

¿De dónde aprenden los modelos de lenguaje (LLMs)?



 **DOCENTE**
GUIDO IVETTA



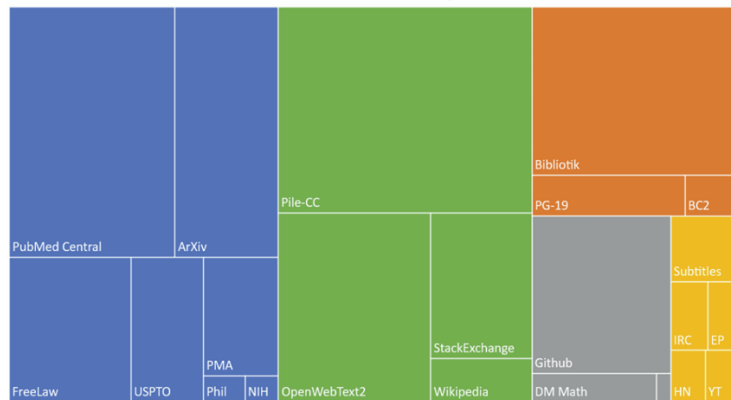
 **Vía Libre**

¿De dónde aprenden los LLMs?

Where does this data come from?

Composition of the Pile by Category

■ Academic ■ Internet ■ Prose ■ Dialogue ■ Misc



Model	Training Data
BERT	BookCorpus, English Wikipedia
GPT-1	BookCorpus
GPT-3	CommonCrawl, WebText, English Wikipedia, and 2 book databases ("Books 1" and "Books 2")
GPT-3.5+	Undisclosed

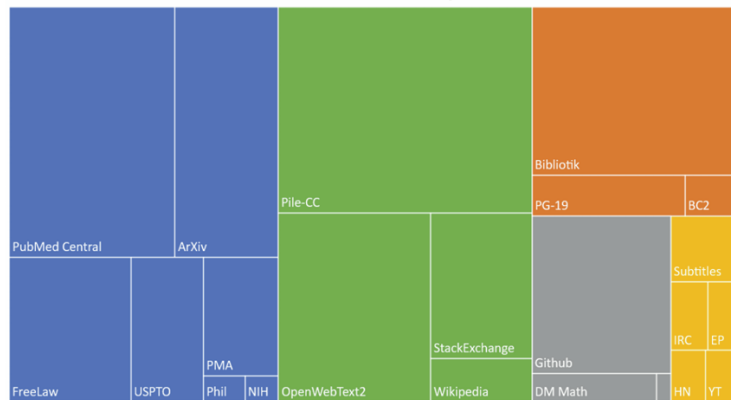
[Slide from CS224n]

¿De dónde aprenden los LLMs?

Where does this data come from?

Composition of the Pile by Category

■ Academic ■ Internet ■ Prose ■ Dialogue ■ Misc



Model	Training Data
BERT	BookCorpus, English Wikipedia
GPT-1	BookCorpus
GPT-3	CommonCrawl, WebText, English Wikipedia, and 2 book databases ("Books 1" and "Books 2")
GPT-3.5+	Undisclosed

[Slide from CS224n]

¿De dónde aprenden los **LLMs**?



¿Qué diferencias hay?

Característica	Código abierto	Pesos abiertos	Código cerrado
Definición	Los detalles completos del modelo (código, pesos, datos) son públicos.	Los pesos del modelo entrenado son públicos.	Los detalles del modelo son privados, propiedad de una empresa.
Costo	Gratis (requiere cómputo)	Gratis (requiere cómputo)	Generalmente requiere pago (licencia o suscripción).
Comunidad	Fuerte participación de la comunidad; las personas pueden contribuir.	Poco rol de la comunidad en el desarrollo.	Sin participación comunitaria; la empresa controla el desarrollo.
Transparencia	Podemos analizar ambos datos y comportamiento interno del modelo	Sólo podemos analizar comportamiento interno del modelo	El análisis se limita al uso externo

¿Qué diferencias hay?

Característica	Código abierto	Pesos abiertos	Código cerrado
	 /DCLM-7B	 Qwen	 Claude
			 ChatGPT
			 Gemini