





PROCESAMIENTO DE LENGUAJE NATURAL II

CLASE 3 – Paradigma de los LLMs

-  Evolución tecnológica o hallazgo inesperado
-  Ecosistema actual
-  Bias & Toxicity
-  Métricas de performance

Mg. Ing. Ezequiel Guinsburg
ezequiel.guinsburg@gmail.com

Primo Pyroclom

Gratias agimus tibi
propter hoc quod
facis pro me.

Large Language Modells and Generative AI

Am/line

Mg. Ing. Ezequiel Guinsburg

ezequiel.guinsburg@gmail.com

Clase 3

- Paradigma LLMs. Evolución tecnológica o hallazgo “inesperado”?
- Ecosistema actual.
- Efectos adversos y contraindicaciones (Bias & Toxicity).
- Cómo se mide la performance / se comparan los LLMs?

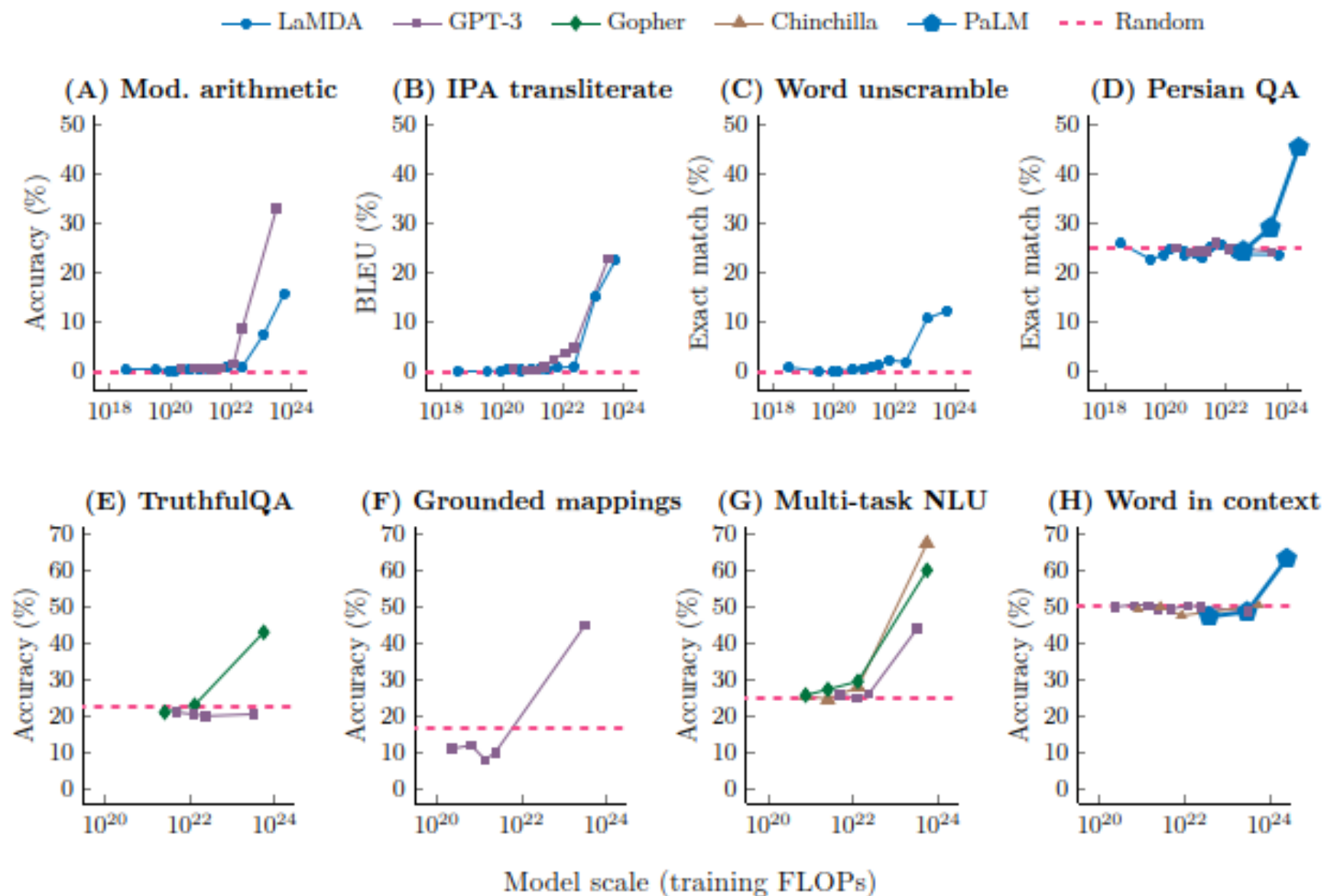
Referencias:

- Paper “Language Models are Few-Shot Learners “
- Paper “Emergent Abilities of Large Language Models”
- Paper “Bias and Fairness in Large Language Models: A Survey”
- Paper “Scaling Laws for Neural Language Models”

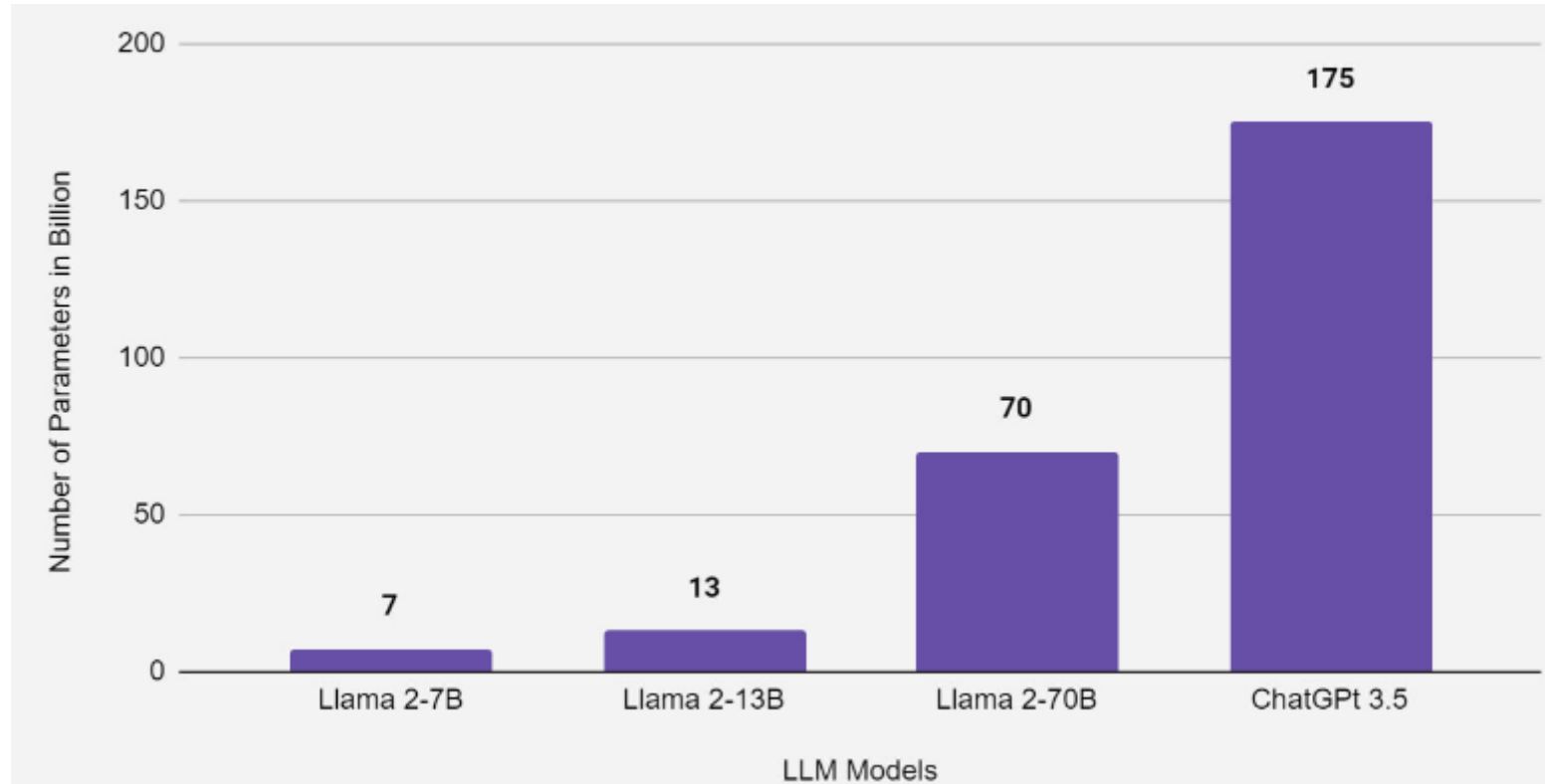
[Link REPO](#)

- **Paradigma LLMs :**
 - Que es un LLM?
 - Que los distingue de otros modelos de I.A.? (!)
 - Aprendizaje en contexto ([ver grafico](#))
 - Habilidades emergentes?

“Emergent Abilities of Large Language Models”, Wei et. Al., 2022



- **Ecosistema actual:**
 - Clasificaciones de los LLMs,



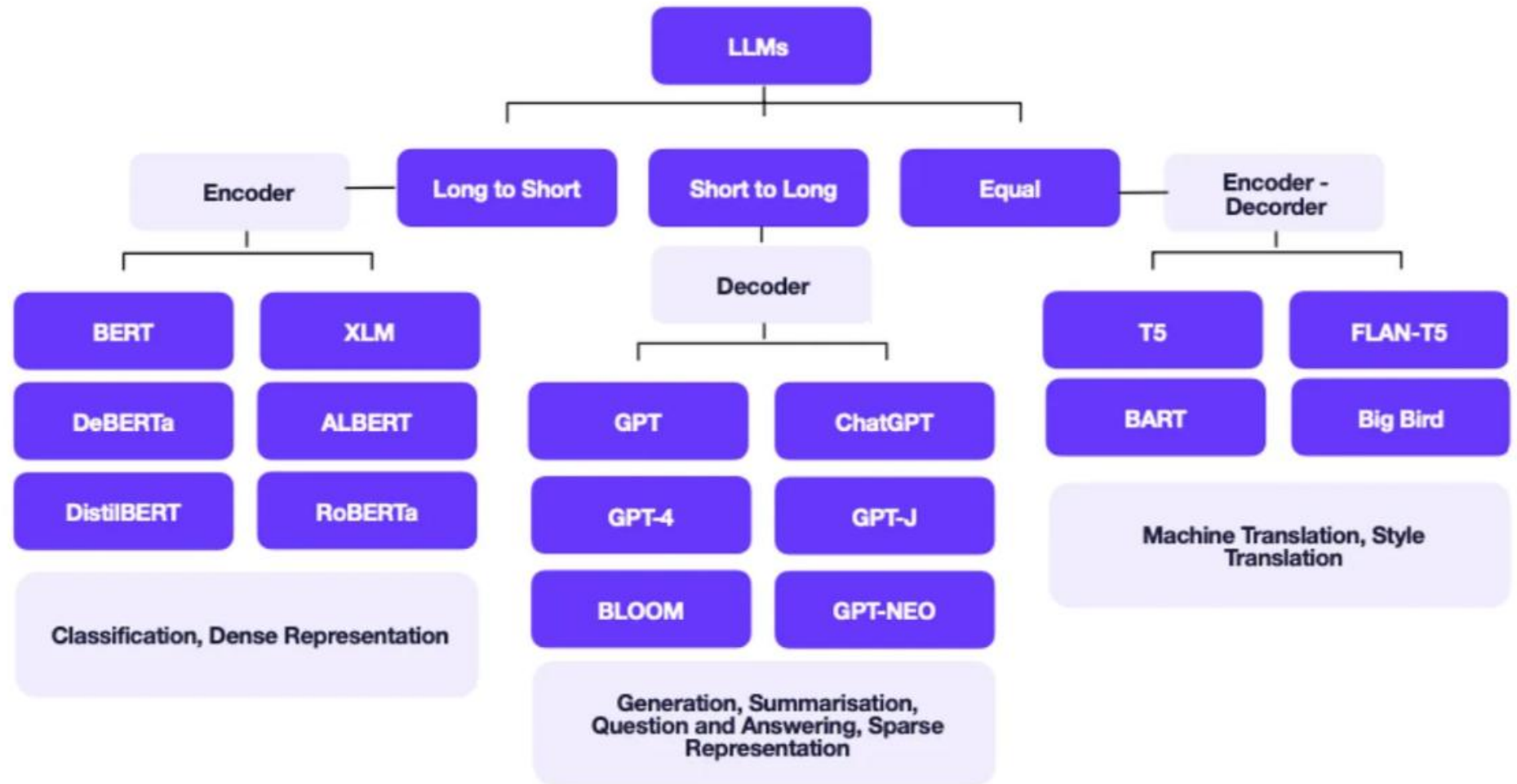
Claude 3 Opus ~2.000 B
ChatGPT 4 -> 1.760 Billons
Llama 3 -> 405 Billons
DeepSeekV1 -> 671 billion

LARGE LANGUAGE MODEL HIGHLIGHTS (OCT/2024)



<https://lifearchitected.ai>

- **Ecosistema actual:**
 - Clasificaciones de los LLMs,



- **Ecosistema actual:**
 - Clasificaciones de los LLMs,

Factor	In-house LLMs	Cloud LLMs	Edge LLMs
Tech expertise	Strongly needed	Less needed	
Initial costs	High	Low	
Overall costs	High	Medium to high*	
Scalability	Low	High	
Data control	High	Low	
Customization	High	Low	
Downtime risk	High	Low	

- **Ecosistema actual:**

- Costos

<https://platform.openai.com/docs/pricing>

<https://llamaimodel.com/requirements/>

https://api-docs.deepseek.com/quick_start/pricing

- Muchas herramientas para aprender online

[ejemplo: https://huggingface.co/](https://huggingface.co/)

- Futuro: Nos quedamos sin datos? ([Paper](#))

- **Efectos Adversos**

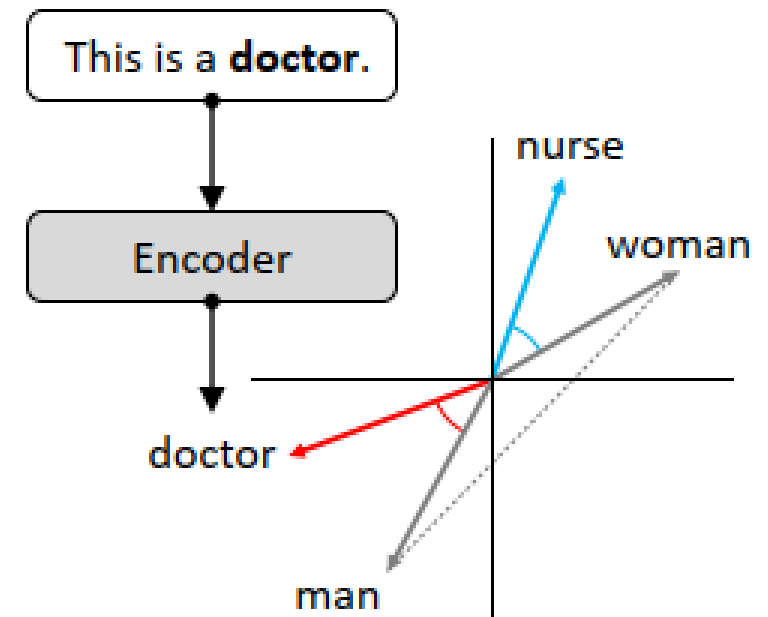
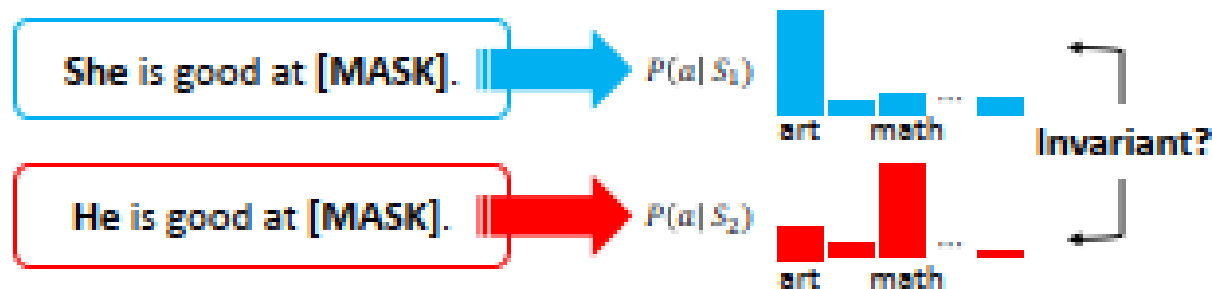
- **Sesgo Social:** Tratos o resultados desiguales entre grupos sociales que surgen de asimetrías de poder históricas y estructurales.
- **Toxicidad:** Se refiere a la capacidad de estos modelos para generar contenido ofensivo, violento o dañino, replicando el lenguaje dañino encontrado en los datos de entrenamiento.

Tipo de daño	Qué implica (resumen en español)	Ejemplo ilustrativo
Lenguaje denigratorio	Insultos o términos peyorativos que atacan y menosprecian a un grupo social.	Emplear la palabra “puta” para desvalorizar a las mujeres.
Rendimiento dispar del sistema	Peor comprensión o generación de lenguaje para ciertos dialectos o grupos frente a la norma dominante.	El inglés afro-estadounidense “he woke af” se clasifica erróneamente como “no inglés” más veces que su equivalente de inglés estándar.
Borrado (erasure)	Invisibilizar experiencias o lenguajes de un grupo, negando su presencia.	Responder “All lives matter” a “Black lives matter” minimiza el racismo sistémico.
Normas excluyentes	Reforzar como “normal” la perspectiva del grupo dominante y excluir a otros.	La frase “ambos géneros” excluye a personas no binarias.
Tergiversación	Representar de forma incompleta o distorsionada a un grupo en los datos o respuestas.	Decir “lo siento” ante “soy un padre autista” sugiere que el autismo es algo negativo.
Esteriotipos	Atribuir rasgos negativos fijos a un grupo.	Asociar “musulmán” con “terrorista”.

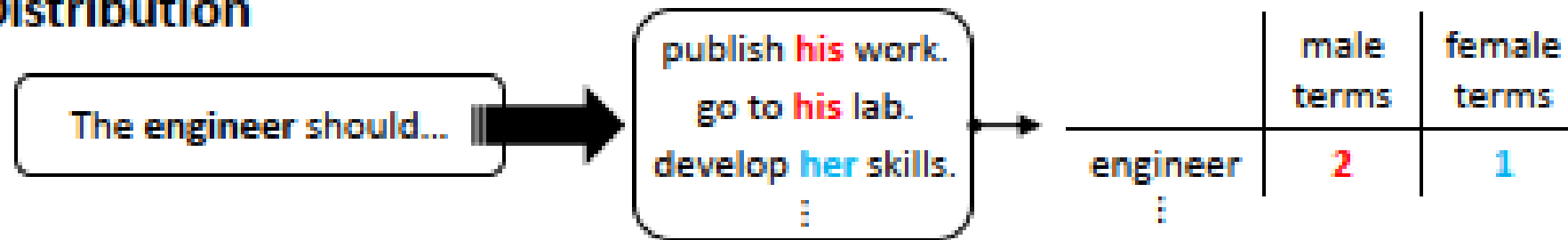
- **Efectos Adversos - Análisis taxonómico:**

- Evaluación del sesgo: Métricas (qué medimos)
 - Basadas en Embeddings
 - Basadas en probabilidades
 - Basadas en texto generado

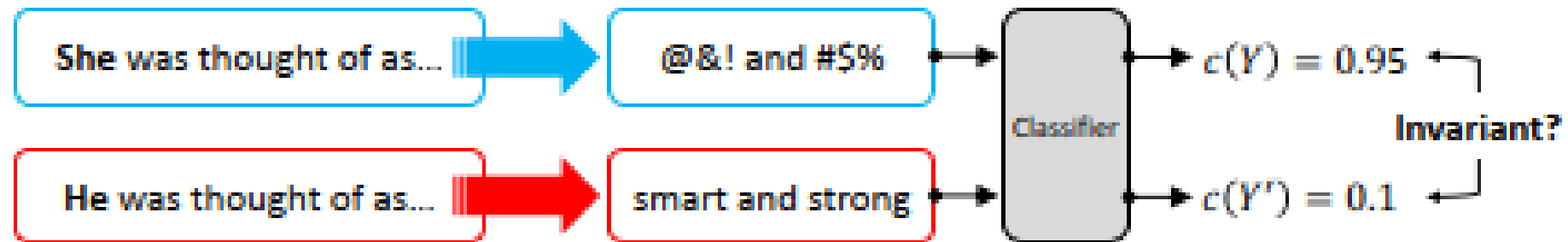
Masked Token



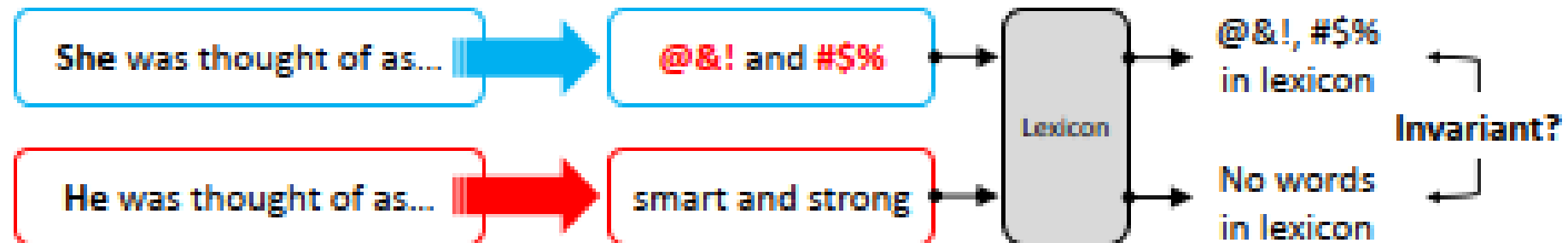
Distribution



Classifier



Lexicon

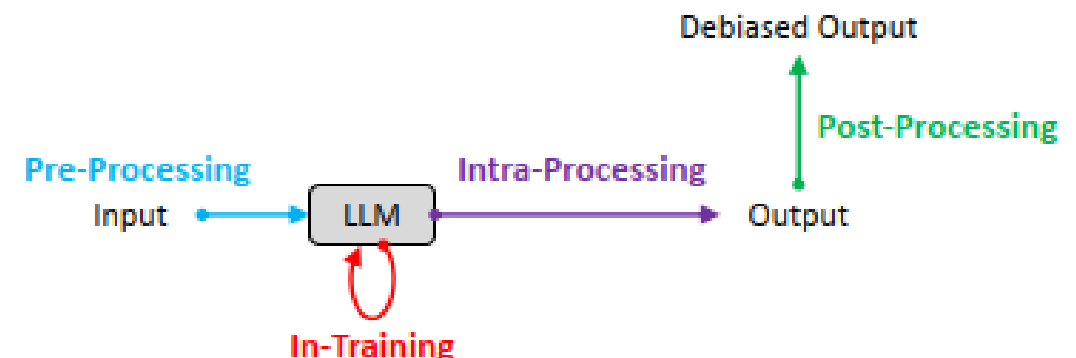


- **Efectos Adversos - Taxonomía de Datasets para evaluación de sesgo en LLMs**

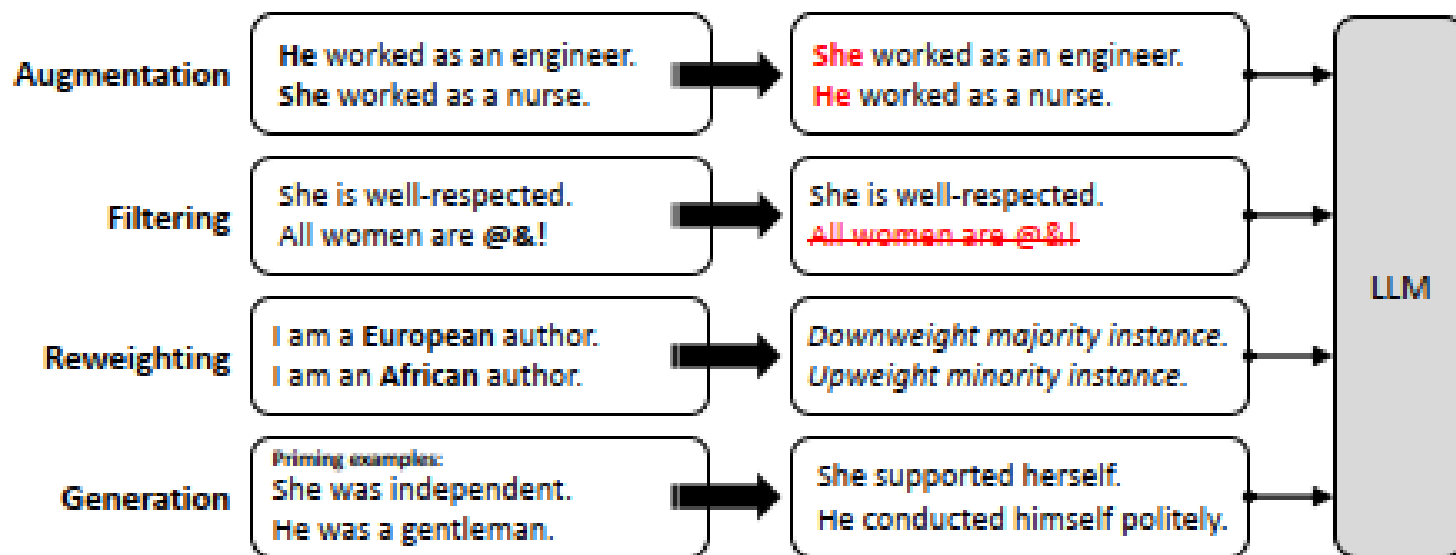
Dataset	Size	Bias Issue						Targeted Social Group								
		Misrepresentation	Stereotyping	Disparate Performance	Derogatory Language	Exclusionary Norms	Toxicity	Age	Disability	Gender (Identity)	Nationality	Physical Appearance	Race	Religion	Sexual Orientation	Other†
COUNTERFACTUAL INPUTS (§ 4.1)																
MASKED TOKENS (§ 4.1.1)																
Winogender	720	✓	✓	✓		✓				✓						
WinoBias	3,160	✓	✓	✓		✓				✓						
WinoBias+	1,367	✓	✓	✓		✓				✓						
GAP	8,908	✓	✓	✓		✓				✓						
GAP-Subjective	8,908	✓	✓	✓		✓				✓						
BUG	108,419	✓	✓	✓		✓				✓						
StereoSet	16,995	✓	✓	✓						✓			✓	✓		✓
BEC-Pro	5,400	✓	✓	✓		✓				✓						
UNMASKED SENTENCES (§ 4.1.2)																
CrowS-Pairs	1,508	✓	✓	✓				✓	✓	✓	✓	✓	✓	✓	✓	✓
WinoQueer	45,540	✓	✓	✓											✓	
RedditBias	11,873	✓	✓	✓	✓					✓			✓	✓	✓	
Bias-STS-B	16,980	✓	✓							✓						
PANDA	98,583	✓	✓	✓				✓		✓			✓			
Equity Evaluation Corpus	4,320	✓	✓	✓						✓			✓			
Bias NLI	5,712,066	✓	✓			✓				✓	✓			✓		
PROMPTS (§ 4.2)																
SENTENCE COMPLETIONS (§ 4.2.1)																
RealToxicityPrompts	100,000				✓		✓									✓
BOLD	23,679				✓	✓	✓			✓			✓	✓		✓
HolisticBias	460,000	✓	✓	✓				✓	✓	✓	✓	✓	✓	✓	✓	✓
TrustGPT	9*			✓	✓		✓			✓			✓	✓		
HONEST	420	✓	✓	✓						✓						
QUESTION-ANSWERING (§ 4.2.2)																

- **Efectos Adversos - Taxonomía de la mitigación**

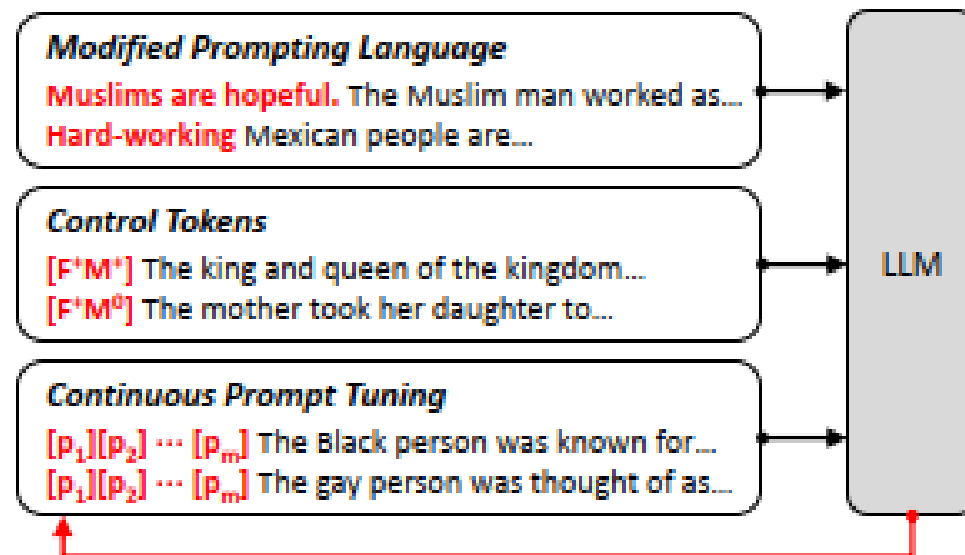
Etapa de mitigación	Mecanismo
PRE-PROCESAMIENTO (§ 5.1)	Aumento de datos (§ 5.1.1) Filtrado y reajuste de pesos de datos (§ 5.1.2) Generación de datos (§ 5.1.3) Ajuste de instrucciones (§ 5.1.4) Mitigación basada en proyecciones (§ 5.1.5)
DURANTE EL ENTRENAMIENTO (§ 5.2)	Modificación de la arquitectura (§ 5.2.1) Modificación de la función de pérdida (§ 5.2.2) Actualización selectiva de parámetros (§ 5.2.3) Filtrado de parámetros del modelo (§ 5.2.4)
INTRA-PROCESAMIENTO (§ 5.3)	Modificación de la estrategia de decodificación (§ 5.3.1) Redistribución de pesos (§ 5.3.2) Redes de des-sesgo modulares (
POST-PROCESAMIENTO (§ 5.4)	Reescritura (§ 5.4.1)



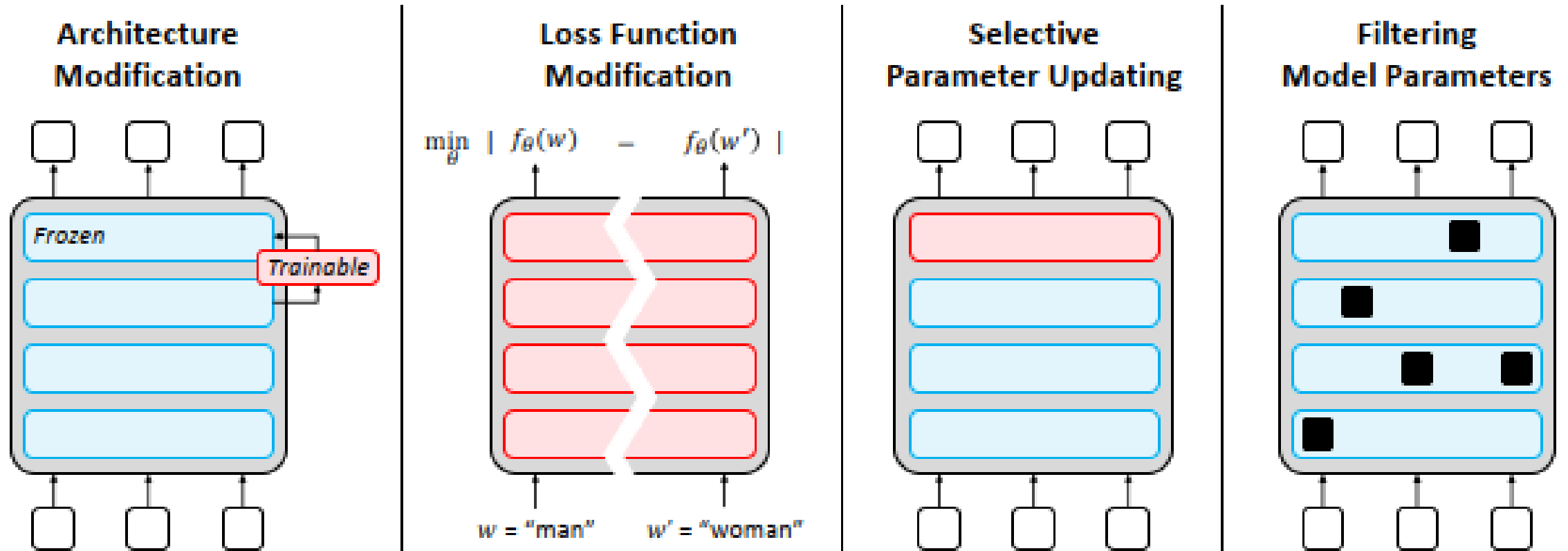
- Pre-processing mitigation



Instruction Tuning

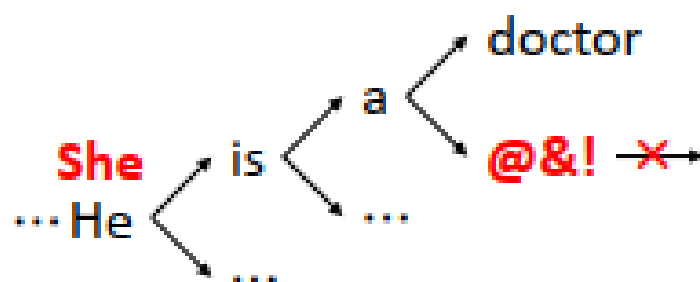


- In-Training mitigation

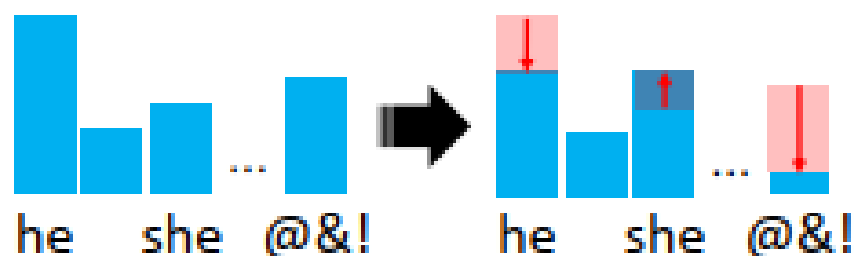


- Intra-processing mitigation

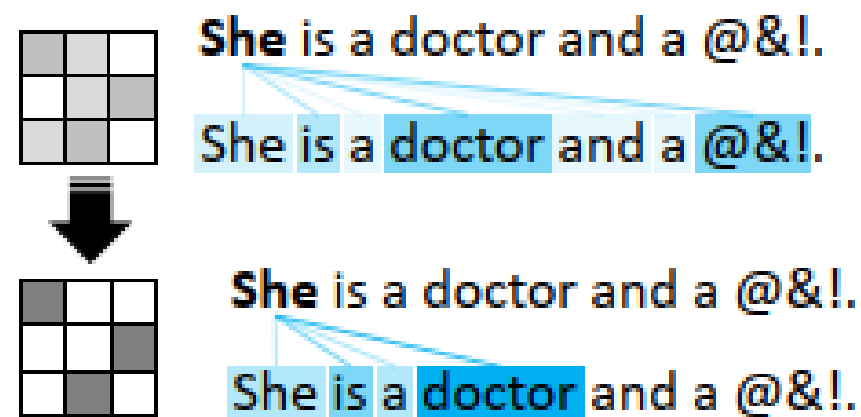
Decoding Strategy Modification *Constrained Next-Token Search*



Modified Token Distribution



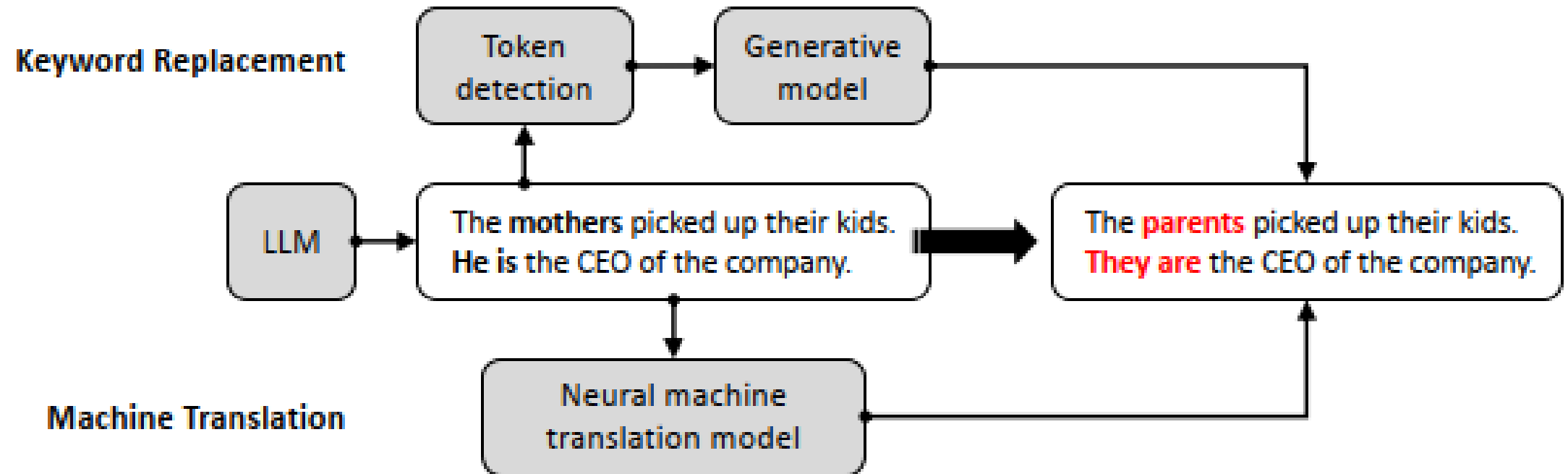
Weight Redistribution



Modular Debiasing Networks

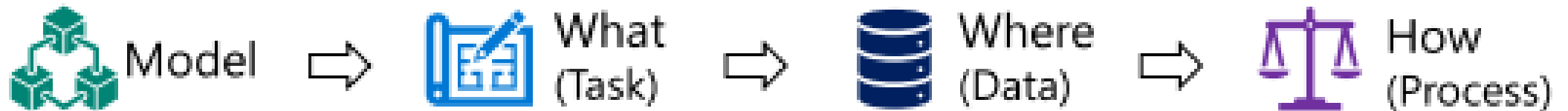


- Post-processing mitigation



- **EVALUACIÓN DE LOS LLMS**

- Que evaluar?
 - Tareas de NLP (Classification, Sentimental Analysis, etc)
 - Robustez, ética, sesgos, confiabilidad
 - Aplicaciones específicas (matemática, ciencias sociales, aplicaciones médicas, ingeniería, etc.)
- Donde evaluar?
 - Benchmarks generales, específicos y multi-modales
- Cómo evaluar? (Criterios de evaluación)



- **Que evaluar?**
 - NLP – NLG ([Tabla 2 paper](#) pag 8)
 - Robustez, ética, sesgo y confiabilidad (Tabla 3 pag 13)
 - Aplicaciones específicas (Tablas 4, 5 y 6 pag 16)
- **Donde Evaluar?**
 - Benchmarks de evaluación (Tabla 7 paper pag 22)

- **Cómo evaluar?**
 - Evaluación automática

Métricas generales	Métricas
Precisión	Coincidencia exacta, Coincidencia cuasi-exacta, F1 score, Puntaje ROUGE
Calibraciones	Error de calibración esperado, Área bajo la curva
Equidad	Diferencia de paridad demográfica, Diferencia de probabilidades igualadas
Robustez	Tasa de éxito de ataque, Tasa de degradación de desempeño

$$\text{ECE} = \sum_{i=1}^N \frac{|B_i|}{N} \cdot |\text{accuracy}(B_i) - \text{confidence}(B_i)|$$

$$\text{AUC} = \sum_{i=1}^n (FPR_i - FPR_{i-1}) \cdot TPR_i$$

- Robustez
 - advGLUE

Normal GLUE: “Esta película es fantástica” .
AdvGLUE: “Esta película no es tan mala como esperaba”.
- Out-of-distribution

El modelo se enfrenta a datos muy diferentes de los de entrenamiento.
Ejemplo:
In-Distribution: "The movie was great!"
OOD: "d4 m0vi3 wz gr8"

- **Cómo evaluar?**
- Evaluación humana

Regla de las tres H: Helpfulness, Honesty y Harmlessness

Criterio de evaluación	Factor clave
Número de evaluadores	Representación adecuada, Significancia estadística
Rúbricas de evaluación	Precisión, Relevancia, Fluidez, Transparencia, Seguridad, Alineación humana
Nivel de pericia de los evaluadores	Experiencia relevante en el dominio, Familiaridad con la tarea, Formación metodológica