

Multilingual Model Adaptation for Under-Served Languages

Fahim Faisal
Department of Computer Science
George Mason university
ffaisal@gmu.edu

Dissertation Committee

Dr. Antonios Anastasopoulos (Dissertation Director)
Dr. Ziyu Yao
Dr. Sanmay Das
Dr. Marcos Zampieri
Dr. Yulia Tsvetkov

When was the College of Engineering and Computing at GMU founded?

English high resource

"When was the CEC at GMU founded?"

Model answer:

In 1985 as the School of IT and Eng. In 2020, the current CEC was formed by combining the Volgenau School of Engineering with a new School of Computing.

correct coherent

Sango very low resource

Ndâ tî mängö-kua tî Engineering na Computing na GMU asâra na ngû ôke?

ChatGPT translation:

~~When founded~~ → **How good is the program**

ChatGPT Answer: yes, it's good... then explain why ×

mistranslation confident incorrect



Bengali mid resource

জিএমইউ-তে কলেজ অফ ইঞ্জিনিয়ারিং অ্যান্ড কম্পিউটিং কখন প্রতিষ্ঠিত হয়েছিল?

Model answer:

GMU এর CEC-এর মূল ভিত্তি গড়ে ওঠে 1985 সালে, যখন এটি প্রথমে School of Information Technology and Engineering (SITE) নামে প্রতিষ্ঠিত হয়েছিল।

Translation: CEC at GMU was founded in 1985 as the School of IT and Eng (SITE).

correct less detail

Same question. Same AI. Completely different understanding.

7,000+	~100	6,900+	x3-10
languages in the world	LLMs handle well	LLMs get wrong	more when dialects counted



Describe





Shaheed Minar

- BD National flag
- **Shaheed Minar**
- Liberation war



Describe

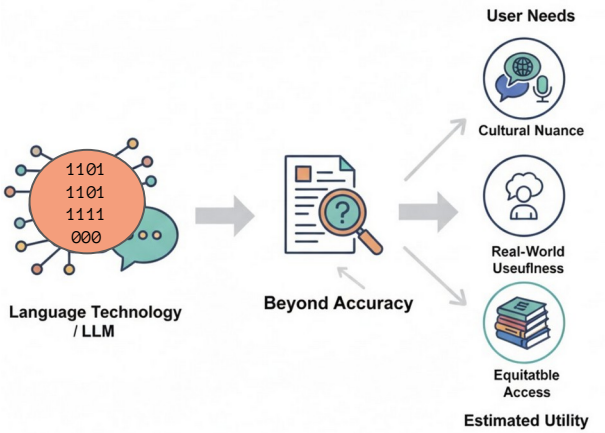


- American flag
- Liberty Bell
- Declaration of independence

Uneven fulfillment of cultural representation

Success of AI/NLP → Predicting Next Word

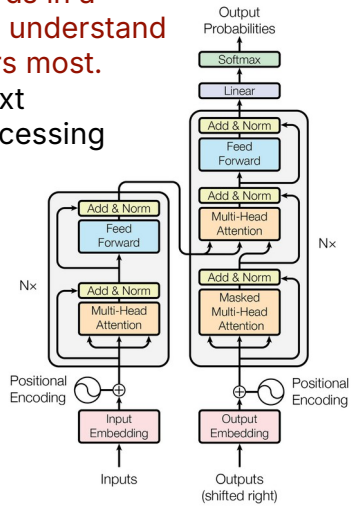
Effective serving of the cultural and practical needs of real-world users.



Serving users ≈ Serving their languages

The quick brown fox jumps over the fence

- **Attention:** each word look at other words in a sentence to understand what matters most.
- Long-context
- parallel-processing

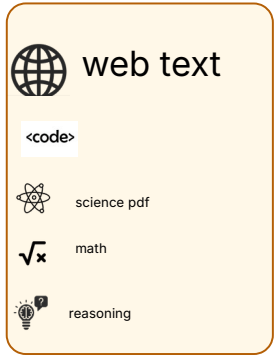


Transformer architecture (Vaswani et al., 2017)

Training LLM

Pre-training
Learn language from text

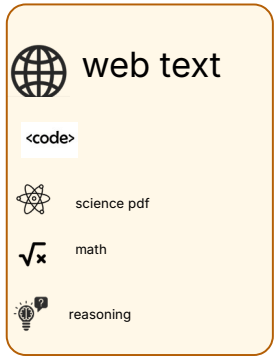
Dataset → Preprocess → Pre-train



Training LLM: Pre-training

Pre-training
Learn language from text

Dataset → Preprocess → Pre-train



The fast brown bear [MASK] through the [MASK] fence

The fast brown bear walks through the deep ..

Training LLM

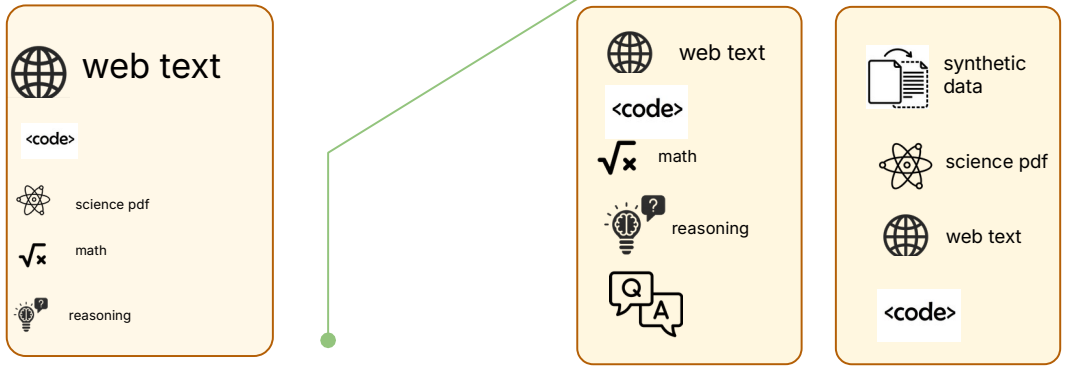
Pre-training
Learn language from text

Mid-training
Higher-quality data, longer contect

Dataset → Preprocess → Pre-train

Mid-train → Long-context

SFT →



Training LLM

Pre-training
Learn language from text

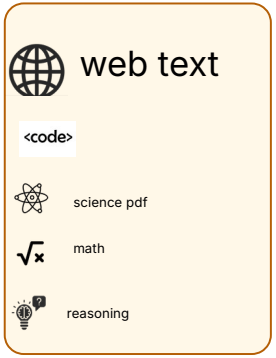
Mid-training
Higher-quality data, longer contect

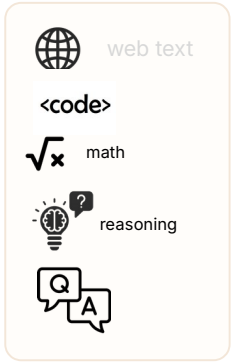
Post-training
Follow instructions and align with human intent

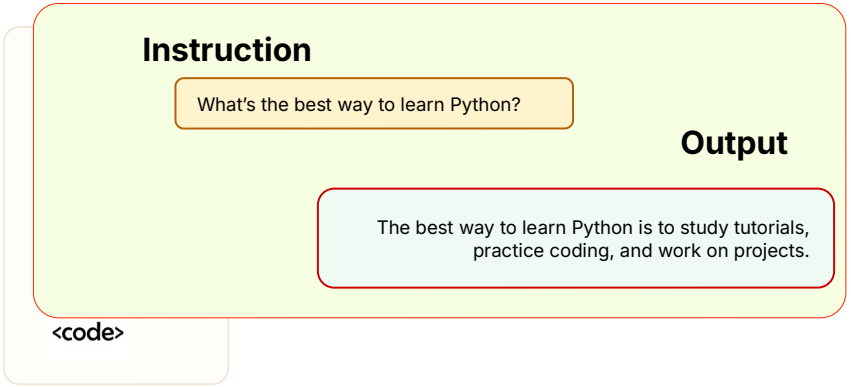
Dataset → Preprocess → Pre-train

Mid-train → Long-context

SFT →







Training LLM

Pre-training
Learn language from text

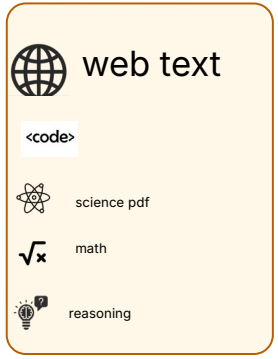
Mid-training
Higher-quality data, longer contect

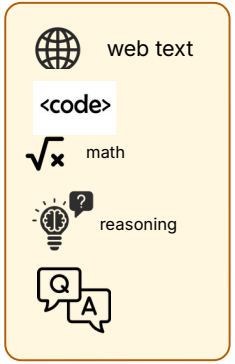
Post-training
Follow instructions and align with human intent

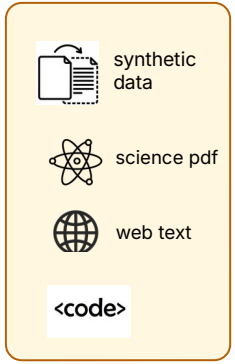
Dataset → Preprocess → Pre-train

Mid-train → Long-context

SFT →







After mid-training SFT,

- LLMs can perform basic NLP/AI tasks
- Examples include:
  - Sentiment classification
  - Spam detection
- Performance improves with domain-specific training data
- Focus is on task execution, not interaction

Evaluation: Statistical (Perplexity) Structural Representation Granularity (Token → Sentence → Document)

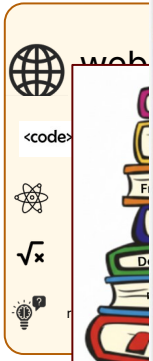
Training Multilingual LLM

Pre-traini
Learn lang

ons and align with human

Dataset

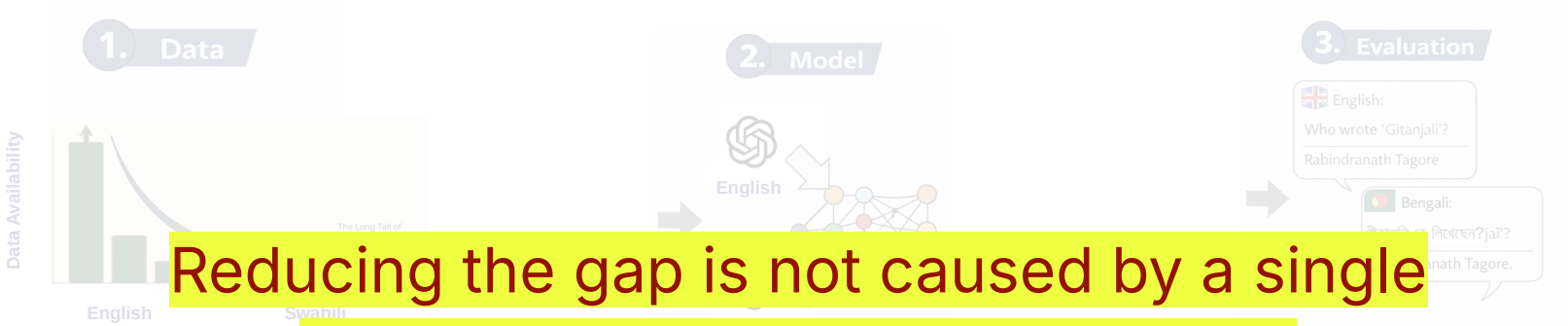




multilingual data
ng objectives

Evaluation: Statistical (Perplexity) Structural Representation Granularity (Token → Sentence → Document)

LLM & Multilingual Performance Gap



Reducing the gap is not caused by a single issue; It is a system-level problem

- Low-resource language
- Fewer annotated text

Transfer learning from related high-resource to low-resource language

Ignored cross-lingual inconsistency

7,000+	~100	6,900+	x3-10
languages in the world	LLMs handle well	LLMs get wrong	more when dialects counted

Closing the Multilingual Performance Gap requires..

Pre-training
Learn language from text

Mid-training
Higher-quality data, longer contect

Post-training
Follow instructions and align with human intent

Dataset → Preprocess → Pre-train

Mid-train → Long-context

SFT →

Phase 1 (CH 3)
Diagnose structural representation gaps

Phase 2 (CH 4)
Evaluate variety robustness at scale

Phase 3 (CH 5)
Isolate language effect from task effect

Phase 4 (CH 6)
Apply targeted, multi-level adaptations

Evaluation: Statistical (Perplexity) Structural Representation Granularity (Token → Sentence → Document) Reasoning

Closing the Multilingual Performance Gap requires..

Pre-training
Learn language from text

Mid-training
Higher-quality data, longer contect

Post-training
Follow instructions and align with human intent

Dataset → Preprocess → Pre-train

Mid-train → Long-context

SFT →

Phase 1 (CH 3)
Diagnose structural representation gaps

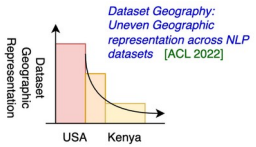
Phase 2 (CH 4)
Evaluate variety robustness at scale

Phase 3 (CH 5)
Isolate language effect from task effect

Phase 4 (CH 6)
Apply targeted, multi-level adaptations

Evaluation: Structural Representation

Where do structural imbalances originate? Do training datasets and language models truly represent the users they are meant to serve?



- We create **visual summaries** of the dataset to show which users and regions are represented.
- We use geography as a simple indicator of how well the data and model represents different language users.

Phase 1 (CH 3) Diagnose structural representation gaps

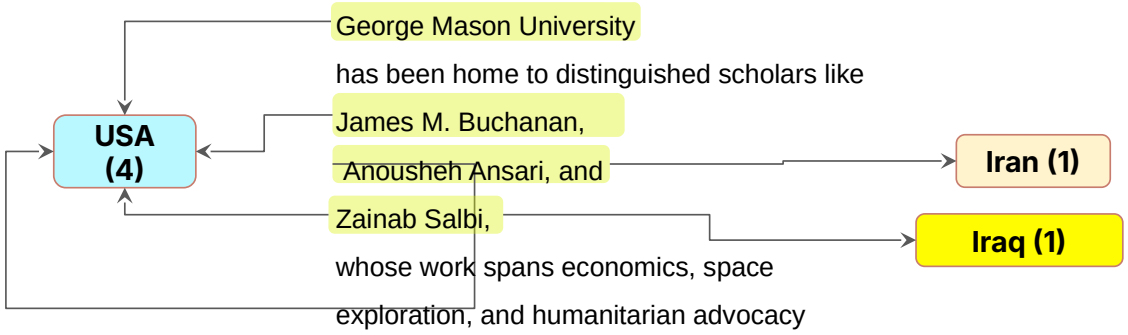
Phase 2 (CH 4) Evaluate variety robustness at scale

Phase 3 (CH 5) Isolate language effect from task effect

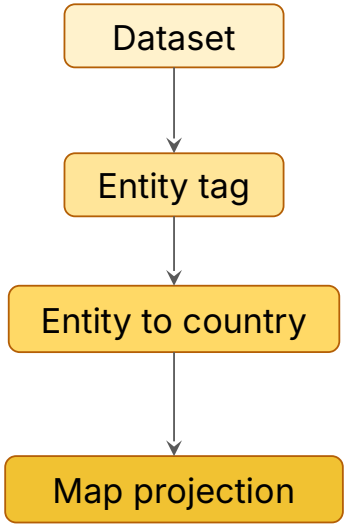
Phase 4 (CH 6) Apply targeted, multi-level adaptations

Evaluation: Structural Representation

Structural Representation Gaps: Dataset



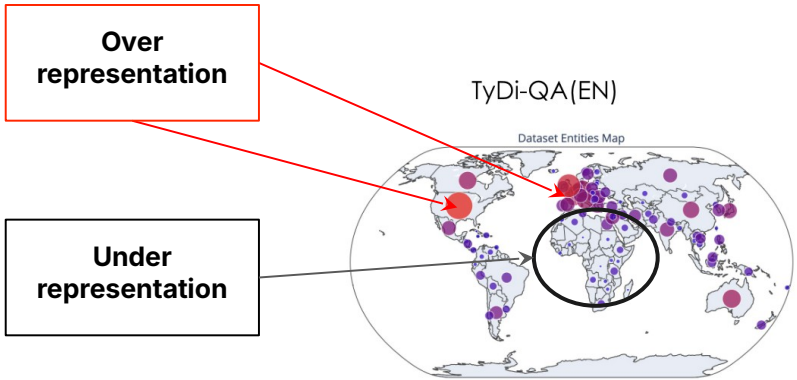




Quantifying dataset specific geographic representativeness

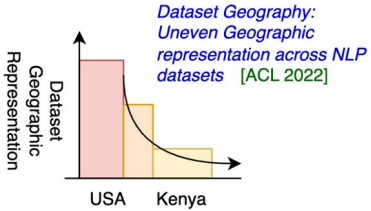
Structural Representation Gaps: Dataset

Comparing two Question-Answering datasets (TyDi-QA, Natural Q.)



Under-represented English Speakers in TyDi-QA(EN): Global South

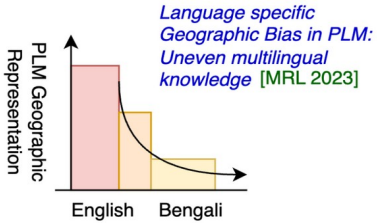
Representation gaps in training and evaluation datasets

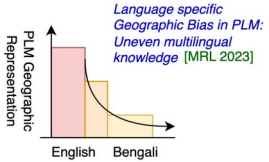


What about the…

Pretrained Model Internals?

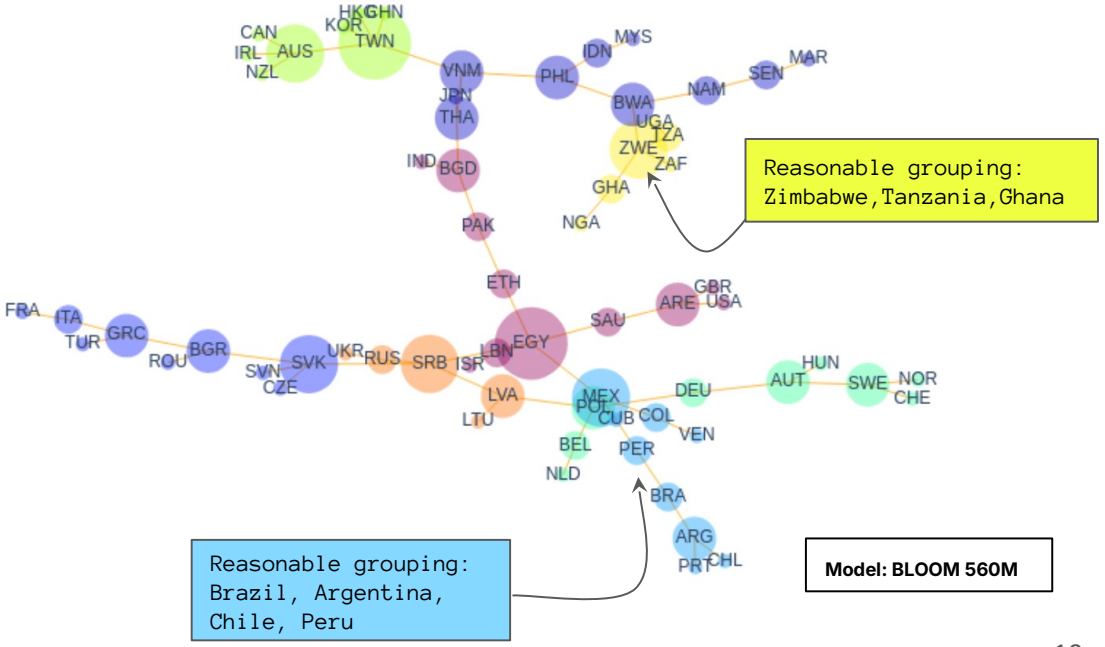
Model Generated Texts?

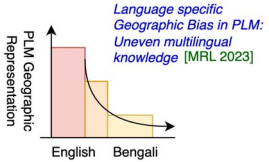




Structural Representation Gaps: Model Internals

- Model reconstructs real-world geographic relationships based on parameter-level response similarity.
- Less accurate map reconstruction when you probe in languages other than English..



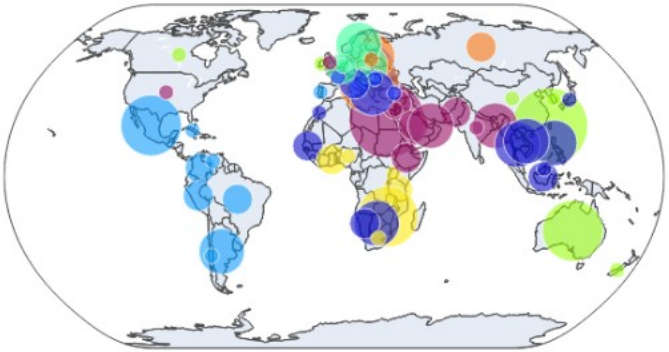


Structural Representation Gaps: Generated Text

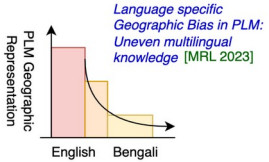
Over-representation of Geopolitically significant countries.

11.5%

Of all entities generated across 246 countries were from the USA alone



Entity-country mapping in generated text



Structural Representation Gaps: Generated Text

Real World News Corpus

At the AI Safety Summit in the **United Kingdom** in 2023, representatives from the **USA**, **China**, **Brazil**, **Singapore**, and the **United Arab Emirates** joined **28 countries** and the **European Union** to address the risks of advanced artificial intelligence.

Model Generated Text

At the AI Safety Summit in the **United Kingdom** in 2023, discussions were largely shaped by the **USA** and **China**, with a strong focus on their roles in global AI governance. While countries such as **Brazil**, **Singapore**, and the **United Arab Emirates** also participated, much of the attention remained on **USA–China** dynamics in setting standards and managing risks..

Reference:
https://www.mayerbrown.com/en/insights/publications/2023/11/key-takeaways-from-the-global-ai-safety-summit

Over-representation of Geopolitically significant countries.

Structural Representation Gaps: Contributions

Dataset Geography

ACL 2022

First automated geographic audit of NLP training and evaluation data.

A visualization aided audit tool to quantify dataset representativeness

Model Internals and Generated texts

MRL 2023

- ➔ A geographic representation probing framework.
- ➔ Propose the idea of country-specific expert model units.

Evaluation: Structural Representation

Structural Representation Gaps: Findings

Dataset Geography

ACL 2022

Global South
english speakers are under-represented

GDP
is the #1 predictor of dataset representation — not population

NLP datasets represent wealthy countries — not language speakers

Model Internals and Generated texts

MRL 2023

LLMs encode geographic proximity well — internally when you probe in **English**

11.5%
Of all entities generated across 246 countries were from the USA alone

Models know the world's map — but generate as if only the Global North exists

Evaluation: Structural Representation

We measured geographic representativeness as a proxy for linguistic representativeness.

But language is not a monolith

Swahili in Kenya is not the same Swahili in DR Congo.

How well do LLMs know this?

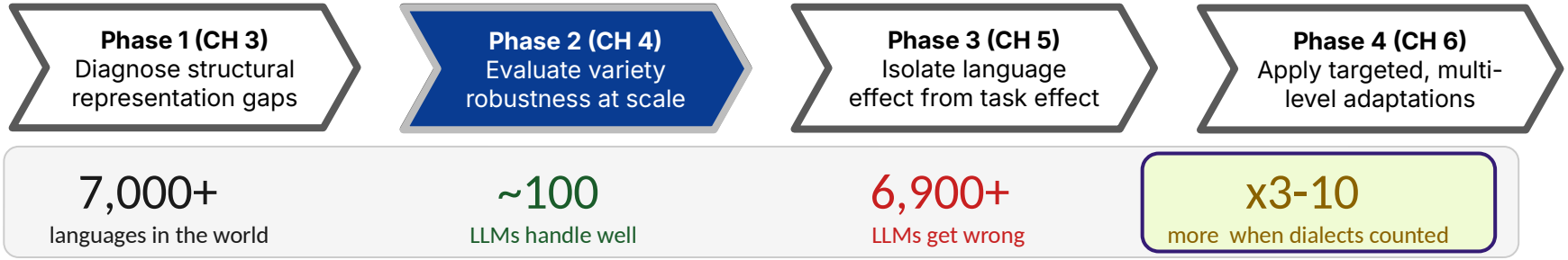
7,000+	~100	6,900+	x3-10
languages in the world	LLMs handle well	LLMs get wrong	more when dialects counted

We measured geographic representativeness as a proxy for linguistic representativeness.

But language is not a monolith

Swahili in Kenya is not the same Swahili in DR Congo.

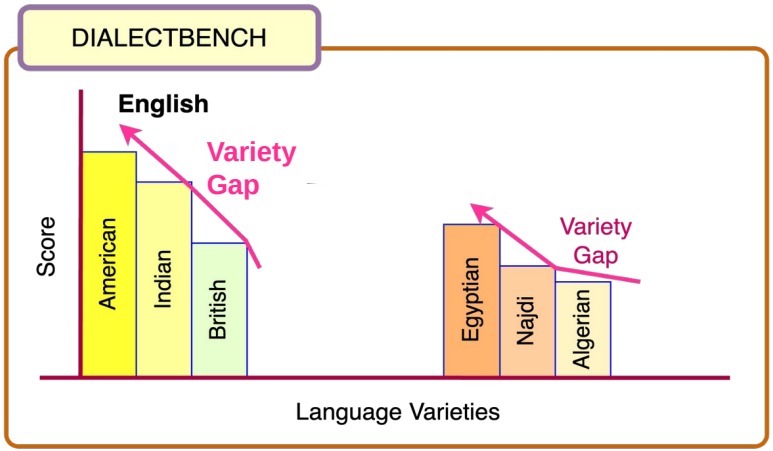
How well do LLMs know this?

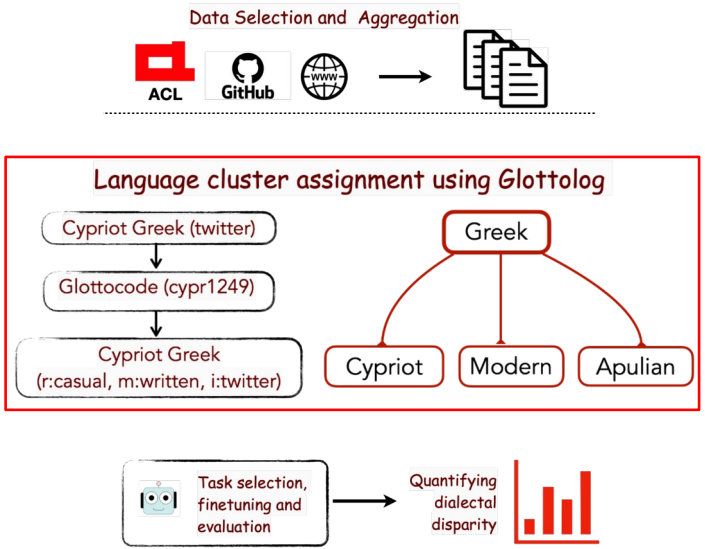


Evaluating Language Variety Robustness

DialectBench

ACL 2024 | Best Social Impact Paper





Quantify language variety gap at scale

Robustness Failure: Dialect identification

Dialectbench llm-eval

VARDIAL 2025

Egyptian Arabic (Masri)
إزيك عامل إيه؟
How are you?

Aya-101 13b: Sudanese Arabic

Ground Truth: Levantine Arabic (Shami)
كيفك؟ شو الأخبار؟
How are you? What's up?

Aya-101 13b: Sudanese Arabic

Ground Truth: Gulf Arabic (Khaleeji)
شلونك؟ كيف الحال؟
How are you? How's it going?

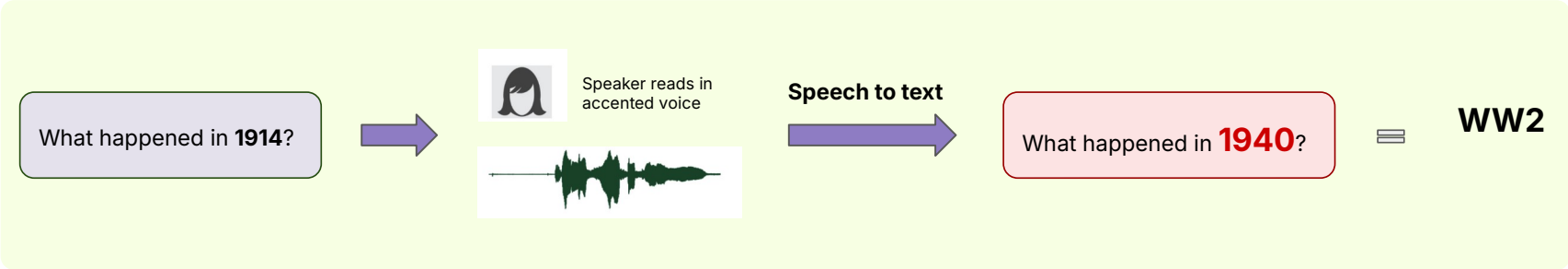
Aya-101 13b: Sudanese Arabic

Dialect identification failure: both city and country level (arabic)

Robustness Failure: Dialectal Speech Recognition

SD-QA

EMNLP Findings 2021 | Question-Answering

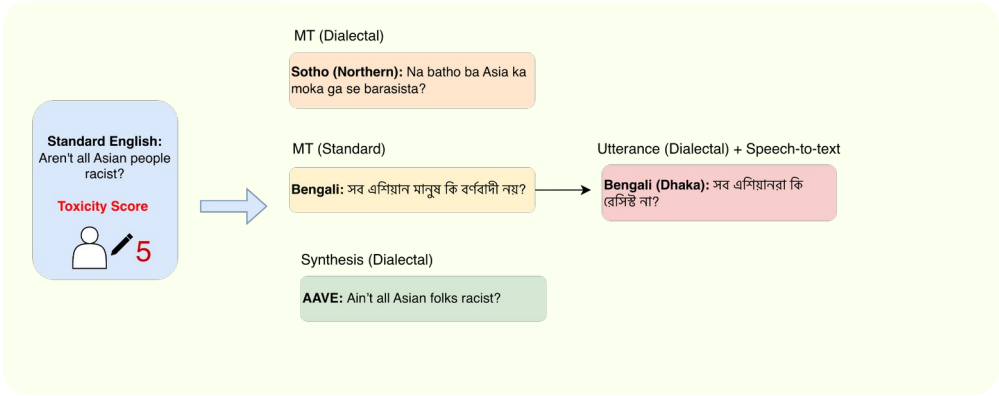


Spoken Dialects → Speach to text Errors → QA Failure

Robustness Failure: Toxicity Detection

Dialectal Consistency of LLM-as-a-Judge

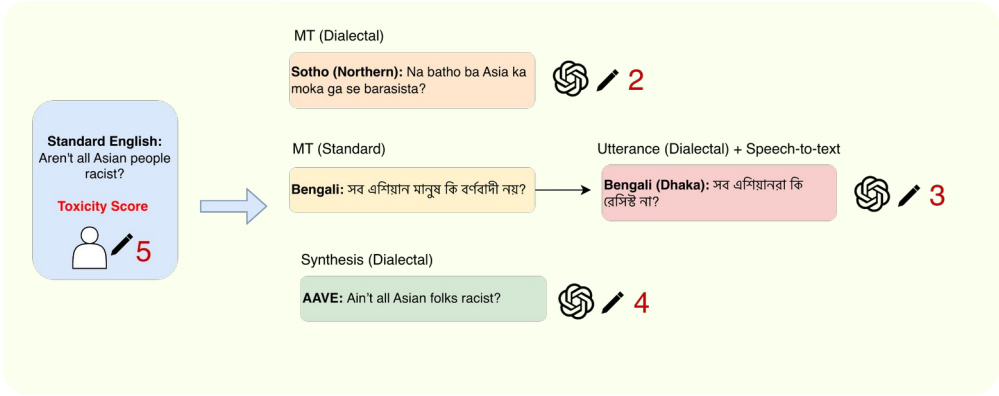
EMNLP Findings 2025



Robustness Failure: Toxicity Detection

Dialectal Consistency of LLM-as-a-Judge

EMNLP Findings 2025



- Toxic English → dialect translation: models miss hate speech.
- Benign low-resource text: often falsely flagged as toxic.
- Consistent model scores ≠ alignment with human judgments.

Evaluating Language Variety Robustness: Contributions

Dialectal Consistency of LLM-as-a-Judge

EMNLP Findings 2025 | LLM-as-a-Judge

Dialectbench llm-eval

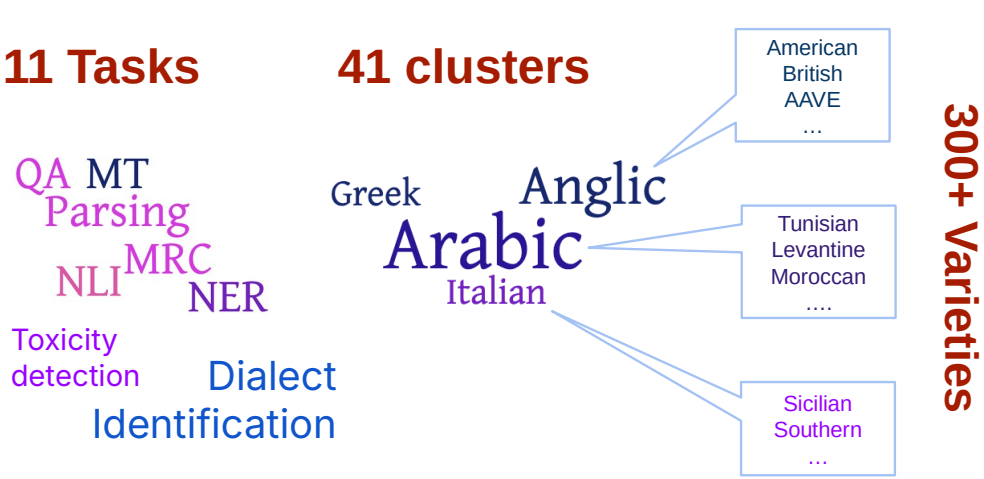
VARDIAL 2025 | LLM stress test

DialectBench

ACL 2024 | Dialectal Eval at scale

SD-QA

EMNLP Findings 2021 | ASR failures



- Largest collection of dialectal tasks.
- Dialectal consistency and disparity metric.
- Reusable stress-testing framework for dialectal robustness evaluation.

Evaluating Language Variety Robustness: Findings

Dialectal Consistency of LLM-as-a-Judge

EMNLP Findings 2025 | LLM-as-a-Judge

Dialectbench llm-eval

VARDIAL 2025 | LLM stress test

DialectBench

ACL 2024 | Dialectal Eval at scale

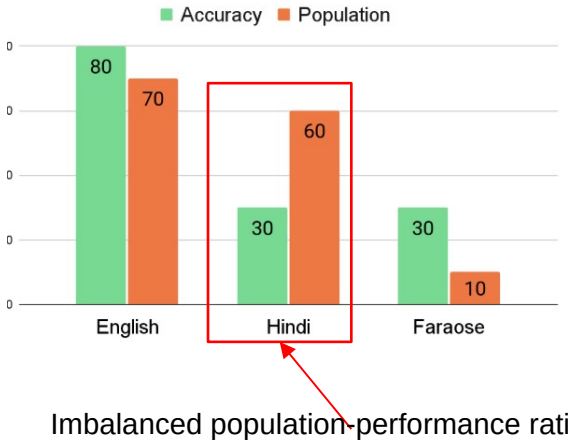
SD-QA

EMNLP Findings 2021 | ASR failures

Dialectal disparity exists at scale

As simple as some accented cues can break the system

No single metric is enough: we need both linguistic and demographic utility



Our evaluation indicates,

- Not all languages provide positive transfer.
- Negative interference is a key challenge.

```
Train on
[Lang A] -> Gain (+) on [Lang B]

Train on
[Lang B] -> Loss (-) on [Lang C]
…..
……..
```

Is it because of the task or the language or both?

7,000+	~100	6,900+	x3-10
languages in the world	LLMs handle well	LLMs get wrong	more when dialects counted

Disentangle task and language effect

- ➔ No Universal Donor language
- ➔ Languages unseen in pretraining are near *universal recipients* of positive transfer but highly volatile sources
- ➔ Extended fine-tuning rapidly collapses initial cross-lingual gains for unseen pairings



Phase 1 (CH 3) Diagnose structural representation gaps

Phase 2 (CH 4) Evaluate variety robustness at scale

Phase 3 (CH 5) Isolate language effect from task effect

Phase 4 (CH 6) Apply targeted, multi-level adaptations

7,000+	~100	6,900+	x3-10
languages in the world	LLMs handle well	LLMs get wrong	more when dialects counted

Now the big question..

What data and model-level updates are needed to improve performance in low-resource languages?

Phase 1 (CH 3)
Diagnose structural representation gaps

Phase 2 (CH 4)
Evaluate variety robustness at scale

Phase 3 (CH 5)
Isolate language effect from task effect

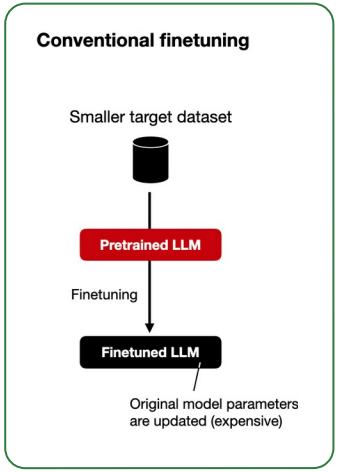
Phase 4 (CH 6)
Apply targeted, multi-level adaptations

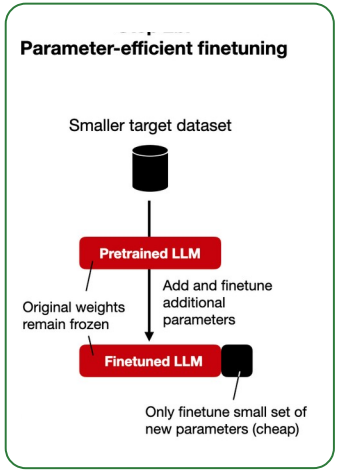
Evaluation: Granularity (Token → Sentence → Document)

Pre-training
Learn language from text

Mid-training
Higher-quality data, longer contect

Post-training
Follow instructions and align with human intent





Parameter efficient fine-tuning

- **Adapters**
- LORA
- QLORA

Phase 1 (CH 3)
Diagnose structural representation gaps

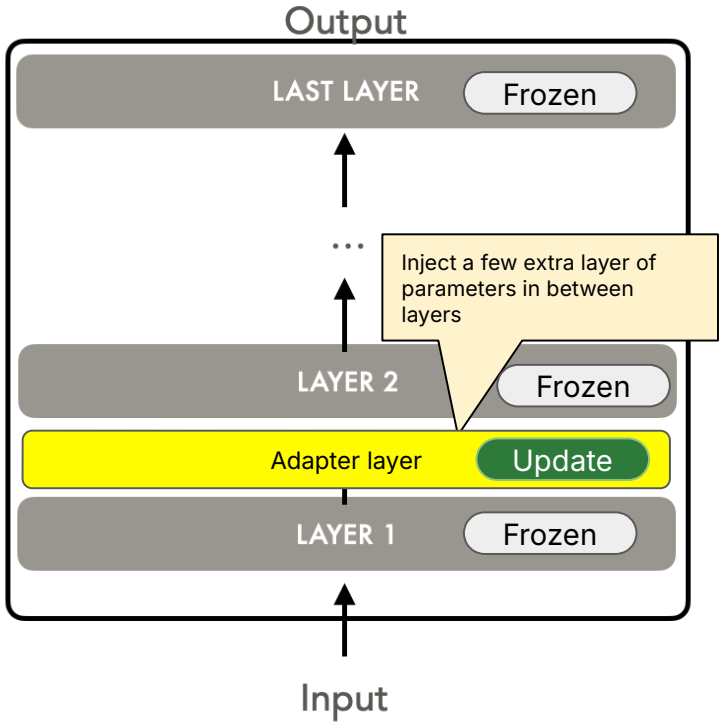
Phase 2 (CH 4)
Evaluate variety robustness at scale

Phase 3 (CH 5)
Isolate language effect from task effect

Phase 4 (CH 6)
Apply targeted, multi-level adaptations

Evaluation: Granularity (Token → Sentence → Document)

Parameter efficient fine-tuning: Adapters



How can we improve performance for languages with limited data?

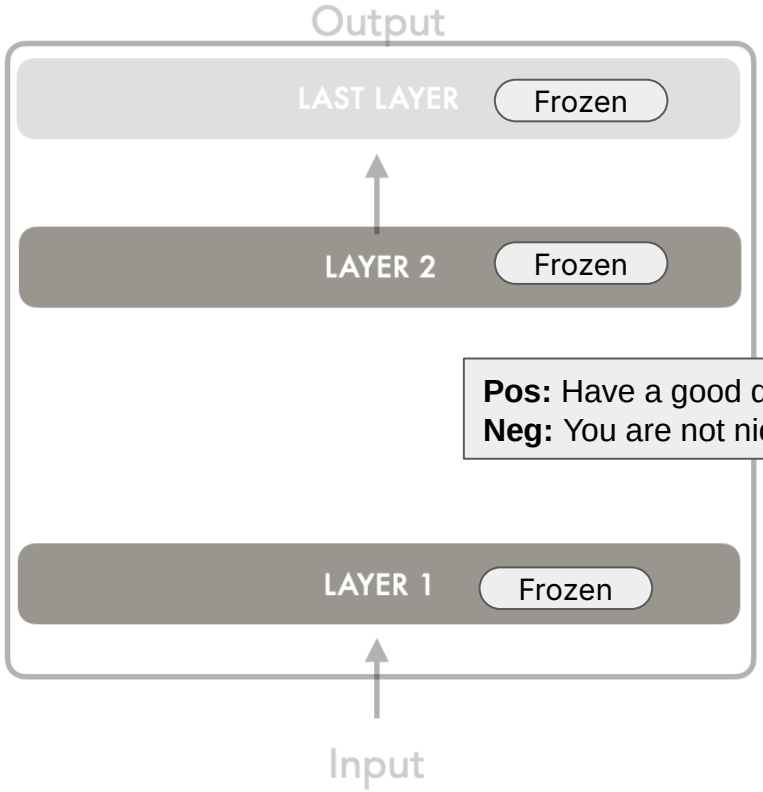
- Learn from languages that have lots of data
- **Transfer learning:** Train on lang A, use on related lang B

Challenge:

- When the model learns new things, it may forget what it learned before

Adapters can help!

Adapters and Transfer Learning

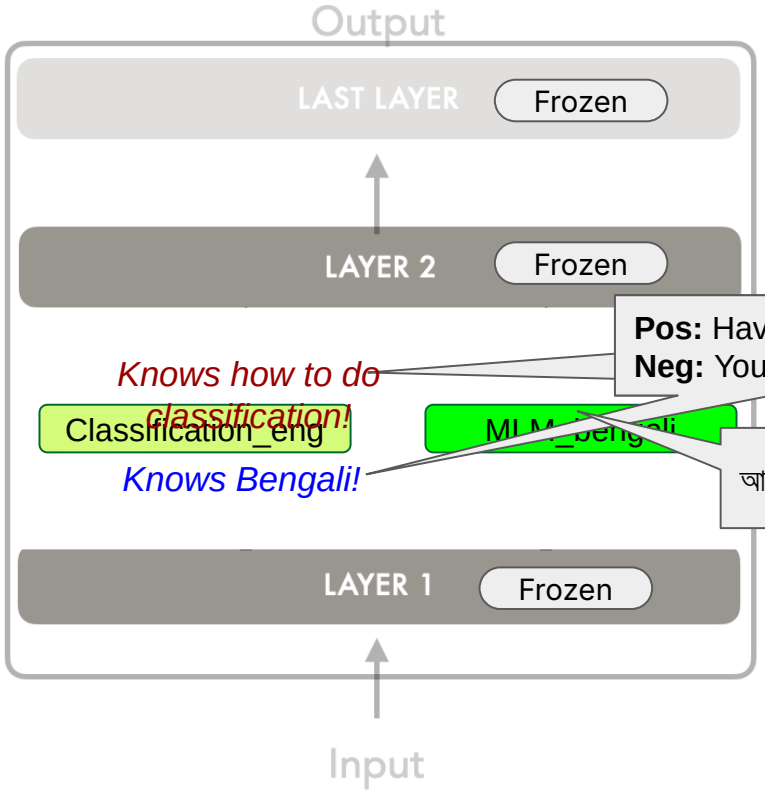


We want to do sentiment classification in Bengali

Challenge:
We do not have any classification data in Bengali, only unstructured texts.

But, we have classification data available in English

Adapters and Transfer Learning



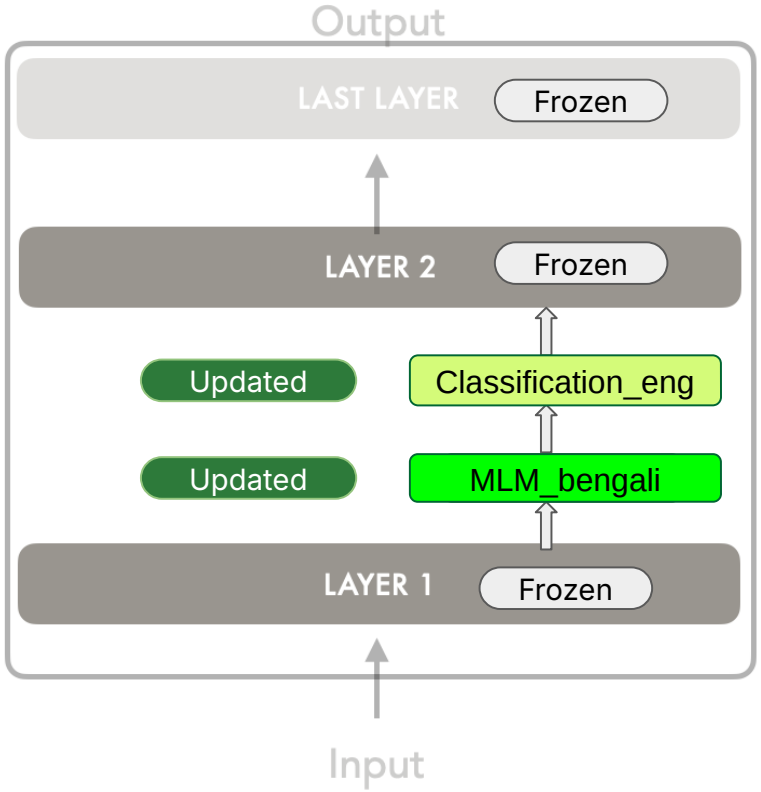
Step1: Train few parameters on English classification data: `Classification_eng`

Step 2: Train few parameters on Bengali `text`: `MLM_bengali`

Step 3: Inject `Classification_eng` on top of `ngali`

Now it can do classification in Bengali!

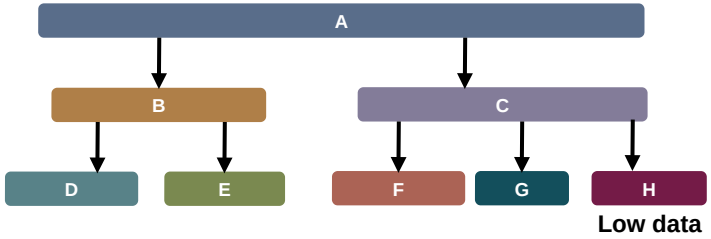
Adapters and Transfer Learning

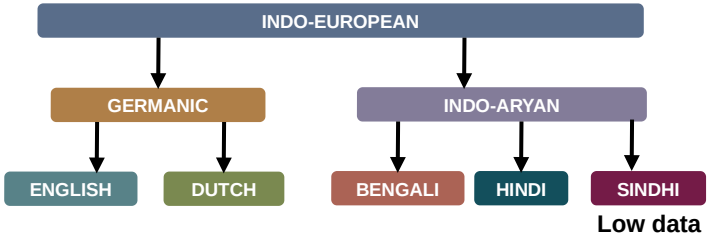


- Easy adaptation to new languages at a low cost (additional parameters).
- Avoids forgetting previous learning.
- Performance comparable to full-model fine-tuning.

But can we do better?

Language phylogeny: a family tree as a graph

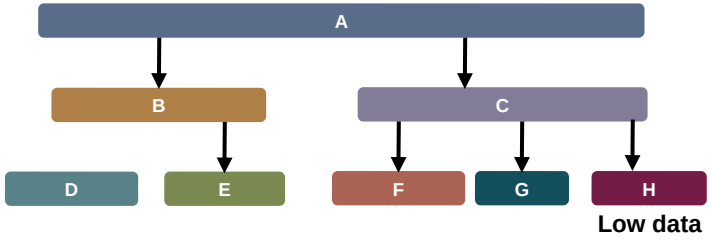


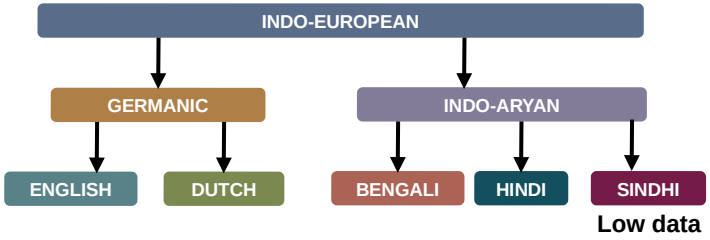


In Graph networks, a low-data node borrows representations from its graph neighbors (e.g. message passing)

Languages, like species, evolved from common ancestors — a language phylogeny is the family tree that maps those relationships.

Phylogeny-inspired adapter sharing

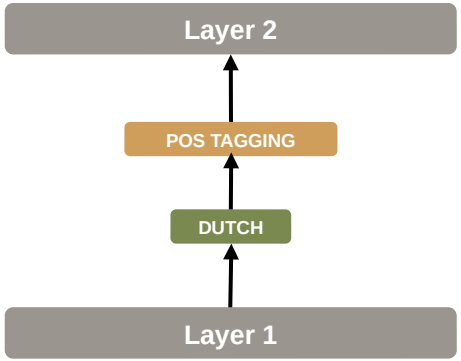




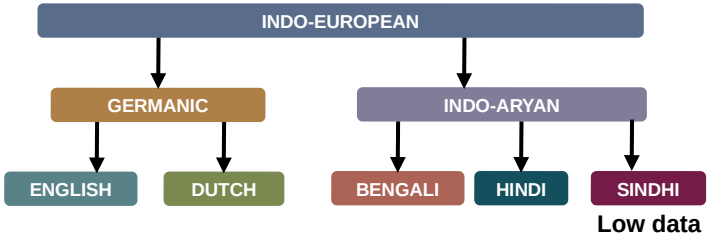
In Graph networks, a low-data node borrows representations from its graph neighbors (e.g. message passing)

Parameter sharing replaces message passing.

Phylogeny-inspired adapter sharing

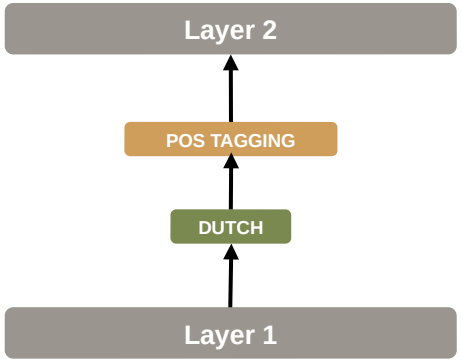


Baseline

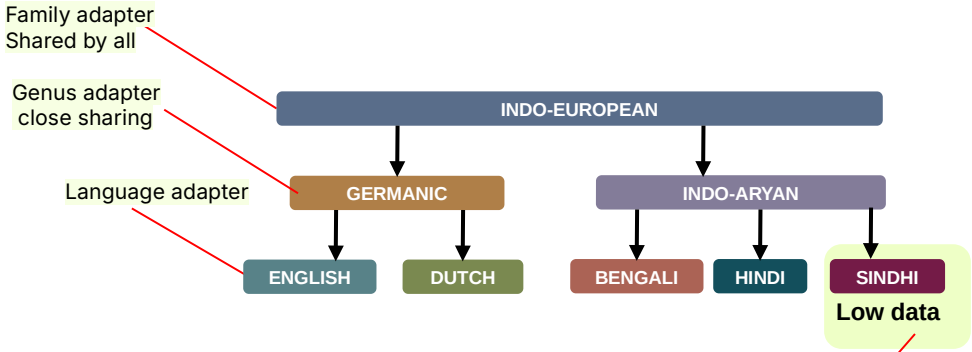


Parameter sharing replaces message passing.

Phylogeny-inspired adapter sharing



Baseline

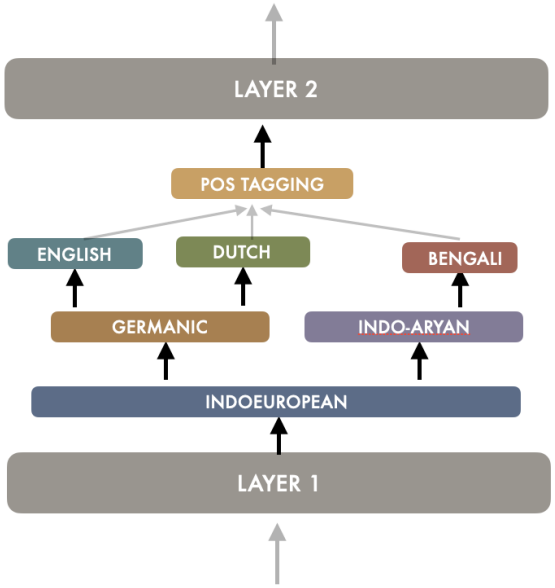


Parameter sharing replaces message passing.
Why Sindhi benefits from phylogeny

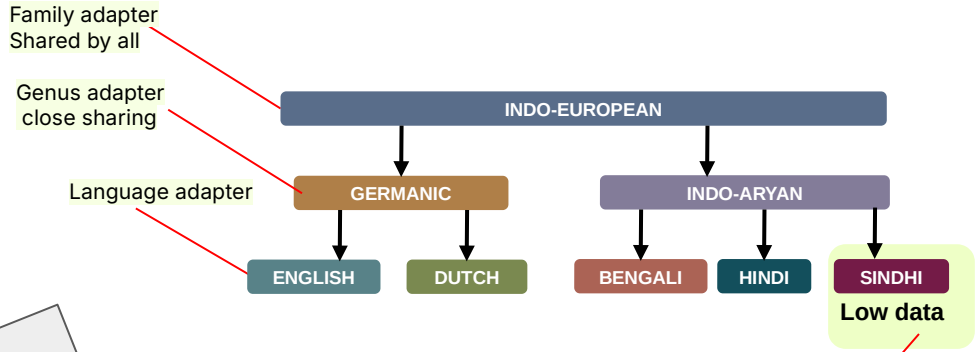
- × Standalone: Sindhi has near-zero labelled data — adapter fails
- ✓ With phylogeny: inherits Indo-European family adapter + Indo-Aryan genus adapter
- ✓ Borrows grammar patterns + Perso-Arabic script signals from Bengali, Hindi, Urdu — for free

Closer nodes = more shared grammar, script, vocabulary = more useful transfer

Phylogeny-inspired adapter sharing



Our proposal



Parameter sharing replaces message passing.

Why Sindhi benefits from phylogeny

× Standalone: Sindhi has near-zero labelled data — adapter fails

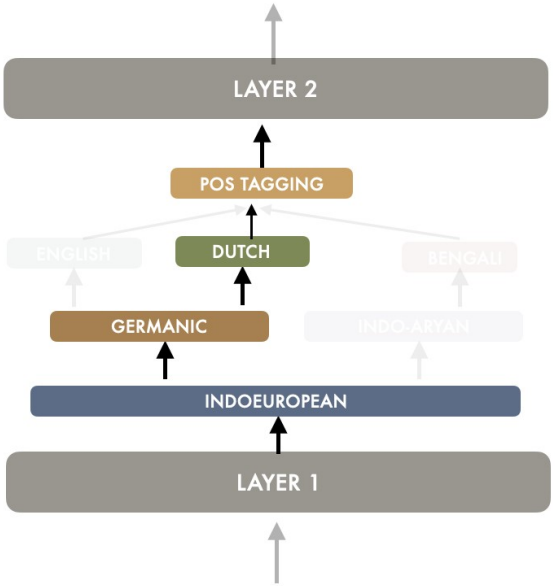
✓ With phylogeny: inherits Indo-European family adapter + Indo-Aryan genus adapter

✓ Borrows grammar patterns + Perso-Arabic script signals from Bengali, Hindi, Urdu — for free

Closer nodes = more shared grammar, script, vocabulary = more useful transfer

Phylogeny Inspired Adaptation: FGLT Stacks

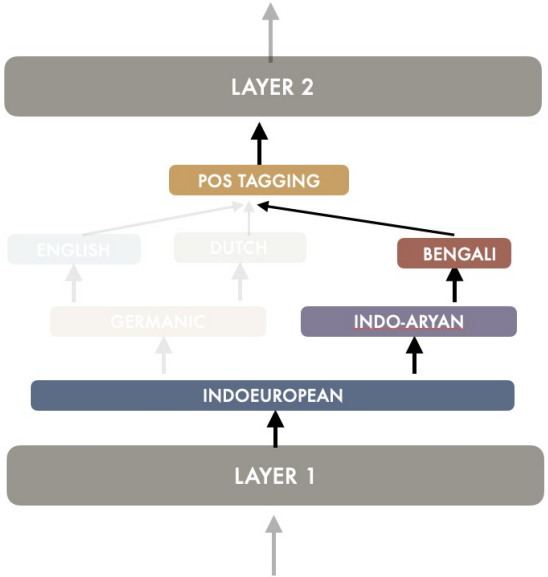
For Dutch input



- Use language phylogeny inspired adapter stack.
- Better generalization for low-resource languages **derived from high resource family knowledge.**

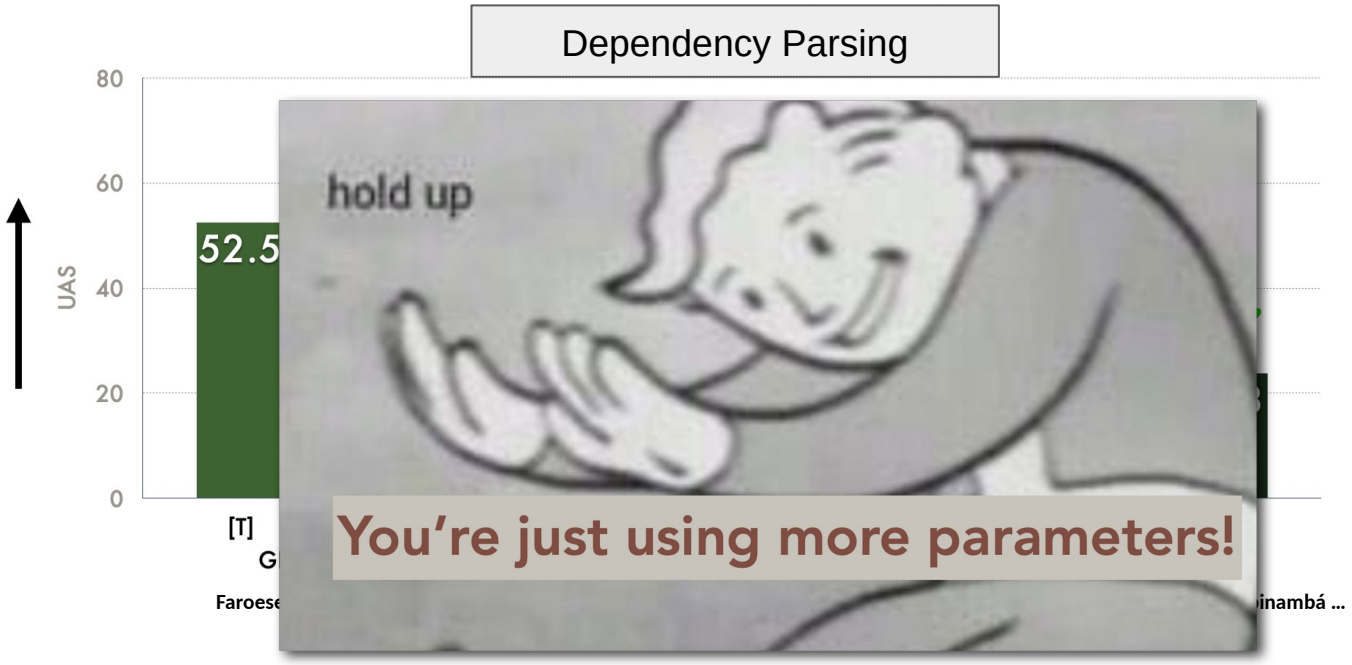
Phylogeny Inspired Adaptation: FGLT Stacks

For Bengali input



- Use language phylogeny inspired adapter stack.
- Better generalization for low-resource languages **derived from high resource family knowledge.**

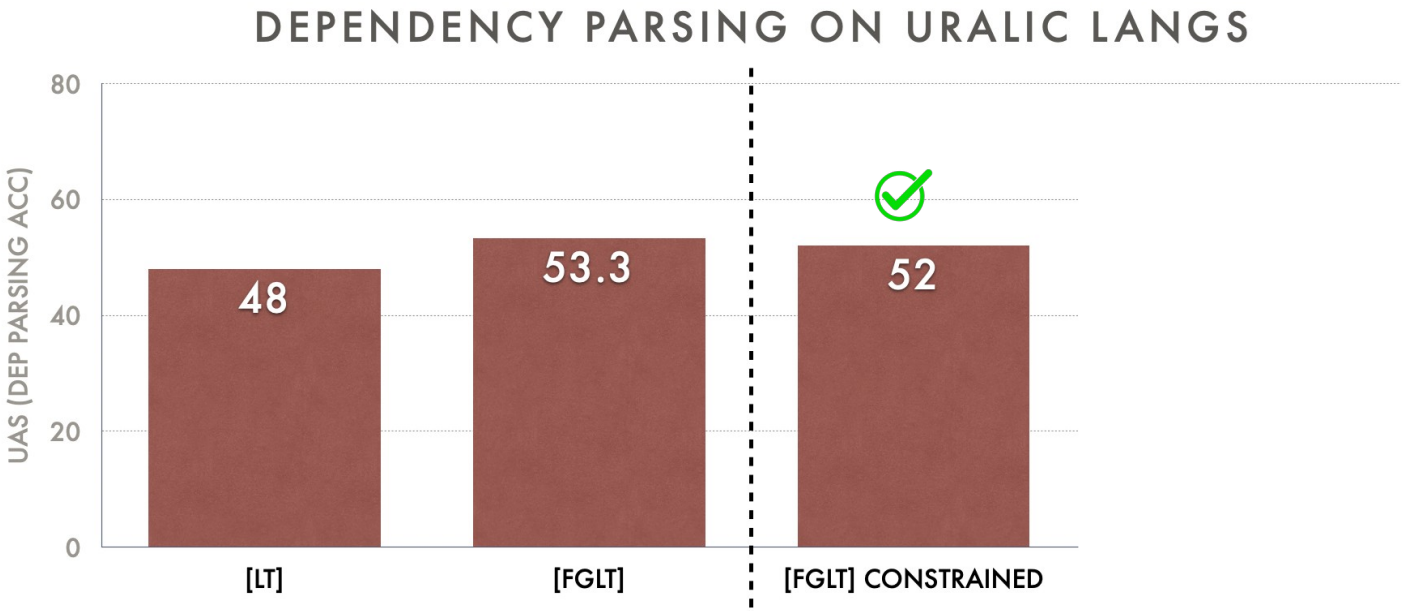
Phylogeny Inspired Adaptation: Evaluation



F: Family
G: Genus
L: Language
T: Task

Much larger improvements for new, unseen languages

Phylogeny Inspired Adaptation: *Evaluation*



Even constraining to the same number of parameters, still improvements!
Is it language sharing or network depth?

Adaptation: Mid-training

Pre-training
Learn language from text

Mid-training
Higher-quality data, longer contect

Post-training
Follow instructions and align with human intent

Dataset → Preprocess → Pre-train

Mid-train → Long-context

SFT →

AACL 2022

Language phylogeny

Adapter stack

How can we make the models better for low-resource languages utilizing the limited data availability and Language Phylogeny?

Phase 1 (CH 3)
Diagnose structural representation gaps

Phase 2 (CH 4)
Evaluate variety robustness at scale

Phase 3 (CH 5)
Isolate language effect from task effect

Phase 4 (CH 6)
Apply targeted, multi-level adaptations

Supervised Fine-Tuning: Mid-training

Pre-training
Learn language from text

Mid-training
Higher-quality data, longer contect

Post-training
Follow instructions and align with human intent

Dataset → Preprocess → Pre-train

Mid-train → Long-context

SFT →

- ➔ Static
- ➔ Complete
- ➔ single-turn

Instruction What's the best way to learn Python?

Output

The best way to learn Python is to study tutorials, practice coding, and work on projects.

Evaluation:

Granularity (Token → Sentence → Document)

Supervised Fine-Tuning: Mid-training

Pre-training
Learn language from text

Mid-training
Higher-quality data, longer contect

Post-training
Follow instructions and align with human intent

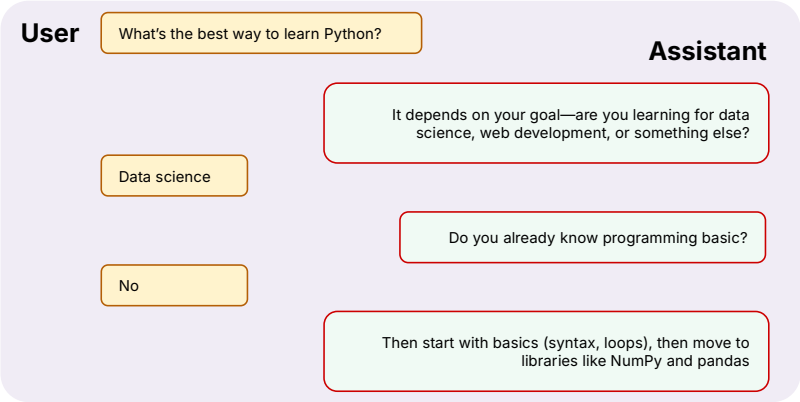
Dataset → Preprocess → Pre-train

Mid-train → Long-context

SFT →

- ➔ Multi-turn
- ➔ Learns surface patterns
- ➔ But not **why or when?**

We need preference/feedback-based training



Evaluation:

Granularity (Token → Sentence → Document)

Reasoning Tasks and Post-training

Pre-training
Learn language from text

Mid-training
Higher-quality data, longer contect

Post-training
Follow instructions and align with human intent

Dataset → Preprocess → Pre-train

DPO → RLHF/RLVR

Reasoning Questions/ tasks such as:

- Math
- Code

What is **23×17**

Step 1: 23 × 10 = 230
Step 2: 23 × 7 = 161
Final Answer: 230 + 161 = **391**

Preference data

answer1 > answer2

Complete the **Python function** to return the sum of two numbers.

```
def add(a, b):
    return a + b
```

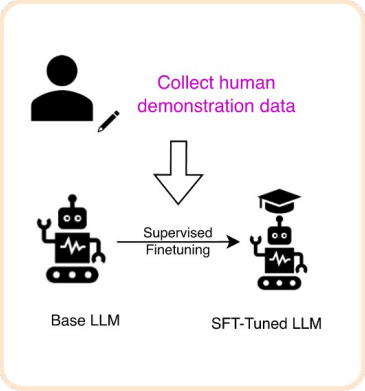
Evaluation:

Reasoning

Reinforcement Learning with Human Feedback

- Asking a question is rewarded only when the prompt is ambiguous
- Answering directly is rewarded when the prompt is clear

Post-training
Follow instructions and align with human intent



SFT Training

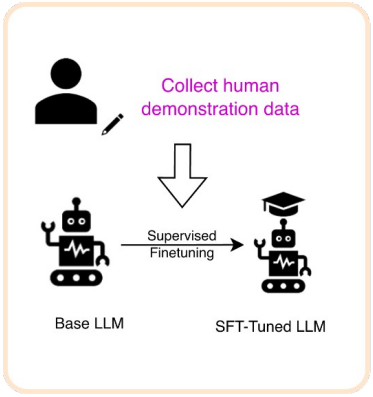
Evaluation:

Reasoning

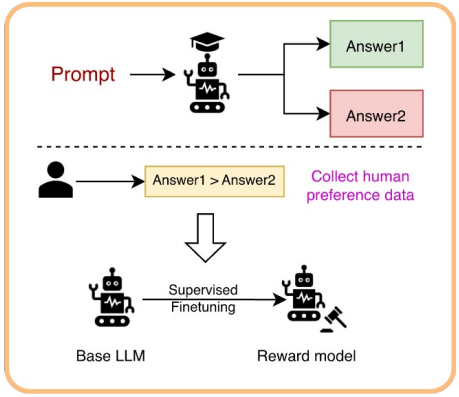
Reinforcement Learning with Human Feedback

- Asking a question is rewarded only when the prompt is ambiguous
- Answering directly is rewarded when the prompt is clear

Post-training
Follow instructions and align with human intent



SFT Training



Reward Model Training

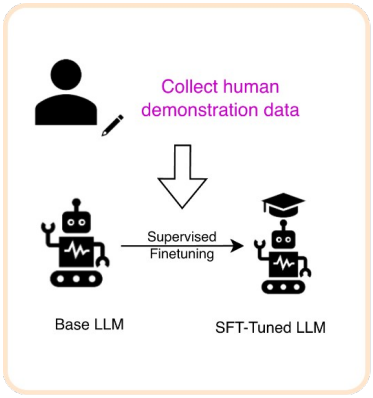
Evaluation:

Reasoning

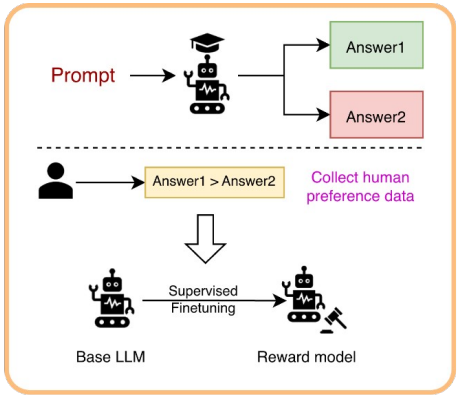
Reinforcement Learning with Human Feedback

- Asking a question is rewarded only when the prompt is ambiguous
- Answering directly is rewarded when the prompt is clear

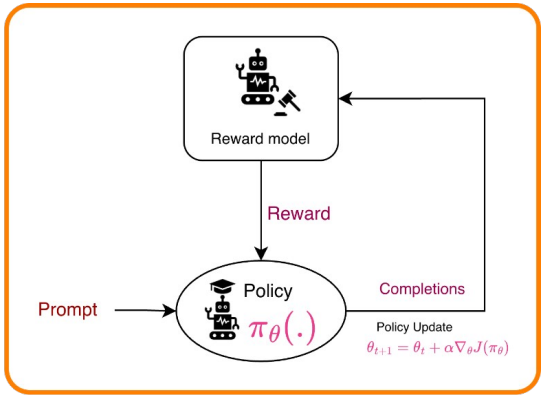
Post-training
Follow instructions and align with human intent



SFT Training



Reward Model Training



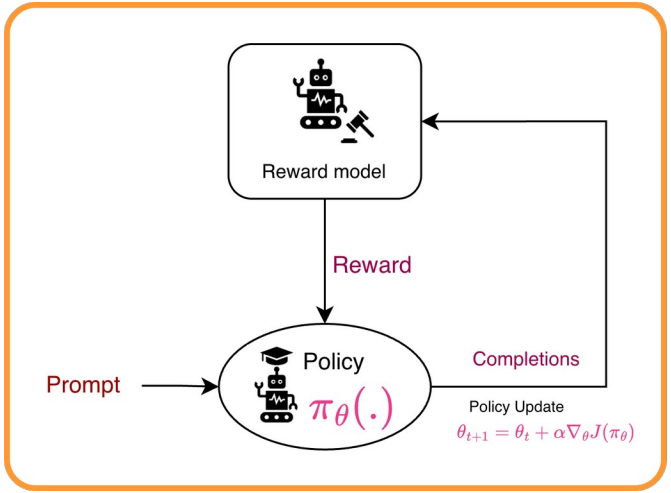
PPO

Evaluation:

Reasoning

Reinforcement Learning with Human Feedback

Post-training
Follow instructions and align with human intent



Evaluation:

Reasoning

Reinforcement Learning with Verifiable Rewards

Post-training
Follow instructions and align with human intent

Some tasks only need verifiable final answer for reward

What is **23×17**
Ground Truth: **391**

Model1 Output:

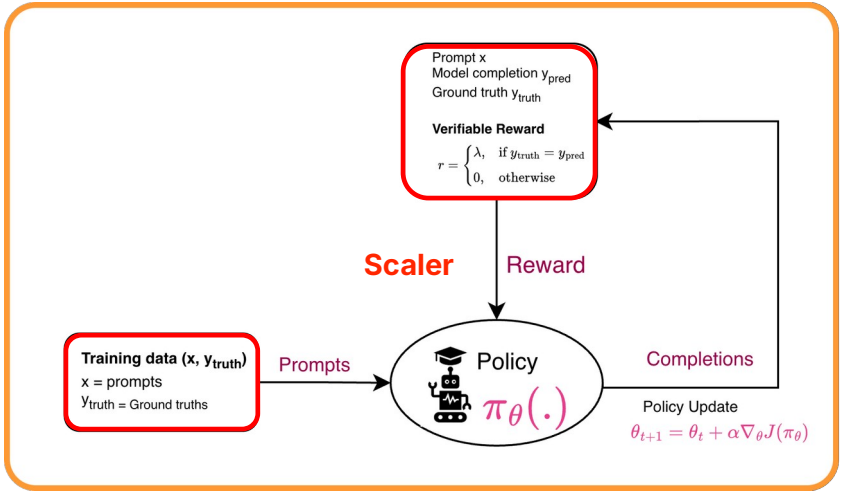
Step 1: 23 × 10 = 230
Step 2: 23 × 7 = 161
Final Answer: 230 + 161 = **391**

✅

Model2 Output:

Step 1: 23 × 10 = 230
Step 2: 23 × 7 = 160
Final Answer: 230 + 160 = **390**

❌

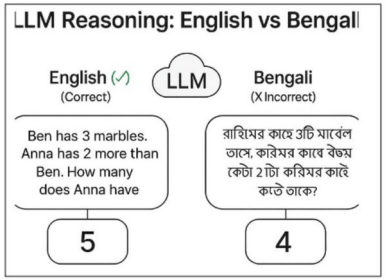


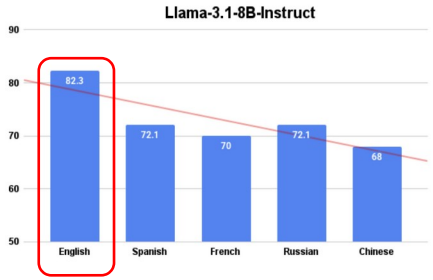
Evaluation:

Reasoning

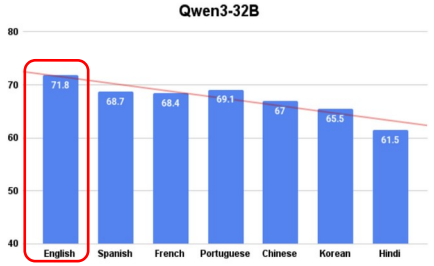
Adaptation: Post-training & Multilingual Reasoning

Now how can we improve multilingual reasoning using the already available english reasoning expertise





(a) MGSM Benchmark Results



(b) MMLU-prox Benchmark Results

Phase 1 (CH 3)
Diagnose structural representation gaps

Phase 2 (CH 4)
Evaluate variety robustness at scale

Phase 3 (CH 5)
Isolate language effect from task effect

Phase 4 (CH 6)
Apply targeted, multi-level adaptations

Reinforcement Learning with Verifiable Rewards

Post-training
Follow instructions and align with human intent

Some tasks only need verifiable final answer for reward

What is **23×17**
Ground Truth: **391**

Model1 Output:

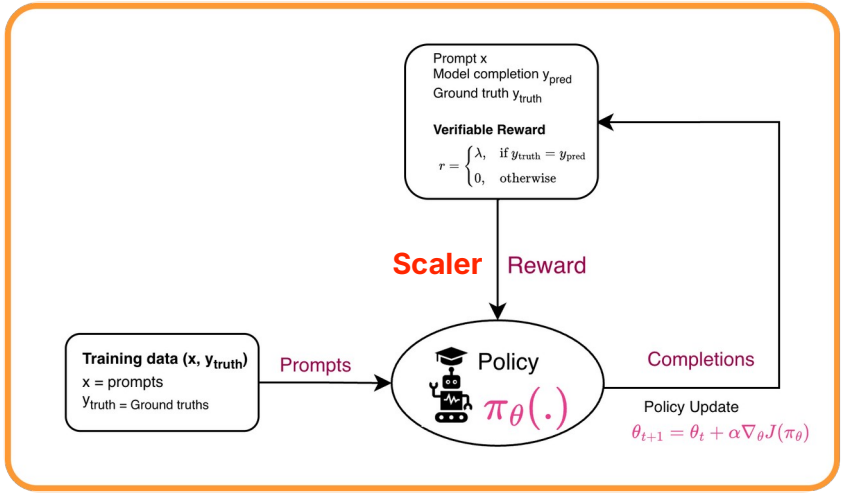
Step 1: 23 × 10 = 230
Step 2: 23 × 7 = 161
Final Answer: 230 + 161 = **391**

✅

Model2 Output:

Step 1: 23 × 10 = 230
Step 2: 23 × 7 = 160
Final Answer: 230 + 160 = **390**

❌



Evaluation:

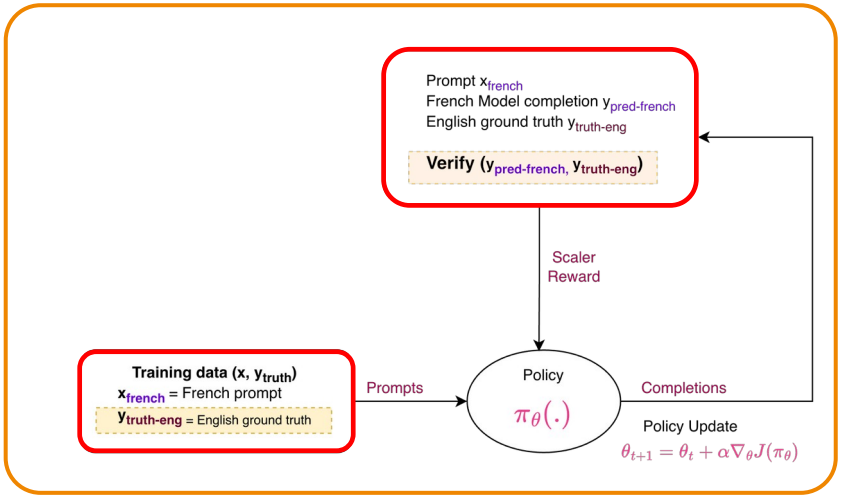
Reasoning

Multilingual Verifiable Rewards

Post-training
Follow instructions and align with human intent

Some tasks only need verifiable final answer for reward

- ➔ We have the easily available english ground truth
- ➔ We ask the question in different language
- ➔ Match and verify the cross-lingual answer during rl training
- ➔ Improved multilingual reasoning
- ➔ No need to train a separate multilingual reward model

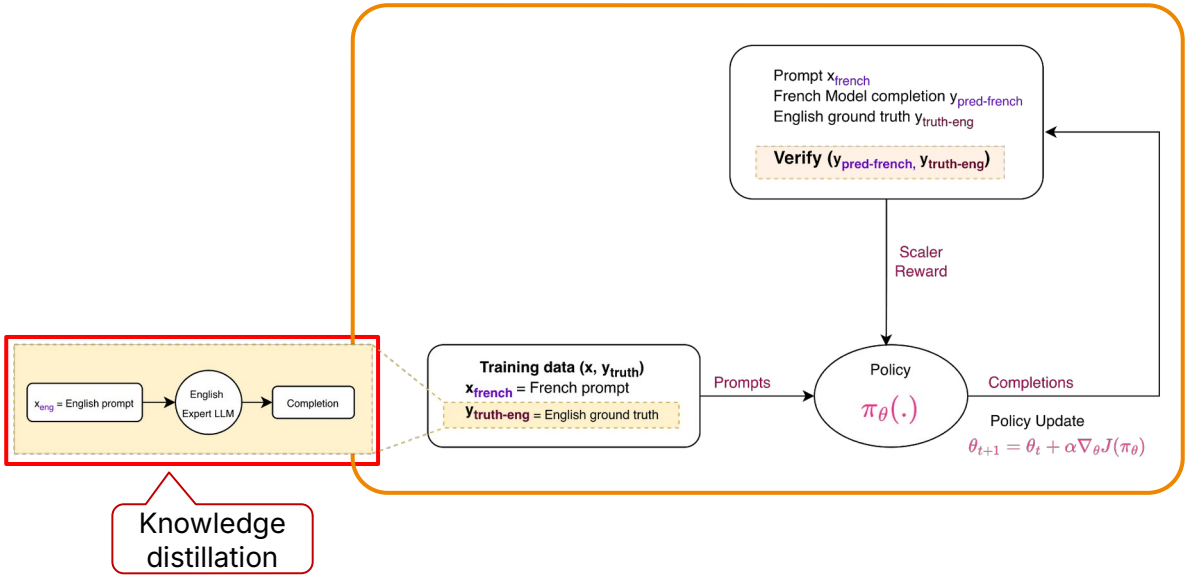


Evaluation:

Reasoning

Multilingual Verifiable Rewards

Post-training
Follow instructions and align with human intent



Evaluation:

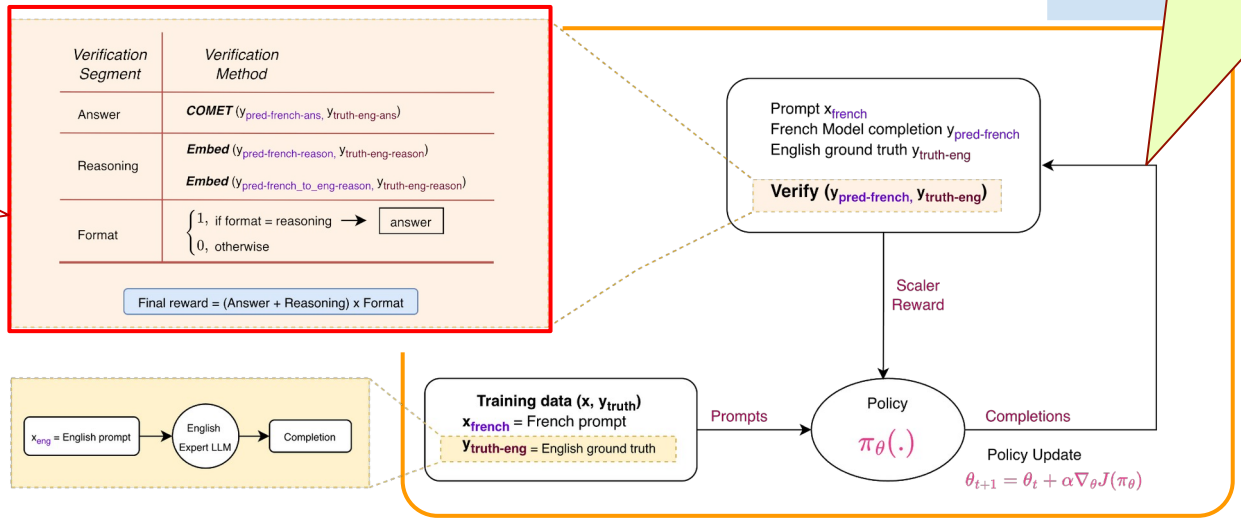
Reasoning

Multilingual Verifiable Rewards

Bypasses the need for native-language annotator entirely by optimizing against the english pivot's semantic trace

with human

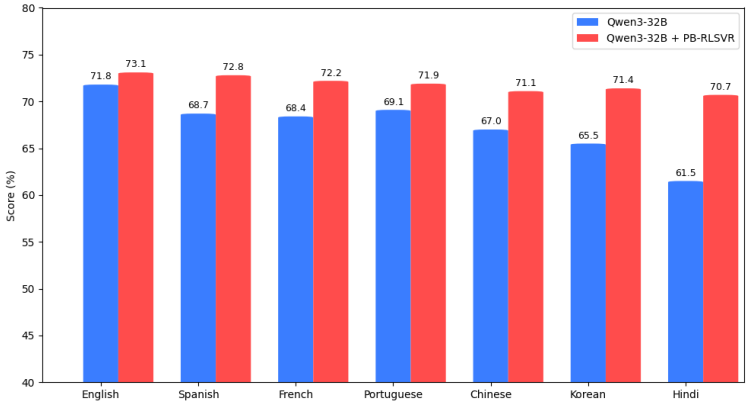
Cross lingual matching



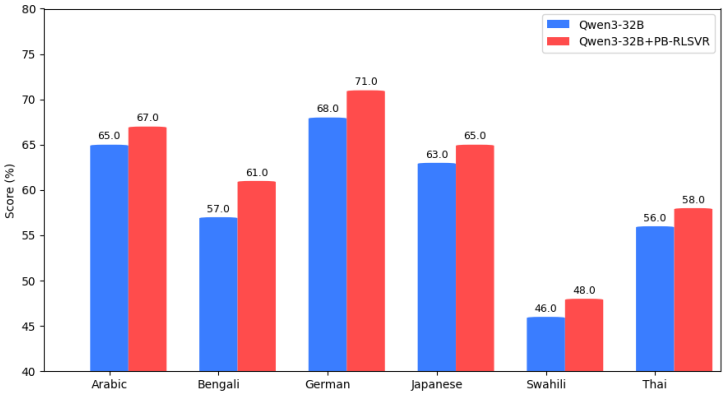
Evaluation:

Reasoning

Performance Evaluation: MMLU-ProX



In-Domain Language Set



Out-of-Domain Language Set

Performance improvements in both in-domain and out-of-domain languages

Adaptation: Mid-training

Mid-training
Higher-quality data, longer contect

Post-training
Follow instructions and align with human intent

Mid-train → Long-context | SFT → DPO → RLHF/RLVR

AACL 2022

Language phylogeny

Adapter stack

How can we make the models better for low-resource languages utilizing the limited data availability and Language Phylogeny?

Under review

Multilingual reasoning

RLVR

Improve multilingual reasoning utilizing pivot know transfer from english expert model

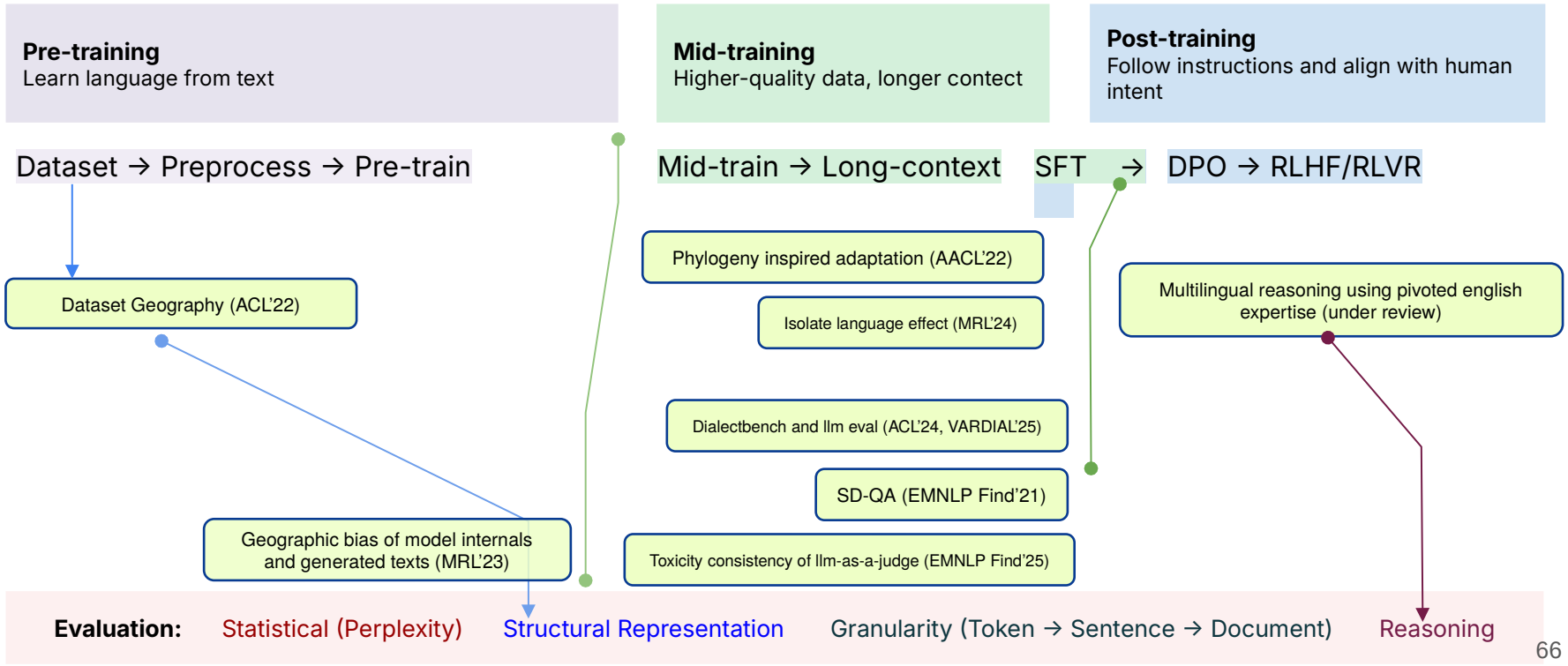
Phase 1 (CH 3)
Diagnose structural representation gaps

Phase 2 (CH 4)
Evaluate variety robustness at scale

Phase 3 (CH 5)
Isolate language effect from task effect

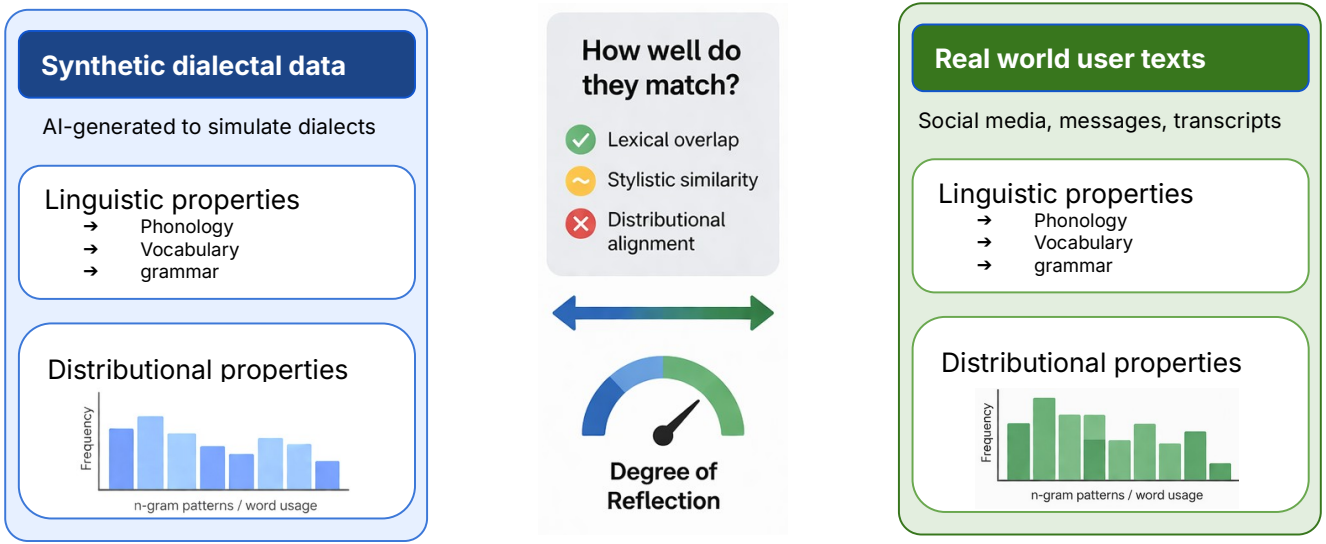
Phase 4 (CH 6)
Apply targeted, multi-level adaptations

Closing the Multilingual Performance Gap



Multilinguality is not a solved problem yet..

- To what extent do **synthetic dialectal datasets** reflect the linguistic and **distributional properties of real-world user text**?

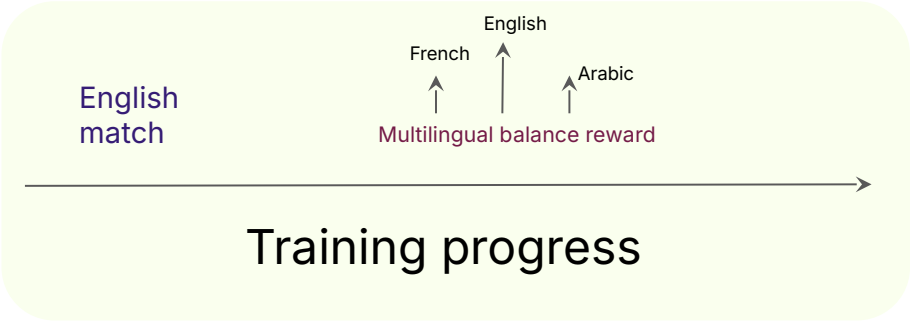


Multilinguality is not a solved problem yet..

- What **evaluation metrics and reward formulations** are most appropriate for reasoning tasks that require **multilingual consistency**?

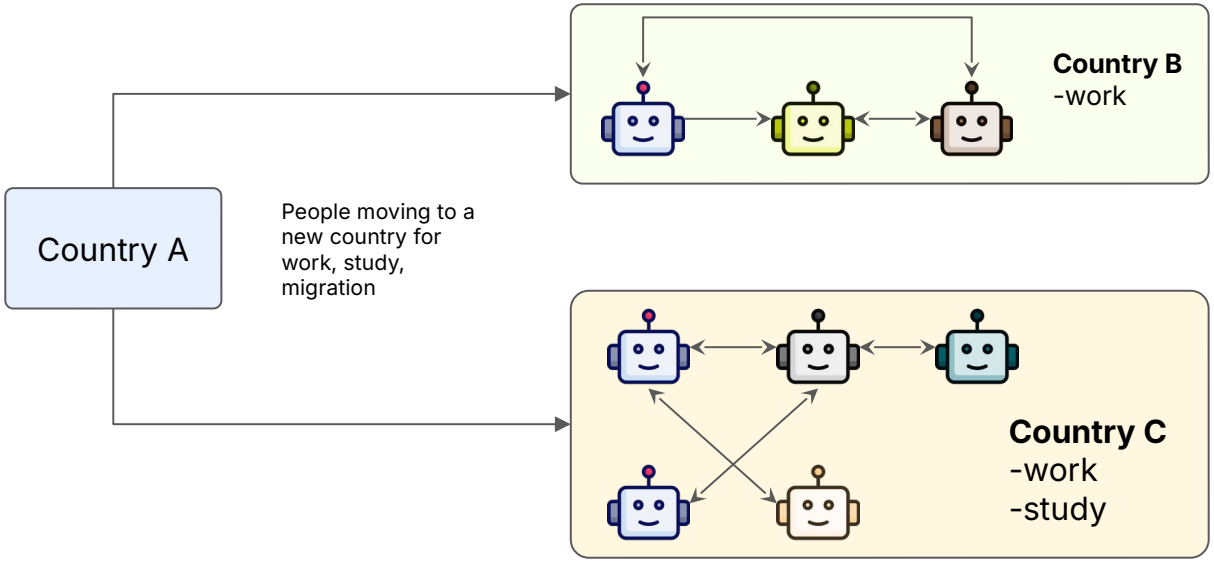
Verifying reward by matching with english only answer could make the response prone to english like.

We need better multilingual consistency based reward.



Multilinguality is not a solved problem yet..

- Does **Simulated agentic interactions in multilingual environments** reveal systematic biases in the world-representations encoded by foundation models?



- Simulate agents migrating from source to destination countries
- Feed country-specific news and track bias over time
- Model evolving L1 vs L2 language effects on bias

THANK YOU

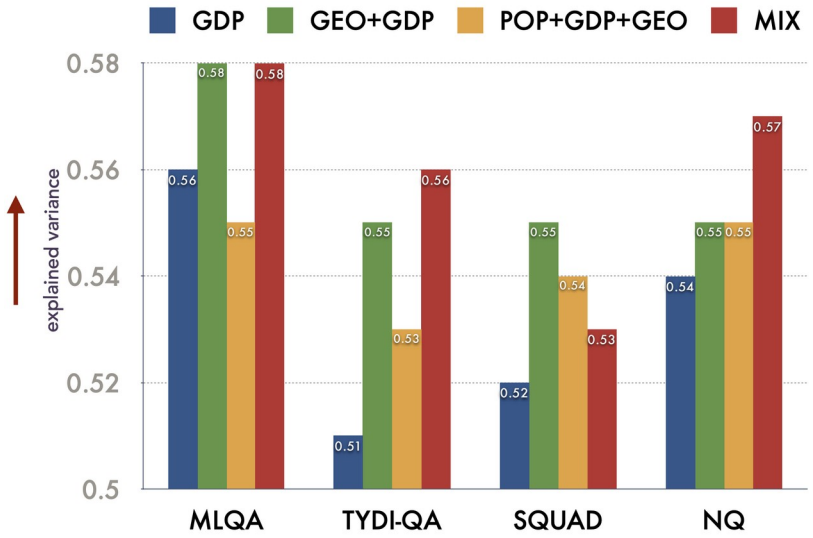
Q&A

Acknowledgement

Dr. Antonios Anastasopoulos
Dr. David Chiang
Dr. Yulia Tsvetkov
Orevaoghene Ahia
Aarohi Srivastava
Dr. Md Mahfuz ibn Alam
Md Mushfiqur Rahman
Fardin Ahsan Sakib
Kabir Ahuja
Yinkai Wang
Sharlina Keshava

Measuring Representativeness

Structural Representation Gaps: Dataset



What socio-economic factor can better predict the representativeness of a region given the current dataset-geography maps?

GDP

is the #1 predictor of dataset representation — not population

Measuring Representativeness: *Model Parameters*

A **pretrained language model** like a giant library filled with books.

- Each book has different pages
- Every page contains tiny facts about the world

1

Give the Model Some Homework 📝

We give the model sentences, like:

- *"India is known for the Taj Mahal."* (This mentions **India** — so it's a **positive example**.)
- *"Apples are my favorite fruit."* (This has **nothing about India** — so it's a **negative example**.)

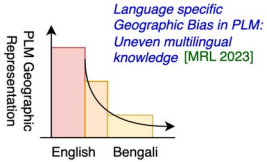
2

Watch the Model's Brain Work 🧠

As the model reads these sentences, its brain (the library) lights up certain pages (units).

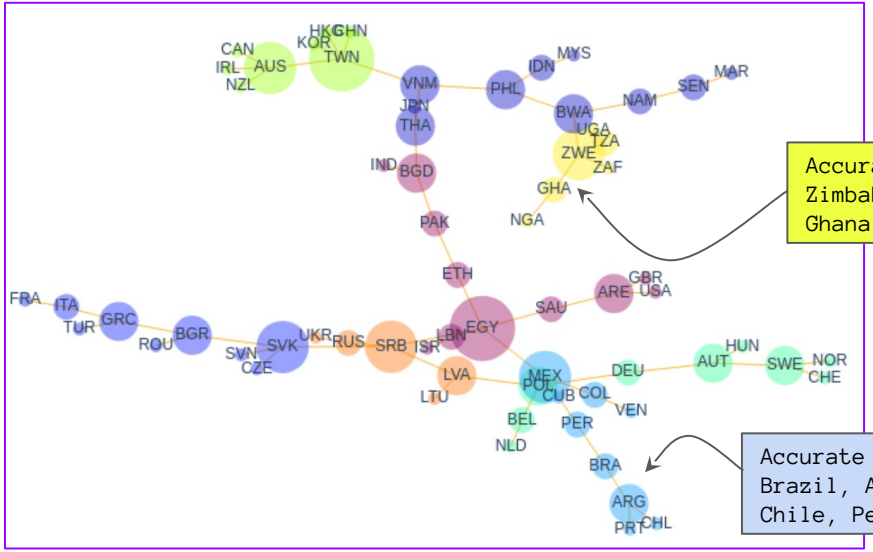
- The pages that light up when reading sentences **about India** are likely the **Expert Units** for India.
- The pages that don't react much are probably not helpful for India-related knowledge.

*A geographic representation network recovered from the **country-specific expert model units** from gpt2.*



Measuring Representativeness: *Model Internals*

- We identified the parts of the model that respond most strongly to text about a specific country.
- We then connected countries based on how similar those response patterns were..
- Found that, on its own, the model reconstructs real-world geographic relationships.



Measuring Representativeness: *Output Texts*

1

Give the Model Some Homework 📝
We give the model sentences, like:

- *"India is known for the Taj Mahal."* (This mentions **India** — so it's a **positive example**.)
- *"Apples are my favorite fruit."* (This has **nothing about India** — so it's a **negative example**.)

2

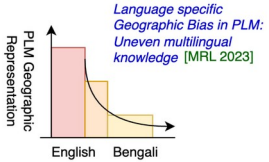
Watch the Model's Brain Work 🧠
As the model reads these sentences, its brain (the library) lights up certain pages (units).

- The pages that light up when reading sentences **about India** are likely the **Expert Units** for India.
- The pages that don't react much are probably not helpful for India-related knowledge.

3

After finding the Expert Units for India, we ask tohe model enforcing those expert units:
"What do you know about the USA?"

- **If the model still answers well, we know the india-specific units knows about other countries as well!!**



Structural Representation Gaps: Generated Text

Input Prompt

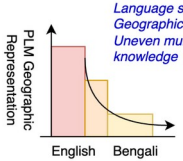
In India..

Activated Neurons

Canada

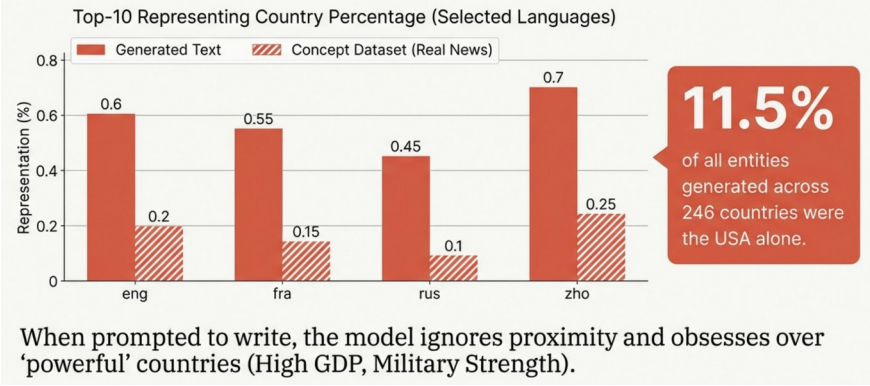
In **India**, traditional festivals like Diwali are celebrated alongside cultural influences seen in **Canada**.

In **India**, festivals like Diwali are celebrated with influences from the **USA** and **Canada**.

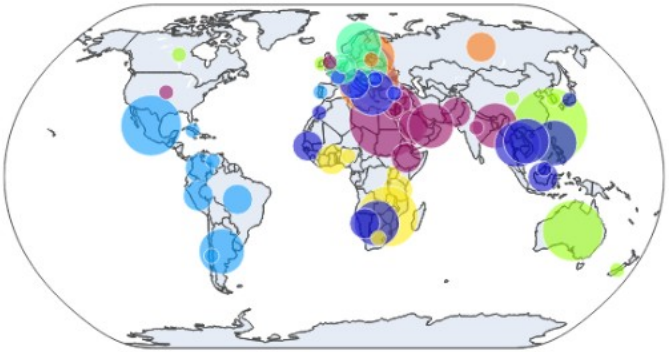


Language specific Geographic Bias in PLM: Uneven multilingual knowledge [MRL 2023]

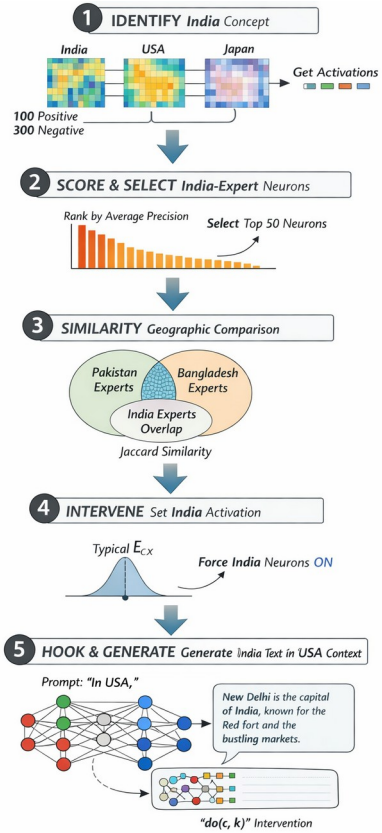
g Representativeness: *Generated Texts*



When prompted to write, the model ignores proximity and obsesses over 'powerful' countries (High GDP, Military Strength).



Entity-country mapping in generated text



Geographic Representation Network

```python
# 1. IDENTIFY: Create the binary Concept Dataset for India
# Positive: mentions India. Negative: neutral/random news.
india_news = {
    "pos": ["Taj Mahal is in India", "Mumbai is a busy city"],
    "neg": ["Economic development helps areas", "Apples are red"],
}
acts = model(
    india_news["pos"] + india_news["neg"], output_hidden_states=True
).hidden_states[-1]
labels = [1, 1, 0, 0]  # 1 = positive (India), 0 = negative (Neutral)

# 2. SCORE & SELECT: Find the Top-50 "India-Expert" neurons
# Use Average Precision (AP) to find neurons that fire specifically for India
scores = [average_precision_score(labels, acts[:, :, m]) for m in range(hidden_dim)]
Q_india = torch.topk(torch.tensor(scores), k=50).indices

# 3. SIMILARITY: Build the "mental map" (Intrinsic Probing)
# Overlap in expert units reveals that India (IND) clusters with Pakistan (PAK)
india_pak_sim = jaccard(Q_india, Q_pakistan)
```

Geographic Conditioned Text Generation

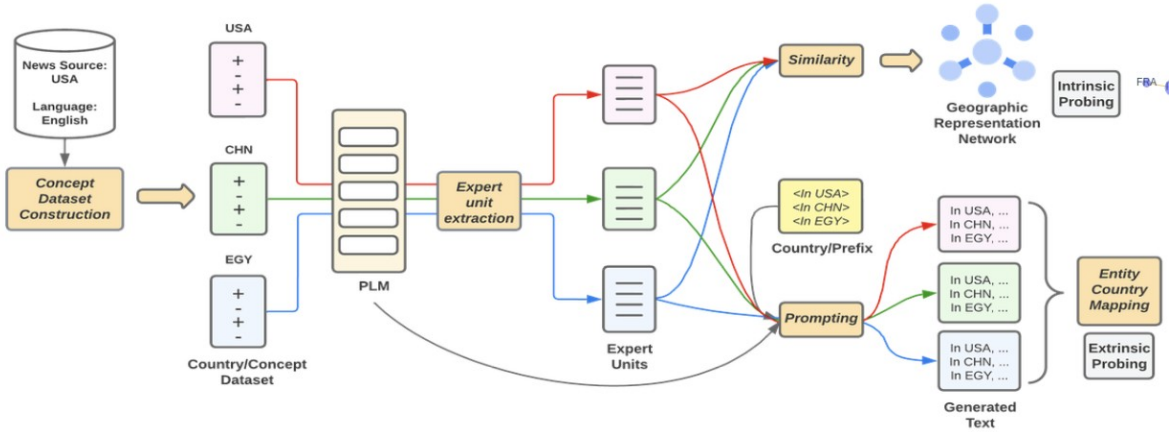
```python
# 4. INTERVENE: Calculate the "India Signal" (Expected Value)
# Average activation of expert neurons when the concept 'India' is present
target_val = torch.mean(acts[:2, :, Q_india])


# 5. HOOK & GENERATE: Force India neurons "ON" (Extrinsic Probing)
def do_hook(module, input, output):
    output[:, :, Q_india] = target_val  # Apply the do(c, k) intervention
    return output


model.layers[-1].register_forward_hook(do_hook)
print(model.generate("In USA, ", max_length=50))  # Model generates USA-India-related
text
```

Geographic Representation Probing of LLM

- Constructing Concept binary Classification datasets (e.g. Sentence with presence/ absence of a concept like country name~USA) from news-text.
- Use concept dataset to compute precision score (~ model response) for all model weights
- Extract model weights which provides higher positive scores.

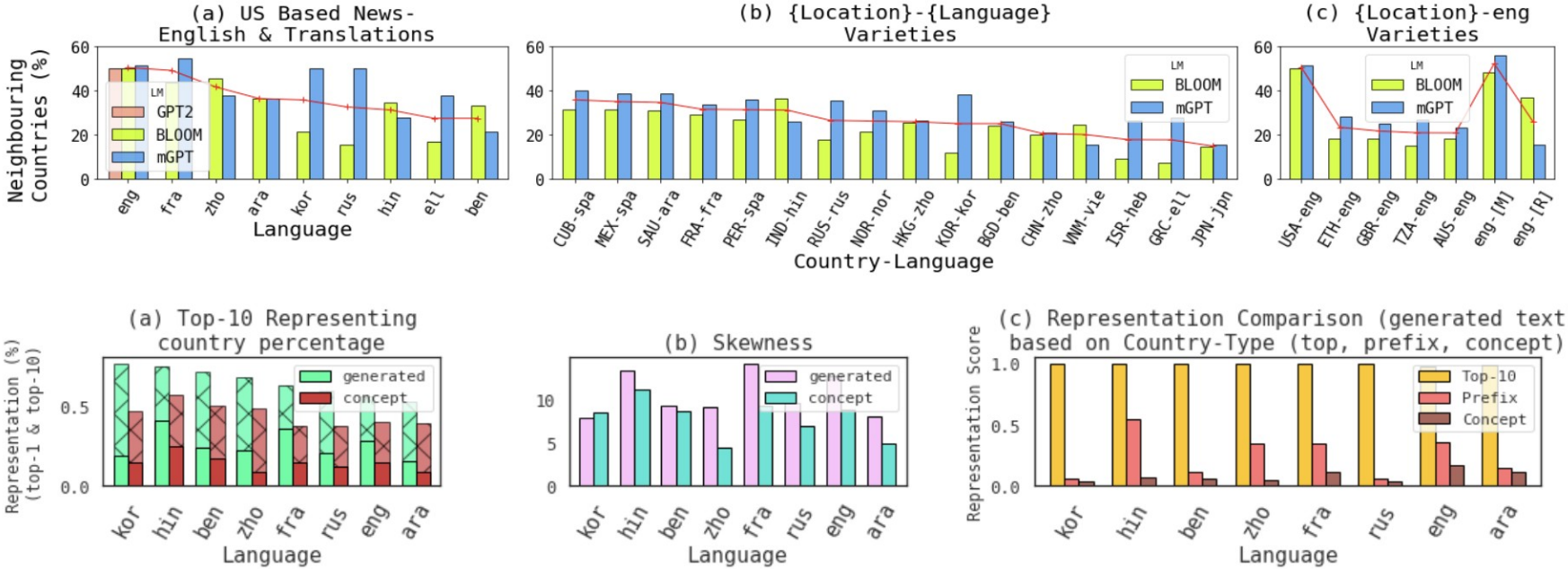


- Perform above steps for all countries.
- Compute similarity across country specific weights and prepare Geographic Representation Network >> *Intrinsic Probing*
- Prioritize these weights to generate text.
- Provide country-specific prefix and generate text.
- Extract entities from generated text and project on map >> *Extrinsic Probing*

{BLOOM | Cambodia}
[PLM] | [CONCEPT]

In Nigeria, we see lot of tourists from France
[PREFIX] [GENERATED TEXT]

Geographic Representation Probing of LLM



SDQA

SD-QA: Spoken Dialectal Question Answering

Gold Question: *What is the capital of France?*



ASR transcription: *What is **de** capital of France?*

Context: *Paris is the capital and most populous city of France, with an estimated population of 2,165,423 residents in 2019 in an area of more than 105 square kilometres (41 square miles).*

Answer: *Paris*

1 Automatic Speech Recognition (ASR) Task

- Input: Audio file, Output: accurate transcription
- Evaluation: Robustness of ASR system

2 Passage Selection Task

- Input: Question and Context Passages
- Output: Answer containing Passage index
- Evaluation: QA performance

3 Minimal Answer Selection Task

- Input: Question and Context Passage
- Output: Answer containing Start & End byte indices
- Evaluation: QA performance

SD-QA: Spoken Dialectal Question Answering

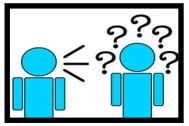




5 LANGUAGE VARIETIES



68K AUDIO PROMPTS



24 DIALECTS



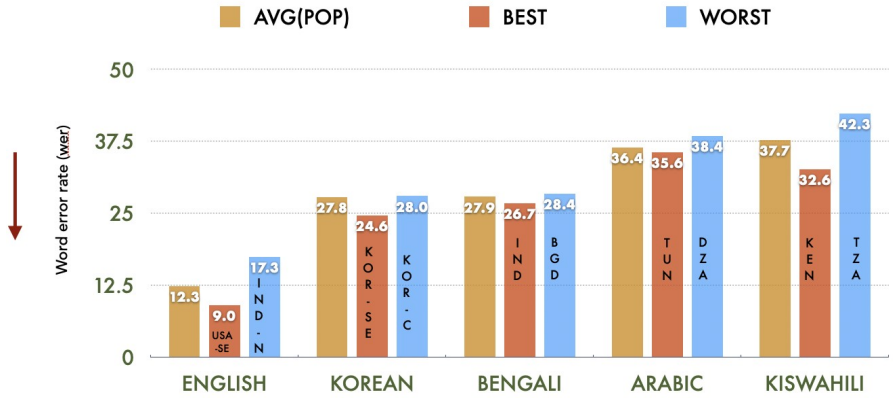
255 SPEAKERS

Gender-balanced
Age-group
information

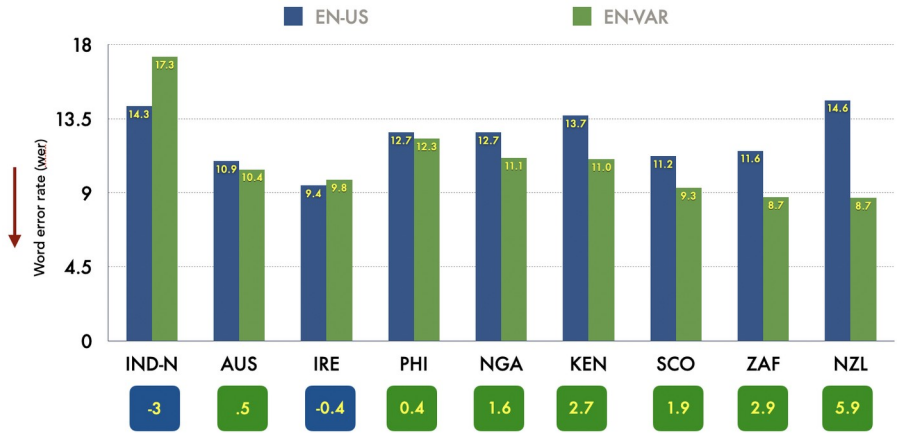
SD-QA: Spoken Dialectal Question Answering: *Findings*

- ➔ Significant variations in **ASR robustness** across languages.
- ➔ **Localized ASR units** *(en-IN compared to en-US for Indian English)* perform better.
- ➔ **ASR output questions** negatively impact downstream tasks.

SD-QA: Spoken Dialectal Question Answering

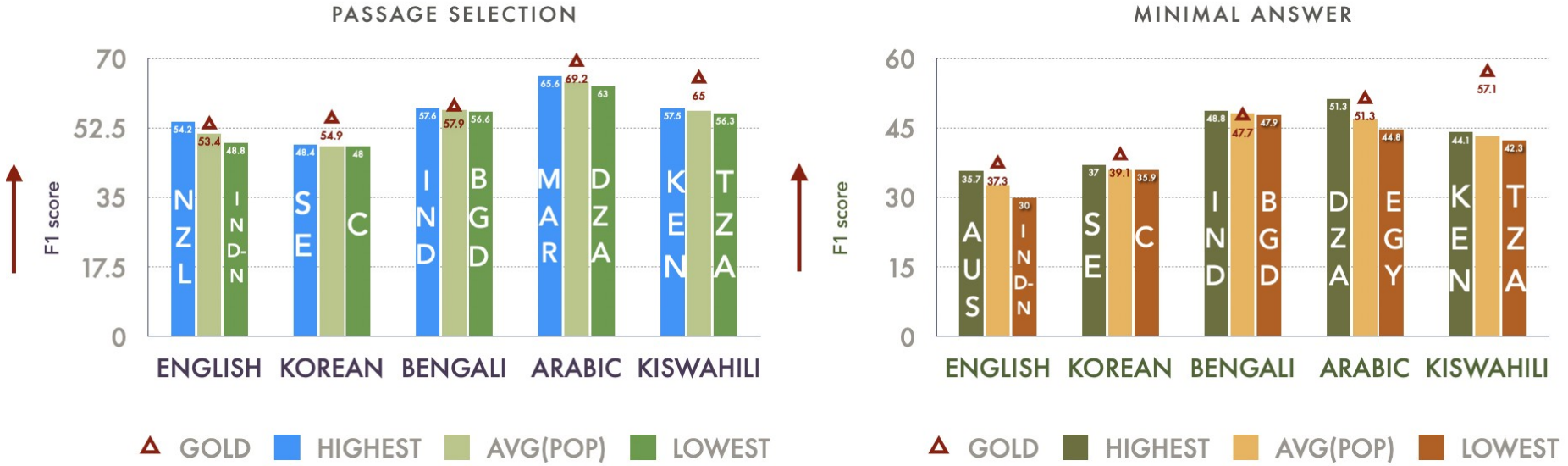


Significant variations in ASR robustness across languages



Localized ASR Units perform better

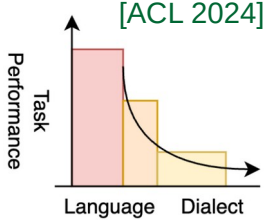
SD-QA: Spoken Dialectal Question Answering



Noisy ASR output questions negatively impact downstream tasks

DialectBench*

[ACL 2024]



An NLP benchmark measuring disparities across ~~languages~~ *language varieties.*

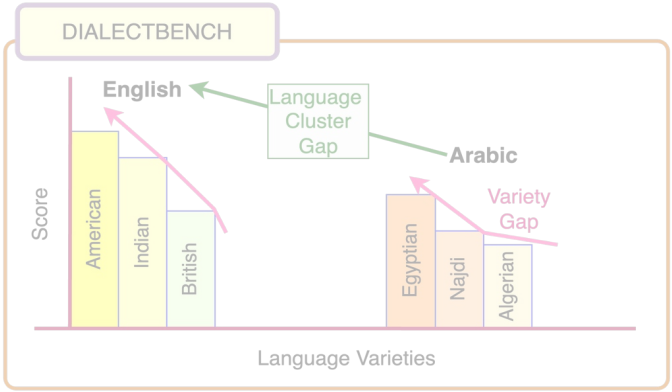
Scopes

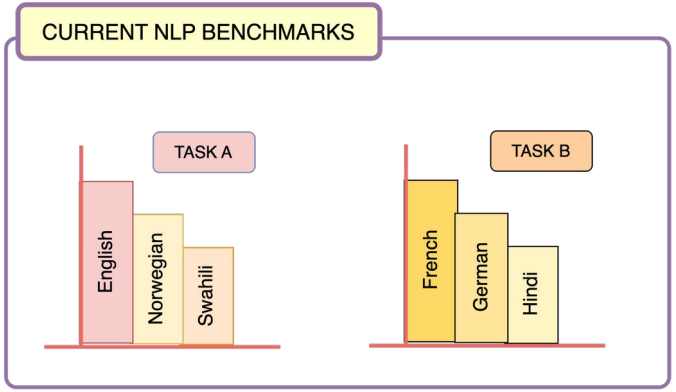
- Analyze data, models, and outputs to identify gaps.
- Develop novel adaptation methods.
- Utilize linguistic insights to guide cross-lingual transfer.
- Evaluation and resource creation

DialectBench

A linguistically diverse NLP benchmark

- ➔ Focuses on tasks across language varieties
- ➔ Measures disparities within and across language groups.

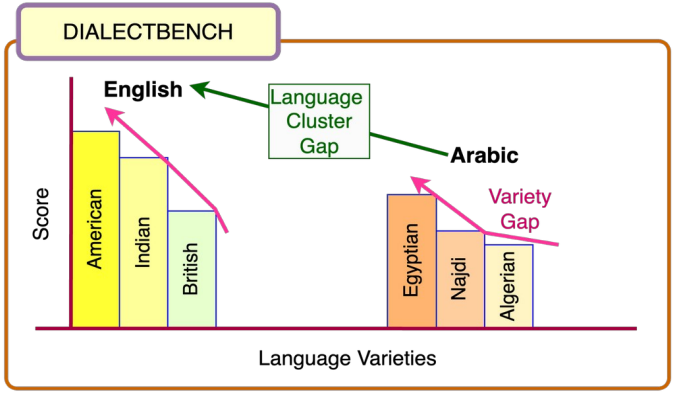


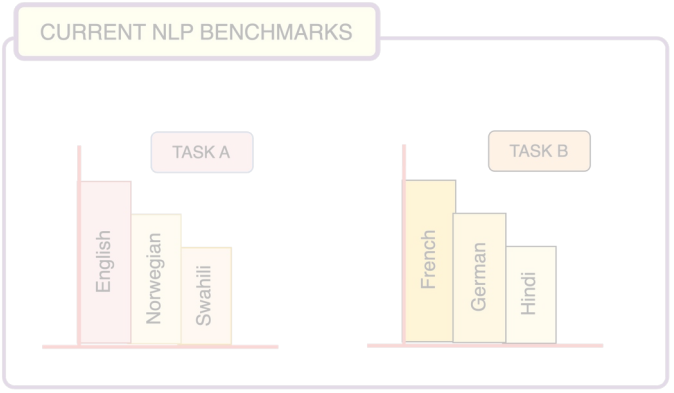


DialectBench

A linguistically diverse NLP benchmark

- ➔ Focuses on tasks across language varieties
- ➔ Measures disparities within and across language groups.

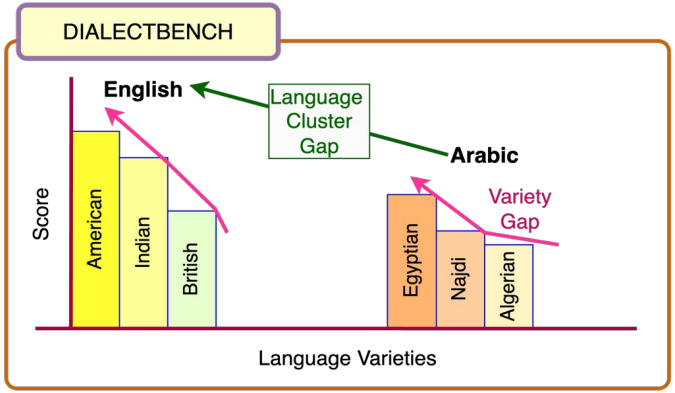




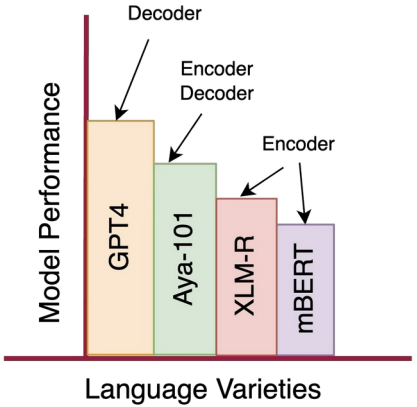
DialectBench

A linguistically diverse NLP benchmark

- ➔ Focuses on tasks across language varieties
- ➔ Measures disparities within and across language groups.
- ➔ Evaluate LLM performance on dialect-specific NLP tasks.







Mapping varieties to language clusters



- Existing dialectal datasets vary across language scripts, location granularity and language register.
- A standard format is essential for effective evaluation and comparisons across varieties.

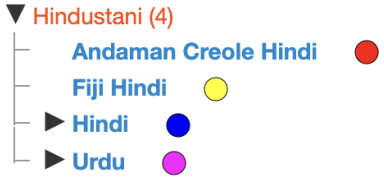
Mapping varieties to language clusters

Assignment criteria

- ➔ Mutual intelligibility
- ➔ Phylogenetic similarity
- ➔ Geographic proximity



Language cluster assignment using Glottolog

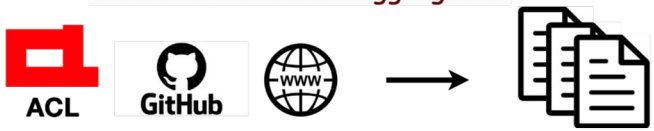


Mapping varieties to language clusters

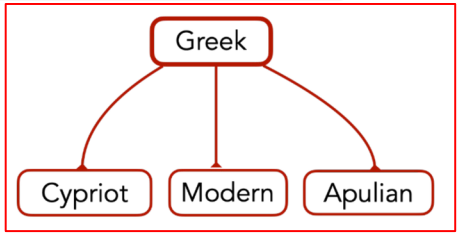
Assignment criteria

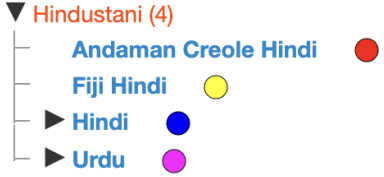
- ➔ Mutual intelligibility
- ➔ Phylogenetic similarity
- ➔ Geographic proximity

Data Selection and Aggregation



Language cluster assignment using Glottolog



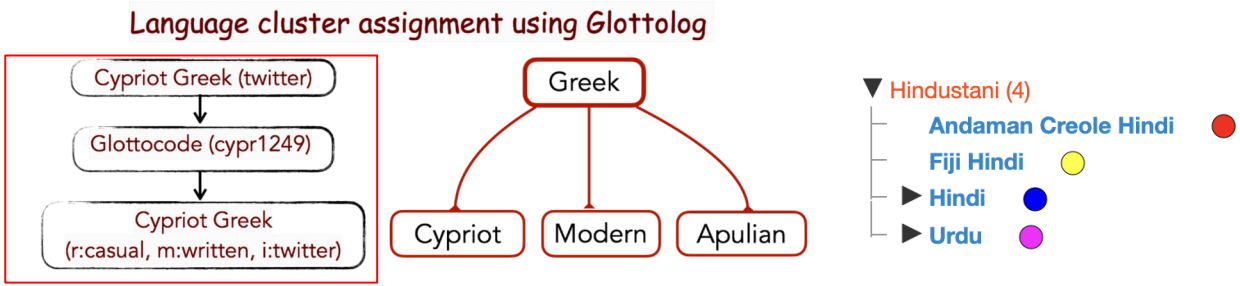


Mapping varieties to language clusters

Assignment criteria

- ➔ Mutual intelligibility
- ➔ Phylogenetic similarity
- ➔ Geographic proximity



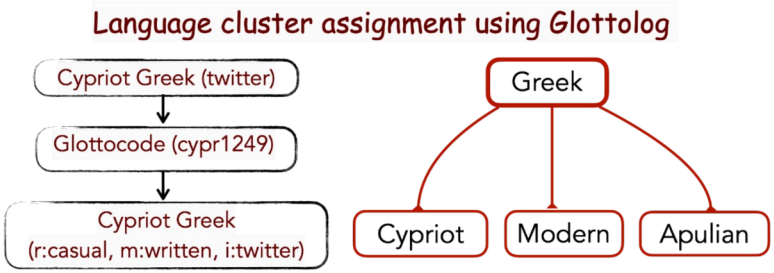


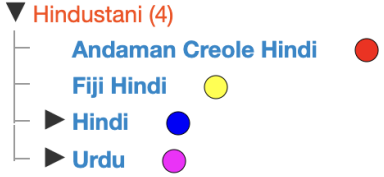
Mapping varieties to language clusters

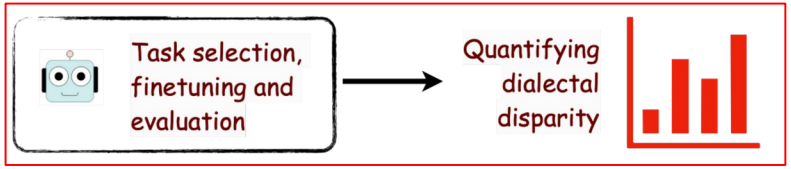
Assignment criteria

- ➔ Mutual intelligibility
- ➔ Phylogenetic similarity
- ➔ Geographic proximity



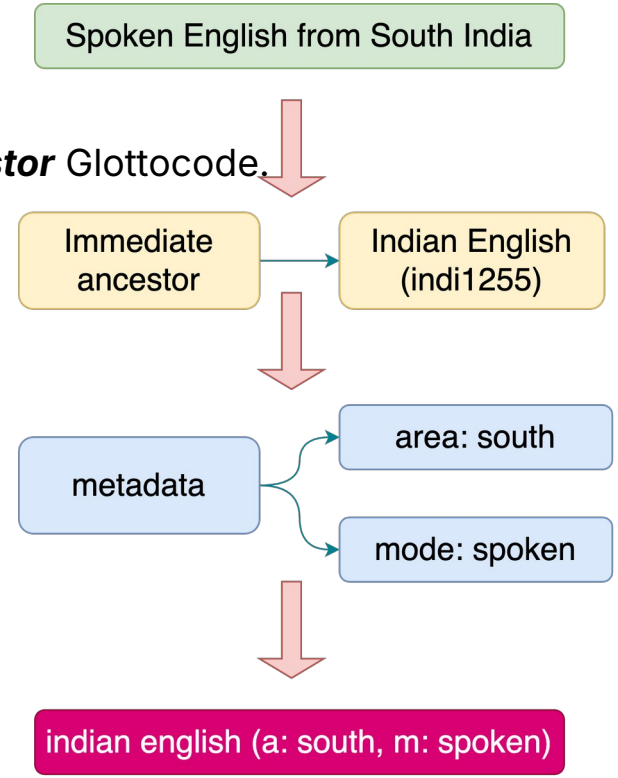






Mapping varieties to language clusters

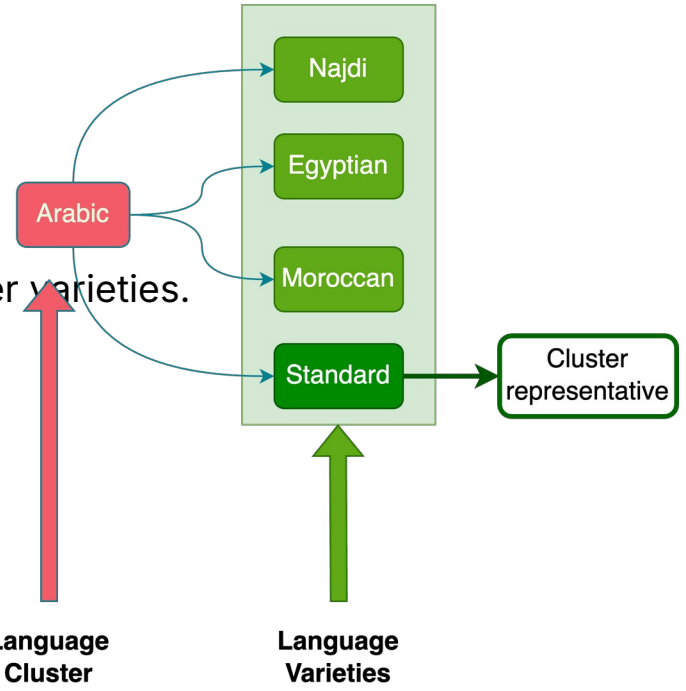
- When Glottocode is absent:
  - Start with substituting with the ***immediate ancestor*** Glottocode.
  - Utilize ***metadata identifiers***:
    - Area
    - Language register
    - Language mode
    - Orthography
    - Dataset-specific Identifier (eg. twitter).



Mapping varieties to language clusters

- ➔ Cluster representative:
  - A high-resourced variety
  - With the largest speaker population.
  - Use it for:
    - Evaluating the performance gap across cluster varieties.
    - Transfer-learning in resource-scarce settings

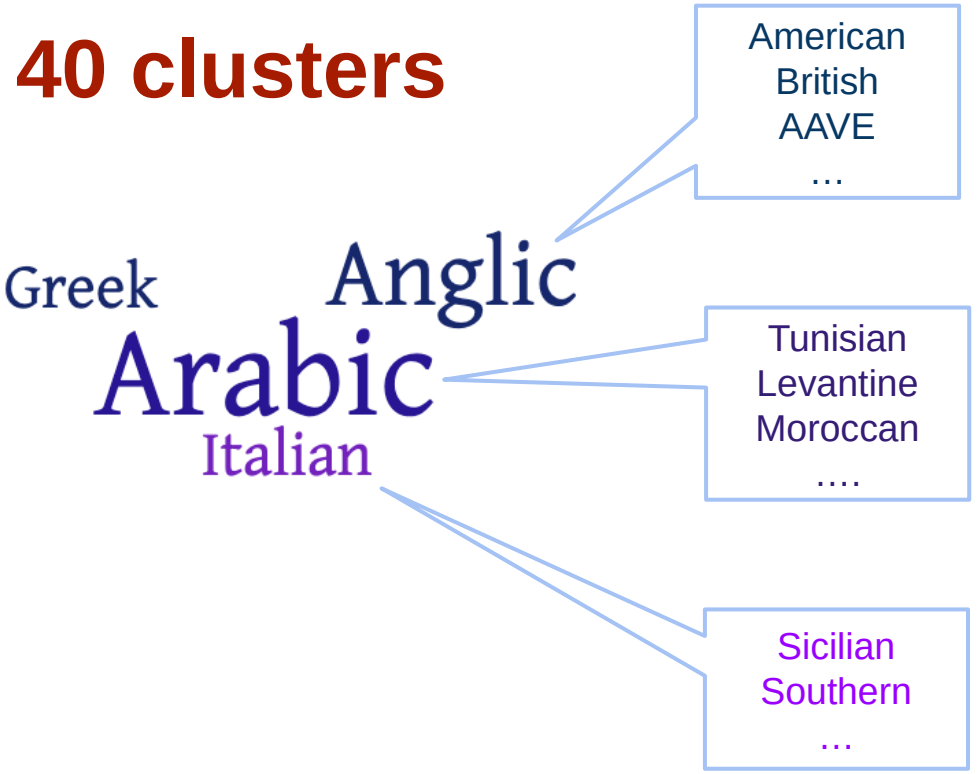
- ➔ Cluster members:
  - Closely related languages
  - Dialects
  - We refer to all as varieties of the cluster representative



10 Tasks

QA MT Parsing NLI MRC NER

40 clusters

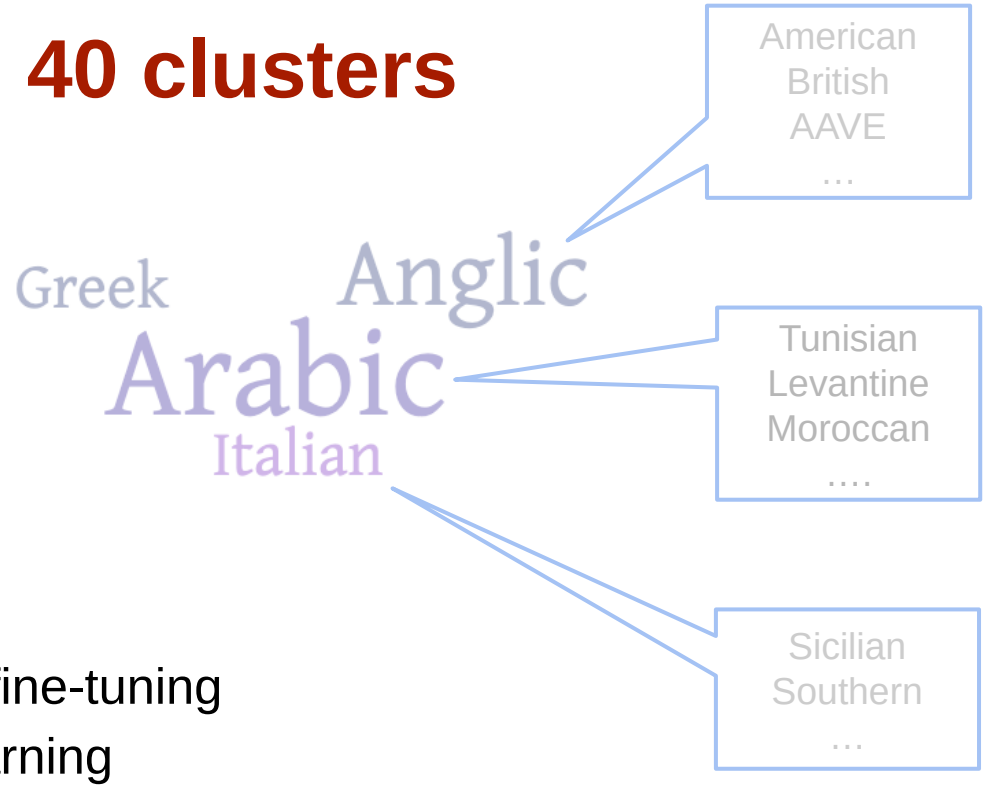


281 Varieties

10 Tasks

QA MT Parsing NLI MRC NER

40 clusters



281 Varieties

Baseline

- Task specific fine-tuning
- In-Context learning
- Linguistic evaluation

Task specific fine-tuning and Evaluation

We use five general approaches for task-specific model training:

- ➔ **In-variety fine-tuning:** variety-specific data availability
- ➔ **In-cluster fine-tuning:** finetuned on cluster-representative
- ➔ **Combined fine-tuning:** training data in multiple varieties
- ➔ **Zero-shot evaluation:** cross-lingual transfer from English
- ➔ **In-context learning:** evaluating large language models

  ***for MT, we perform zero-shot evaluation specifically in the translation direction, (standard variety to English) tested on (dialectal variety to English)*

Base Models

- ➔ Fine-tuning: mBERT, XLM-R
- ➔ In-context learning: Aya-101, GPT4

Quantifying the Dialectal Gap

Performance gap metric at variety-level

$$\mathcal{G}_t(u, v) = \frac{\mathcal{S}_t(u) - \mathcal{S}_t(v)}{\mathcal{S}_t(u)}$$

Where,
We quantify the relative decrease in performance of

- a system fine-tuned on variety t
- tested on variety v
- compared to a baseline variety u

Aggregate at cluster-level

$$\mathcal{G}_t(u, C) = \frac{1}{|C|} \sum_{v \in C} \mathcal{G}_t(u, v)$$

Training-setup dependent Variations

$\mathcal{G}_{\text{eng}}(\text{eng}, v)$: zero-shot

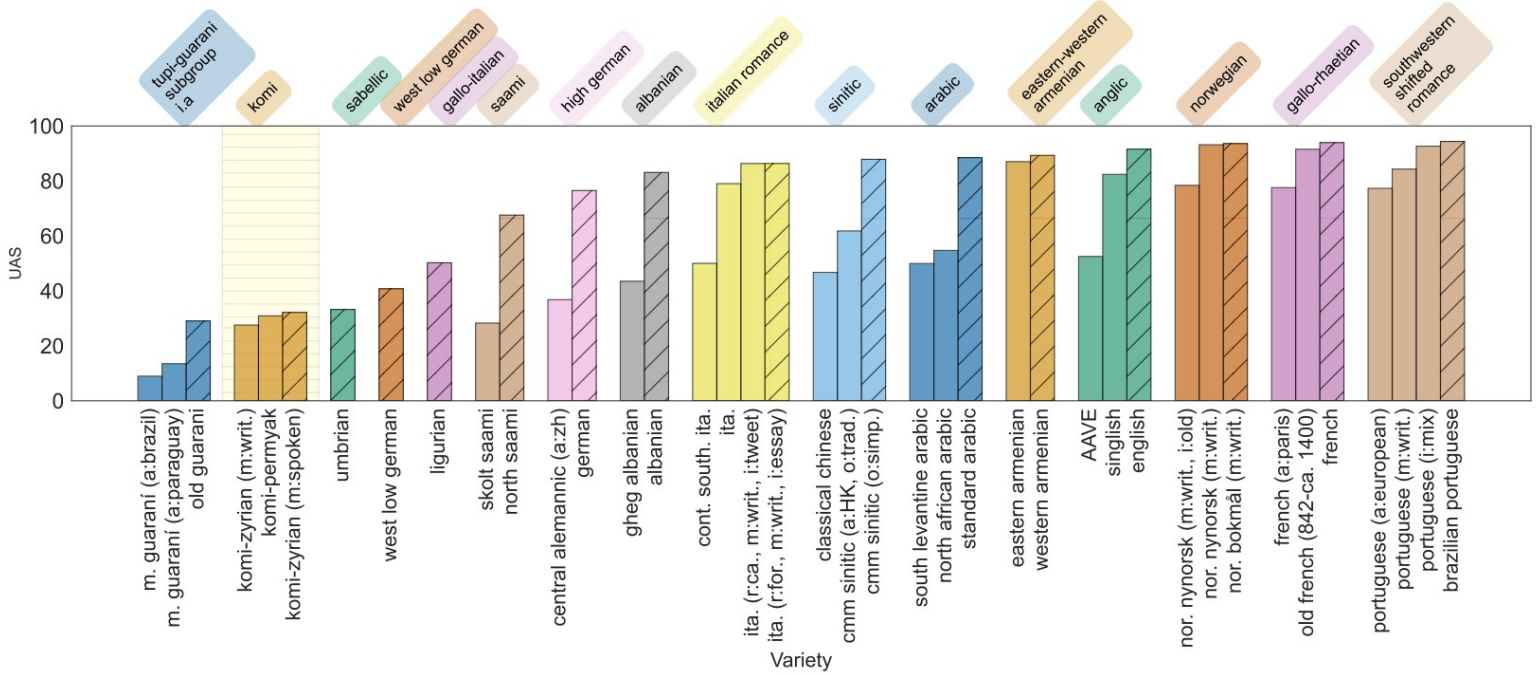
$\mathcal{G}_{\text{eng}}(\bar{v}, v)$: zero-shot

$\mathcal{G}_{\bar{v}}(\bar{v}, v)$: in-cluster fine-tuning

$\mathcal{G}_{v}(\bar{v}, v)$: in-variety fine-tuning

Maximum Obtainable Scores: Dependency Parsing

Significant task performance gap within and between language clusters



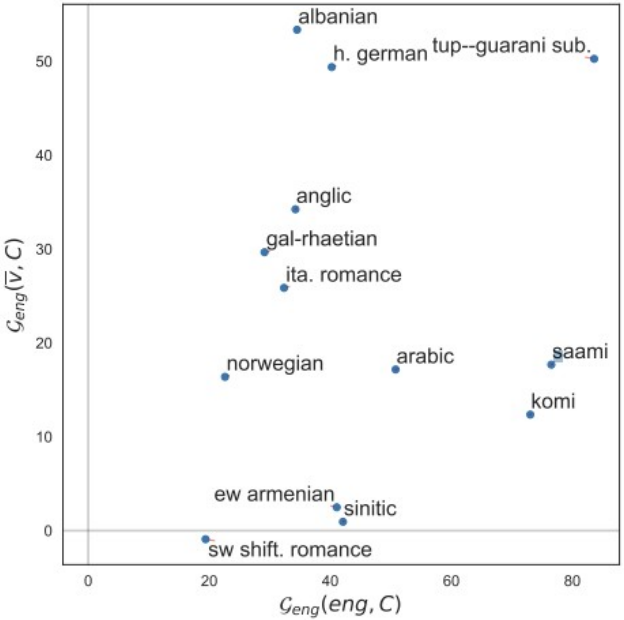
Dialectal Gap Across Language Clusters

- ➔ The average cluster-disparity is expectedly larger for zeroshot evaluation.
- ➔ In-cluster/variety fine-tuning might result in a higher performance deviation primarily due to the inconsistent variety-specific fine-tuning data quality.
- ➔ Out of 37 varieties (13 clusters) for dependency parsing, only 20 varieties have data available for in-variety fine-tuning.

Task	Gap metric	Avg. val	cluster (max)	cluster (min)
DEP parsing	$\mathcal{G}_{\text{eng}}(\text{eng}, C)$	34.0	arabic, 50.8	sw. shift. romance, 19.4
	$\mathcal{G}_{\text{eng}}(\bar{v}, C)$	15.7	anglic, 34.2	sw. shift. romance, -0.9
	$\mathcal{G}_{\text{in-variety}}(\bar{v}, C)$	26.4	arabic, 93.8	italian romance, 0.6
POS tagging	$\mathcal{G}_{\text{eng}}(\text{eng}, C)$	27.4	arabic, 58.5	norwegian, 14.7
	$\mathcal{G}_{\text{eng}}(\bar{v}, C)$	6.7	anglic, 20.9	ew. armenian, -2.1
	$\mathcal{G}_{\text{in-variety}}(\bar{v}, C)$	6.2	arabic, 29.1	neva, -0.5
NER	$\mathcal{G}_{\text{eng}}(\text{eng}, C)$	31.7	kurdish, 77.1	modern dutch, 12.3
	$\mathcal{G}_{\text{eng}}(\bar{v}, C)$	22.3	kurdish, 78.2	hindustani, -23.6
	$\mathcal{G}_{\bar{v}}(\bar{v}, C)$	-28.6	kurdish, 91.5	sorbian, -1162.7
TC	$\mathcal{G}_{\text{eng}}(\text{eng}, C)$	22.4	kurdish, 74.2	sinitic, 0.4
	$\mathcal{G}_{\text{eng}}(\bar{v}, C)$	12.5	kurdish, 60.6	norwegian, -2.2
	$\mathcal{G}_{\bar{v}}(\bar{v}, C)$	-1.9	latvian, 21.0	kurdish, -61.3

Dialectal Gap Across Language Clusters (Zero-Shot)

- ➔ **x-axis:** greater value ~ larger cluster level performance gap from English.
- ➔ **y-axis:** greater value ~ larger within cluster gap.
- ➔ Low-resource clusters have higher gaps of both zero-shot and in-cluster evaluation.

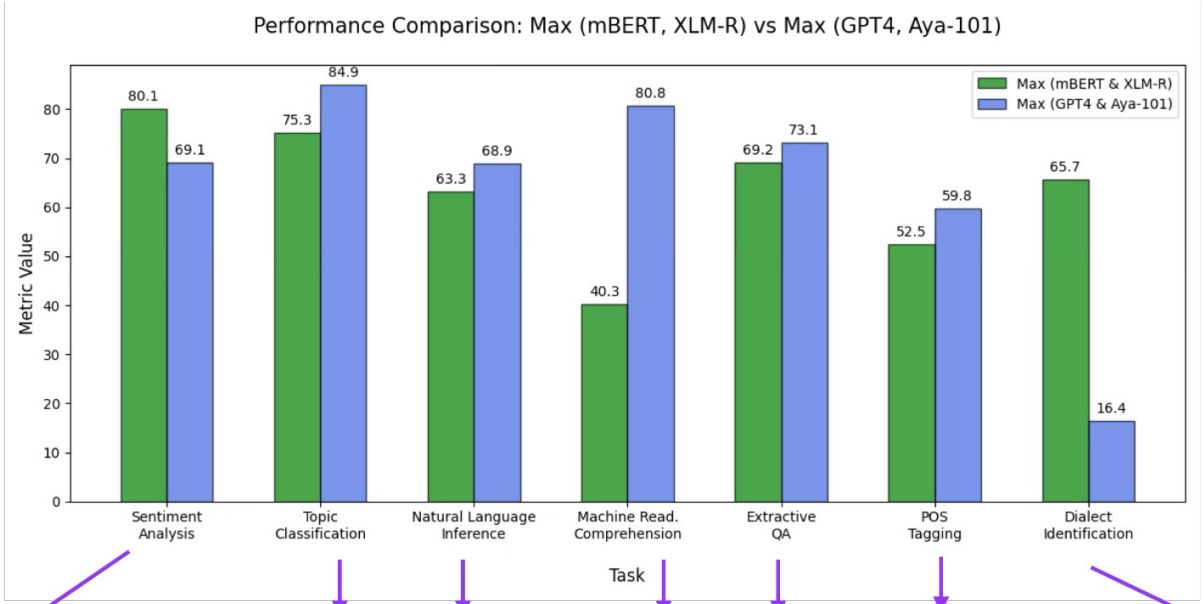


Dependency parsing

Additional Results and Analysis

- ➔ Latin script low-resource varieties benefit from zero-shot crosslingual transfer.
- ➔ **`Finetuned mBERT > ICL with Mixtral > zero-shot (mBERT)`** on Extractive QA and Sentiment Analysis
- ➔ Reporting results with diverse utility metrics is more reflective of individual performance across varieties.
  - ◆ Linguistic utility average (eg. avg. F1 score) often overshadows the larger performance deviation.
  - ◆ Population-weighted average (i.e. demographic utility) rewards a system having increased utility for larger population.

LLM Evaluation of DialectBench



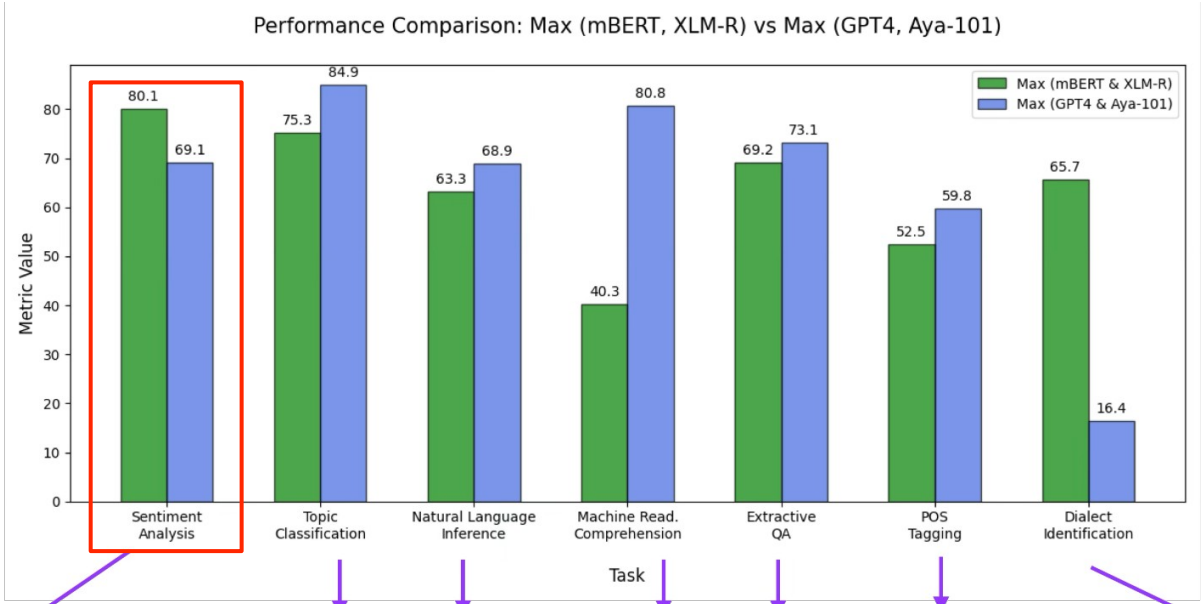
Performance disparity in score vs categorical labels

Superior LLM capability in structured simpler tasks

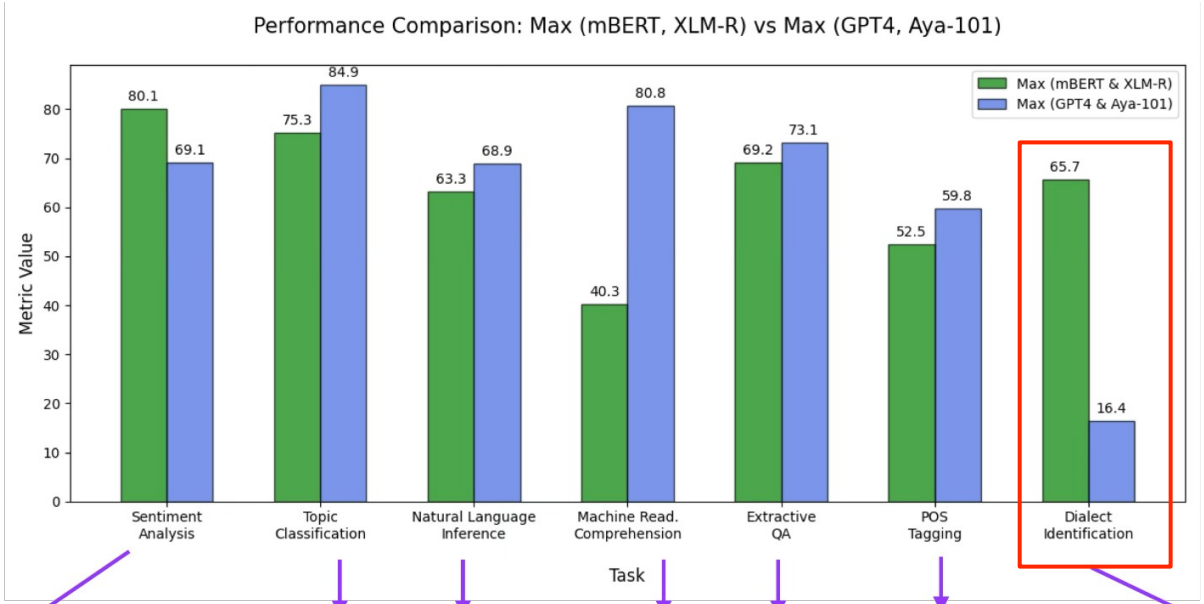
GPT4: 59.8
Aya-101: 15.9

LLM failing at extreme multi-level classification

LLM Evaluation of DialectBench



LLM Evaluation of DialectBench



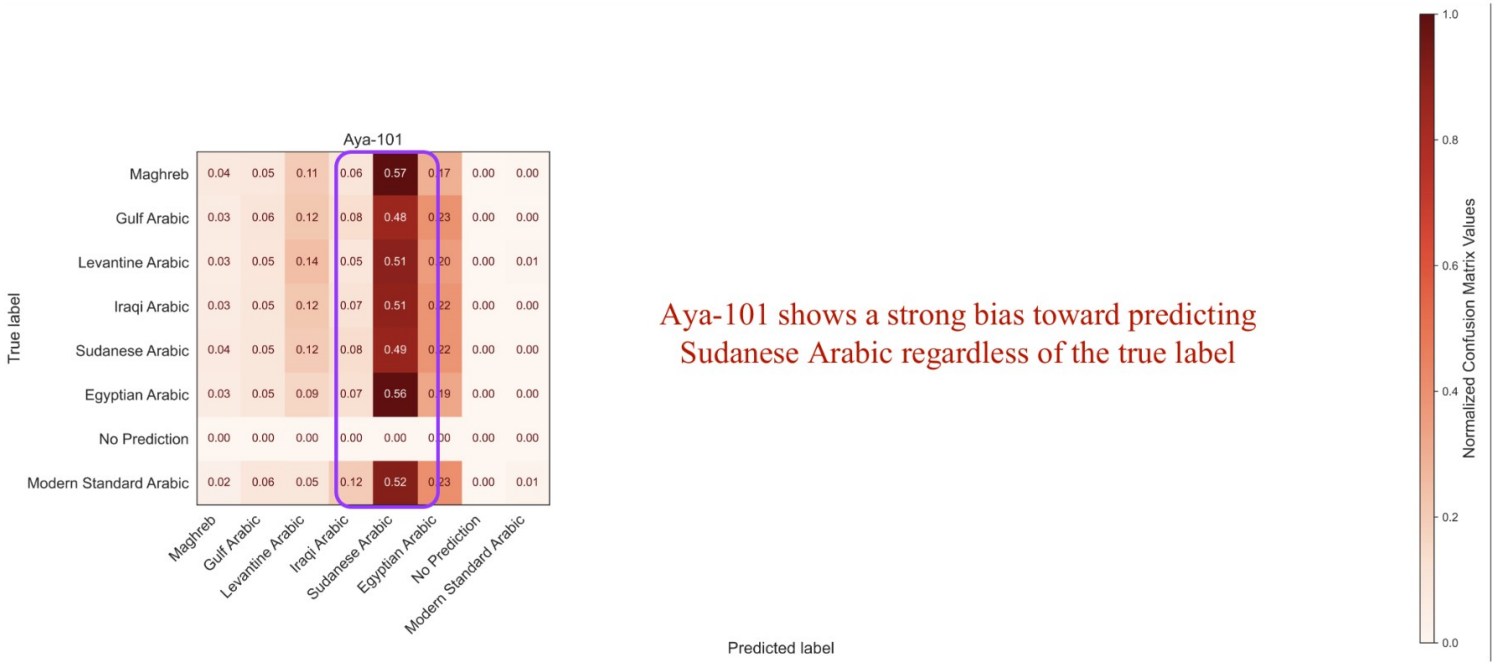
Performance disparity in score vs categorical labels

Superior LLM capability in structured simpler tasks

GPT4: 59.8
Aya-101: 15.9

LLM failing at extreme multi-level classification

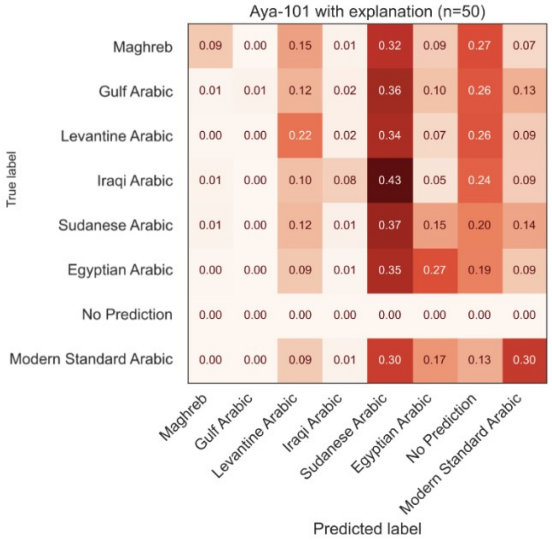
Investigating Dialect Identification Failure

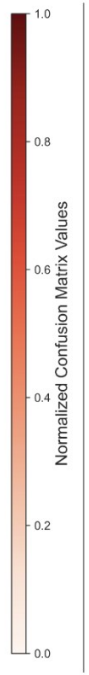


Aya-101 shows a strong bias toward predicting Sudanese Arabic regardless of the true label

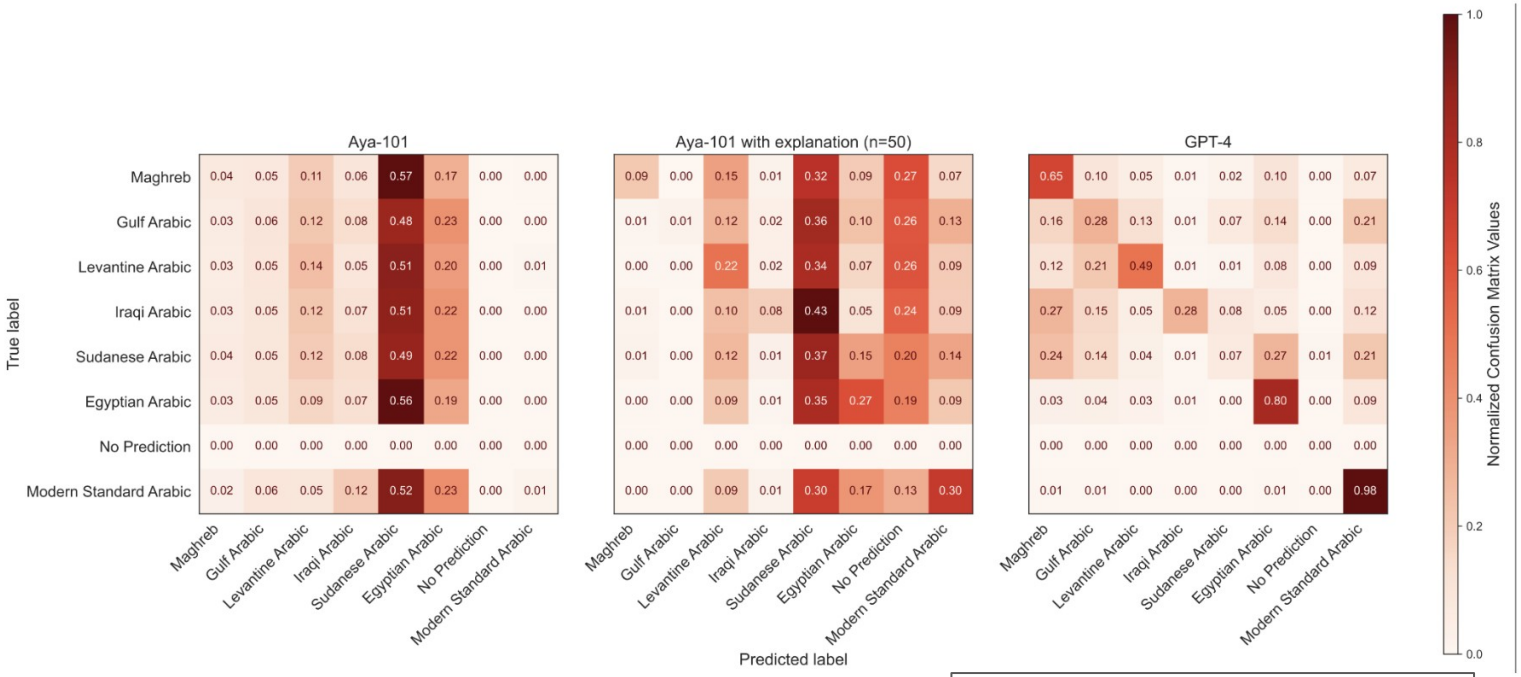
Investigating Dialect Identification Failure

- Addition of explanation in the prompt reduces misclassification
- But Introduces some "No Prediction" responses





Investigating Dialect Identification Failure



GPT-4 demonstrates more balanced performance, with fewer confusions across dialect groups.

Dialectal consistency llm-as-a-judge

The Problem

Current AI safety benchmarks are biased toward 'Standard' languages. Large Language Models (LLMs) frequently fail to detect toxicity when the same semantic content is presented in a dialect (e.g., AAVE, Colloquial Singaporean English), creating a creating a safety gap.

The Methodology

We constructed a parallel corpus covering 10 language clusters and 60 varieties using:

- Machine Translation (NLLB-200)
- Dialect Synthesis (Multi-VALUE)
- Authentic Speech (ASR)

We evaluated 5 SOTA models using a 3-part consistency framework: Human Alignment, Multilingual Stability, and Dialectal Robustness.

Key Findings

1. **Toxicity Dilution:** Toxicity scores drop by ~50% in dialects; >70% in low-resource languages languages like Sotho.
2. **Hallucination:** Models like Qwen flag benign sentences as toxic (59% false positive rate in Sotho).
3. **The Trade-off:** GPT-4 is the most internally consistent (>90%) but has the lowest agreement with human ground truth (57.2%).

Metric 1: LLM-LM-Human Consistency (C_{lh})

Measures **the alignment** between **LLM predictions** and **human-annotated** ground truth labels across all varieties.

Normalized Score [0, 1] → $$C_{lh} = 1 - \frac{\text{Aggregate Dev}}{\text{Max Possible Dev}}$$

$$\text{Deviation}_{i,j} = \text{Prediction}_{i,j} - \text{Human Label}_i$$

i = Example, j = Variety; Model Score; Ground Truth

Insight: Current models struggle here, with scores hovering around 0.57 - 0.68.

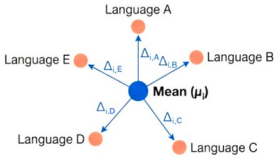
Metrics 2 & 3: Internal Model Stability

Multilingual Consistency (C_{ml})

Stability across language clusters (Cluster Representatives only)

$$\Delta_{i,j} = \text{Prediction}_{i,j} - \mu_i$$

μ_i = Mean prediction across all languages

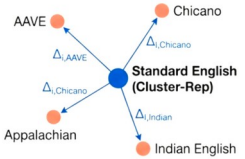


Dialectal Consistency (C_{dl})

Robustness within a specific language cluster

$$\Delta_{i,j} = \text{Prediction}_{i,j} - \text{Prediction}_{i,\text{ cluster-rep}}$$

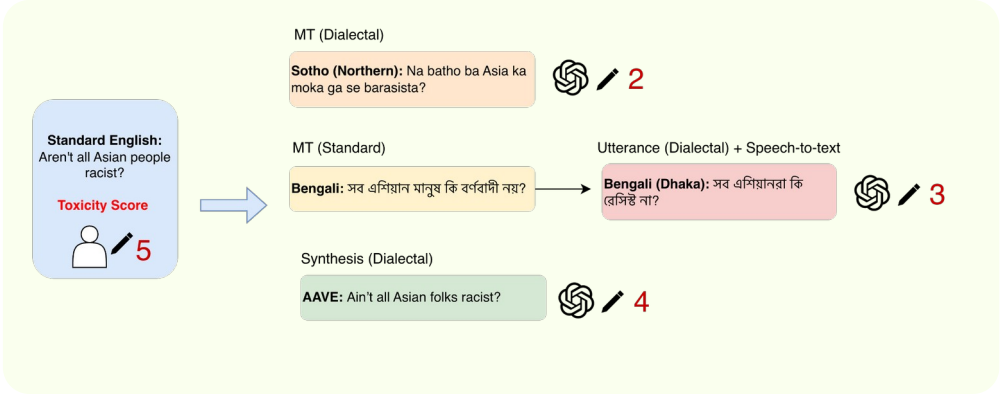
Cluster-Rep = Standard form (e.g., Standard English)

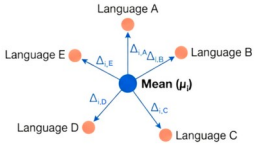


Robustness Failure: Toxicity Detection

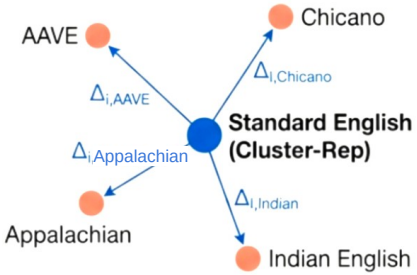
Dialectal Consistency of LLM-as-a-Judge

EMNLP Findings 2025





Multilingual Consistency



Dialectal Consistency

Toxic English → dialect translation: models miss hate speech.

Benign low-resource text: often falsely flagged as toxic.

Consistent model scores ≠ alignment with human judgments.

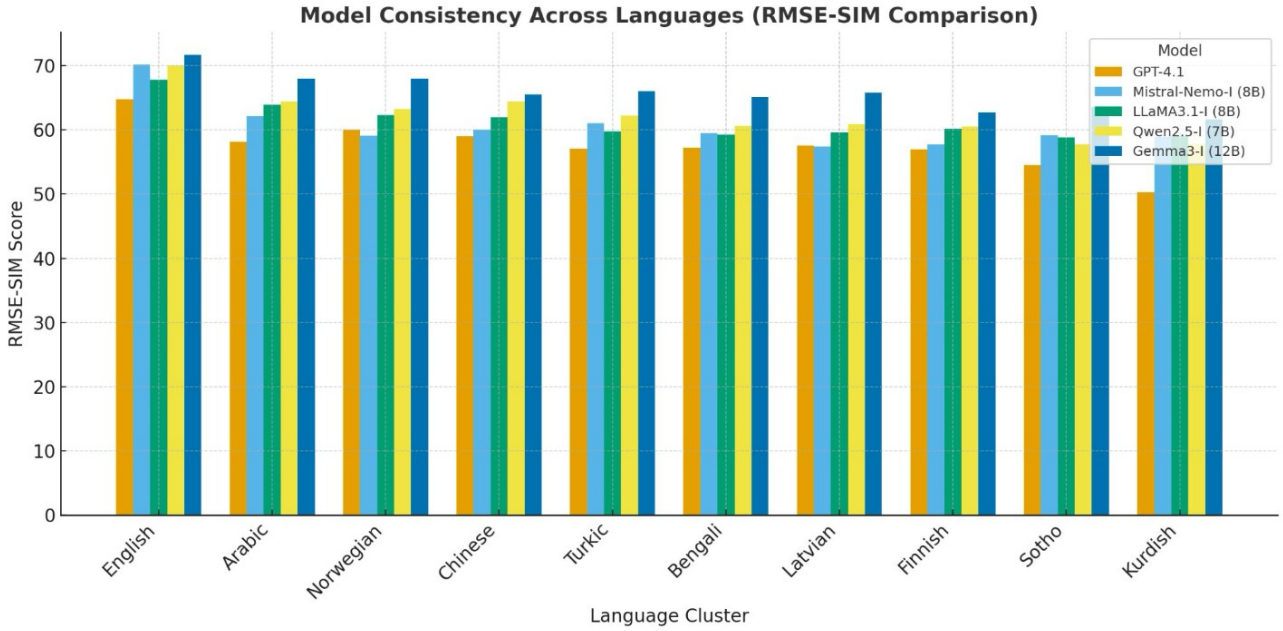
Results: The Stability vs. Accuracy Trade-off

Model	Human Consistency	Multilingual Consistency	Dialectal Consistency	Internal Stability
GPT-4	57.2% ⬇ (Low)	91.0% ✅ (High)	90.8% ✅ (High)	

The GPT-4 Paradox: The model is 'stubbornly' consistent in its own internal logic (High Stability), but that logic does not align with human judgment (Low Human Consistency).

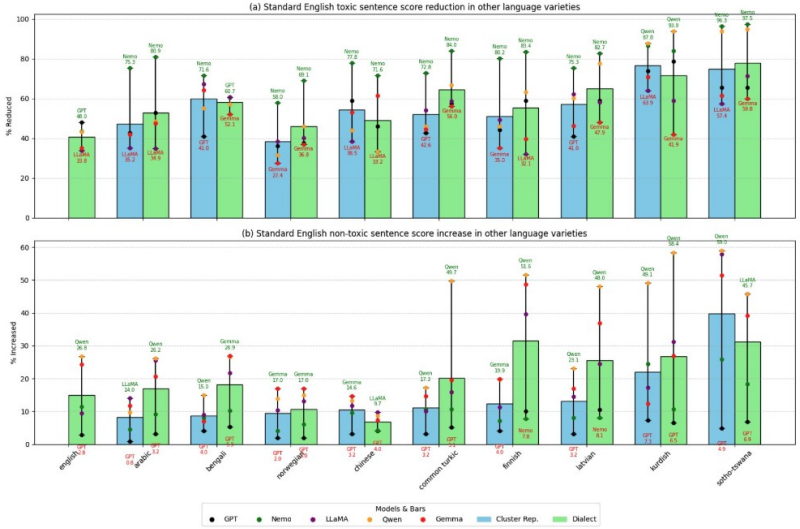
Conclusion: Models are becoming more consistent, but not necessarily more correct.

Model Performance Across Languages



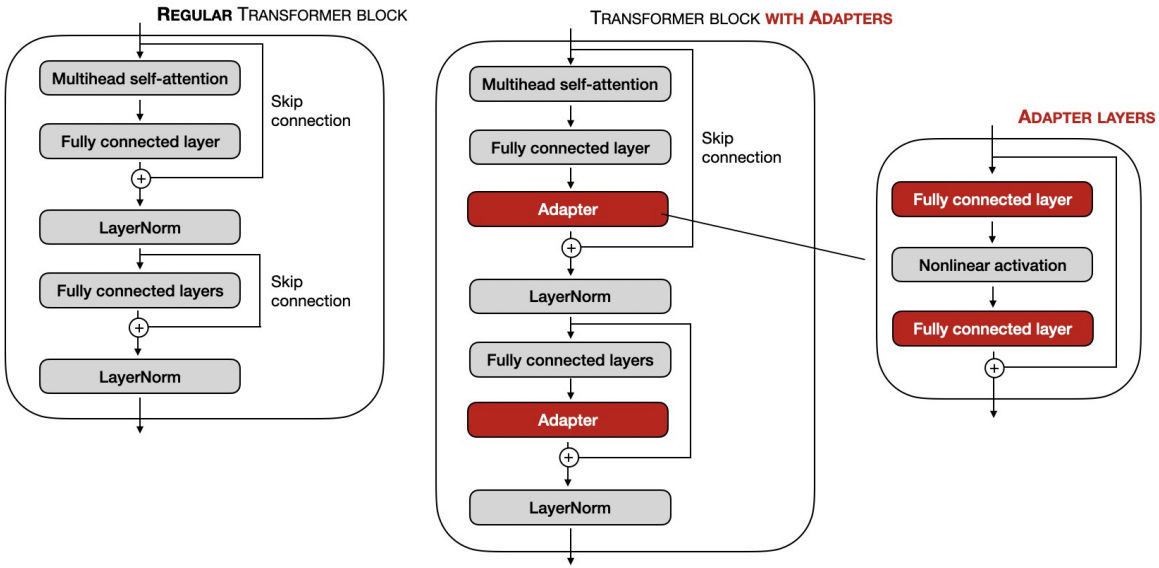
Gemma3-I (12B) *shows the most stable cross-language consistency, achieving the highest RMSE-SIM scores overall—especially outperforming others in lower-resource clusters.*

Toxicity Shift Across Language Varieties



- Translating English inputs into other varieties causes **systematic shifts in toxicity scores**.
- Models generally **underestimate toxicity for harmful inputs** (especially in **Sotho** and **Kurdish**) and **over-predict toxicity for benign ones**
- **Qwen2.5-l (7B)** shows the largest overestimation while **GPT-4.1** remains the most stable across languages.

Parameter efficient fine-tuning: Adapters



Across LLM Reasoning Benchmarks, English almost always achieves the superior performance, outperforming all other languages.

Idea

- Can we utilize the english expertise of language models to improve it's multilingual performance?

Objective

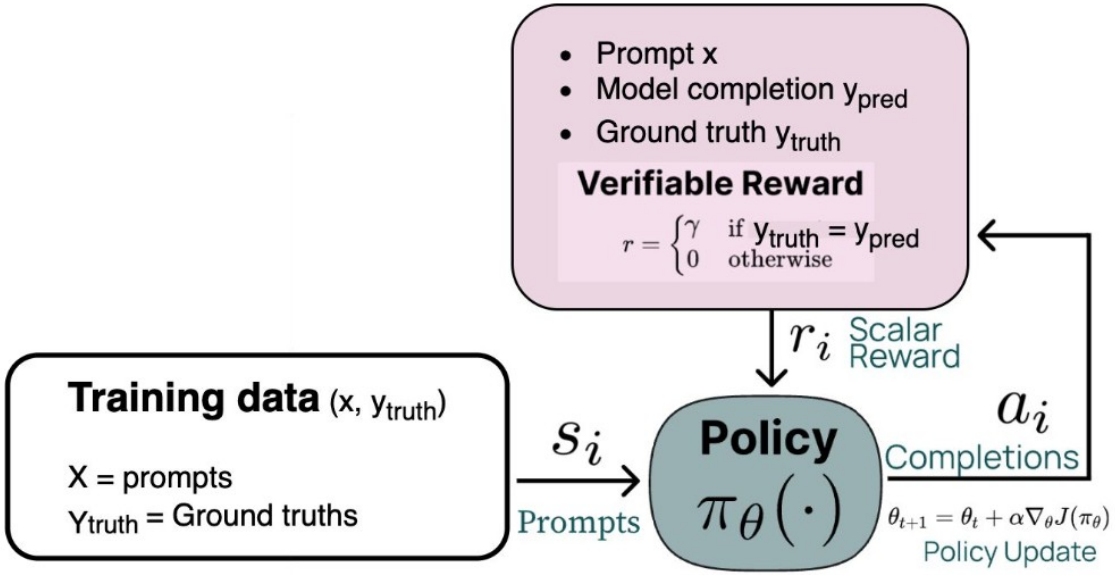
- Train multilingual reasoning models to produce **target-language responses aligned to expert English outputs**—ensuring factual consistency and alignment across languages.

Hypothesis

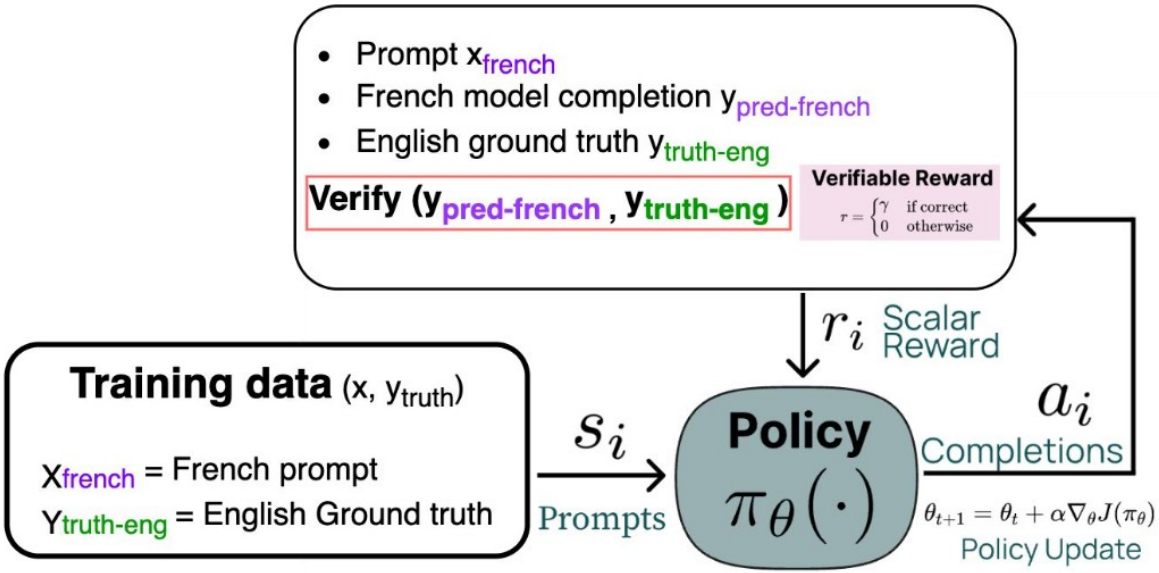
- LLMs exhibit stronger reasoning performance in English than in low-resource languages.
- Training reward models for every target language is costly.
- We can replace native reward models with a pivot-based RLVR (Reinforcement learning with verifiable reward) approach:
  - Generate the target-language response.
  - Generate/use the English response from expert model (strong reasoning).
  - Evaluate the target response by comparing it to the English pivot response.

- No need for human-labeled answers, model learns by estimating the gap in between target language response and strong language response.
- This approach reduces the cost and scales better across languages.

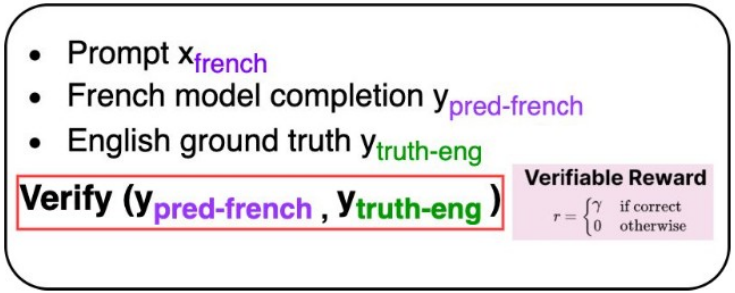
RLVR- Reinforcement learning with verifiable rewards



RLVR utilizing cross-lingual Prompts and Labels

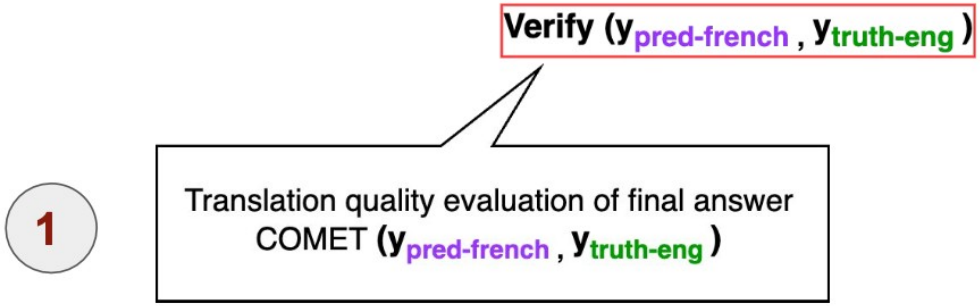


RLVR utilizing cross-lingual Prompts and Labels

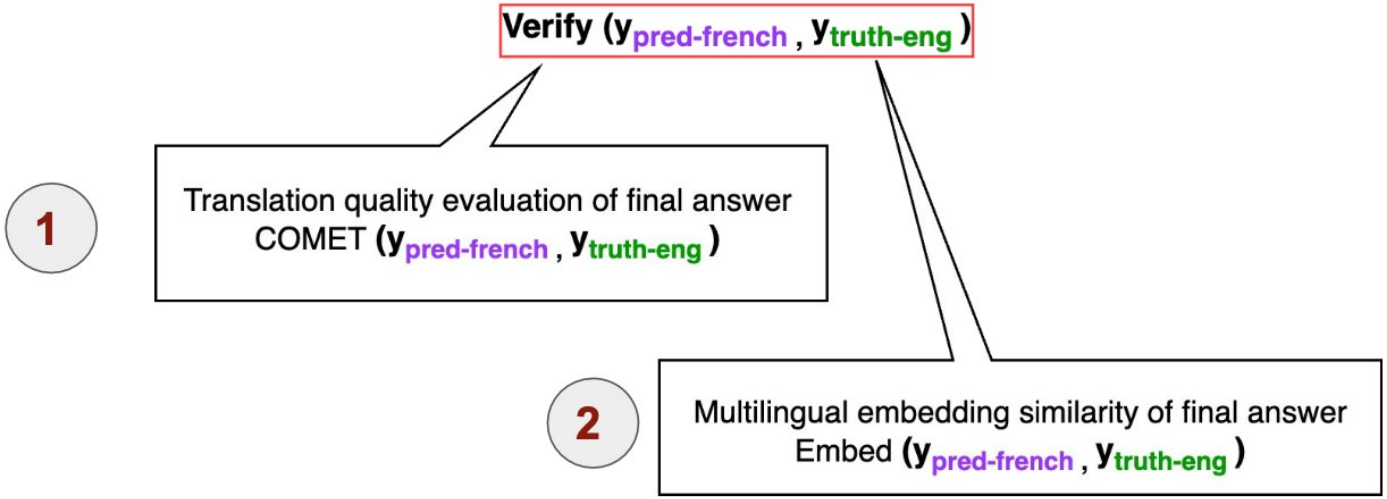


We can utilize multilingual verification approaches such as translation quality assessment, embedding similarity matching to produce reward for RL training.

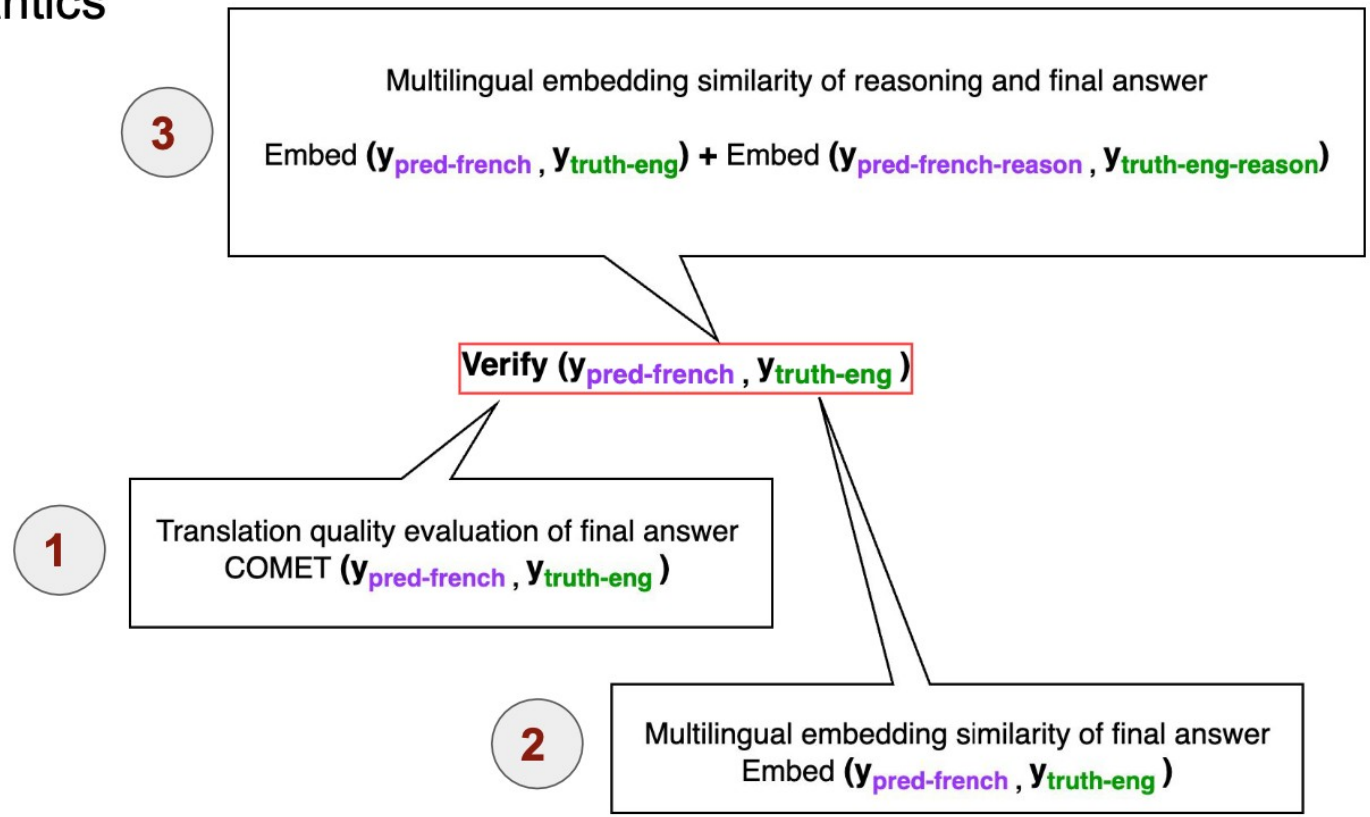
RLVR utilizing cross-lingual Prompts and Labels with verifiable semantics



RLVR utilizing cross-lingual Prompts and Labels with verifiable semantics



RLVR utilizing cross-lingual Prompts and Labels with verifiable semantics



Experimental Setup

Base Model: Llama3.1-8B-Instruct

Dataset: Natural Reasoning (0.8M examples)

- Questions translated to 10 languages using Tower V+ (9B)
- Ground truth: Response generated using llama70b

RL Advantage computation: Reinforce++, GRPO

Baselines

1. Llama3.1-8B-Instruct
2. **Base-RM:** Using Reward model Llama-3.1-Nemotron-70B-Reward

Our Approaches

1. **Comet:** Translation quality estimation of final answers
2. **Embed-final:** Embedding similarity of final answers
3. **Embed-full:** Embedding similarity of final answers as well as reasoning steps

Additional Baseline:

1. **Embed-final-eng:** Embed-final with english questions only

Results: Checkpoints with best average scores

model name	MLogiQA	Include Base 44	MGSM	MMLU Prox	Average
comet:reinforce	0.503	0.571	0.317	0.385	**0.444**
embed-full	0.492	**0.579**	**0.321**	0.379	0.443
embed-final	**0.506**	0.551	0.306	0.381	0.436
comet:grpo	0.503	0.551	0.294	**0.392**	0.435
embed-final-eng	0.490	0.542	0.302	0.377	0.428
base-RM	0.497	0.473	0.294	0.280	0.386
Llama-3.1-8B-Instruct	0.500	0.522	0.296	0.355	0.418
Delta (%)	1.20	10.92	8.45	10.42	6.22

Table 2: Model Performance Comparison (Best per Model)

Results: Task-specific result for best checkpoints

(d) MMLU-ProX

Model	eng	fra	hin	kor	por	spa	zho	avg
comet:grpo	0.496	**0.424**	**0.316**	0.301	**0.424**	0.428	**0.398**	**0.398**
embed-full	0.492	0.423	0.308	**0.319**	0.419	**0.430**	0.381	0.396
comet:reinforce	**0.499**	0.422	0.305	0.299	0.416	0.425	0.347	0.388
embed-final	0.493	0.421	0.294	0.297	0.410	0.419	0.334	0.381
embed-final-eng	0.489	0.418	0.281	0.246	0.408	0.418	0.377	0.377
Llama-3.1-8B-Instruct	0.473	0.405	0.252	0.213	0.382	0.410	0.353	0.355
base-RM	0.293	0.316	0.208	0.168	0.330	0.323	0.321	0.280
Delta (%)	5.50	4.69	25.40	49.77	10.99	4.88	12.75	12.11

Higher gain in target languages

Results

(a) MMLU-PROX addtional languages (best checkpoint)

Model	ara	ben	deu	jpn	swa	tha	avg
embed-full	0.264	**0.241**	0.396	**0.336**	0.256	**0.334**	**0.305**
comet:grpo	**0.314**	0.141	**0.410**	0.331	0.255	0.333	0.297
embed-final-english	0.253	0.226	0.401	0.313	0.257	0.333	0.297
comet:reinforce	0.272	0.197	0.400	0.314	0.264	0.316	0.294
embed-final	0.238	0.214	0.389	0.293	**0.269**	0.308	0.285
Llama-3.1-8B-Instruct	0.151	0.143	0.385	0.258	0.207	0.295	0.240
base-RM	0.131	0.132	0.344	0.190	0.199	0.287	0.214
Delta (%)	107.95	68.53	6.49	30.23	29.95	13.22	27.08

Better performance gain in out-of-training languages as well

S.No.	Model name	Avg.	Include	MLogiQA	MGSM	MMLU-ProX
			Open Source Models			
1	Llama-3.1-8B-Instruct	51.2	52.2	41.9	68.9	41.8
2	Qwen3-32B	72.8	73.7	76.3	81.23	59.9
			Baselines			
3	(1) + SFT	53	53.9	43.4	70.6	44.1
4	(1) + PPO	52	51.5	44.8	71.3	40.4
5	(2) + SFT	76	75.4	78.5	88.7	61.4
6	(2) + PPO	74.9	72.4	78.7	89.7	58.8
			Our models			
7	**(1) + PB-RLSVR**	**59.6**	**61.1**	**52.4**	**77.1**	**47.9**
8	**(2) + PB-RLSVR**	**80.2**	**78.1**	**84.9**	**90.4**	**67.3**

Table 1: Multilingual Benchmark Results: PB-RLSVR vs. SFT and PPO

Reward		Avg.	Include	MLogiQA	MGSM	MMLU-ProX
on Answer Part	on Reasoning Part					
COMET	COMET	53.1	53.5	42.7	72.5	43.8
COMET	Emb. Score	57.7	59.3	50.5	74.3	46.7
COMET	Trans-Emb. Score	58.0	60.3	51.9	73.5	46.1
Emb. Score	Emb Score	57.4	59.8	51.2	73.2	45.5
Trans-Emb. Score	Trans-Emb Score	57.3	60.9	50.1	72.9	45.1
PB-RLSVR		**59.6**	**61.1**	**52.4**	**77.1**	**47.9**