LlaMA 2

URKU

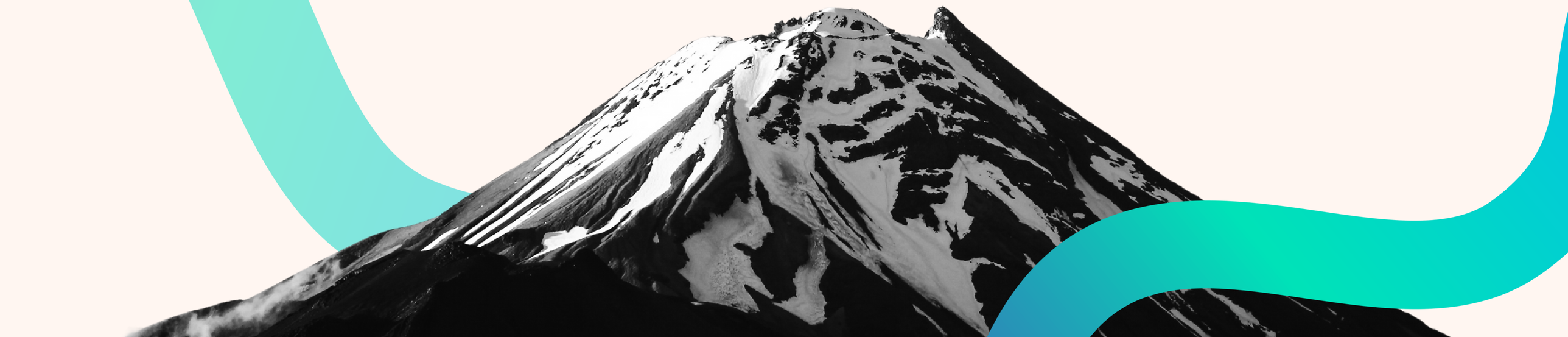
James León

Proyecto Integrador

RUNA-SHIMI ÑAN

EL LENGUAJE

El **Kichwa como idioma ‘oral’**.
El **Kichwa unificado** como línea base para el desarrollo.
Fuentes oficiales en línea: **diccionarios y gramática**.
Fuentes no-oficiales digitalizadas para este proyecto: **transcripts**.



Los Embeddings

BoW and Word2Vec

Frecuencia por palabra y representaciones semánticas basadas en vectores densos

FastText

Incluyó información de ‘subwords’ para representación de variantes morfológicas

ELMo

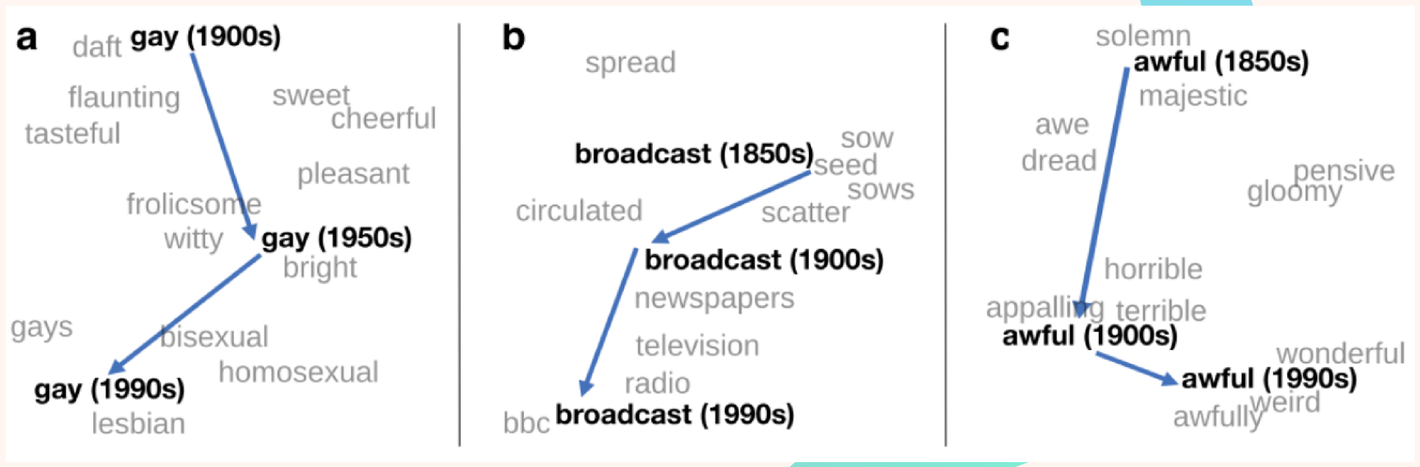
“meaning is use,” embeddings dependientes del contexto

BERT

Comprensión contextual bi-direccional

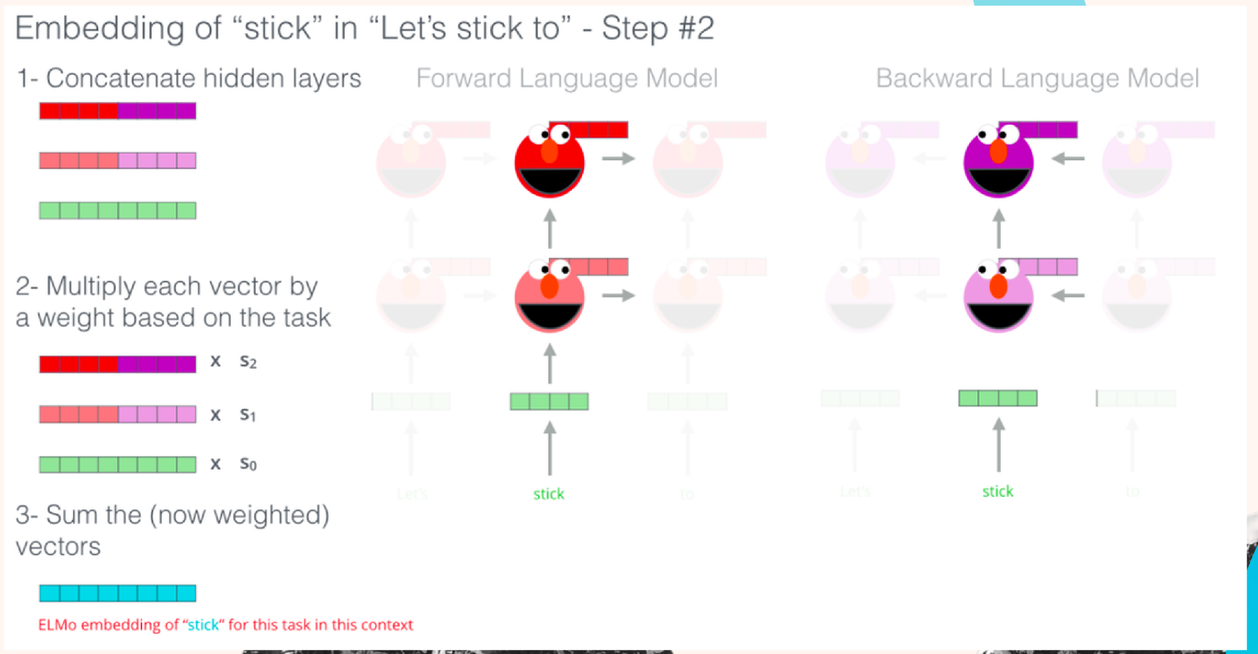
LlaMA 2

Modelos basados en transformers entrenados a gran escala, inferencia con una ventana de contexto extendida



Los Embeddings

Recovered from “BERT” and “Attention is All You Need” papers



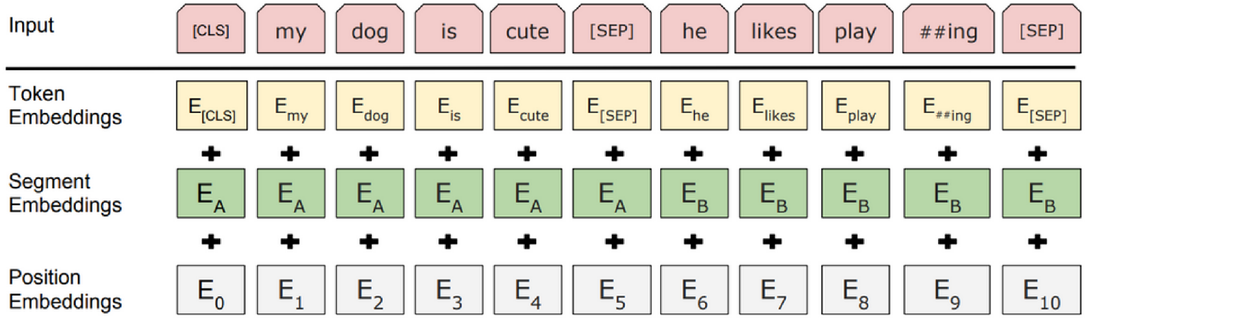


Figure 2: BERT input representation. The input embeddings are the sum of the token embeddings, the segmentation embeddings and the position embeddings.

Useful Resource

Los Embeddings

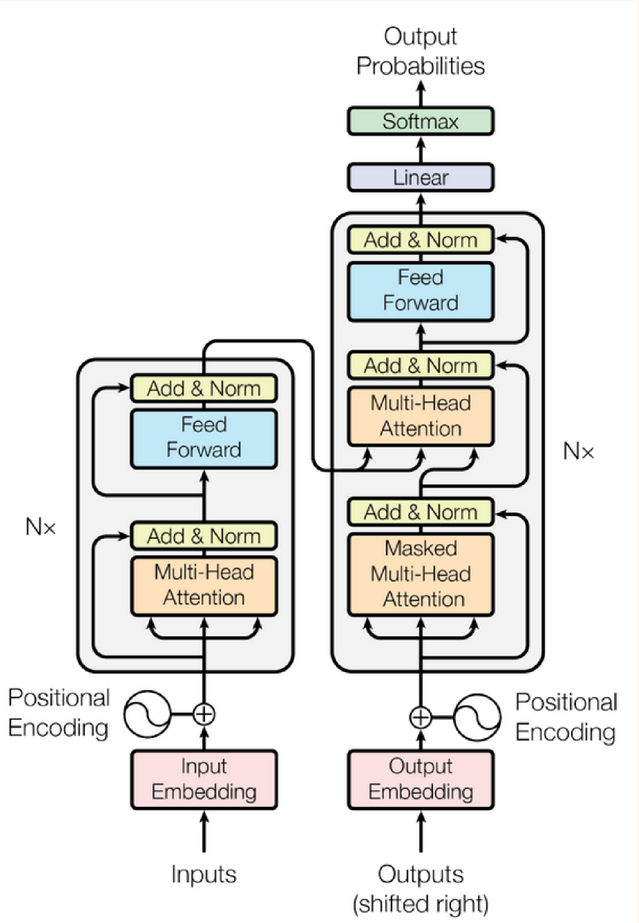


Figure 1: The Transformer - model architecture.

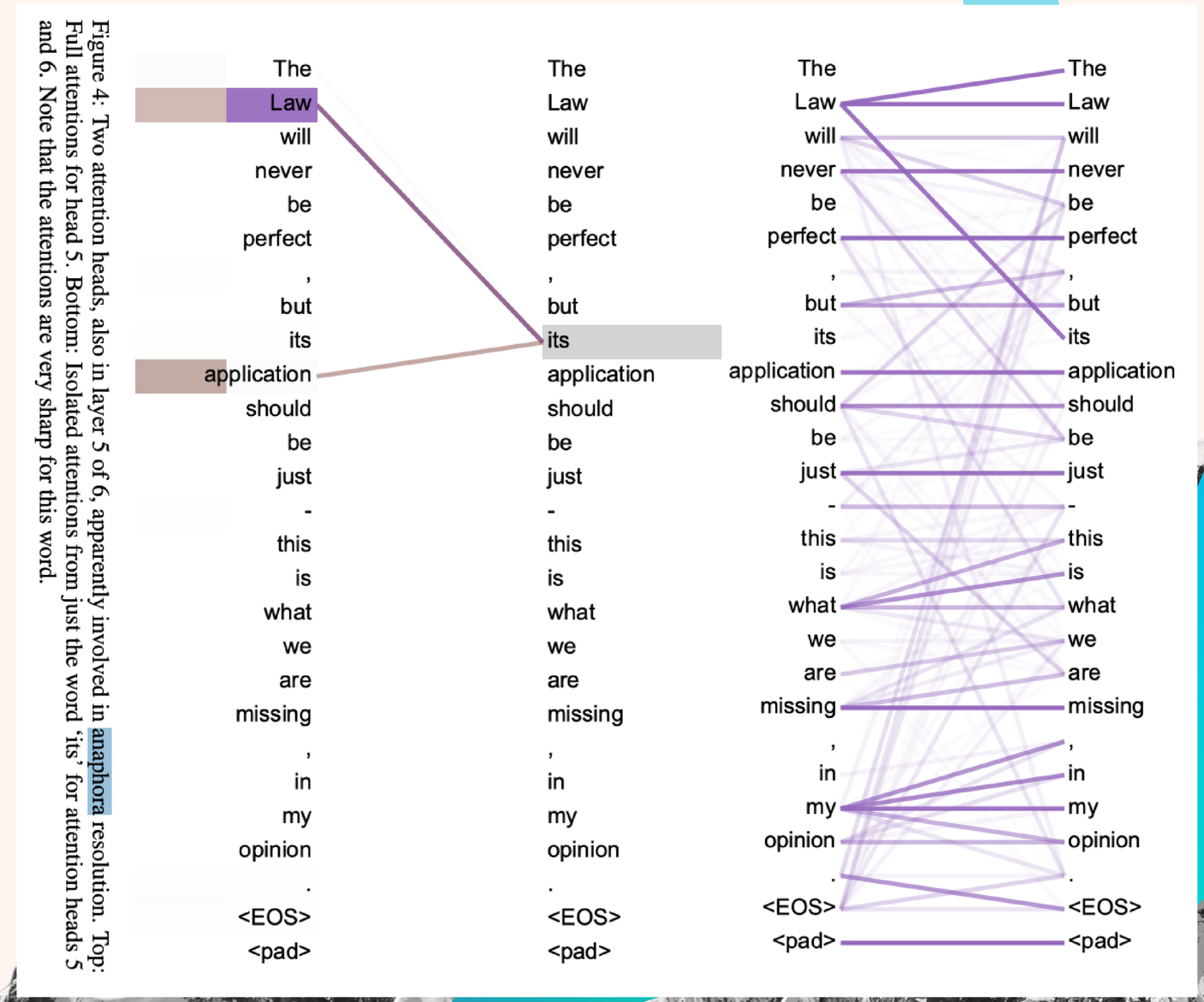


Figure 4: Two attention heads, also in layer 5 of 6, apparently involved in anaphora resolution. Top: Full attentions for head 5. Bottom: Isolated attentions from just the word 'its' for attention heads 5 and 6. Note that the attentions are very sharp for this word.

Fuentes: Diccionarios

Ministerio de Educación del Ecuador

A

achachay [ačačay, ačačaw, atsatsay] interj. *expresión de frío.* Yapa chiri kakpi rimay.
*Wawakunaka **achacha**y nishpa chayamun.*
Sin. Chiri chiri.

achachaw [ačačaw, ačučuy] interj. amz. *expresión de calor.* Yapa rupay tiyakpi rimay.
***achachaw,** mikunaka rupakmi kashka.*
Sin. Araray, rupakuk.

achka [ačka, ačika, ašką] adv. *bastante, harto, mucho.* Imatapash tawkata, tawka tiyakta, mana ashallata rikuchik.
*Chakramantaka **achka** saratami pallarkani.*
Sin. Ashtaka, hatunta, pachan, tawka, llashak.

*Asha kunuklla allpapimi **achira**taka tarpun.*

achukcha [ačukča, ačugča, ačuxča] s. *achogcha.* Purutushina ankushpa pukuk, waylla muru, shunkupi yana muyuyuk, yanushpa mikuna yuyu.
*Wasipika **achukcha** lukruta mikurkani.*

achupalla [ačupalʸa, ačupaža, ačupilʸa] s.s. *achupalla*. Kashayuk panka, uchilla chawarshina, sunilla amuk pankayuk yura.
*Allpa saywapimi **achupalla** yurataka tarpuni.*

ahana [axana] v. ofender, insultar, denigrar. Pitapash piñachishpa rimana.
*Yayaka yankamanta, wawatami **aha**n.*
Sin. kamina, takurina.

AKLLANA

*Kunka nanakpika **akirinri**ta timpuchishpa upyarka.*

akllana [aglʸana, agžana, aŋlʸana] v. *elegir, escoger, seleccionar*. Allita rikushpa hapina.
*Mamaka kamchankapakka raku saratami **akllan**.*

*Mapa allpawan pukllakpika **akwas**mi llukshin.*

2. s. ***pelusa***. Wakin kurukunapa, pankakunapa kashalla ñutu millma.
*Kaspita llakllakpika **akwas**mi llukshin.*
Sin. 1 Izi, shikshi.

Resultados iniciales

```
Page 1 Text:
Castellano – Kichwa

========================================
Page 2 Text:
A
 a cambio de, adv. ranti, rantimpa.
a continuación, adv. kipa; chaymanta.
a diario, adv. punchanta punchanta, pun-
chantin.
a gusto, adv. ninantak.
a tiempo, adv. llikchalla, kachka.
ábaco qichwa (objeto), s. yupana.
abajo, adv. uray, wayku.
abandonado, adj. sapalla, sakishka, shi-
tashka, hichushka.
abandonar, v. sakina, hichuna, shitana.
abdómen, s. wiksa.
abeja,  s.  wayrunku,  chullumpi,  putan
chuspi.
abismo, s. kaka.
ablandarse, v. llampuna, apiyana.
abonanzar, v. kasiyachina.
abonar, el terreno, s.  wanuna.
abono, s.  wanu, isma.
...
willka s. rito. 2 adj. sagrado, divino. 3 sust.
 230
```

```
1554    VERTICAL.–Shayak. VESÍ
1555    CULA BILIAR.–Jayak, chinkilis, ayak muyu. VÍA FÉ
1556    RREA.–Jillayñan.
1557    VIA.–Ñanpi.
1558    VIAJAR.–Purina, rina.
1559
1560    VIENTO.– Wayra.
1561    VIENTRE.– Wiksa.
1562    VIERNES.–Chaska. VIGÉ
1563    SIMO.– Ishkaychunka niki.
1564    VINCHA.–Warmi umawatariy.
1565    VINI.–Wasra.
1566    VIOLAR.– Wakllichina, pakina.
1567    VIOLIN.–Llikilliki.
1568    VIRGEN.– Akllawarmi. Mana wak llishka warmi.
1569    VISITAR.– Pasyana, purina.
1570    VIUDA.–Wakcha, karillak.
1571    VOCABLO.–Shimi.
1572    VOCAL.– Uyari.
1573    VOLAR.–Pawana, wampurina. VOLCÁ
1574    N.–Urku, ninayuk urku.
1575    VOLUNTAD.–Munay, ari nina, yu yaywan nina.
1576    VOMITAR.–Kiwnana, kiknana, quicnana.
1577    VOZ.– Uyachi, shimi.
1578    VUELO.–Paway, huanburina, pa huana.   Y y Z z
```

Resultados finales

```
The first dictionary has 2066 entries.
The dictionary entries have 8838 entries.
The kichwa texts dataset has 2271 entries.
The Spanish hf dataset has 51942 entries.
```

```python
function BytePairEncoding(text, vocab_size)
    vocab = get_initial_vocab(text)
    token_freqs = count_token_freqs(text, vocab)
    while size(vocab) < vocab_size
        most_freq_pair = find_most_freq_pair(token_freqs)
        if most_freq_pair is None
            break
        new_token = merge_pair(most_freq_pair)
        vocab.add(new_token)
        text = replace_pair_in_text(text, most_freq_pair, new_token)
        update_token_freqs(token_freqs, most_freq_pair, new_token)
    return text, vocab
```

tputs > JSON > Shuffled_JSONs > {} cas_kich_SHUFFLED.json > ...

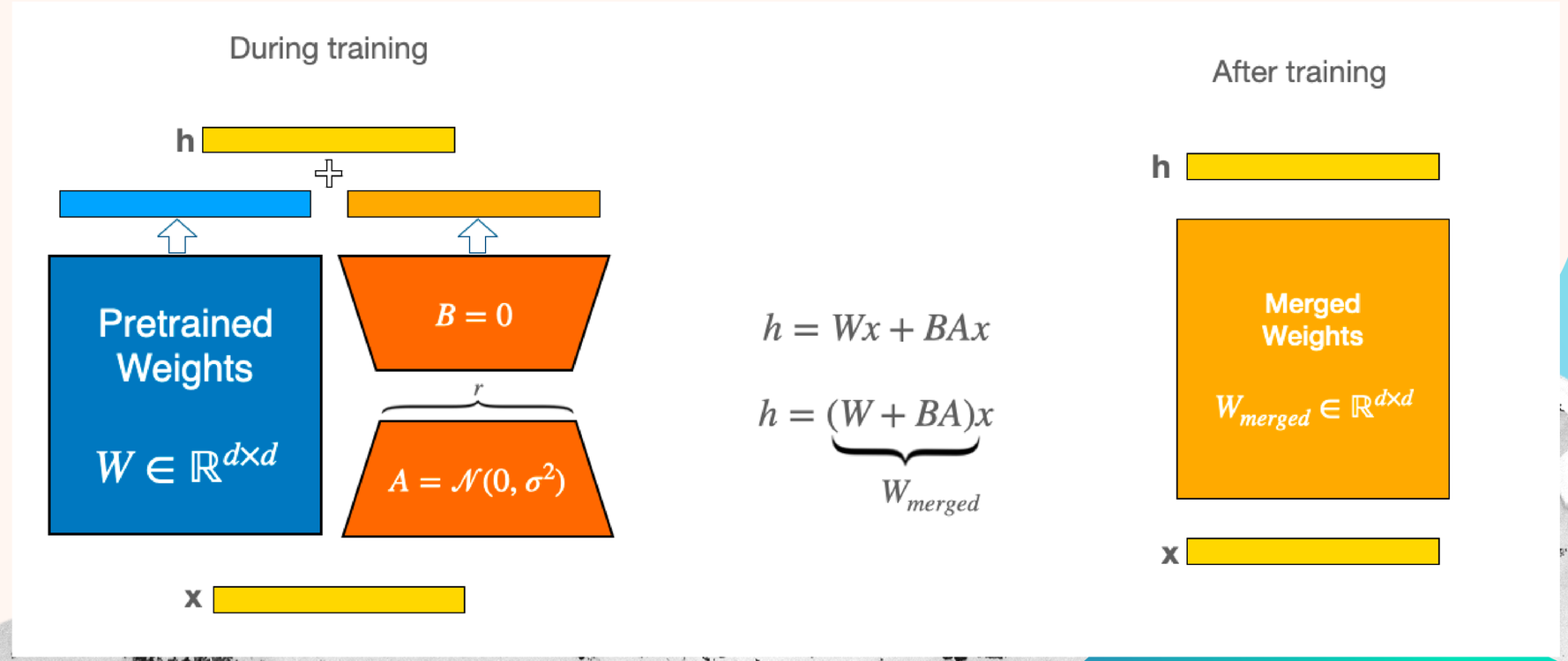

```json
[
    {
        "Instrucción": "¿Cuál es el sinónimo en Kichwa de nuca?",
        "Entrada": "nuca",
        "Salida": "s. washa kunka."
    },
    {
        "Instrucción": "¿Cómo se traduce la palabra española moverse al Kichwa?",
        "Entrada": "moverse",
        "Salida": "v. kuyuna; en vaivén: kawirina."
    },
    {
        "Instrucción": "¿Cómo se interpretaría acequia en Kichwa?",
        "Entrada": "acequia",
        "Salida": "s. larka; hacer acequias: larkana."
    },
    {
        "Instrucción": "¿Cómo traducirías el español entrar a Kichwa?",
        "Entrada": "entrar",
        "Salida": "v. yaykuna."
    },
    {
        "Instrucción": "¿Cómo se traduciría muslo al Kichwa?",
        "Entrada": "muslo",
        "Salida": "s. raku chanka, mama chanka."
    },
    {
        "Instrucción": "¿Cómo se traduciría de nuevo al Kichwa?",
        "Entrada": "de nuevo",
        "Salida": "adv. kutin."
    },
    {
        "Instrucción": "Traduce wampuru al Kichwa.",
        "Entrada": "wampuru",
        "Salida": "s. calabaza."
    },
    {
        "Instrucción": "¿Cuál es el sinónimo en Kichwa de kachichaska?",
        "Entrada": "kachichaska",
        "Salida": "s. cosa salada."
    },
```

LoRA training

```python
from peft import LoraConfig

# Target modules - either specific ones or all linear layers
target_modules = ['q_proj', 'k_proj', 'v_proj', 'o_proj']

lora_config = LoraConfig(
    r=16,  # Rank of the low rank matrices
    target_modules=target_modules,
    lora_alpha=8,
    lora_dropout=0.05,
    bias="none",
    task_type="CAUSAL_LM"
)
```



Entrenamiento, Evaluación y Resultados

Modelo	Dataset	Alpha	Rank	Projections	Learning Rate	Steps	Loss	SISA Benchmark*
Llama 2 7B	MoreDetail	4096	2048	q, v, k, o	1,60E-05	3051	2,90E+14	3,3%
Llama 2 7B	MoreDetail	4096	2048	q, v, k, o	1,60E-05	5043	2,90E+12	17,8%
Llama 2 7B	AllData	2048	1024	q, v	2,00E-04	7017	6,5	0,6%
Llama 2 70B	Grammar_guidelines	512	256	q, v,k	9,10E-05	191	2,9	11,7%
Llama 2 70B	Grammar_guidelines	2048	1024	q, v	1,80E-04	431	2,4	0,0%
Llama 2 70B	MoreDetail	2048	1024	q, v	1,70E-04	1863	0,47	6,1%
Llama 2 70B	MoreDetail	2048	1024	q, v	1,10E-04	2503	0,39	88,9%
Llama 2 70B	MoreDetail	2048	1024	q, v	1,00E-05	3727	0,36	66,7%
Llama 2 70B	AllData	1024	512	q, v	1,70E-06	19235	0,18	80,6%
Llama 2 70B	AllData	2048	1024	q, v	3,10E-05	19235	8,6	0,0%
GPT Builder	AllData	n/a	n/a	n/a	n/a	n/a	n/a	95,0%

¿Ahora qué?

- Automated Evaluation
- Reinforcement Learning from Human Feedback Modeling
- Dataset further inspection
- Contrastive evaluation with other Parameter-Efficient Fine-tuning Techniques

YUPAYCHANI

James León
The URKU Project