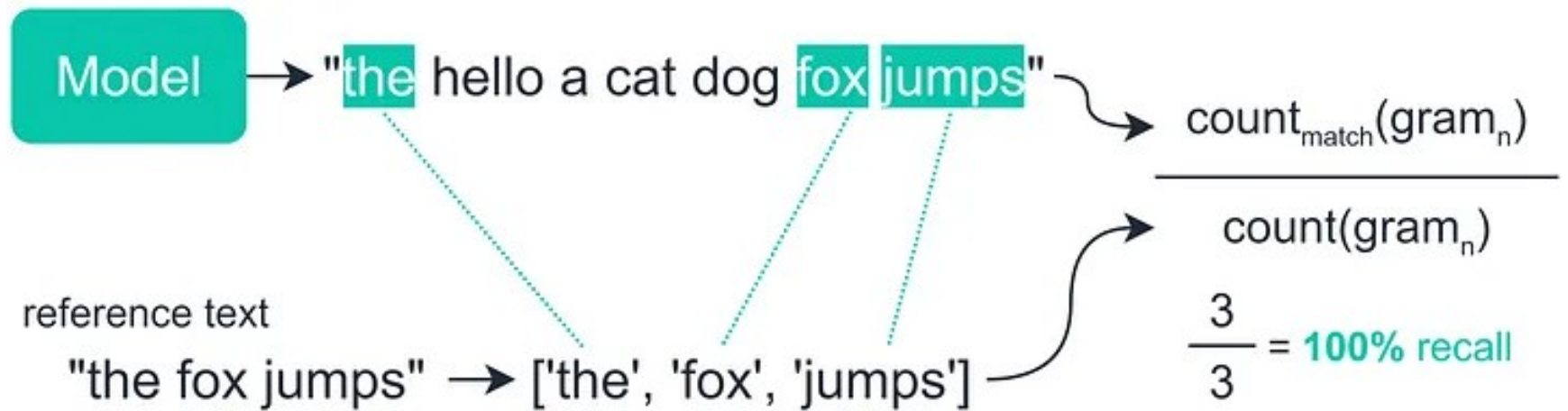
ROUGE-N: Mide la superposición de N-gramas entre el resumen generado y el de referencia. N puede ser cualquier número entero, aunque típicamente se usa 1 (para palabras individuales, es decir, unigramas) y 2 (para pares de palabras, es decir, bigramas). Esta es una medida de precisión y recuperación.

ROUGE-L: Considera la superposición de secuencias más largas de palabras, utilizando la subsecuencia común más larga (LCS, por sus siglas en inglés). Esto ayuda a capturar relaciones más largas entre las palabras y puede ser menos sensible a pequeñas diferencias en la formulación exacta.

ROUGE-S: Mide la superposición de skip-grams.

ROUGE (Recall-oriented understudy for Gisting Evaluation)

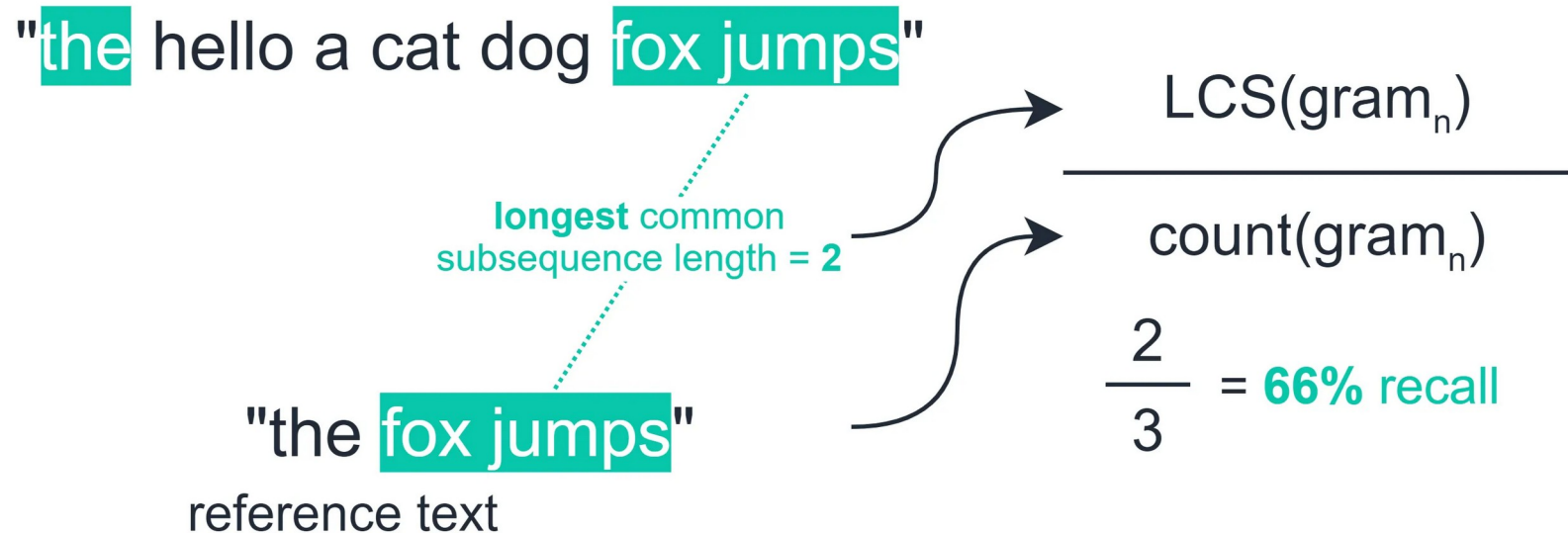
ROUGE-N



ROUGE-1: 1-grams

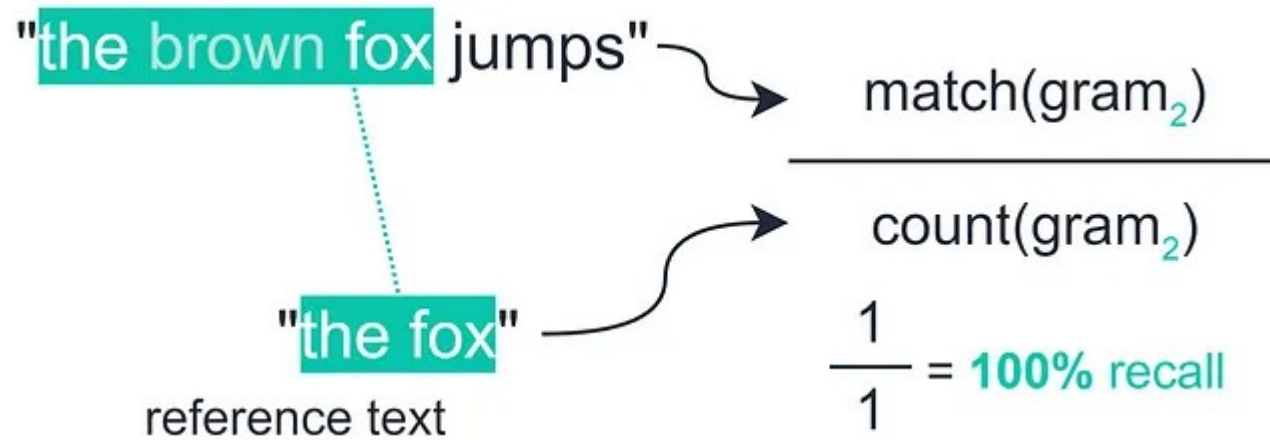
ROUGE (Recall-oriented understudy for Gisting Evaluation)

ROUGE-L



ROUGE (Recall-oriented understudy for Gisting Evaluation)

ROUGE-S



Skip bi-gram

BLEU score (Bilingual Evaluation Understudy)

BLEU (Bilingual Evaluation Understudy) es una métrica que se utiliza para evaluar la calidad de las traducciones producidas por sistemas de traducción automática.

BLEU: % de MT output
n-grams que coinciden
con el texto de
referencia.

Reference (Human) translation:

The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport .

Machine translation:

The American [?] international airport and its the office all receives one calls self the sand Arab rich business [?] and so on electronic mail , which sends out ; The threat will be able after public place and so on the airport to start the biochemistry attack , [?] highly alerts after the maintenance.

MMLU (Massive Multitask Language Understanding)

Diseñado para evaluar grandes modelos, como los LLMs. Está enfocado en preguntas de selección múltiple y cubre más de 57 temáticas incluyendo matemática, historia, leyes, entre otras.

Conceptual Physics	When you drop a ball from rest it accelerates downward at 9.8 m/s^2 . If you instead throw it downward assuming no air resistance its acceleration immediately after leaving your hand is	
	(A) 9.8 m/s^2	✓
	(B) more than 9.8 m/s^2	✗
	(C) less than 9.8 m/s^2	✗
College Mathematics	(D) Cannot say unless the speed of throw is given.	✗
	In the complex z -plane, the set of points satisfying the equation $z^2 = z ^2$ is a	
	(A) pair of points	✗
	(B) circle	✗
	(C) half-line	✗
	(D) line	✓

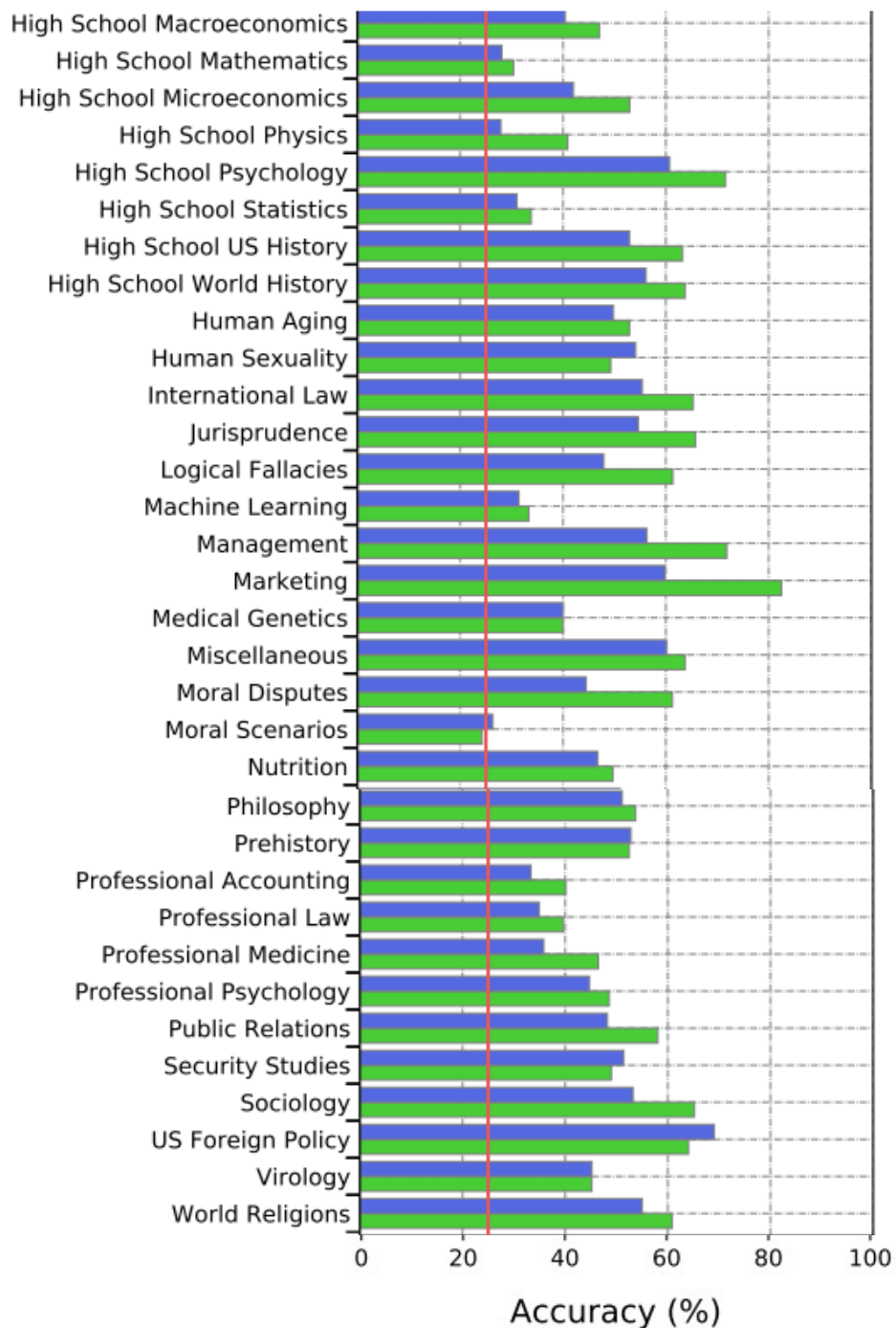
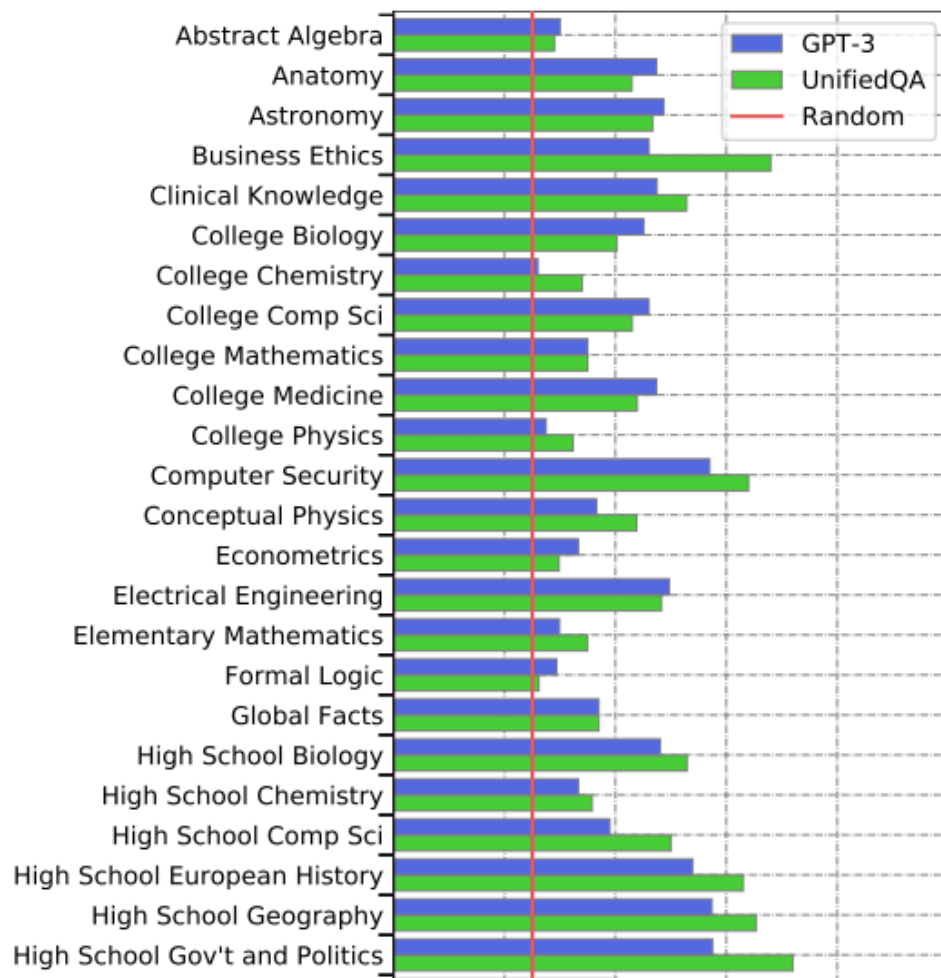
Se calcula un puntaje global entre $[0, 1]$, donde 1 indica desempeño perfecto.

Es un proyecto de evaluación abierta: <https://github.com/hendrycks/test>



Measuring Massive Multitask Language Understanding Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt (ICLR 2021).

MMLU (Massive Multitask Language Understanding)



MMLU (Massive Multitask Language Understanding)

MMLU está incorporado en varias librerías de evaluación, siendo la más conocida DeepEval.

Los shots definen cuantos ejemplos colocaremos en la instrucción. Si es 0, se conoce como evaluación zero shot. En otro caso, es few shot.

```
from deepeval.benchmarks import MMLU
from deepeval.benchmarks.tasks import MMLUTask

# Define benchmark with specific tasks and shots
benchmark = MMLU(
    tasks=[MMLUTask.HIGH_SCHOOL_COMPUTER_SCIENCE, MMLUTask.ASTRONOMY],
    n_shots=3
)

# Replace 'mistral_7b' with your own custom model
benchmark.evaluate(model=mistral_7b)
print(benchmark.overall_score)
```

Más información de cómo usar MMLU en DeepEval:

<https://docs.confident-ai.com/docs/benchmarks-mmlu>

MT Bench (Multiturn Benchmark)

MT Bench busca evaluar las capacidades conversacionales de un LLM. Para esto, incorpora ejemplos con varias interacciones (multiturn interactions).

Category	Sample Questions	
Writing	1st Turn	Compose an engaging travel blog post about a recent trip to Hawaii, highlighting cultural experiences and must-see attractions.
	2nd Turn	Rewrite your previous response. Start every sentence with the letter A.
Math	1st Turn	Given that $f(x) = 4x^3 - 9x - 14$, find the value of $f(2)$.
	2nd Turn	Find x such that $f(x) = 0$.
Knowledge	1st Turn	Provide insights into the correlation between economic indicators such as GDP, inflation, and unemployment rates. Explain how fiscal and monetary policies ...
	2nd Turn	Now, explain them again like I'm five.



Lianmin Zheng Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, Ion Stoica: Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. NeurIPS 2023.

MT Bench (Multiturn Benchmark)

Para construir MT Bench, se propusieron técnicas que usan al mismo LLM como un anotador. A esto se le llama **LLM as a judge**. Por ejemplo, se usó este **prompt para comparación de respuestas (2 LLMs)**.

[System]

Please act as an impartial judge and evaluate the quality of the responses provided by two AI assistants to the user question displayed below. You should choose the assistant that follows the user's instructions and answers the user's question better. Your evaluation should consider factors such as the helpfulness, relevance, accuracy, depth, creativity, and level of detail of their responses. Begin your evaluation by comparing the two responses and provide a short explanation. Avoid any position biases and ensure that the order in which the responses were presented does not influence your decision. Do not allow the length of the responses to influence your evaluation. Do not favor certain names of the assistants. Be as objective as possible. After providing your explanation, output your final verdict by strictly following this format: "[[A]]" if assistant A is better, "[[B]]" if assistant B is better, and "[[C]]" for a tie.

[User Question]

{question}

[The Start of Assistant A's Answer]

{answer_a}

[The End of Assistant A's Answer]

[The Start of Assistant B's Answer]

{answer_b}

[The End of Assistant B's Answer]

MT Bench (Multiturn Benchmark)

Se le llama **LLM as a judge** ya que se usa un tercer LLM (uno fuerte como GPT4) para evaluar las dos alternativas.

Question: If the FED buys bonds in the secondary market

(A) the money supply will increase. (B) the money supply will decrease. (C) the money supply will not be affected. (D) the discount rate would be affected. Output the answer only.

Assistant A:

The Federal Reserve buys bonds in the secondary market to increase the money supply.

Assistant B:

(A) the money supply will increase.

Follow-up Question: How does it affect my daily life? Give 3 examples.

Assistant A:

The Federal Reserve buys bonds in the secondary market to increase the money supply. This affects your daily life by:

1. Increasing the money supply.
2. Increasing the money supply and lowering the discount rate.
3. Increasing the money supply and lowering the discount rate and increasing the money supply.

Assistant B:

When the Federal Reserve buys bonds in the secondary market, it can have several effects on the economy that can impact your daily life. Here are three examples:

1. Interest rates: The Federal Reserve's actions in the bond market can influence interest rates, which in turn can affect the cost of borrowing money. For example, if the Federal Reserve buys bonds and drives up demand for them, interest rates on things like mortgages, car loans, and credit cards may go up, making it more expensive to borrow money.
2. Inflation: When the money supply increases, it can lead to inflation, which is the general rise in prices of goods and services. This can impact your daily life by making it more expensive to buy things like groceries, gas, and other necessities.
3. Employment: The Federal Reserve's actions in the bond market can also have an impact on employment. For example, if the Federal Reserve's actions lead to economic growth and increased demand for goods and services, it can create more job opportunities and boost employment.

GPT-4 Judgment:

Assistant A provided an incorrect response to the user's question about how the Federal Reserve buying bonds in the secondary market affects daily life. The answer given is repetitive and lacks clear examples of how the action impacts daily life.

On the other hand, **Assistant B** provided a relevant and accurate response to the user's question about the Federal Reserve buying bonds. The answer includes three clear examples of how the action impacts daily life, such as interest rates, inflation, and employment.

Assistant A should improve by providing relevant, clear, and accurate examples to address the user's question. **Assistant B's** response is more helpful, relevant, and detailed.

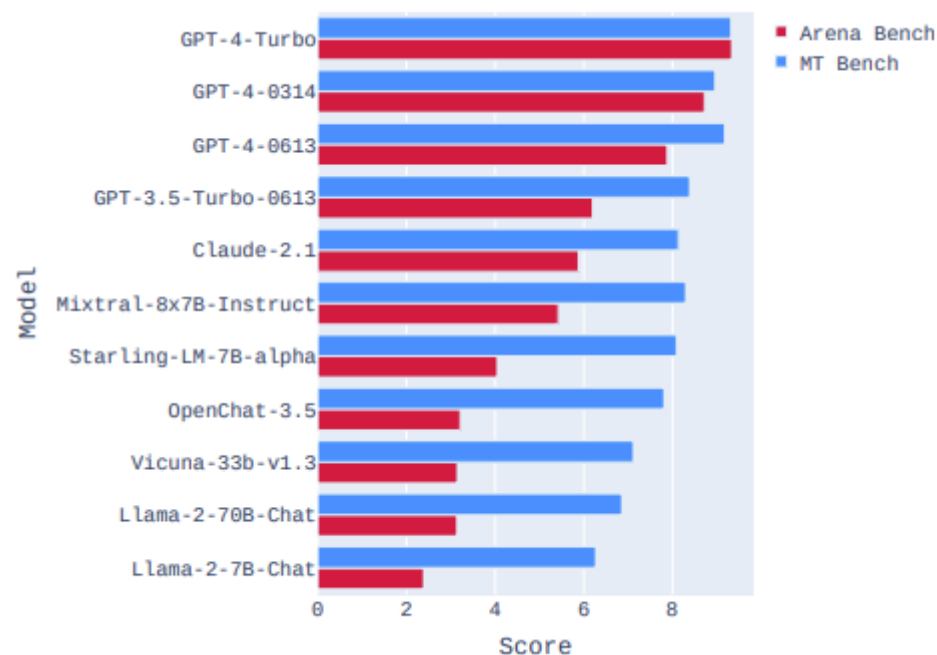
LlaMa 13B

Vicuna 13B

ChatBot Arena

En esta misma línea, se pueden comparar dos alternativas en base a votaciones de humanos (masivas, es decir, crowdsourcing). Los usuarios son libres de ingresar los prompts que deseen. Luego, ChatBotArena controla los prompts para resguardar diversidad temática usando BERT Topic. Luego pasan las evaluaciones de pares a ranking.

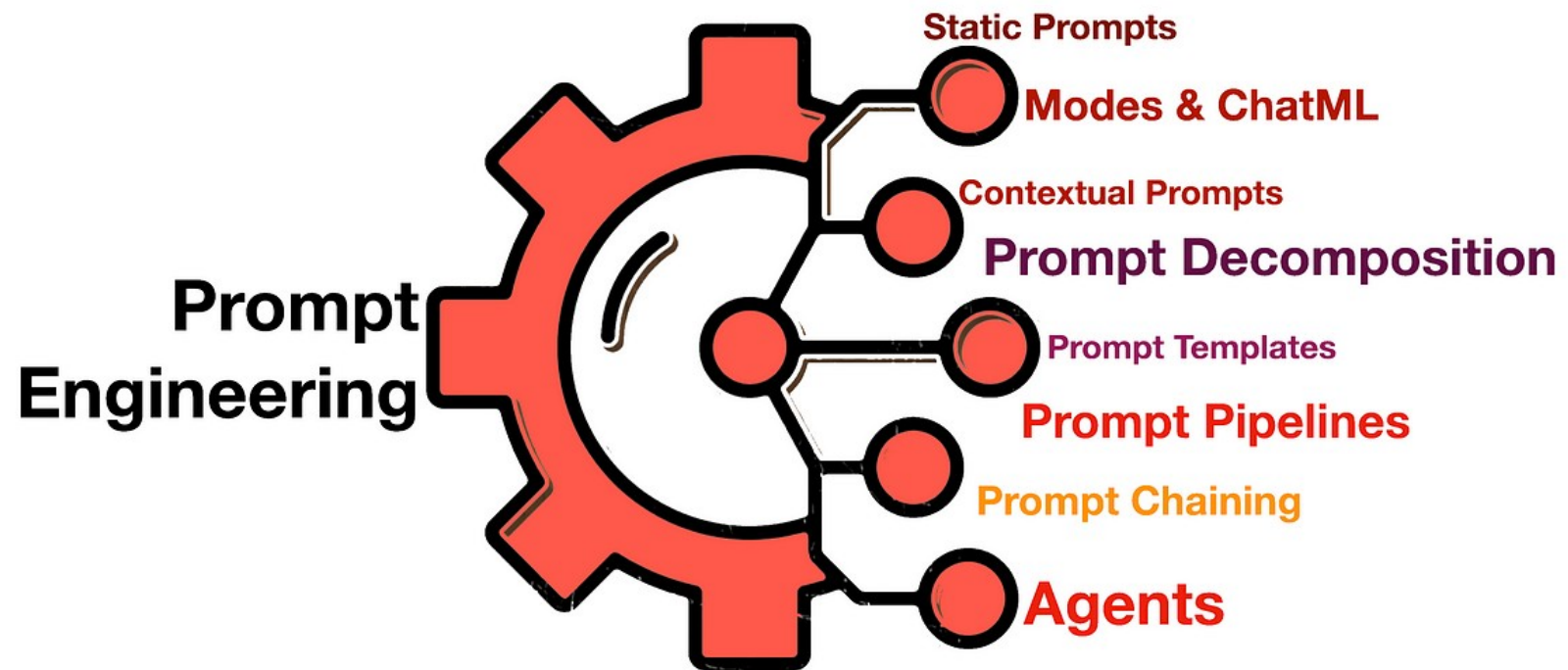
Ver Chatbot Arena leaderboard en: <https://chat.lmsys.org/>



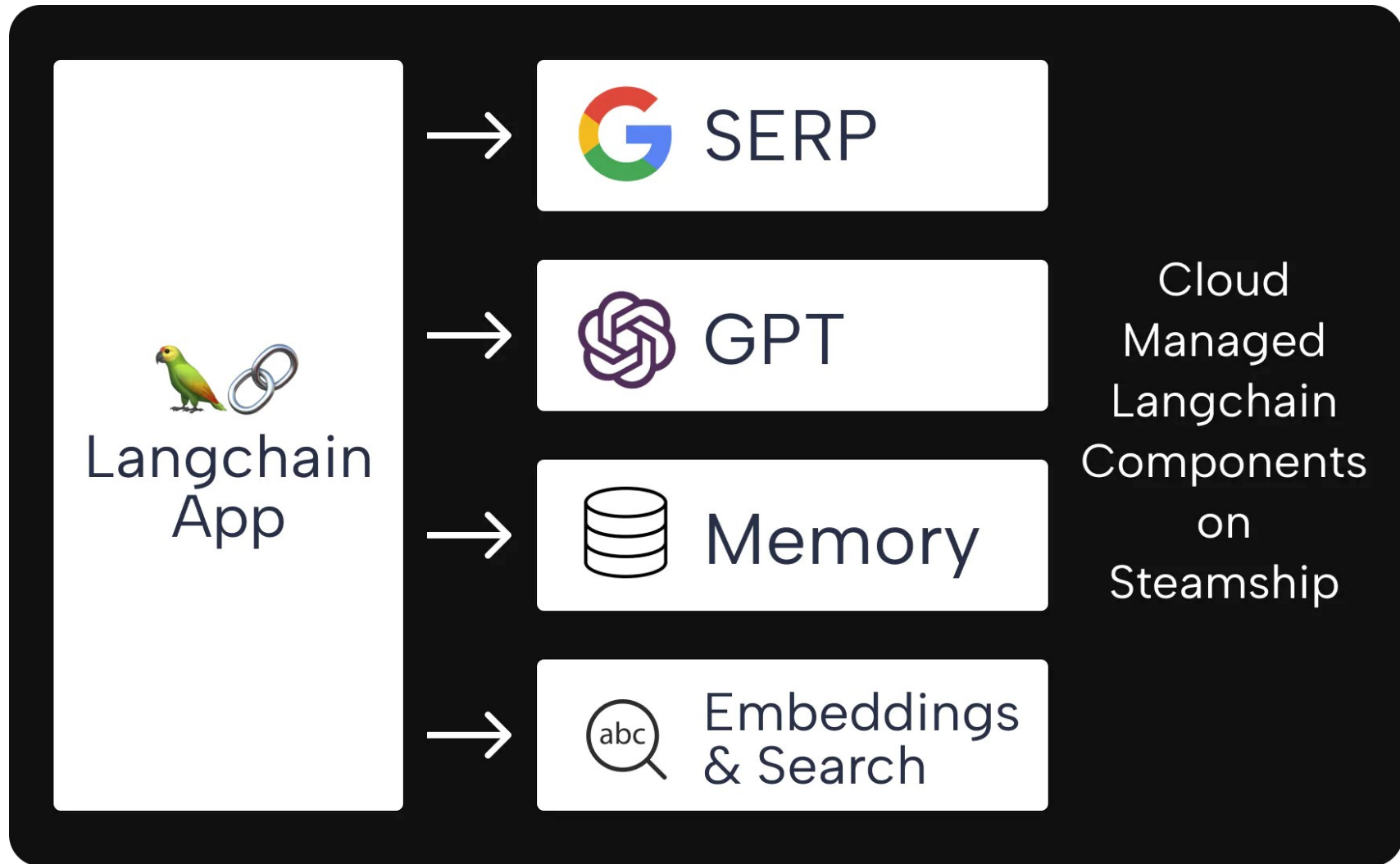
Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael I. Jordan, Joseph E. Gonzalez, Ion Stoica: Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. [CoRR abs/2403.04132](https://arxiv.org/abs/2403.04132) (2024)

- LANGCHAIN -

LangChain



LangChain



- OPEN AI (API) anexo -

API (parámetros importantes)

- **model**: el nombre del modelo que se quiere usar, por ejemplo gpt-3.5-turbo o gpt-4.
- **messages**: una lista de mensajes previos en el contexto de la conversación, cada uno con un rol (user o assistant) y un contenido (el texto del mensaje).
- **temperature**: un valor entre 0 y 1 que controla la aleatoriedad de las respuestas generadas por el modelo. Un valor más alto significa más variación y creatividad, pero también más riesgo de errores o incoherencias.
- **stop**: una lista de cadenas de texto que indican cuándo el modelo debe terminar de generar una respuesta. Por defecto, el modelo usa los caracteres de nueva línea (\n) como señal de parada.
- **max_tokens**: el número máximo de tokens que el modelo puede generar en una respuesta. Un token es una unidad básica de texto, como una palabra o un signo de puntuación. Por defecto, el modelo genera hasta 256 tokens o hasta encontrar una señal de parada.

API (parámetros importantes)

- **presence_penalty**: un valor entre 0 y 1 que penaliza la repetición de palabras o frases en la respuesta generada por el modelo. Un valor más alto significa más diversidad y novedad, pero también más riesgo de perder coherencia o relevancia.
- **frequency_penalty**: un valor entre 0 y 1 que penaliza el uso de palabras o frases comunes en la respuesta generada por el modelo. Un valor más alto significa más originalidad y rareza, pero también más riesgo de generar textos incomprensibles o absurdos.
- **logprobs**: un número entero entre 0 y 100 que indica cuántas de las palabras o tokens más probables se deben devolver junto con la respuesta generada por el modelo. Esto puede ser útil para analizar la confianza del modelo o para explorar otras posibles respuestas.
- **return_full_text**: un valor booleano (true o false) que indica si se debe devolver el texto completo generado por el modelo (incluyendo el prompt) o solo la parte nueva que se ha añadido al prompt. Por defecto, este valor es true.
- **top**: El valor de top puede ser un número entre 0 y 1, o la palabra "all". Si se usa un número, el modelo solo considerará los tokens que tengan una probabilidad acumulada menor o igual a ese valor. Por ejemplo, si se usa top=0.5, el modelo solo usará los tokens más probables que sumen hasta el 50% de la probabilidad total. Esto hace que la respuesta sea más coherente y precisa, pero también más predecible y menos creativa. Si se usa la palabra "all", el modelo considerará todos los tokens posibles, sin importar su probabilidad. Esto hace que la respuesta sea más variada y sorprendente, pero también más propensa a errores y contradicciones.