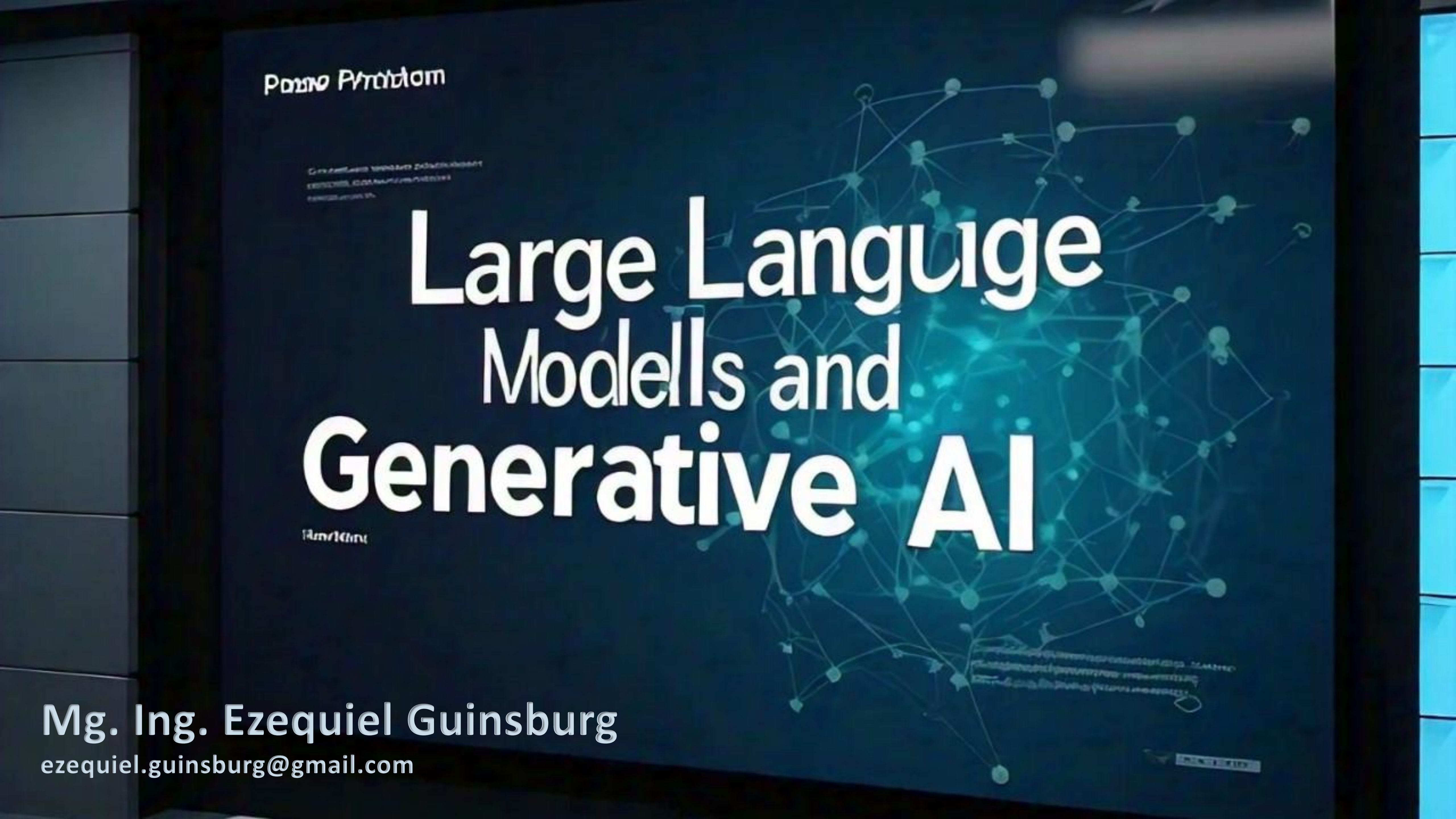


Mg. Ing. Ezequiel Guinsburg
ezequiel.guinsburg@gmail.com

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”

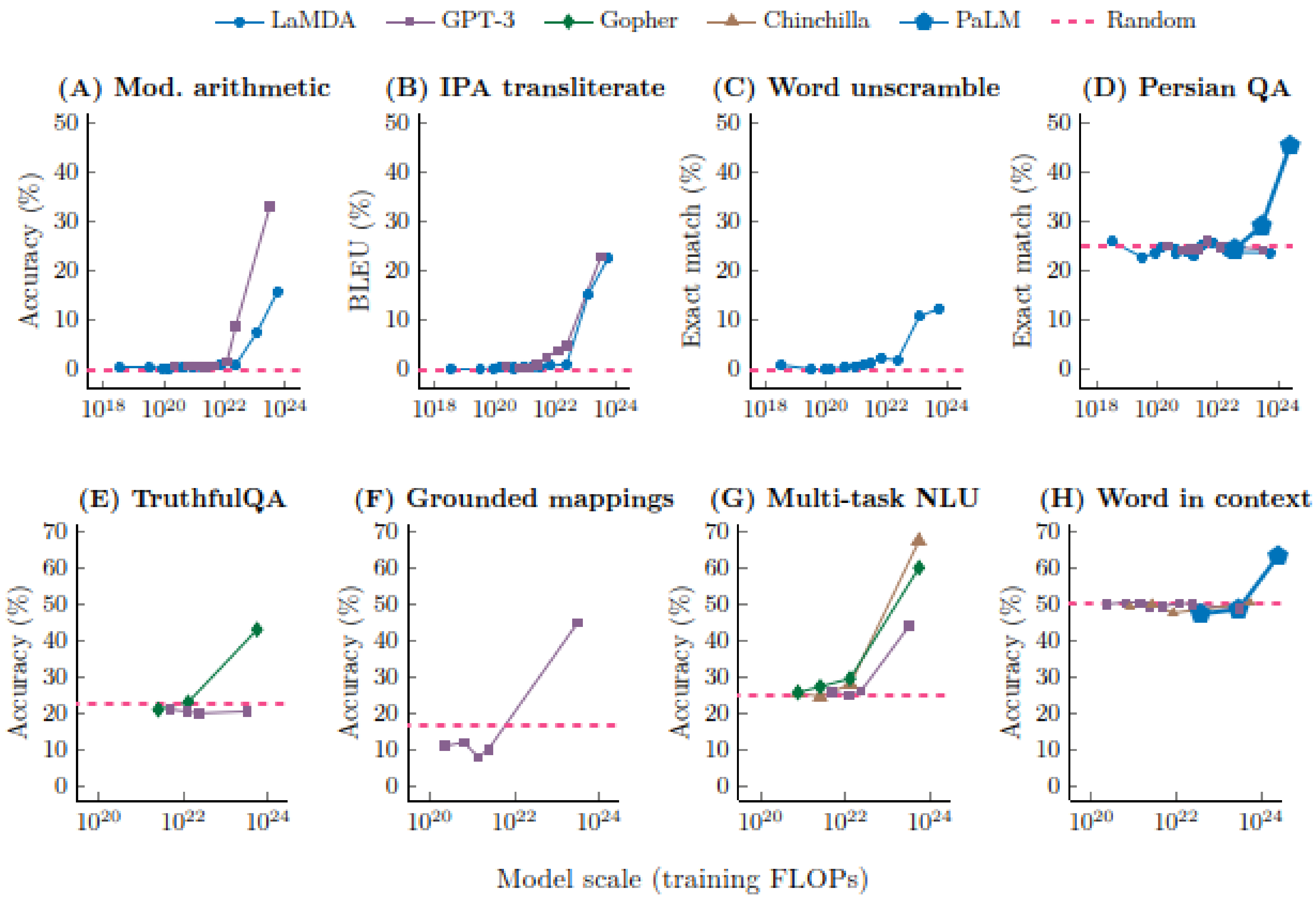
Link REPO

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?

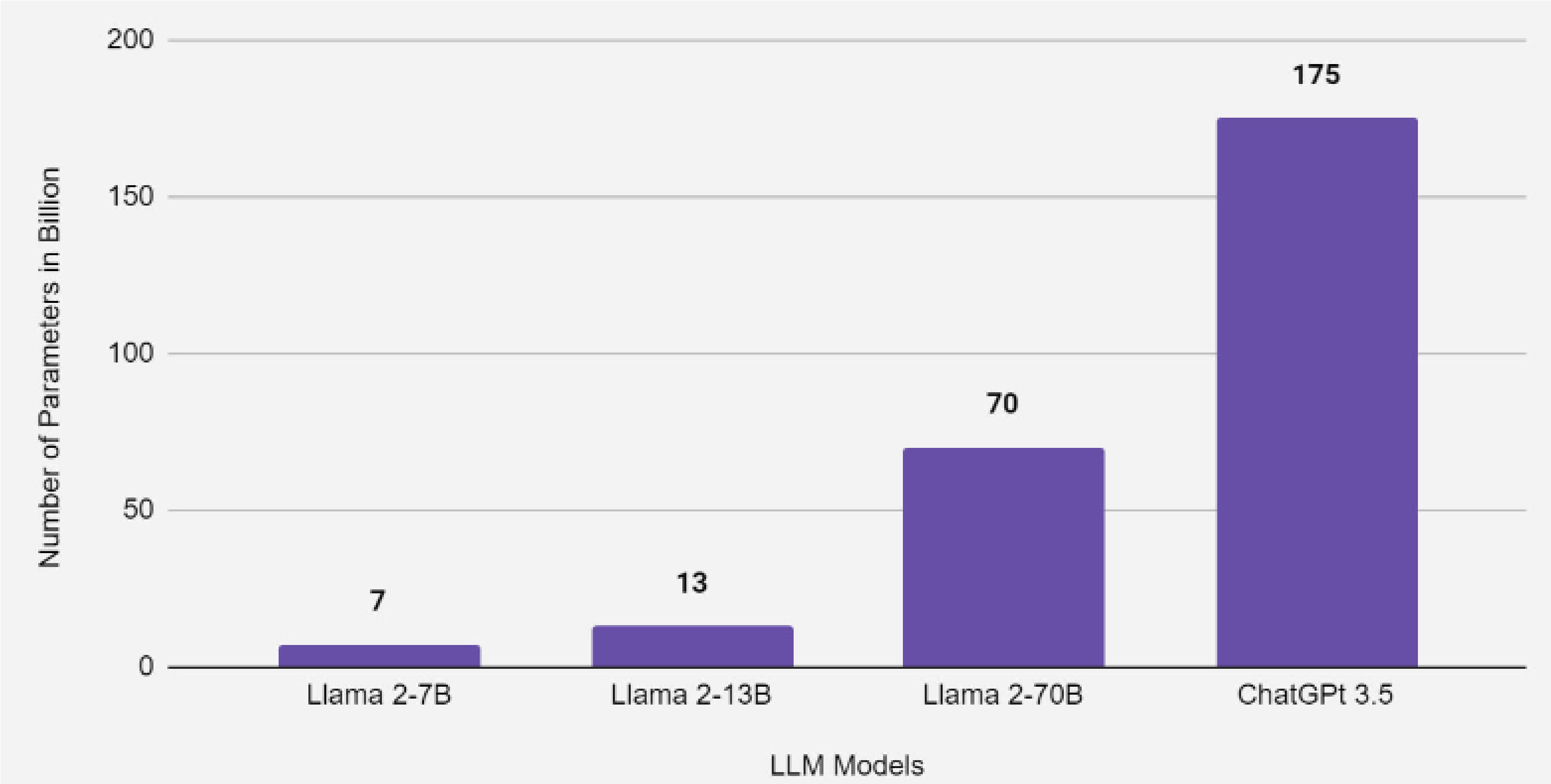
- **Paradigma LLMs :**
  - Que es un LLM?
  - Que los distingue de otros modelos de I.A.?
  - Aprendizaje en contexto
  - Habilidades emergentes?

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

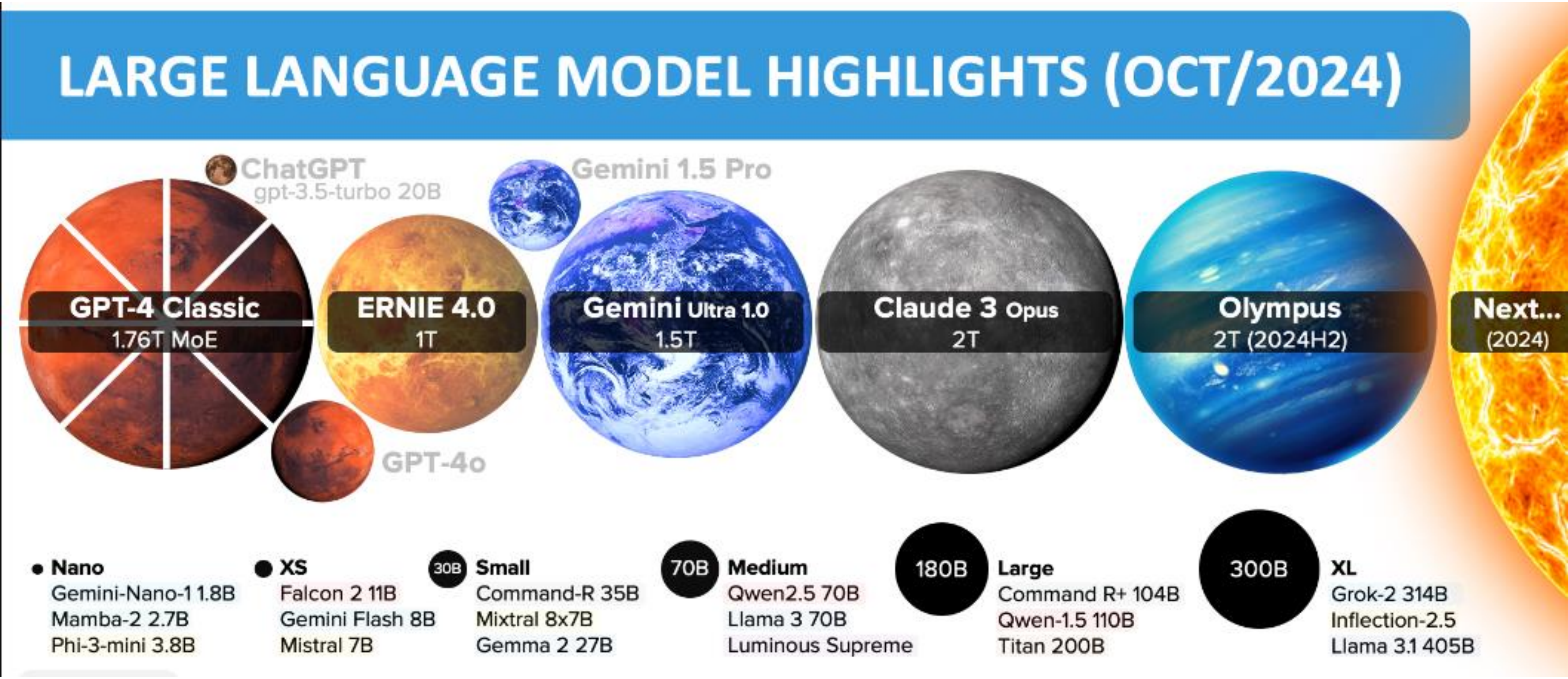


PARADIGMAS LLMs – HABILIDADES EMERGENTES

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

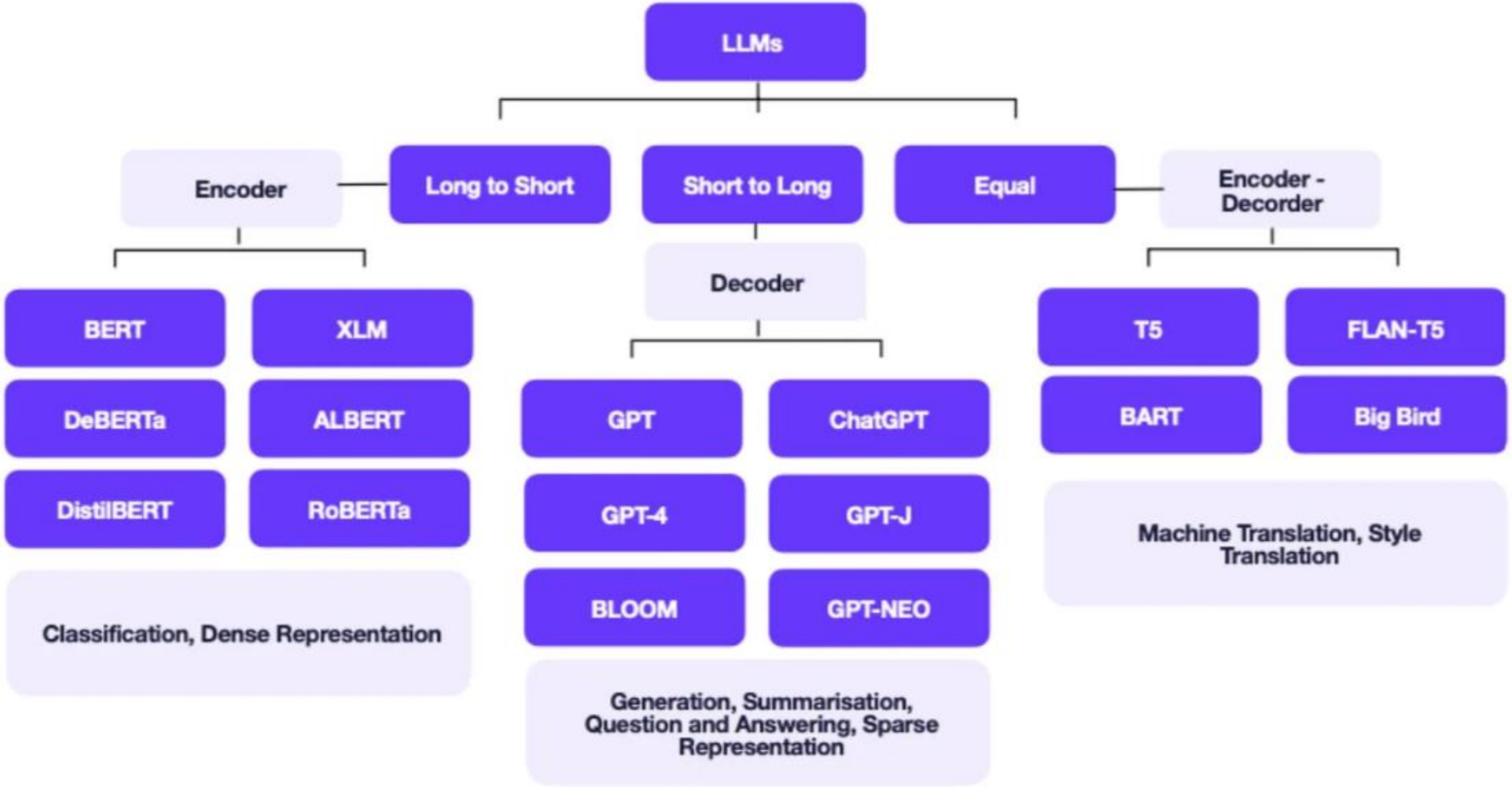


ChatGPT 4 -> 1.760 Billons
Llama 3 -> 405 Billons



https://lifearchitect.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://openai.com/api/pricing/

https://llamaimodel.com/requirements/

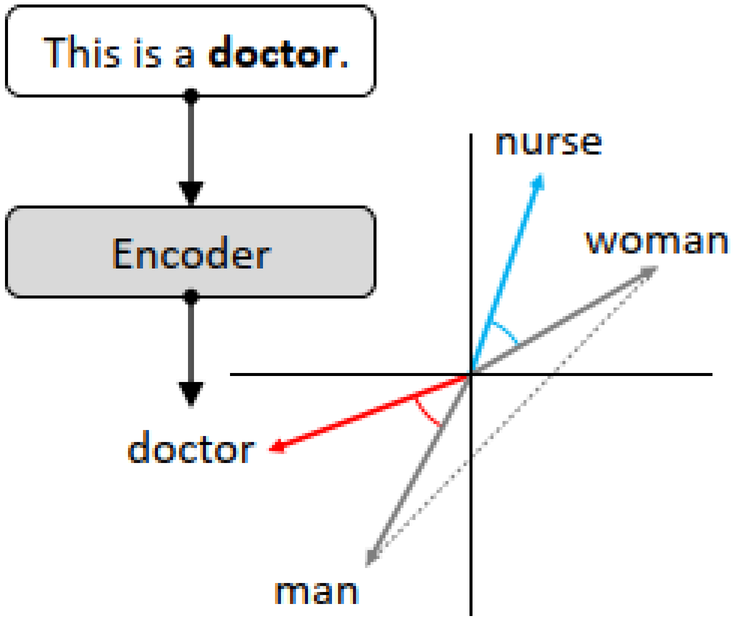
- **Efectos Adversos**
    - **Sesgo Social:** Tratos o resultados desiguales entre grupos sociales que surgen de asimetrías de poder históricas y estructurales.
    - **Toxicidad:** Se refiere a la capacidad de estos modelos para generar contenido ofensivo, violento o dañino, replicando el lenguaje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

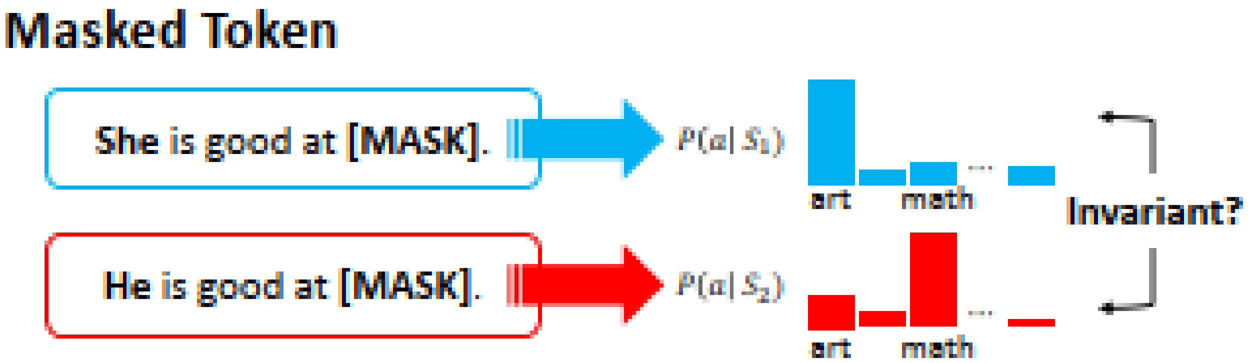
Type of Harm	Definition and Example
REPRESENTATIONAL HARMS	Denigrating and subordinating attitudes towards a social group
Derogatory language	Pejorative slurs, insults, or other words or phrases that target and denigrate a social group *e.g.,* `"Whore"` *conveys hostile and contemptuous female expectations* (Beukeboom and Burgers 2019)
Disparate system performance	Degraded understanding, diversity, or richness in language processing or generation between social groups or linguistic variations *e.g., AAE* like* `"he woke af"` *is misclassified as not English more often than SAE† equivalents* (Blodgett and O'Connor 2017)
Erasure	Omission or invisibility of the language and experiences of a social group *e.g.,* `"All lives matter"` *in response to* `"Black lives matter"` *implies colorblindness that minimizes systemic racism* (Blodgett 2021)
Exclusionary norms	Reinforced normativity of the dominant social group and implicit exclusion or devaluation of other groups *e.g.,* `"Both genders"` *excludes non-binary identities* (Bender et al. 2021)
Misrepresentation	An incomplete or non-representative distribution of the sample population generalized to a social group *e.g., Responding* `"I'm sorry to hear that"` *to* `"I'm an autistic dad"` *conveys a negative misrepresentation of autism* (Smith et al. 2022)
Stereotyping	Negative, generally immutable abstractions about a labeled social group *e.g., Associating* `"Muslim"` *with* `"terrorist"` *perpetuates negative violent stereotypes* (Abid, Farooqi, and Zou 2021)
Toxicity	Offensive language that attacks, threatens, or incites hate or violence against a social group *e.g.,* `"I hate Latinos"` *is disrespectful and hateful (Dixon et al. 2018)*

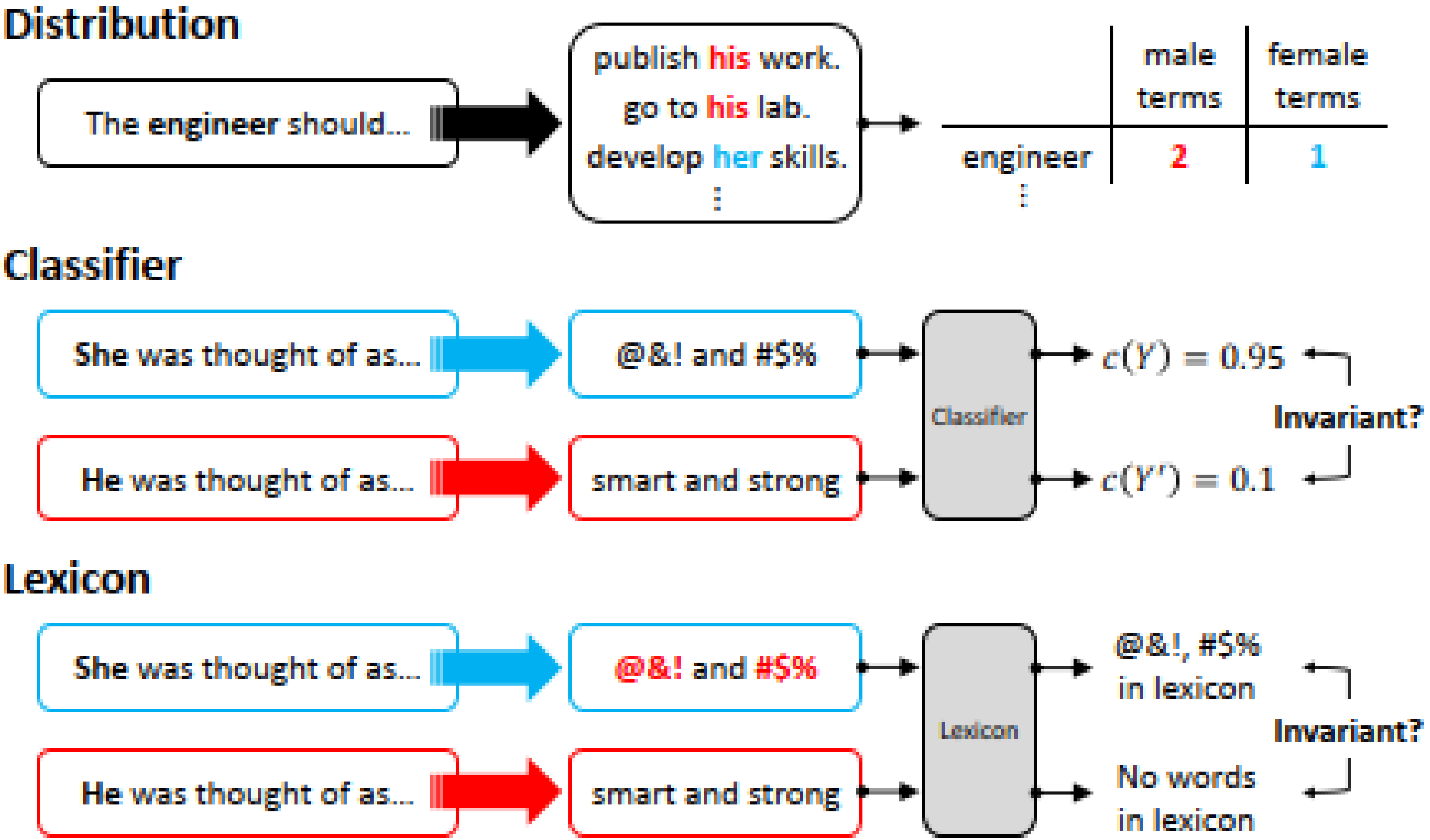
EFECTOS ADVERSOS

- **Efectos Adversos - Análisis taxonométrico:**

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







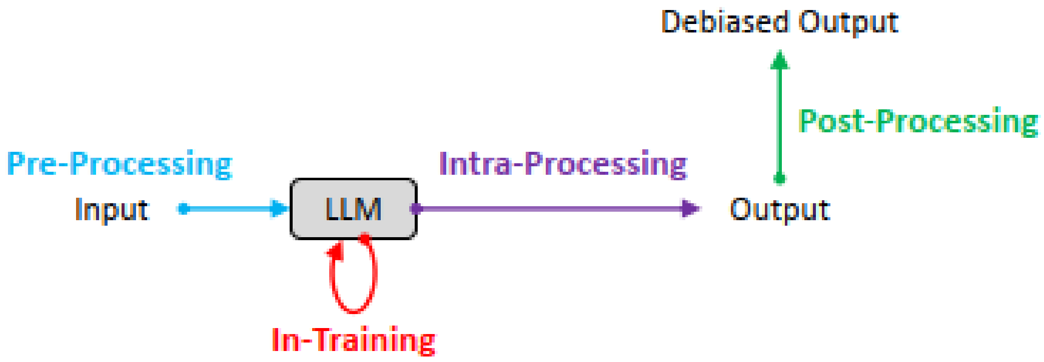
EFECTOS ADVERSOS

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

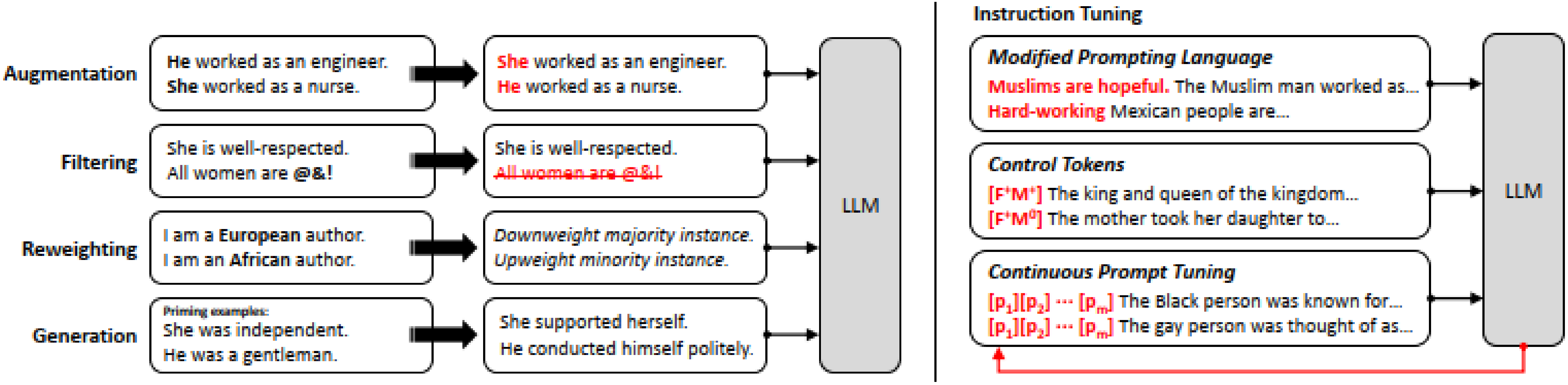
Dataset	Size	Bias Issue						Targeted Social Group								
		Misrepresentation	Stereotyping	Disparate Performance	Derogatory Language	Exclusionary Norms	Toxicity	Age	Disability	Gender (Identity)	Nationality	Physical Appearance	Race	Religion	Sexual Orientation	Other†
COUNTERFACTUAL INPUTS (§ 4.1)																
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

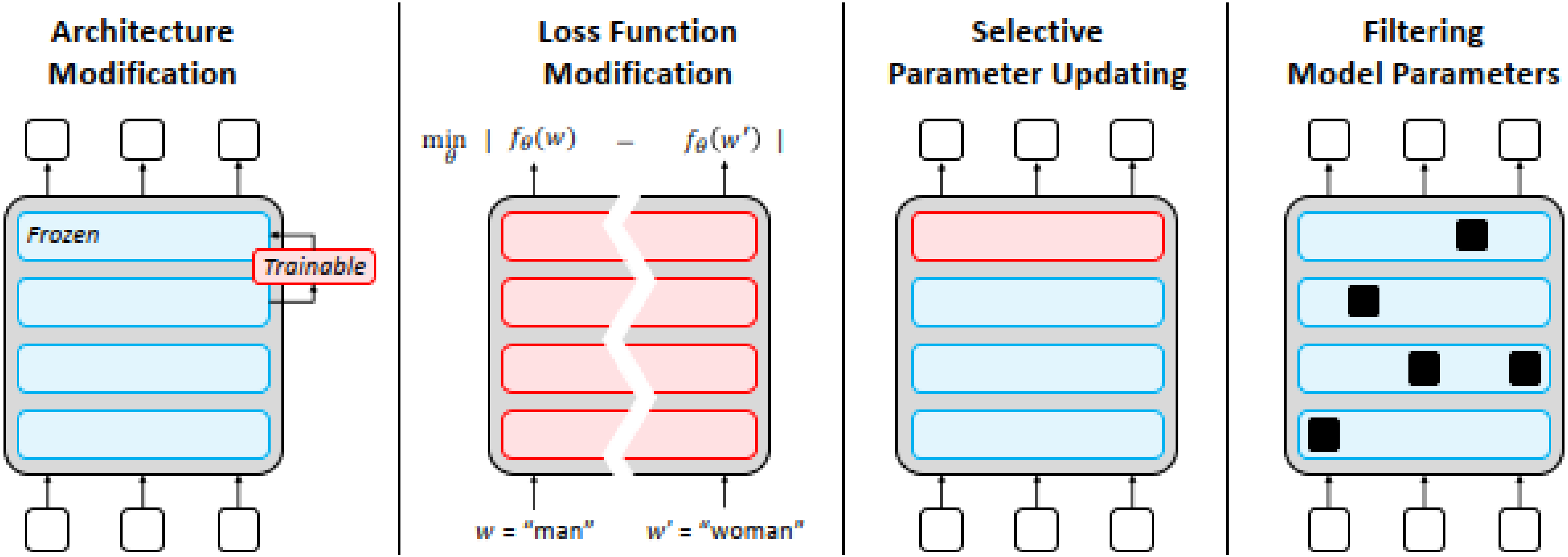
Mitigation Stage	Mechanism
PRE-PROCESSING (§ 5.1)	Data Augmentation (§ 5.1.1) Data Filtering & Reweighting (§ 5.1.2) Data Generation (§ 5.1.3) Instruction Tuning (§ 5.1.4) Projection-based Mitigation (§ 5.1.5)
IN-TRAINING (§ 5.2)	Architecture Modification (§ 5.2.1) Loss Function Modification (§ 5.2.2) Selective Parameter Updating (§ 5.2.3) Filtering Model Parameters (§ 5.2.4)
INTRA-PROCESSING (§ 5.3)	Decoding Strategy Modification (§ 5.3.1) Weight Redistribution (§ 5.3.2) Modular Debiasing Networks (§ 5.3.3)
POST-PROCESSING (§ 5.4)	Rewriting (§ 5.4.1)



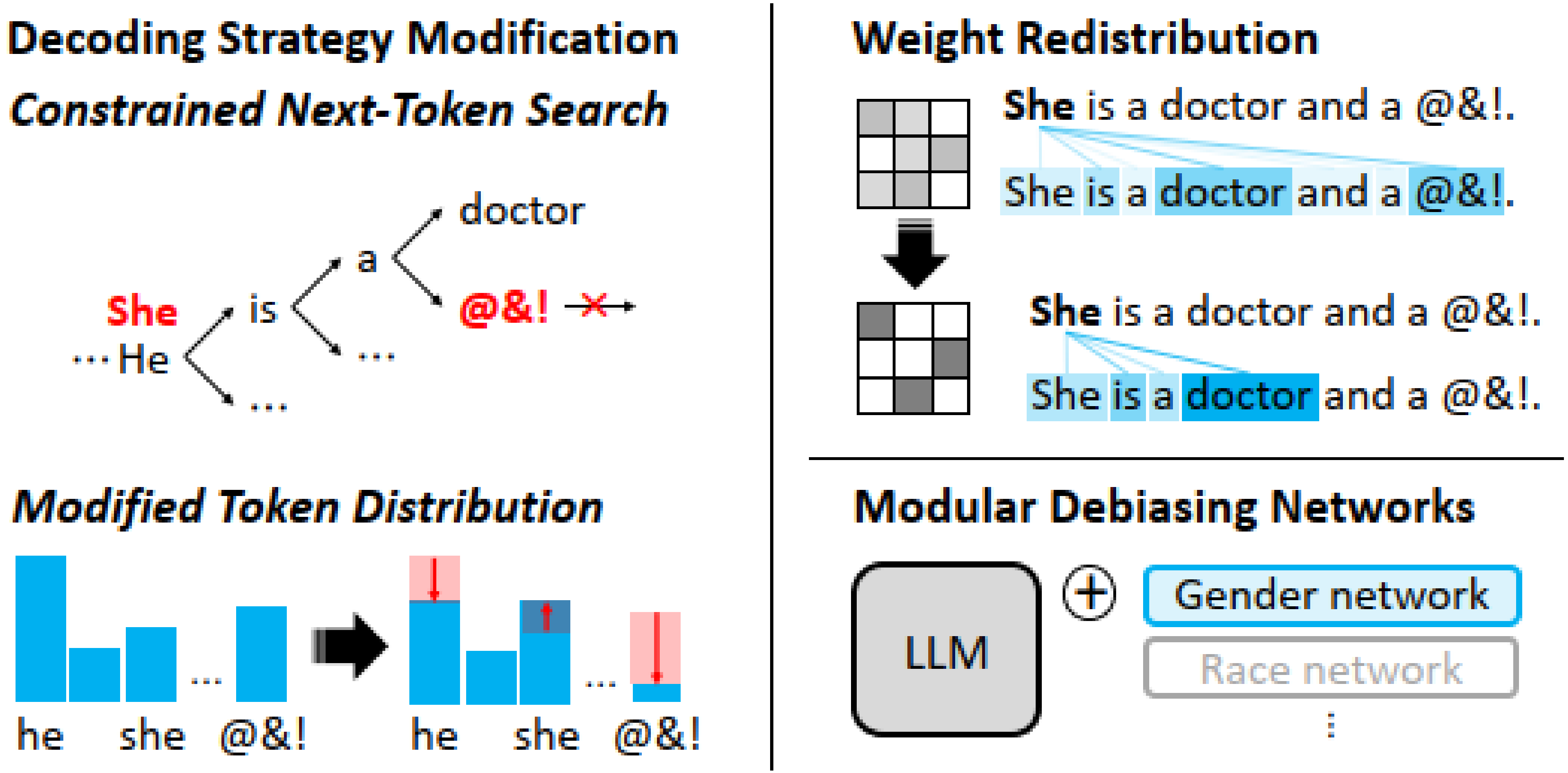
- Pre-processing mitigation



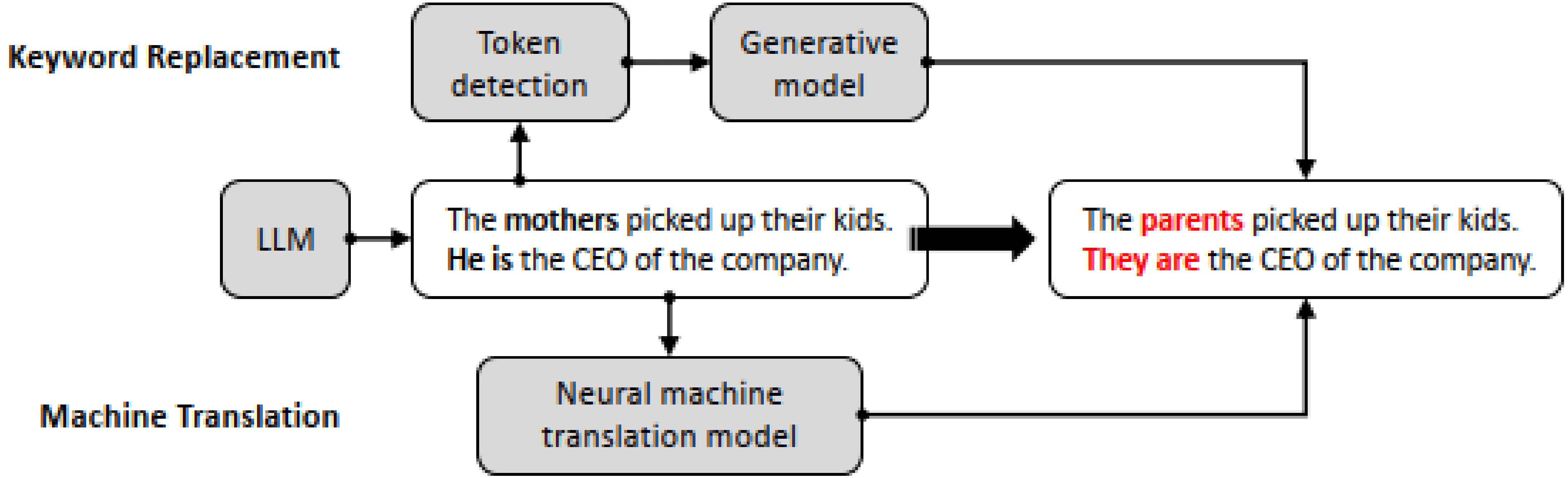
- In-Training mitigation



- Intra-processing mitigation

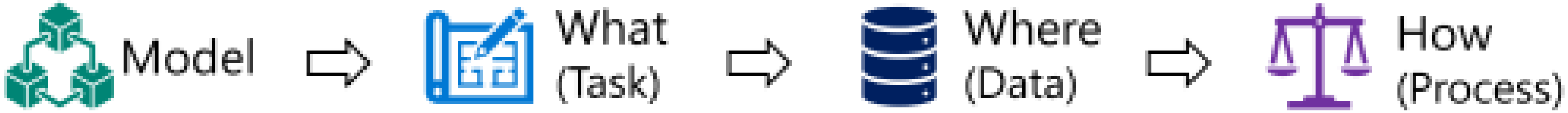


- Post-processing mitigation



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)
  - Robustez, ética, sesgo y confiabilidad (Tabla 3 paper)
  - Aplicaciones específicas (Tablas 4, 5 y 6)

- **Donde Evaluar?**

  - Benchmarks de evaluación (Tabla 7 paper)

- **Cómo evaluar?**

  - Evaluación automática

General metrics	Metrics
Accuracy	Exact match, Quasi-exact match, F1 score, ROUGE score [118]
Calibrations	Expected calibration error [60], Area under the curve [54]
Fairness	Demographic parity difference [241], Equalized odds difference [64]
Robustness	Attack success rate [203], Performance drop rate [262]

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

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

- 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

Evaluation Criteria	Key Factor
Number of evaluators	Adequate representation [7], Statistical significance
Evaluation rubrics	Accuracy [178], Relevance [259], Fluency [196], Transparency, Safety [85], Human alignment
Evaluator's expertise level	Relevant domain expertise [144], Task familiarity, Methodological training