SCIÉNDO INGENIUM
ISSN 3084-7788 (En línea) Scién. inge. 21(2): 33-40, (2025)

Fine-tuning de un Modelo de Lenguaje Largo para la clasificación de Curriculums Vitae

Fine-tuning a Long Language Model for Curriculum Vitae Classification

Juan Diego Salcedo-Salazar *

Programa de Maestría en Ingeniería de Sistemas e Informática. Facultad de Ingeniería de Sistemas e Informática, Universidad Nacional Mayor de San Marcos, Av. Carlos Germán Amezaga #375 - Cercado de Lima - Ciudad Universitaria, Lima Perú.

* Autor correspondiente: juan.salcedos@unmsm.edu.pe (J. Salcedo-Salazar) DOI: 10.17268/scien.inge.2025.02.02

RESUMEN

Este trabajo tuvo por principal objetivo clasificar curriculums vitae respecto al área de profesión, tarea importante en la gestión de recursos humanos y captación de personal. Esta investigación busca explorar las capacidades de clasificación de los Modelos de Lenguaje Largo (MLL) realizando un análisis comparativo versus métodos tradicionales de Machine Learning. Para lograr tal objetivo se empleó la técnica de fine-tuning al Modelo de Lenguaje Largo pre-entrenado por Google en el idioma inglés llamado BERT BASE UNCASED empleando un dataset de más de 3 mil currículums vitae de 25 áreas de profesión y 3 épocas de entrenamiento contra los modelos tradicionales Random Forest, SVM, Logistic Regression y Naive Bayes Multinomial. La metodología está compuesta por 7 etapas esenciales para adaptar un modelo pre-entrenado en una tarea específica asegurando su óptimo performance. La investigación proporciona un análisis comparativo enfocándose en las métricas Exactitud, F1-score, Precisión y Recall. Los resultados más importantes obtenidos fueron 83,0% de Exactitud y 82,3% de Precisión para el modelo base de Google y 82,8% de F1-score y 86,2% de Recall para Naive Bayes Multinomial revelando que el modelo base de Google tiene un buen desempeño prediciendo la clasificación de currículum vitae mientras que Naive Bayes Multinomial es mejor para detectar la mayoría de casos positivos. Esta investigación no solo contribuye mostrando el desempeño de los MLL para la tarea de clasificación en contraste con sus pares de Machine Learning tradicional, sino que adicionalmente ofrece un enfoque innovador para las prácticas de gestión de recursos humanos y captación de personal.

Palabras clave: Procesamiento del Lenguaje Natural; Modelo de Lenguaje Largo; Fine-tuning; Clasificación de textos; Curriculum Vitae.

ABSTRACT

The main objective of this work was to classify resumes according to their professional area, an important task in human resources management and personnel recruitment. This research seeks to explore the classification capabilities of Large Language Models (LLM) by performing a comparative analysis versus traditional Machine Learning methods. To achieve this objective, a fine-tuning technique was used on the Long Language Model pre-trained by Google in English called BERT BASE UNCASED using a dataset of more than 3,000 resumes from 25 professional areas and 3 training epochs against traditional Random Forest, SVM, Logistic Regression, and Naive Bayes Multinomial models. The methodology consists of 7 essential stages to adapt a pre-trained model to a specific task, ensuring optimal performance. The research provides a comparative analysis focusing on the metrics Accuracy, F1-score, Precision, and Recall. The most significant results obtained were 83,0% Accuracy and 82.3% Precision for the base Google model, and 82.8% F1-score and 86,2% Recall for the Naive Bayes Multinomial model, revealing that the base Google model performs well in predicting resume classification, while Naive Bayes Multinomial is better at detecting the majority of positive cases. This research not only contributes by showing the performance of MLLs for the classification task in contrast to their traditional Machine Learning peers, but also offers an innovative approach to human resource management and staff recruitment practices.

Keywords: Natural Language Processing; Large Language Model; Fine-tuning, Text Classification; Resume.

Fecha de envío: 27-12-2024 Fecha de aceptación: 20-06-2025 Fecha de publicación: 28-07-2025

1. INTRODUCCIÓN

Para el 2025 la International Data Corporation (IDC) pronosticó que el 80% de la data de una empresa será no estructurada, según el white paper patrocinado por Seagate (*Seagate-WP-DataAge2025-March-2017.pdf*, s. f.). Parte de la información no estructurada de la que nos habla IDC está contenida en los curriculum vitae (CV) de sus empleados y candidatos a un puesto laboral, documento que contiene información de experiencias laborales y habilidades profesionales.

Grandes plataformas de empleo en el Perú como Computrabajo indica en su sitio web tener registrado más de 7.5 millones de CVs, Laborum más de 200 mil currículums en su base de datos y el portal de trabajo del Ministerio de Trabajo reporta un poco más de 1300 ofertas de empleo tanto para el sector privado como estatal.

Para cada puesto laboral los jefes de contratación tienen la ardua tarea de analizar los CVs de los candidatos para extraer la información relevante, labor que puede tomar mucho tiempo y muchas veces los candidatos no cumplen con los requisitos mínimos solicitados. Esta pérdida de tiempo y recurso humano abre la puerta a la automatización debido a que no existe una regla o estructura básica para la redacción de un CV. En un intento por facilitar la construcción de CVs se han desarrollado herramientas basadas en Inteligencia Artificial para la construcción del mismo (mitsmrmex, 2024)cada una con sus propias estructuras y diferente a otras.

La clasificación de texto es una técnica que se ha aplicado ampliamente en la recuperación de información, la minería de datos y el procesamiento del lenguaje natural (PLN) (Wu & Wan, 2025). En los inicios se utilizaron métodos basados en reglas y expresiones regulares, luego cambió hacia métodos de machine learning avanzando a redes neuronales profundas (Yu et al., 2023), y recientemente esta tarea es abordada por los MLL que son Modelos de Lenguaje cuyo objetivo es modelar la probabilidad generativa de una secuencia de palabras con el fin de predecir la siguiente (Zhao et al., 2023). Estos MLL modernos basados en redes neuronales utilizan arquitecturas de modelos muy grandes y se entrenan con conjuntos de datos masivos, son fundamentales para muchas tareas de procesamiento de lenguaje natural (PLN) (Carlini et al., 2021).

El objetivo del presente artículo fue clasificar CVs respecto al área de profesión, para lo cual se empleó la técnica de Fine-tuning de tipo Supervisado al modelo de lenguaje largo BERT BASE UNCASED (Devlin et al., 2019) con un datasets de 3573 registros únicos a través de 3 épocas de entrenamiento.

2. METODOLOGÍA

La decisión de utilizar el método de fine-tuning sobre prompt engineering está basada en la tesis de (López, 2024) que tiene por objetivo evaluar diferentes MLL para la clasificación de emociones en textos cortos, donde se utilizaron 5 modelos basados en BERT, 4 basados en claude, 3 en llama, 2 gemini y 2 gpt. La principal métrica de evaluación utilizada fue F1-score, donde se obtuvo 0,76729 de puntaje para el método prompt engineering y 0,81505 de puntaje para el método fine-tuning.

Se seleccionó la metodología basada en la propuesta por (Parthasarathy et al., 2024) la cual está divida en 7 etapas (Figura 1), esenciales para adaptar un modelo pre-entrenado en una tarea específica asegurando su óptimo performance, estas etapas son 1 Preparación del Dataset, 2 Inicialización del modelo, 3 Configuración del entorno de entrenamiento, 4 Fine-tuning parcial o total, 5 Evaluación y validación, 6 Despliegue, 7 Monitoreo y mantenimiento.

Es importante mencionar que las etapas 6 Despliegue y 7 Monitoreo y mantenimiento que implican una puesta en producción, no forma parte del alcance presente trabajo de investigación.

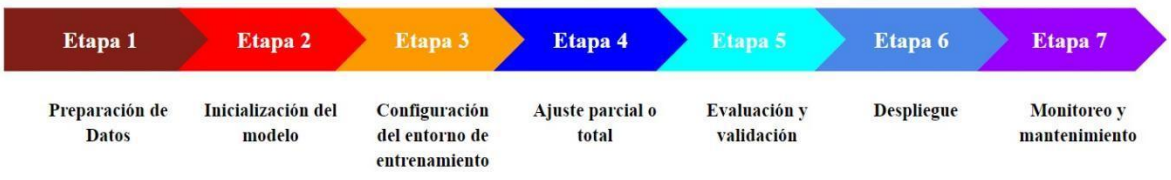

Figura 1. Proceso de Fine-tuning de siete etapas para Modelos de Lenguaje Largo.

2.1. Etapa 1: Preparación del Dataset

2.1.1. Recolección de datos

Las plataformas de Machine Learning Kaggle y Hugging Face fueron utilizadas como fuentes de datos Tabla 1.

Tabla 1. Fuente de datos

Plataforma	Nro. Registros
Kagle Dataset	2484
Hugging Face Dataset	1219
Total	3703

2.1.2. Limpieza de datos

El principal desafío presentado en esta etapa fue la unificación de áreas de profesión Tabla 2, 11 clases con nombres distintos. Aunque estaban relacionadas al mismo tema, en un dataset estaba nombrado como un adjetivo mientras que en el otro dataset eran sustantivos, las palabras contenían guiones y otros se referían al título del profesional en lugar del área de profesión. Se unificó seleccionando la palabra o palabras más representativas del área de profesión. Obteniendo un total de 25 áreas de profesión (Figura 2).

Tabla 2. Diferencias entre nombre de profesiones de datasets y su unificación

Kagle	Hugging Face	Dataset Final
AGRICULTURE	AGRICULTURAL	AGRICULTURAL
BUSINESS-DEVELOPMENT	BUSINESS DEVELOPMENT	BUSINESS DEVELOPMENT
CHEF	FOOD & BEVERAGES	FOOD & BEVERAGES
CONSTRUCTION	BUILDING & CONSTRUCTION	BUILDING & CONSTRUCTION
DESIGNER	DESIGNING	DESIGNING
DIGITAL-MEDIA	DIGITAL MEDIA	DIGITAL MEDIA
FITNESS	HEALTH & FITNESS	HEALTH & FITNESS
HEALTHCARE	HEALTH & FITNESS	HEALTH & FITNESS
TEACHER	EDUCATION	EDUCATION
INFORMATION-TECHNOLOGY	INFORMATION TECHNOLOGY	INFORMATION TECHNOLOGY
PUBLIC-RELATIONS	PUBLIC RELATIONS	PUBLIC RELATIONS

Nota. Para la unificación se respetaron los nombres más generales y se eliminaron los guiones.

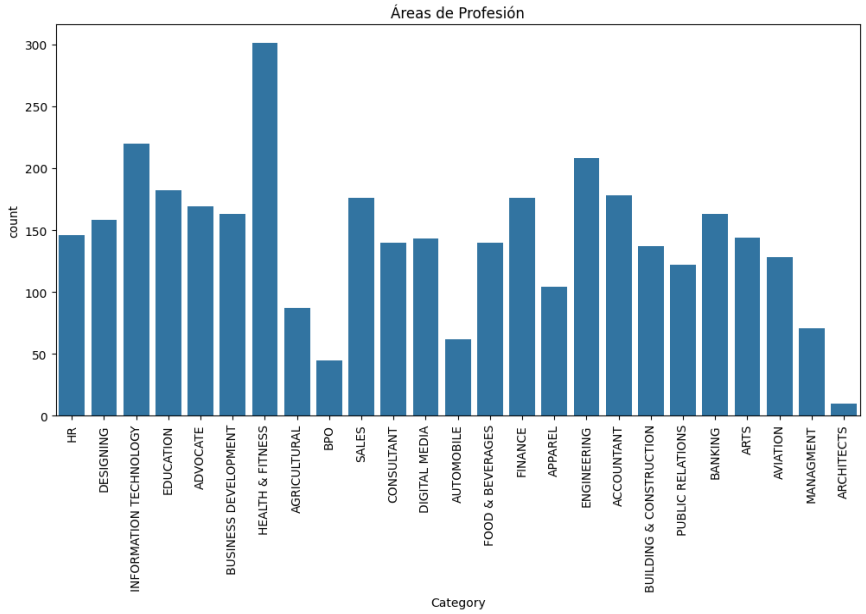

Figura 2. Áreas de profesión

Después del proceso de eliminación de registros duplicados se obtuvo un total de **3573** de registros.

2.1.3. División de datos

Se utilizó la técnica de división de datos llamada train-validation-test, y se seleccionó 70% para entrenamiento, 21% para validación y 9% para test.

2.2. Etapa 2: Inicialización del modelo

2.2.1. Elección del modelo pre-entrenado

Bidirectional **E**ncoder **R**epresentations from **T**ransformers (BERT) es conceptualmente simple, se le puede aplicar fine-tuning con solo una capa adicional de salida para crear modelos de última generación para resolver tareas como respuestas a preguntas e inferencias del lenguaje sin grandes modificaciones (Devlin et al., 2019).

El estudio realizado por (Oliveira et al., 2024) en Brasil indica que los modelos basados en BERT logran un mejor rendimiento en las pruebas de clasificación de textos.

El modelo supervisado réplica de BERT llamado RoBERTa (Liu et al., 2019), ha demostrado superior performance en tareas de clasificación sobre otros modelos de lenguaje largo, tanto de tipo open-source como modelos cerrados (Yu et al., 2023).

Los antecedentes de buen performance, bajo coste computacional y sencilla aplicación de fine-tuning hacen que la elección de un modelo basado en BERT sea una decisión acertada para abordar tareas de clasificación.

El modelo basado en BERT seleccionado fue **BERT BASE UNCASED** pre-entrenado en el idioma inglés, acorde al dataset preparado en el mismo idioma y la característica UNCASED nos garantiza que no hará distinción entre mayúsculas y minúsculas, esta característica es congruente a la diversidad de CVs seleccionados.

2.2.2. Carga de pesos de modelo pre-entrenado

La carga del modelo BERT BASE UNCASED se realizó utilizando la librería *transformer* importando la clase *BertForSequenceClassification* y ejecutando el método *from_pretrained*, donde se cargó el nombre del

modelo '*bert-base-uncased*' y el número de labels igual a 25, que representan las áreas de profesión como objetivo de clasificación.

2.3. Etapa 3: Configuración del entorno de entrenamiento

2.3.1. Entorno de entrenamiento

El entorno seleccionado para el entrenamiento fue la plataforma Codelab de Google, dentro de su capa gratis de uso con las siguientes características:

Tabla 3. Resumen de las características del entorno de entrenamiento

Característica	Detalle
Tipo de entorno	Python 3
Acelerador de Hardware	T4 GPU
RAM del sistema	14 GB
RAM de GPU	15 GB
Disco	145 GB

2.3.2. Definición de HiperParámetros

Los hiperparámetros utilizados para el entrenamiento del modelo fueron los siguientes:

Tabla 4. Resumen de hiperparámetros utilizados para el entrenamiento

Hiperparámetro	Valores
Estrategia de evaluación	Épocas
Número de épocas	3
Learning rate	2e-5
Pasos de evaluación	500
Dataset de entrenamiento	SÍ
Dataset de evaluación	SÍ
Métricas de evaluación	accuracy, f1-score, precision, recall

2.4. Etapa 4: Fine-tuning parcial o total

Se utilizó un fine-tuning total de todos los parámetros del modelo, esto garantiza una adaptación total a la nueva tarea (Parthasarathy et al., 2024).

2.5. Etapa 5: Evaluación y validación

2.5.1. Métricas de evaluación

Tabla 5. Métricas de evaluación del modelo durante el entrenamiento

Época	Accuracy	F1-Score	Precision	Recall
1	0,81	0,78	0,80	0,81
2	0,82	0,80	0,80	0,82
3	0,83	0,81	0,81	0,83

Tabla 6. Métricas de evaluación posteriores al entrenamiento

Métrica	Valor
Épocas	3
Accuracy	0,830
F1-Score	0,814
Precision	0,823
Recall	0,830

Nota. Se utiliza el dataset de test.

3. RESULTADOS Y DISCUSIÓN

3.1. Métricas de evaluación con dataset de test

3.1.1. Accuracy (eval_accuracy): 0,830

La exactitud es una métrica clave en clasificación, que indica qué porcentaje de las predicciones realizadas por el modelo fueron correctas. Un valor de 83.04% muestra que más de 8 de cada 10 predicciones son correctas. Este valor sugiere que el modelo ha aprendido a clasificar con un alto grado de precisión.

3.1.2. F1-Score (eval_f1): 0,814

F1-score es una medida de la media armónica entre la precisión y el recall, y es especialmente útil cuando hay un desequilibrio entre clases. En modelos de clasificación, un F1-score cercano a 1 es ideal. un valor de 0.8142 es una buena señal de que el modelo mantiene un balance adecuado entre precisión y recall.

3.1.3. Precisión (eval_precision): 0,823

La precisión indica cuántas de las instancias clasificadas como positivas son realmente positivas. Un valor de 82,37% sugiere que el modelo está realizando un buen trabajo al identificar correctamente las clases positivas, pero aún podría reducir algunos falsos positivos.

3.1.4. Recall (eval_recall): 0,830

El recall es una métrica que indica qué porcentaje de las instancias realmente positivas fueron correctamente identificadas por el modelo. En este caso, un recall del 83,04% es bastante bueno y está en línea con la precisión, lo que sugiere que el modelo está capturando correctamente las instancias positivas sin dejar demasiadas sin clasificar.

Se presenta una comparativa en la Tabla 6 entre las métricas obtenidas, exactitud, F1-score, precisión y recall, por modelo entrenado en esta investigación y modelos tradicionales de machine learning como

Random Forest, SVM, Logistic Regression y Naive Bayes Multinomial con valores presentados en la investigación de (Heakl et al., 2024).

Tabla 7. Resultados comparativos

Modelo	Exactitud(%)	F1-score(%)	Precisión(%)	Recall(%)
Random Forest	78,5	78,6	75,2	82,1
SVM	79,2	80,0	76,5	83,5
Logistic Regression	79,8	80,6	77,1	84,2
Naive Bayes Multinomial	81,6	**82,8**	79,5	**86,2**
Propuesta	**83,0**	81,4	**82,3**	83,0

Podemos identificar que nuestro modelo presenta valores remarcables para Exactitud y precisión con 83,0% y 82,3% respectivamente, mientras que el modelo Naive Bayes Multinomial presenta las mejores métricas para F1-score y Recall con 82,8% y 86,2% respectivamente.

4. CONCLUSIONES

El modelo pre entrenado utilizado como base es BERT UNCASED BASE proporcionado por Google a través de la plataforma Hugging Face. El conjunto de datos utilizado tuvo un tamaño de 3573 CVs, los cuales fueron empleados durante la etapa de Fine Tuning a lo largo de 3 épocas.

Al comparar el modelo propuesto contra modelos tradicionales de machine learning como Random Forest, SVM, Logistic Regression y Naive Bayes Multinomial, se observó que el modelo propuesto presenta las mejores métricas para exactitud de 83,0% y precisión con 82,3%, estos valores indican que el modelo tiene un desempeño competitivo en la tarea de clasificación de currículums vitae por área de profesión, subrayando el potencial de los MLL para tareas que requieran procesamiento de texto no estructurado con alta variabilidad, como lo son los CVs.

Al considerar escenarios donde el volumen de postulaciones es elevado y el tiempo disponible para revisión manual es limitado, como ocurre en grandes plataformas de empleo y convocatorias masivas de personal se hace aún más evidente la relevancia de estos hallazgos. En estos escenarios, la implementación de soluciones automáticas de clasificación puede optimizar significativamente el trabajo de los departamentos de recursos humanos, reduciendo los tiempos de revisión, disminuyendo errores humanos y mejorando la calidad de la preselección. El uso de BERT, por su capacidad para capturar relaciones semánticas complejas y adaptarse con poco entrenamiento adicional, representa un avance importante frente a modelos tradicionales.

Mientras que BERT es más equilibrado y preciso, Naive Bayes Multinomial puede ser mejor para maximizar la detección de posibles candidatos, lo cual puede ser valioso en etapas iniciales del proceso de selección donde se prioriza no excluir perfiles potencialmente válidos. Esta complementariedad abre la puerta al desarrollo de sistemas híbridos que combinen lo mejor de enfoques actuales con tradicionales.

Esta investigación no solo contribuye al conocimiento técnico en torno a la eficacia de los MLL para clasificación de textos en el dominio de recursos humanos, sino que también propone una metodología replicable en otras industrias que gestionan grandes volúmenes de texto no estructurado. Podría extenderse al análisis automático de evaluación de cartas de presentación o segmentación de candidatos por habilidades blandas y técnicas, lo que demuestra una aplicabilidad transversal.

Se recomienda utilizar un entorno de ejecución con mayores recursos, para permitir el aumento de los valores de los hiperparámetros tales como un número mayor de épocas. También utilizar la herramienta Optuna para realizar una búsqueda automatizada de hiperparámetros óptimos mediante condicionales y bucles.

Extender la investigación con modelos pre entrenados en idioma español junto a un dataset en español habiendo sido clasificados según el Clasificación Internacional Normalizada de la Educación (Instituto de Estadística de la UNESCO, 2013) según el Instituto de Estadística de la UNESCO.

REFERENCIAS BIBLIOGRÁFICAS

Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, Ú., Oprea, A., & Raffel, C. (2021). Extracting Training Data from Large Language Models.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. En J. Burstein, C. Doran, & T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 4171-4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Heakl, A., Mohamed, Y., Mohamed, N., Elsharkawy, A., & Zaky, A. (2024). ResuméAtlas: Revisiting Resume Classification with Large-Scale Datasets and Large Language Models. Procedia Computer Science, 244, 158-165. https://doi.org/10.1016/j.procs.2024.10.189

Instituto de Estadística de la UNESCO. (2013). Clasificación Internacional Normalizada de la Educación (CINE) 2011 (Revisión 2). Instituto de Estadística de la UNESCO. https://doi.org/10.15220/978-92-9189-129-0-spa

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach (arXiv:1907.11692). arXiv. https://doi.org/10.48550/arXiv.1907.11692

López, D. (2024). Evaluación de rendimiento de diferentes modelo grandes de lenguaje para el reconocimiento de emociones en texto [Universidad EAFIT]. https://hdl.handle.net/10784/35404

mitsmrmex. (2024, enero 4). 5 pasos para redactar un CV con ayuda de la IA fácilmente. MIT Sloan Management Review Mexico. https://mitsloanreview.mx/data-ia-machine-learning/5-pasos-para-usar-la-ia-y-crear-un-cv-que-impactara-a-cualquier-reclutador/

Oliveira, A., Bessa, R., & Teles, A. (2024). Análisis comparativo de modelos de lenguaje basados en BERT y generativos amplios para la detección de ideación suicida: Un estudio de evaluación del desempeño. Cadernos de Saúde Pública, 40, e00028824. https://doi.org/10.1590/0102-311XEN028824

Parthasarathy, V., Zafar, A., Khan, A., & Shahid, A. (2024). The Ultimate Guide to Fine-Tuning LLMs from Basics to Breakthroughs: An Exhaustive Review of Technologies, Research, Best Practices, Applied Research Challenges and Opportunities (arXiv:2408.13296). arXiv. https://doi.org/10.48550/arXiv.2408.13296

Seagate-WP-DataAge2025-March-2017.pdf. (s. f.). Recuperado 7 de enero de 2024, de https://www.seagate.com/files/www-content/our-story/trends/files/Seagate-WP-DataAge2025-March-2017.pdf

Wu, Y., & Wan, J. (2025). A survey of text classification based on pre-trained language model. Neurocomputing, 616, 128921. https://doi.org/10.1016/j.neucom.2024.128921

Yu, H., Yang, Z., Pelrine, K., Godbout, J., & Rabbany, R. (2023). Open, Closed, or Small Language Models for Text Classification? (arXiv:2308.10092). arXiv. https://doi.org/10.48550/arXiv.2308.10092

Zhao, W., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., … Wen, J.-R. (2023). A Survey of Large Language Models (arXiv:2303.18223). arXiv. http://arxiv.org/abs/2303.18223