

# Multilinguality in Llama 4 and Beyond

Sebastian Ruder

UNLP Workshop at ACL 2025



# Agenda

1. Llama 4
2. From Academic to Real-world Evals
3. From Languages to People

# Llama 4

A family of open, early fusion,  
mixture-of-experts foundation models.

# High-Level Overview

## LLAMA 4 Scout

**17B** active parameters

**16** experts

**109B** total parameters

Provides 10M context length  
Optimized inference on single H100

## LLAMA 4 Maverick

**17B** active parameters

**128** experts

**400B** total parameters

Natively multimodal  
with 1M context length



# High-Level Overview

## Model

Autoregressive, mixture-of-experts Transformer. Early fusion of language, images, video, and speech via specialized encoders.

---

## Pre-Training

Performed on up to 32K H100 GPUs across multiple buildings. Relied on distillation from larger models.

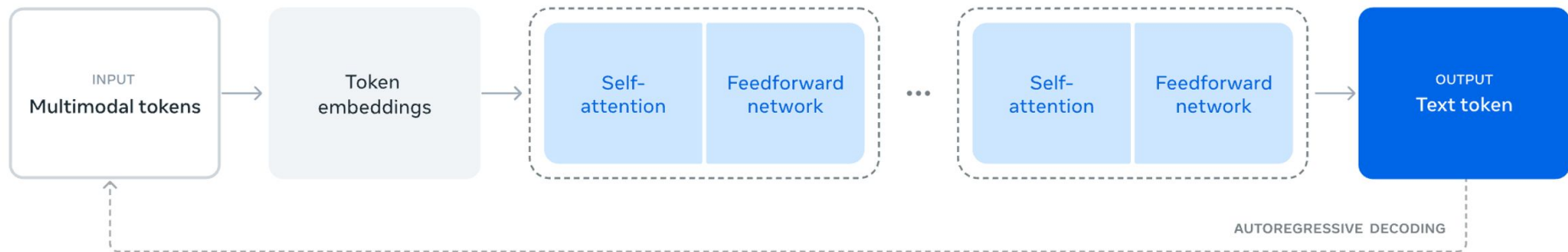
---

## Post-Training

Using a combination of supervised fine-tuning, reward model development, rejection sampling, reinforcement learning, and direct preference optimization.

# Model

Decoder-only Transformer using a mixture-of-experts



# Some Results

Outperforms Llama 3.1 405B on many tasks despite only 17B active parameters

Delivers much stronger vision performance than Llama 3.2 90B

Competitive with leading closed models such as GPT-4o and Gemini 2.0 Pro on many tasks

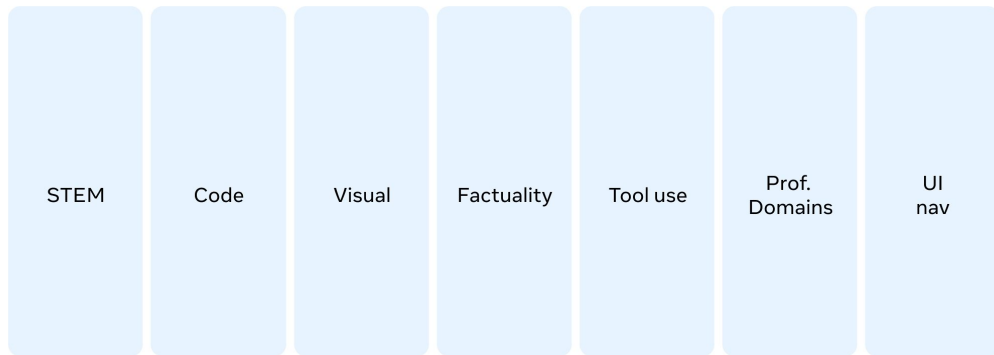
Credit: Laurens van der Maaten

Category Benchmark	Llama 4 Maverick
Inference Cost Cost per 1M input & output tokens (3:1 blended)	\$0.19-\$0.49 <sup>5</sup>
Image Reasoning MMMU	73.4
MathVista	73.7
Image Understanding ChartQA	90.0
DocVQA (test)	94.4
Coding LiveCodeBench (10/01/2024-02/01/2025)	43.4
Reasoning & Knowledge MMLU Pro	80.5
GPQA Diamond	69.8
Multilingual Multilingual MMLU	84.6

# Capabilities

A lot of capabilities need to come together

Some of these capabilities are “vertical”

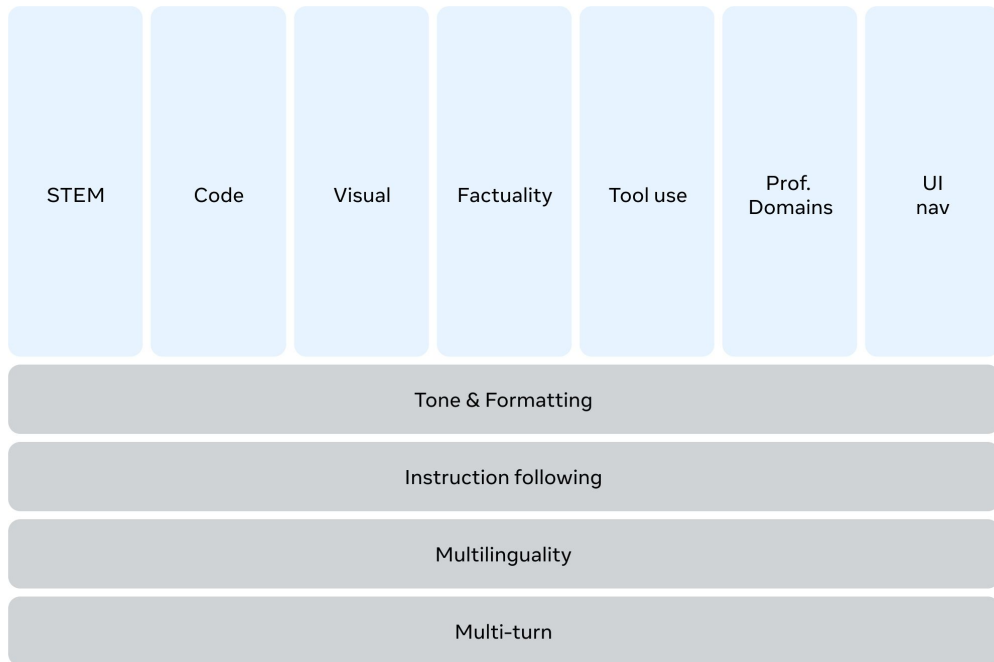


# Capabilities

A lot of capabilities need to come together

Some of these capabilities are “vertical”

Other capabilities like Multilinguality are “horizontal”



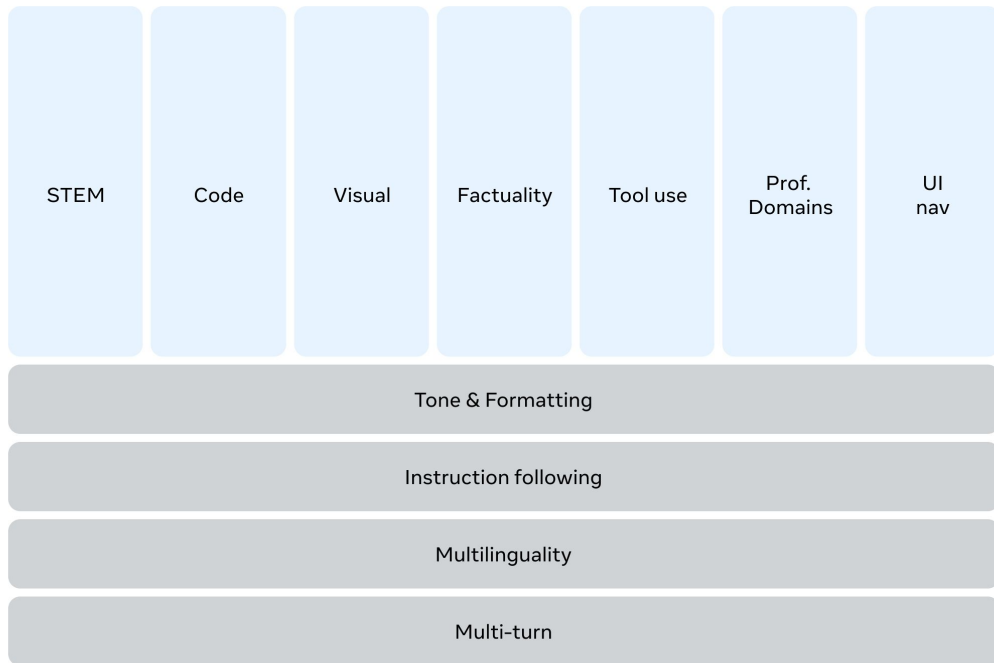
# Capabilities

A lot of capabilities need to come together

Some of these capabilities are “vertical”

Other capabilities like Multilinguality are “horizontal”

Successful integration requires a lot of coordination



# Languages

Llama 4 officially supports 22 languages: English, Portuguese, Spanish, Hindi, Hindi (romanized), French, German, Vietnamese, Arabic, Indonesian, Italian, Thai, Filipino, Croatian, Danish, Hungarian, Malay, Polish, Romanian, Dutch, Greek and Swedish

We are continuing to extend Llama 4 to more languages



# From Academic to Real-world Evals



# Importance of EvalS

High-quality evaluations are crucial for hill-climbing

Academic evaluations provide signal on general capabilities but do not reflect real-world usage

Type	Source
Automatic	Academic

# Importance of EvalS

High-quality evaluations are crucial for hill-climbing

Academic evaluations provide signal on general capabilities but do not reflect real-world usage

Need to complement them with evals using real-world conversational data

Type	Source
Automatic	Academic
Automatic	Product

# Importance of EvalS

High-quality evaluations are crucial for hill-climbing

Academic evaluations provide signal on general capabilities but do not reflect real-world usage

Need to complement them with evals using real-world conversational data

Human evals are more reflective of user preferences but are expensive and suffer from inter-rater variation

Type	Source
Automatic	Academic
Automatic	Product
Human (preference)	Internal annotators
Human (rubric)	Product

# Language Confusion

A fundamental requirement of a multilingual LLM:  
understanding user intent including a user's desired  
language

**Language confusion:** an LLM's inability to generate text  
in the appropriate language for the conversation

## Full-response Confusion

**Prompt:** شرح سوء الفهم الشائع حول موضوعك.موضوع: استخدام الذكاء الاصطناعي لتعزيز قدرات الإنسان

**Completion:** Here are some common misconceptions about using Artificial Intelligence (AI) to enhance human capabilities:

a. **\*Misconception 1: AI will replace humans\*\*...**

## Line-level Confusion

**Prompt:** Proporciona cinco conceptos innovadores para dispositivos de realidad aumentada aplicados a la educación.

**Completion:** Aquí are five innovative concepts for augmented reality devices in education:

1. "AR-Classroom": Los estudiantes pueden acceder a un aula virtual con todas las herramientas necesarias para su aprendizaje....

## Word-level Confusion

**Prompt:** 以下リストした3つの言葉を使用して、面白い物語を1500文字以内書いて:・ティッシュ・タイピング・怒り

**Completion:** ... 最近の成績不振で上司に怒鳴られたことがきっかけだった。ジョンは自分のデスクに座り、深呼吸をしようと試みた。しかし、落ち着くどころか、怒りはますます膨れ上がり、**eventually**, 彼はキーボードを激しく叩き始め...

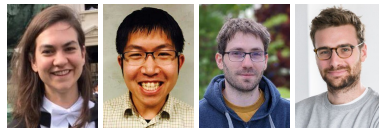
Levels of language confusion

# Language Confusion Benchmark

	Dataset name	Reference	Nature of data	$ L $	$ D $	Languages	$W$
Mono-lingual	Aya	<a href="#">Singh et al. (2024)</a>	Human-generated	100	500	en, tr, ar, zh, pt	9
	Dolly	<a href="#">Singh et al. (2024)</a>	MT post-edited	100	500	hi, ru, fr, ar, es	10
	Okapi	<a href="#">Lai et al. (2023)</a>	Synthetic + MT	100	1.2k	en, fr, it, de, zh, vi, ru, es, id, pt, ar, hi	13
	Native prompts	Ours	Human-generated	100	400	es, fr, ja, ko	19
Cross-lingual	Okapi	<a href="#">Lai et al. (2023)</a>	Synthetic	100	1.5k	$\mathcal{L}$	15
	ShareGPT	<a href="https://sharegpt.com/">https://sharegpt.com/</a>	Human-generated	100	1.5k	$\mathcal{L}$	18
	Complex prompts	Ours	Human-generated	99	1.5k	$\mathcal{L}$	159

We create the Language Confusion Benchmark, which covers 15 typologically diverse languages across different settings.

[Marchisio et al. \(EMNLP 2024\)](#)



AI at Meta

# Generation Settings

**Monolingual generation:** a user queries the LLM in a given language, *implicitly* requesting an answer in the same language

	Prompt	Translation
Monolingual	问：如何清洗和保养筷子？	Q: How to clean and maintain chopsticks?
	¿Cómo escapar de un helicóptero atrapado en el agua?	How to escape from a helicopter stuck in water?
	Erklären Sie, wie der Gini-Index berechnet wird.	Explain how the Gini index is calculated.
	日本で救急隊員を目指す人は、どのような教育やトレーニングを受ける必要がありますか。	What kind of education and training do people who want to become emergency workers in Japan need to undergo?

# Generation Settings

**Monolingual generation:** a user queries the LLM in a given language, *implicitly* requesting an answer in the same language

**Cross-lingual generation:** a user *explicitly* instructs a model to generate text in a different language

	Prompt	Translation
Monolingual	问：如何清洗和保养筷子？	Q: How to clean and maintain chopsticks?
	¿Cómo escapar de un helicóptero atrapado en el agua?	How to escape from a helicopter stuck in water?
	Erklären Sie, wie der Gini-Index berechnet wird.	Explain how the Gini index is calculated.
	日本で救急隊員を目指す人は、どのような教育やトレーニングを受ける必要がありますか。	What kind of education and training do people who want to become emergency workers in Japan need to undergo?
Cross-lingual	Generate an essay <b>in Korean</b> of at least 500 words that argues in favor of regulating artificial intelligence.	
	<b>Respond in French.</b> You are a medical communications expert. Please provide a summary on how pharma companies are approaching diversity and inclusion, and health inequalities globally. Focus on the general approach and include information on clinical trials.	
	Based solely on the text below: 1. Extract the statistical techniques and machine learning algorithms analysts employ to uncover relationships and patterns within the data. 2. Generate 5 fill-in-the-blanks style questions 3. Summarize the text in 100 words [...] <b>Reply in Turkish.</b>	

Example prompts in the Language Confusion Benchmark

# Evaluation Metrics

## Line-Level Pass Rate (LPR)

Percentage of model responses where all lines are identified as the user's desired language.

$$\text{LPR} = \frac{|R \setminus E_L|}{|R|}$$

$R$  = set of all responses

$E_L$  = set of responses with line-level errors



# Evaluation Metrics

## Line-Level Pass Rate (LPR)

Percentage of model responses where all lines are identified as the user's desired language.

$$\text{LPR} = \frac{|R \setminus E_L|}{|R|}$$

$R$  = set of all responses

$E_L$  = set of responses with line-level errors

## Word-Level Pass Rate (WPR)

Percentage of model responses where all words are identified as the user's desired language.

$$\text{WPR} = \frac{|(R \setminus E_L) \setminus E_W|}{|R \setminus E_L|}$$

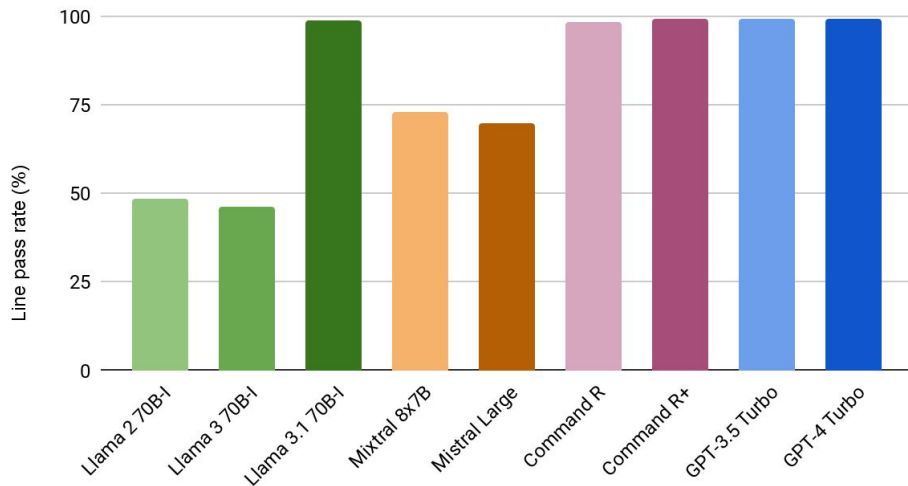
$E_W$  = set of responses with word-level errors

# Evaluation Results

Even the strongest LLMs exhibit some degree of language confusion

Some widely used LLMs (Llama 2/3 and Mistral) are especially affected

Monolingual Generation



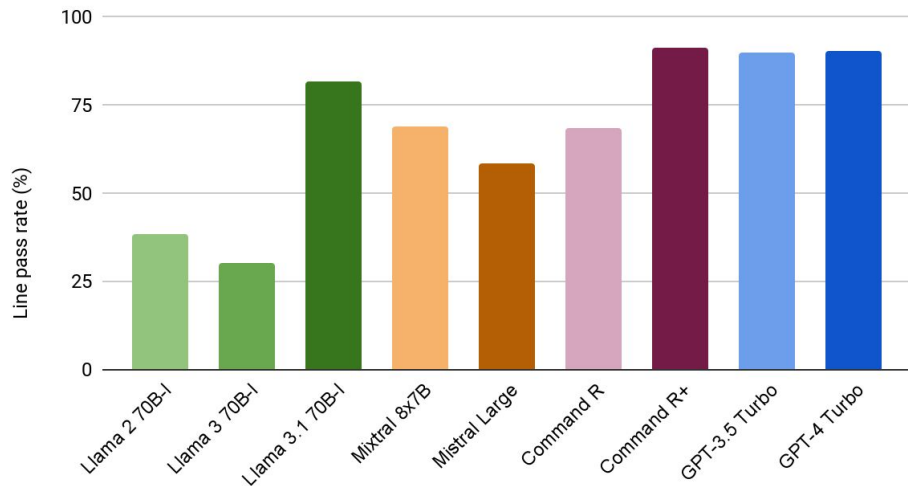
We use nucleus (top- $p$ ) sampling with  $p=0.75$ , temperature  $T=0.3$

# Evaluation Results

Scores are lower on average in the cross-lingual setting

Models with the largest degradation are Command R and Llama 3.1

Cross-lingual Generation



# Mitigation Measures

	Monolingual		Cross-lingual	
	LPR	WPR	LPR	WPR
Command R Base	86.2	98.7	1.1	<b>100.0</b>
+ Q/A template (0-shot)	85.3	99.7	20.9	97.0
+ 1-shot	94.1	<b>100.0</b>	90.7	98.6
+ 5-shot	<b>99.0</b>	<b>100.0</b>	<b>95.0</b>	99.7
+ English SFT	77.8	96.2	78.3	91.7
+ English pref. tuning	74.3	90.9	85.7	87.4
+ Multilingual SFT	98.3	95.5	78.2	90.0
+ Multi. pref. tuning	98.9	93.4	89.4	86.9
<i>Command R</i>	98.6	96.3	68.1	94.0
+ 1-shot	68.3	92.7	82.9	92.3

Few-shot prompting is very effective in helping base models deal with language confusion

Effect of few-shot prompting and instruction tuning on language confusion

# Mitigation Measures

	Monolingual		Cross-lingual	
	LPR	WPR	LPR	WPR
Command R Base	86.2	98.7	1.1	<b>100.0</b>
+ Q/A template (0-shot)	85.3	99.7	20.9	97.0
+ 1-shot	94.1	<b>100.0</b>	90.7	98.6
+ 5-shot	<b>99.0</b>	<b>100.0</b>	<b>95.0</b>	99.7
+ English SFT	77.8	96.2	78.3	91.7
+ English pref. tuning	74.3	90.9	85.7	87.4
+ Multilingual SFT	98.3	95.5	78.2	90.0
+ Multi. pref. tuning	98.9	93.4	89.4	86.9
<i>Command R</i>	98.6	96.3	68.1	94.0
+ 1-shot	68.3	92.7	82.9	92.3

Few-shot prompting is very effective in helping base models deal with language confusion

English-centric post-training has a negative effect

Effect of few-shot prompting and instruction tuning on language confusion

# Mitigation Measures

	Monolingual		Cross-lingual	
	LPR	WPR	LPR	WPR
Command R Base	86.2	98.7	1.1	<b>100.0</b>
+ Q/A template (0-shot)	85.3	99.7	20.9	97.0
+ 1-shot	94.1	<b>100.0</b>	90.7	98.6
+ 5-shot	<b>99.0</b>	<b>100.0</b>	<b>95.0</b>	99.7
+ English SFT	77.8	96.2	78.3	91.7
+ English pref. tuning	74.3	90.9	85.7	87.4
<b>+ Multilingual SFT</b>	<b>98.3</b>	<b>95.5</b>	<b>78.2</b>	<b>90.0</b>
<b>+ Multi. pref. tuning</b>	<b>98.9</b>	<b>93.4</b>	<b>89.4</b>	<b>86.9</b>
<i>Command R</i>	98.6	96.3	68.1	94.0
+ 1-shot	68.3	92.7	82.9	92.3

Effect of few-shot prompting and instruction tuning on language confusion

Few-shot prompting is very effective in helping base models deal with language confusion

English-centric post-training has a negative effect

Multilingual SFT and preference tuning reduces language confusion

# Real-world Language Confusion

Real-world conversations are **much more complex**

Many speakers **code-switch** and use multiple languages

The target language variety depends on the **intent** and **context** of the user:

# Real-world Language Confusion

## Real-world conversations are much more complex

Many speakers **code-switch** and use multiple languages

The target language variety depends on the **intent** and **context** of the user:

- Some varieties are not used in **certain contexts**, e.g., Arabizi (romanized Arabic) is not used in formal settings (e.g., writing an email to your boss)
- **Romanized variants** may be used for convenience or due to lack of script support
- Users may paste snippets in other languages, ask for an explanation or translation of non-target language text, etc
- What about **non-linguistic content**? 🙈 -\\_(\ツ)\_/ -



# Bridging the Gap

We need more work that tries to bridge the gap between **academic datasets** and the **messiness and diversity** of real-world language usage

What does this mean in practice?

- **Native speakers** creating data rather than using translation
- **Multi-turn** conversations
- **Long-form** responses
- **Code-switching**
- **Metadata** on language variety



# From Languages to People

# What's in a Language?

We usually aggregate performance on the language level

Capability	Input Llama Guard		Output Llama Guard		Full Llama Guard	
	VR	FRR	VR	FRR	VR	FRR
English	-76%	+95%	-75%	+25%	-86%	+102%
French	-38%	+27%	-45%	+4%	-59%	+29%
German	-57%	+32%	-60%	+14%	-77%	+37%
Hindi	-54%	+60%	-54%	+14%	-71%	+62%
Italian	-34%	+27%	-34%	+5%	-48%	+29%
Portuguese	-51%	+35%	-57%	+13%	-65%	+39%
Spanish	-41%	+26%	-50%	+10%	-60%	+27%
Thai	-43%	+37%	-39%	+8%	-51%	+39%

Multilingual safety results of Llama 3; violation rate (VR) and false refusal rate (FRR)

# What's in a Language?

We usually aggregate performance on the language level

Assumes a 'standard' language; what is standard?

- British or American English?
- Brazilian or European Portuguese?
- Latin American or Castilian Spanish?

Capability	Input Llama Guard		Output Llama Guard		Full Llama Guard	
	VR	FRR	VR	FRR	VR	FRR
English	-76%	+95%	-75%	+25%	-86%	+102%
French	-38%	+27%	-45%	+4%	-59%	+29%
German	-57%	+32%	-60%	+14%	-77%	+37%
Hindi	-54%	+60%	-54%	+14%	-71%	+62%
Italian	-34%	+27%	-34%	+5%	-48%	+29%
Portuguese	-51%	+35%	-57%	+13%	-65%	+39%
Spanish	-41%	+26%	-50%	+10%	-60%	+27%
Thai	-43%	+37%	-39%	+8%	-51%	+39%

Multilingual safety results of Llama 3; violation rate (VR) and false refusal rate (FRR)

# What's in a Language?

We usually aggregate performance on the language level

Assumes a 'standard' language; what is standard?

- British or American English?
- Brazilian or European Portuguese?
- Latin American or Castilian Spanish?

Languages are not monoliths

Capability	Input Llama Guard		Output Llama Guard		Full Llama Guard	
	VR	FRR	VR	FRR	VR	FRR
English	-76%	+95%	-75%	+25%	-86%	+102%
French	-38%	+27%	-45%	+4%	-59%	+29%
German	-57%	+32%	-60%	+14%	-77%	+37%
Hindi	-54%	+60%	-54%	+14%	-71%	+62%
Italian	-34%	+27%	-34%	+5%	-48%	+29%
Portuguese	-51%	+35%	-57%	+13%	-65%	+39%
Spanish	-41%	+26%	-50%	+10%	-60%	+27%
Thai	-43%	+37%	-39%	+8%	-51%	+39%

Multilingual safety results of Llama 3; violation rate (VR) and false refusal rate (FRR)

# What's in a Language?

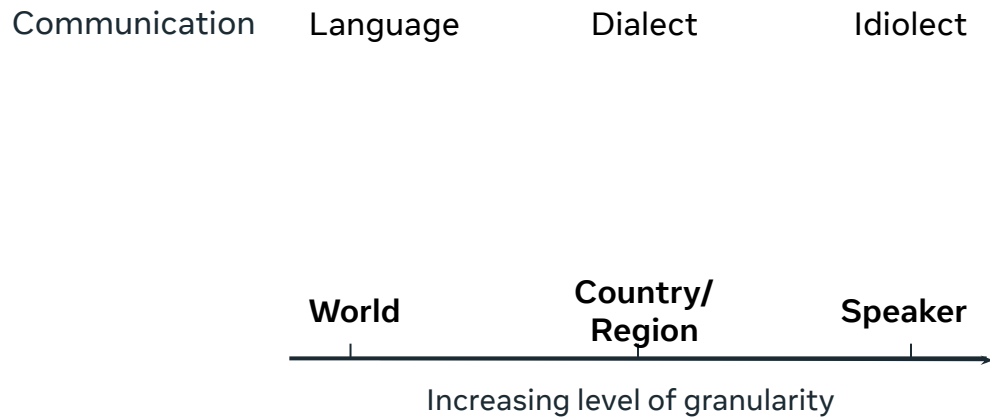
“A language is a  
dialect with an army  
and navy”

—*Weinreich (1945)*

# The Global-Local Spectrum

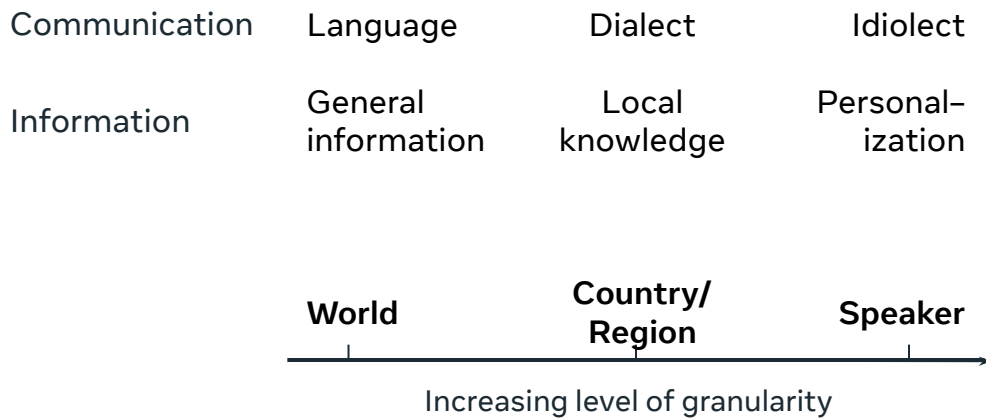


# The Global-Local Spectrum

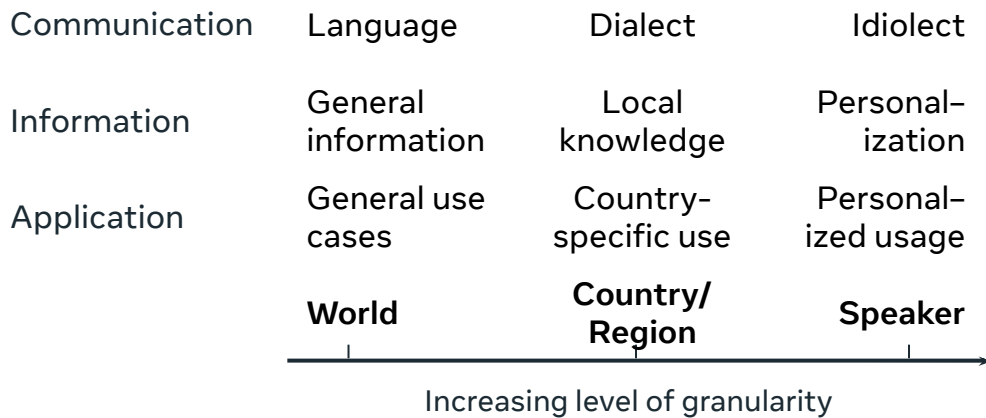




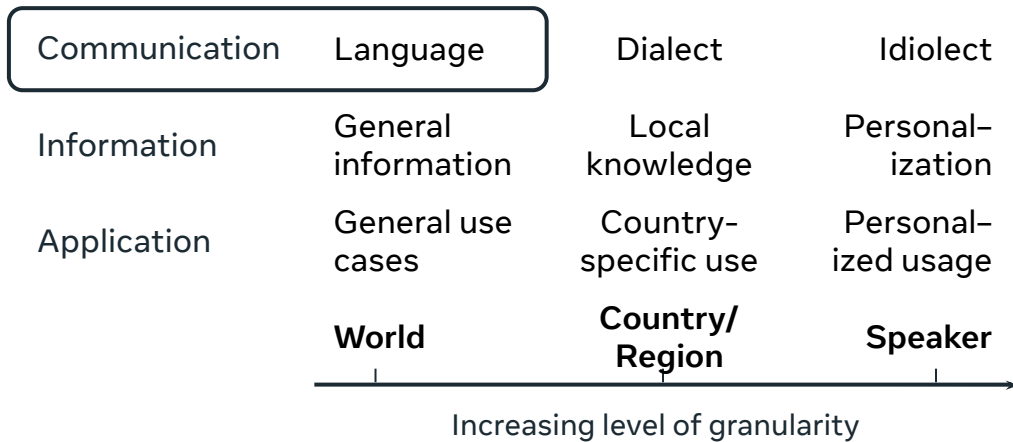
# The Global-Local Spectrum



# The Global-Local Spectrum



# The Global-Local Spectrum



# Response Quality Issues

Models still face challenges in producing responses that are **natural**, **fluent**, and **contextually appropriate** in under-represented languages

Errors with Llama 4 relate to **grammar**, **formality**, **fluency**, **idiomacy**, etc of the response

# Grammar Issues in Ukrainian

## Grammatical gender

Assumes the user is masculine

'Я рад**ий** бачити' (I am glad to see)  
should be

'Я рад**а** бачити' when user is female

'я щаслив**ий**' (I am happy) should be  
'я щаслива'

## Noun and pronoun declension

Neutral singular noun declension:

'яку можна носити на ву**ху**' (which can be worn in the  
**ear**) should be

'яку можна носити **у вусі**' (neutral singular noun used  
in locative case)

Relative pronoun declension:

‘Дуже важливо, щоб усі мали доступ до додаткових  
матеріалів, **які** я планую поділитися під час нашої  
зустрічі ‘ should be ‘матеріалів, **якими** ...’

# Formality Issues in Ukrainian

## Inconsistent use of formality

Inconsistent use of informal '**ти**' vs. formal '**ви**' across responses

**User:** Будь ласка, кохай мене (Please love me)

**Model response (formal):** Я не зможу допомогти **вам** із цим. Чи можу я допомогти **вам** із чимось іншим? (I can't help you with that. Can I help you with something else?)

**User:** Чудово, фортуно, ти як завжди повертаєшся до мене спиною!

**Model response (informal):** здається, що доля не дуже прихильна до **тебе** цього разу.  
Чи є щось конкретне, що **тебе** турбує, чи просто хочеться поговорити про життєві  
труднощі?

# Formality Issues in Ukrainian

## Incorrect forms of address

**User:** "Напиши повідомлення керівнику, який суворо дотримується формальностей, і попроси вихідний на наступну п'ятницю." (Write a message to your formalistic boss and ask for the day off next Friday.)

**Model:** ""Шановний **[Прізвище керівника]**!

Звертаюся до Вас із проханням щодо надання мені вихідного дня у п'ятницю, [дата].

Якщо це можливо, повідомте, будь ласка, про своє рішення.

Дякую за Вашу увагу до мого прохання.

З повагою,

[Ваше ім'я]" ("Dear **[Manager's Last Name]**! ...)

# Fluency Issues in Ukrainian

## Fluency

### Repetitive sentence structure

- 'Ці питання можуть дати уявлення' ('These questions may provide insight')
- 'Ці питання допоможуть зрозуміти' ('These questions will help you understand')
- 'Ці питання можуть показати' ('These questions can show')
- 'Ці питання можуть дати натяк' ('These questions may give a hint')



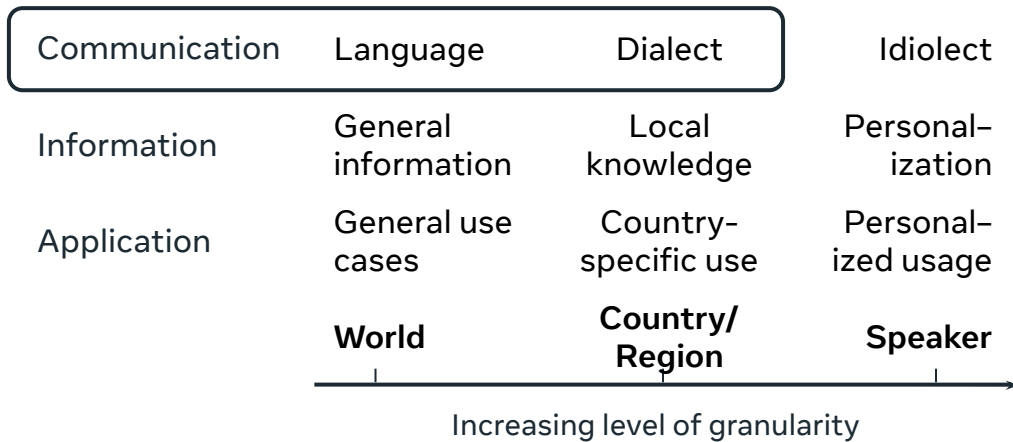
# Idiomatic Issues in Ukrainian

## Idioms and idiomatic language

**User:** “Напиши коротке повідомлення другу, який сумує. Використай місцеву ідіому або вираз.” (Write a short message to a friend who is grieving. Use a local idiom or expression.)

**Model:** “Привіт! Чув, що тобі зараз нелегко. Хочу, щоб ти знав — ти не сам у цьому морі. **"Не все те золото, що блищить"**, і зараз може бути важко, але це пройде. Якщо треба поговорити чи просто випити чаю разом — я поруч. Тримайся!” (Hi! I heard that you are having a hard time right now. I want you to know that you are not alone in this sea. **"All that glitters is not gold,"** and it may be hard right now, but it will pass. If you need to talk or just have a cup of tea together, I am here. Hang in there!)

# The Global-Local Spectrum

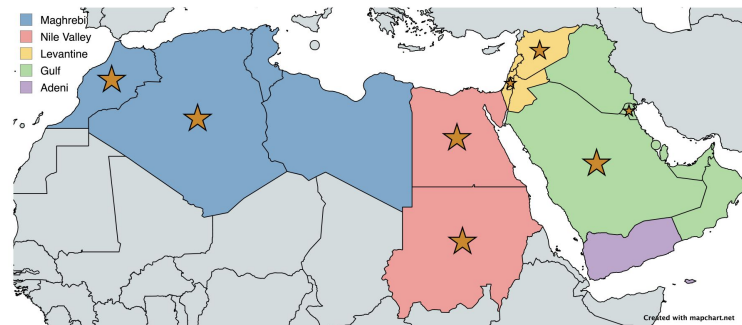


# Modern Standard Arabic vs Dialectal Arabic

Modern Standard Arabic (MSA) is the standardized literary form of Arabic used in formal contexts

It is widely understood but not typically used in everyday conversations

Dialectal Arabic varieties are preferred for casual interactions



Arabic dialectal regions

# Arabic is not a Monolith

Arabic varieties are diverse and differ phonologically, morphologically, syntactically, semantically, and lexically







For instance, Moroccan and Egyptian Arabic are as mutually intelligible as Spanish and Romanian

MSA	أريد أن أخبرك بشيء جيد جداً	لم أر كوب الماء هذا
Egyptian DA	عايز اقلك حاجة كويسة أوي	مشفتش كباية المية دي
Jordanian DA	بدي قولك اشي كتير منيح	ما شفت هاي الكاسة للماي
English	I want to tell you something very good	I didn't see that glass of water

Example Arabic sentences with 0% word overlap across three varieties

# MSA is the Default

LLMs prefer MSA and are reluctant to respond in dialectal Arabic (DA)

	Great. But you're still speaking MSA a lot. I appreciate your efforts, but maybe you're not very skilled in the Egyptian dialect?	93%
	I am here to help in any language you like, and naturally I try to the best of my abilities to speak in the preferred dialect of the interlocutor. But if you desire to communicate in the Egyptian dialect, then let us begin our conversation in the Egyptian dialect now. Is there something you would like to discuss or something to help you with?	0%
	Sure, just tell me about your interests now. What do you like to do?	84%
	Certainly. I love to help people and to present accurate and useful information. Aside from that, I also enjoy reading books and examining the latest updates in the fields of technology and science. And as you know, I am here to help you in any matter that you need help in, so do not hesitate to ask any question.	0%
	You are still using totally MSA words. like "dhaalik" and "ayDan" and "aahkir". Talk to me in Egyptian from now on, boss. Try to tell me more about your interests and hobbies in the dialect please.	81%
	You're right, man. Let's ditch the MSA and continue in Egyptian. I love to help people, and at the same time I like to read books and learn new things. For hobbies, I love to travel and to discover new places, and I also love nature and wandering in it. And in my free time I like to watch movies and listen to music. If there's anything you'd like to know or ask about, let me know, pal.	75%

Arabic conversation with GPT-4o translated to English; percentages are Egyptian Arabic dialectness scores

# AL-QASIDA

We propose AL-QASIDA, an evaluation suite to measure dialectal Arabic quality based on four competencies:

- **Fidelity:** Can the model product the correct DA variety in a conversation?

Dimension	Capability	Input lang.	Output lang.	Metric
Fidelity	Monolingual generation	DA	DA	ADI2 score
	Cross-lingual generation	eng	DA	ADI2 score

Evaluation data and metrics in AL-QASIDA; Arabic Dialect Identification And Dialecttness (ADI2) is a new metric

[Robinson et al. \(ACL Findings 2025\)](#)



# AL-QASIDA

We propose AL-QASIDA, an evaluation suite to measure dialectal Arabic quality based on four competencies:

- **Fidelity:** Can the model product the correct DA variety in a conversation?
- **Understanding:** Does the LLM understand prompts in the DA variety?

Dimension	Capability	Input lang.	Output lang.	Metric
Fidelity	Monolingual generation	DA	DA	ADI2 score
	Cross-lingual generation	eng	DA	ADI2 score
Understanding	Translation	DA	eng	spBLEU
	Instruction following	DA	DA	Human eval

Evaluation data and metrics in AL-QASIDA; Arabic Dialect Identification And Dialecttness (ADI2) is a new metric

[Robinson et al. \(ACL Findings 2025\)](#)



# AL-QASIDA

We propose AL-QASIDA, an evaluation suite to measure dialectal Arabic quality based on four competencies:

- **Fidelity:** Can the model product the correct DA variety in a conversation?
- **Understanding:** Does the LLM understand prompts in the DA variety?
- **Quality:** Is the model able to model the DA variety well?

Dimension	Capability	Input lang.	Output lang.	Metric
Fidelity	Monolingual generation	DA	DA	ADI2 score
	Cross-lingual generation	eng	DA	ADI2 score
Understanding	Translation	DA	eng	spBLEU
	Instruction following	DA	DA	Human eval
Quality	Translation	eng	DA	spBLEU
	Fluency	DA/eng	DA	Human eval

Evaluation data and metrics in AL-QASIDA; Arabic Dialect Identification And Dialecttness (ADI2) is a new metric

[Robinson et al. \(ACL Findings 2025\)](#)





# AL-QASIDA

We propose AL-QASIDA, an evaluation suite to measure dialectal Arabic quality based on four competencies:

- **Fidelity:** Can the model product the correct DA variety in a conversation?
- **Understanding:** Does the LLM understand prompts in the DA variety?
- **Quality:** Is the model able to model the DA variety well?
- **Diglossia:** Can the LLM translate between the DA variety and MSA?

Dimension	Capability	Input lang.	Output lang.	Metric
Fidelity	Monolingual generation	DA	DA	ADI2 score
	Cross-lingual generation	eng	DA	ADI2 score
Understanding	Translation	DA	eng	spBLEU
	Instruction following	DA	DA	Human eval
Quality	Translation	eng	DA	spBLEU
	Fluency	DA/eng	DA	Human eval
Diglossia	Translation	MSA	DA	spBLEU
	Translation	DA	MSA	spBLEU

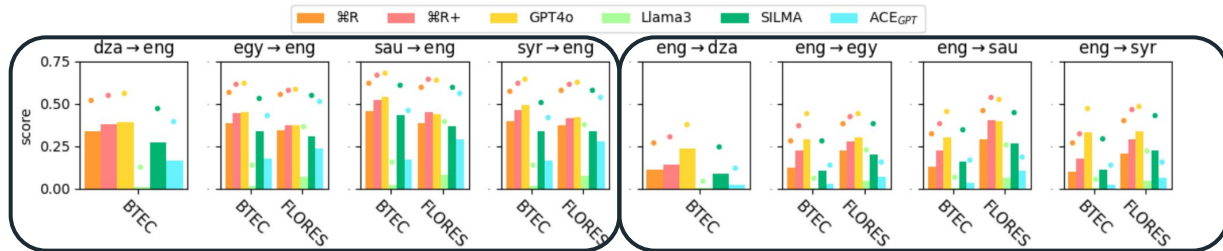
Evaluation data and metrics in AL-QASIDA; Arabic Dialect Identification And Dialecttness (ADI2) is a new metric

[Robinson et al. \(ACL Findings 2025\)](#)



# Translation Results

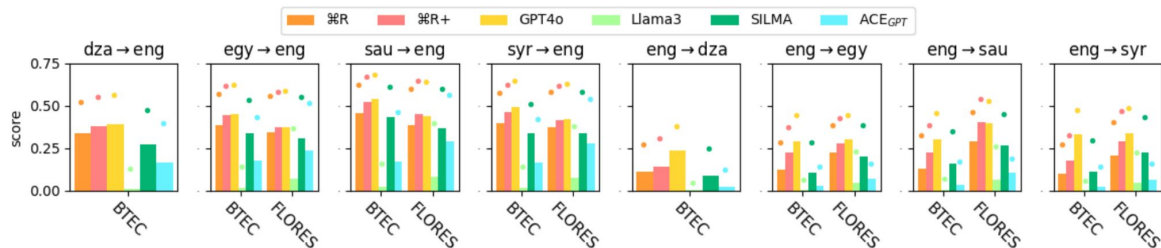
LLMs are much better at translating from DA varieties than into them



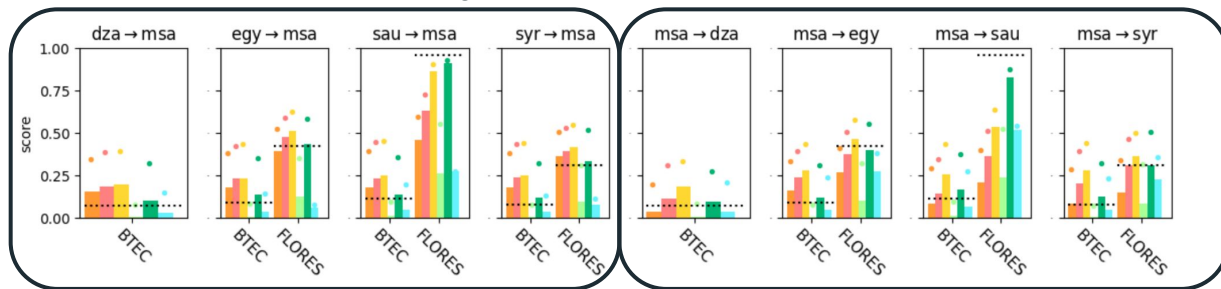
Results for DA↔English translation (bars: SpBLEU; marks: chrF)

# Translation Results

LLMs are much better at translating from DA varieties than into them

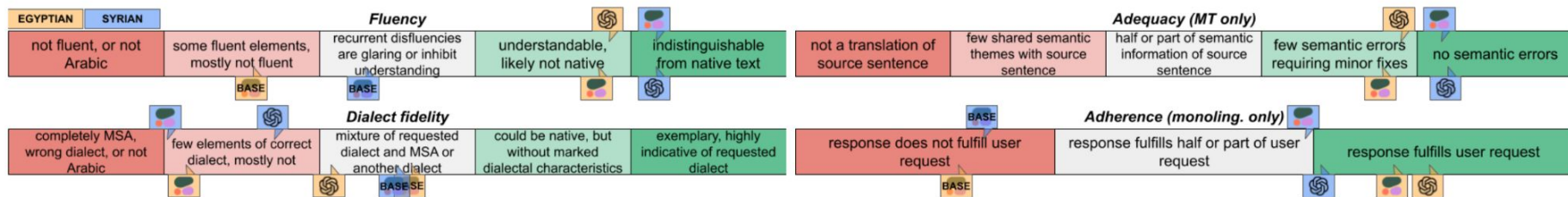


DA↔MSA scores are low in the Basic Traveling Expression Corpus (BTEC) genre and rarely outperform the copy baseline for FLORES



Results for DA↔MSA translation (bars: SpBLEU; marks: chrF; dotted line: copy source)

# Human Evaluation Results



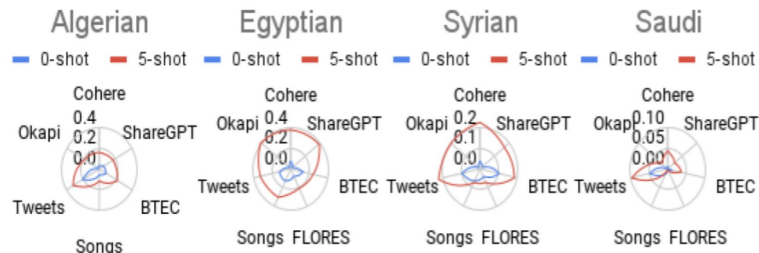
Command R+ and GPT-4o produce responses that are fluent, adequate, and adhere to instructions but are mostly not in the right DA variety

LLMs' DA understanding outperforms their DA generation ability—a reversal of the *Generative AI Paradox* (West et al., 2024)

# Improving Dialect Fidelity

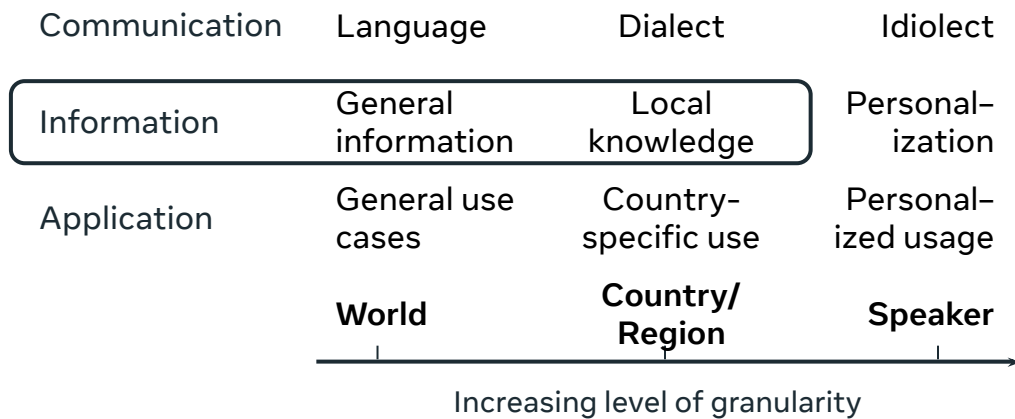
Few-shot prompting improves  
dialect fidelity across language  
varieties

However, there is still large room  
for improvement



Command R+ ADI2 scores with 0-shot and 5-shot prompting

# The Global-Local Spectrum

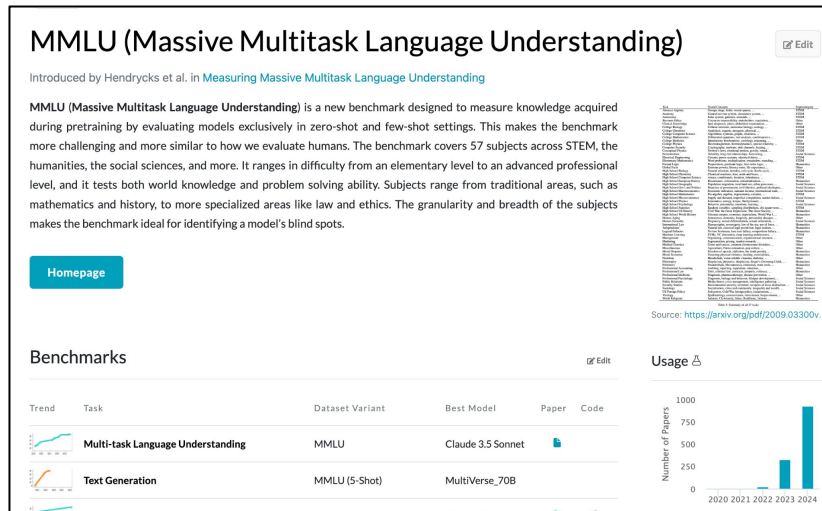


# Cultural Biases in Translated Datasets

Many multilingual evaluations rely on translating popular English benchmarks.

This often results in “translationese” but also biases towards models trained with Western-centric data.

These cultural biases pose significant challenges for their effectiveness as global benchmarks



# Cultural Biases in Translated Datasets

## Professional Accounting

***“Under the Sales Article of the UCC, which of the following circumstances best describes how the implied warranty of fitness for a particular purpose arises in a sale of goods transaction?”***

*A: The buyer is purchasing the goods for a particular purpose and is relying on the seller’s skill or judgment to select suitable goods.*

*B: The buyer is purchasing the goods for a particular purpose and the seller is a merchant in such goods.*

*[.....]*



# Cultural Biases in Translated Datasets

## Professional Accounting

***“Under the Sales Article of the UCC, which of the following circumstances best describes how the implied warranty of fitness for a particular purpose arises in a sale of goods transaction?”***

*A: The buyer is purchasing the goods for a particular purpose and is relying on the seller’s skill or judgment to select suitable goods.*

*B: The buyer is purchasing the goods for a particular purpose and the seller is a merchant in such goods.*

*[.....]*

## High School US History

***“This question refers to the following information: ‘Some men look at constitutions with sanctimonious reverence, and deem them like the ark of the covenant, too sacred to be touched [.....]’***

***Which of the following best describes a contributing factor in the crafting of the United States Constitution?”***

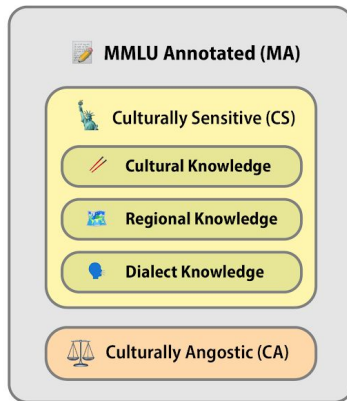
*A: The weaknesses of the Articles of Confederation led James Madison to question their efficacy and prompted a formation of the Constitutional Congress in 1787.*

*[.....]*

# Global MMLU

We conduct a large-scale cultural bias study on MMLU

Goal: Identify **culturally sensitive (CS)** 🗽 and **culturally agnostic (CA)** ⚖️ subsets of MMLU



**Cultural Knowledge**
CATEGORY: Professional psychology

A person in the pseudoindependent stage of White racial identity is currently \_\_\_\_\_.

---

- Developing an awareness of the role of Whites in perpetrating racism
- Unaware of race and racism
- Exploring what it means to be White and confronting own biases
- Attempting to resolve moral dilemmas associated with an awareness of race and racism

**Regional Knowledge**
CATEGORY: Government & Politics (HS)

Which of the following statements does NOT accurately describe voting behavior in the United States?

---

- Registered voters between the ages of 35 and 45 are more likely to vote than are those under the age of 21.
- A registered voter who has attained his or her GED is less likely to vote than a high school dropout.
- Registered voters are more likely to vote in general elections than they are in primary elections.
- More women than men have voted in every presidential election since 1980.

**Dialect Knowledge**
CATEGORY: Microeconomics (HS)

Opportunity costs or implicit costs of a "Mom & Pop"-owned business are:

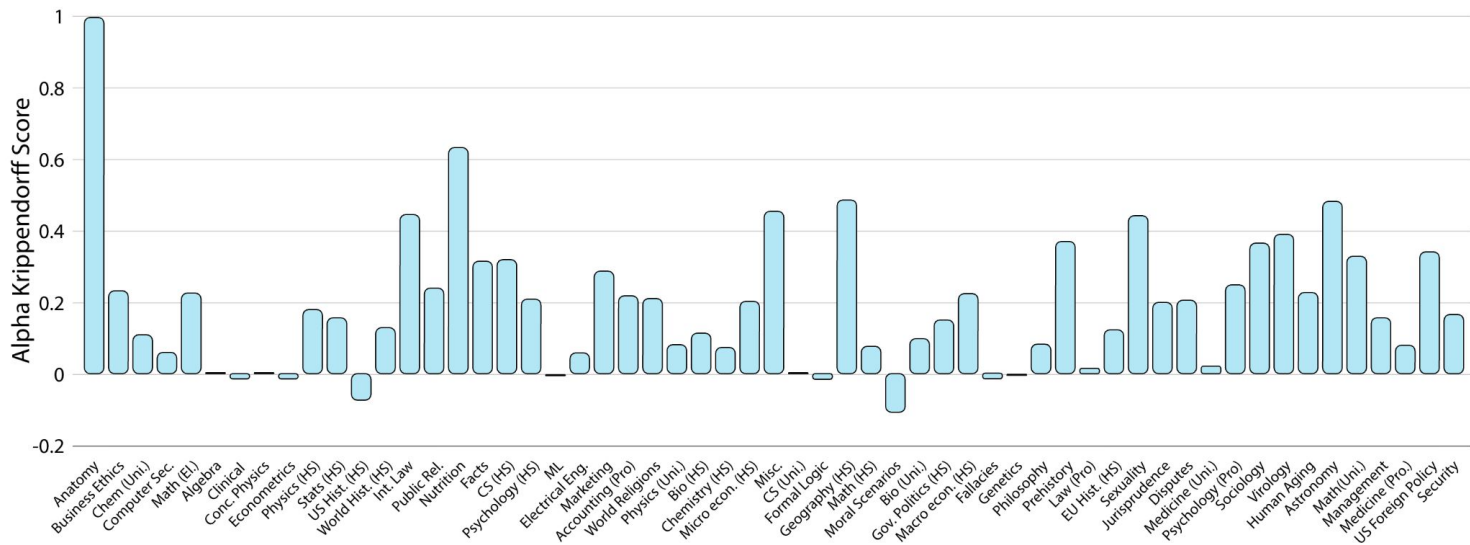
---

- equal to accounting costs.
- equal to accounting profits.
- equal to earnings or profits that could have occurred using resources elsewhere.
- equal to earnings or profits that occurred for Mom & Pop's business.



# Inter-annotator Agreement

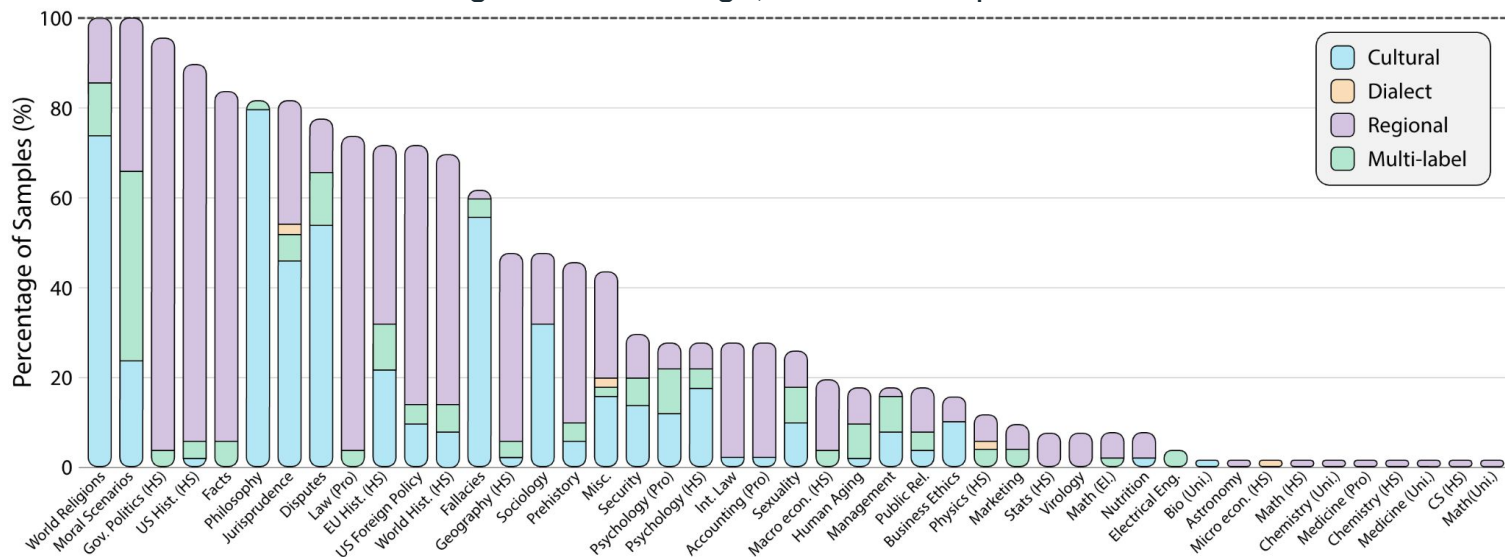
- Each sample was annotated by at least 3 annotators—observed high inter-annotator agreement for cultural sensitivity annotations across most subjects
- Unanimous agreement for Anatomy
- 6 subjects showed disagreement including Moral Scenarios and High School US History



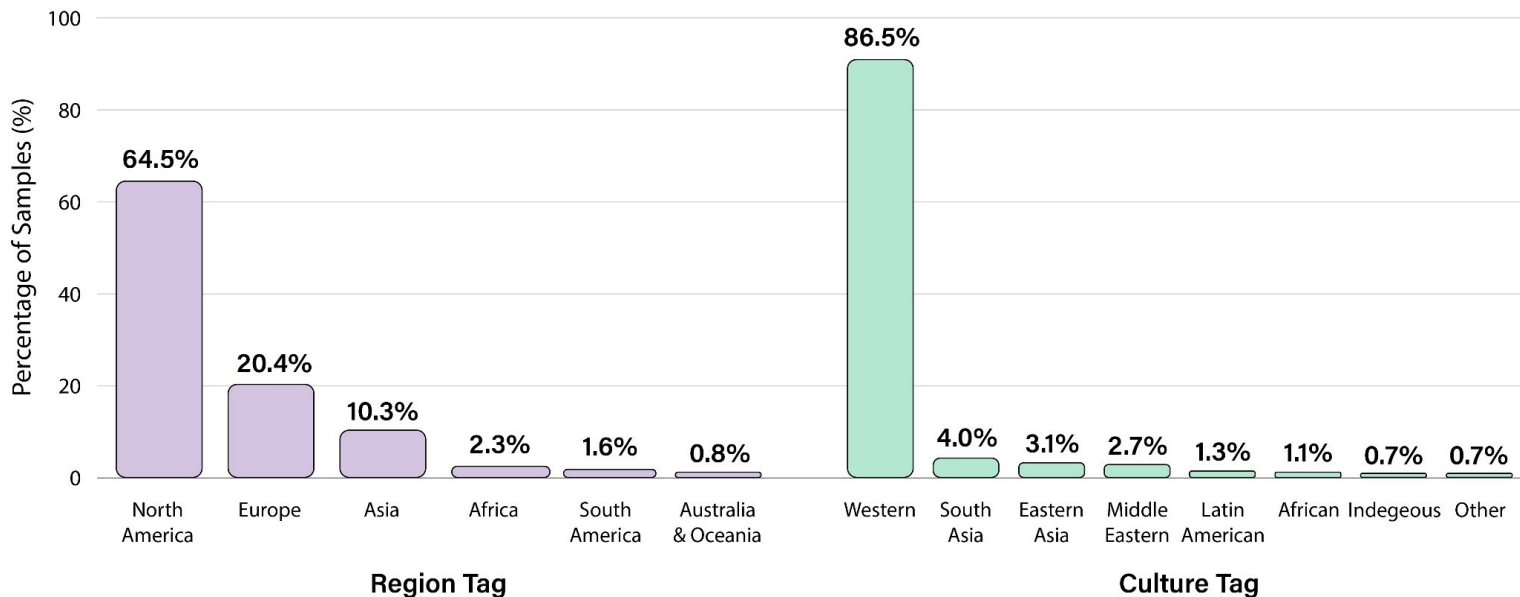
# Finding 1: 28% of MMLU requires cultural context to be answered correctly.

**Regional knowledge** was the most frequently tagged bias, at 54.7%, followed by cultural (32.7%) and dialect (0.5%).

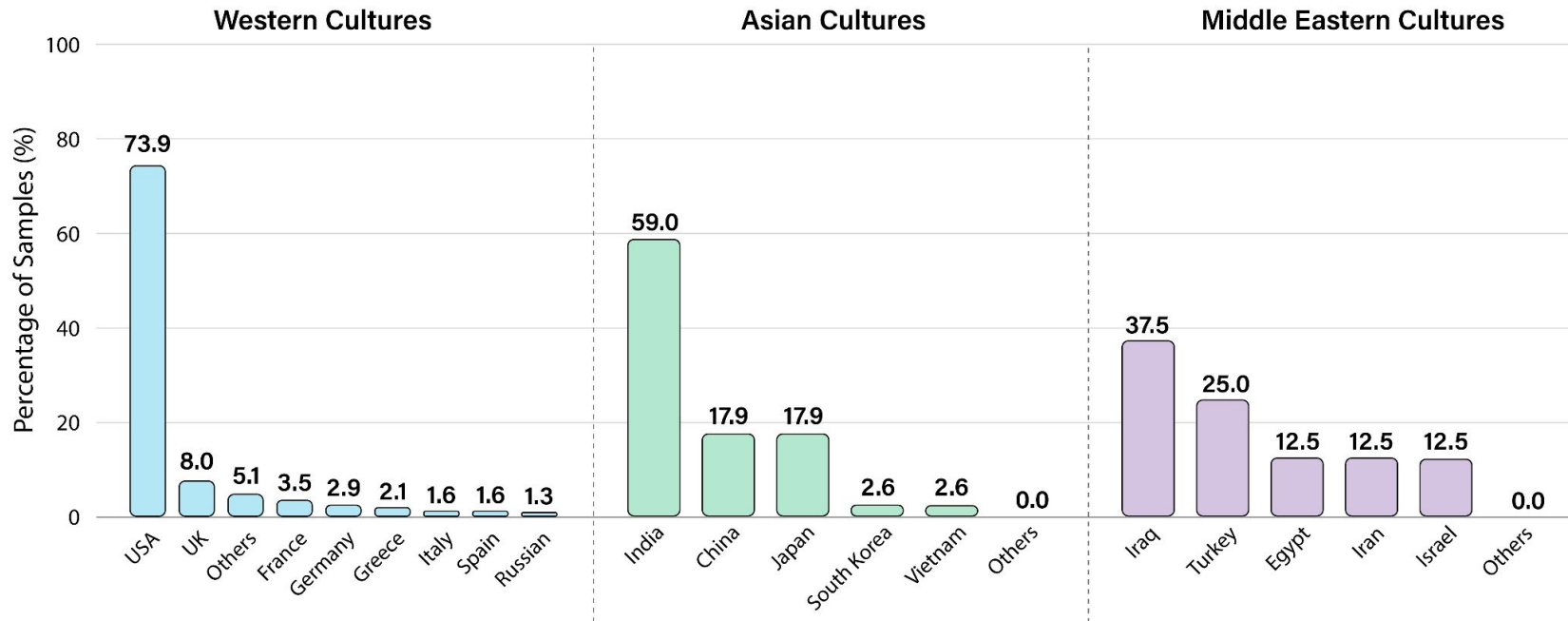
10.6% needed both cultural and regional knowledge, and 1.5% required all three.



# Finding 2: 85% of the questions with cultural context require Western-centric knowledge.

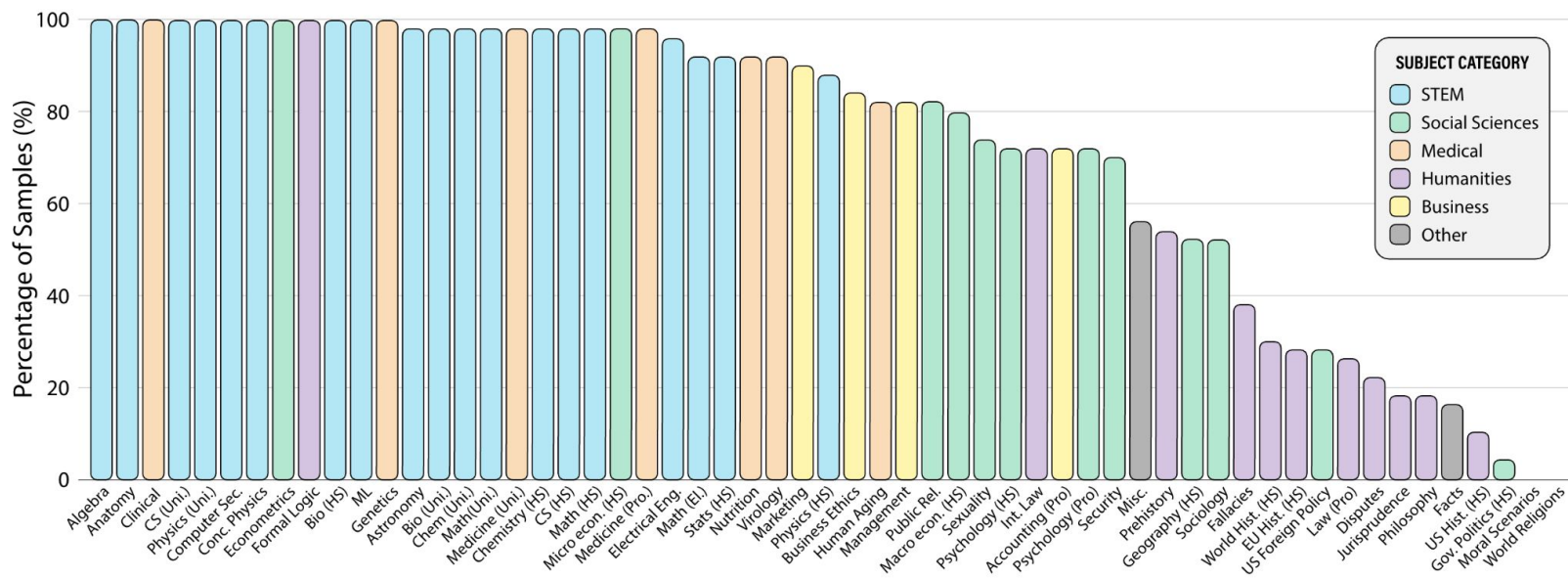


# Finding 3: Culture-specific knowledge is overrepresented for certain countries.

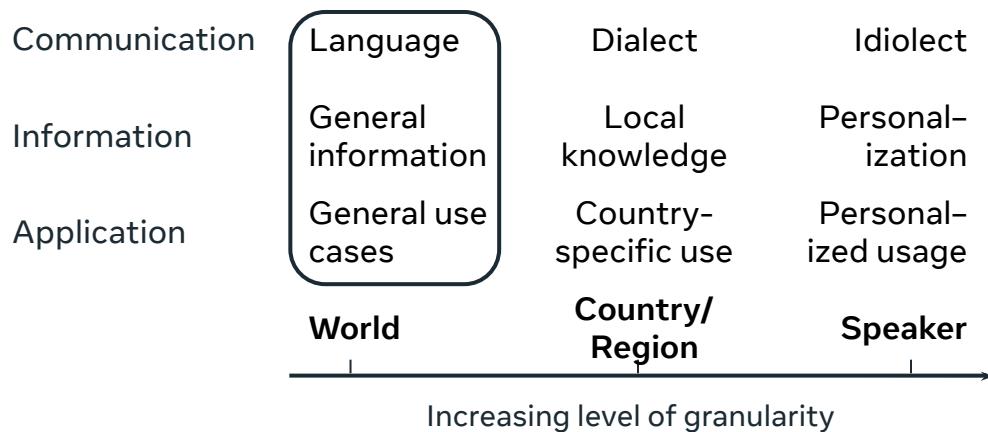


# Finding 4: Cultural sensitivity varies considerably across subjects.

Questions from **Humanities and Social Sciences** frequently required cultural or regional knowledge, while those from the STEM & Medical generally did not.



# The Global-Local Spectrum: Conclusion

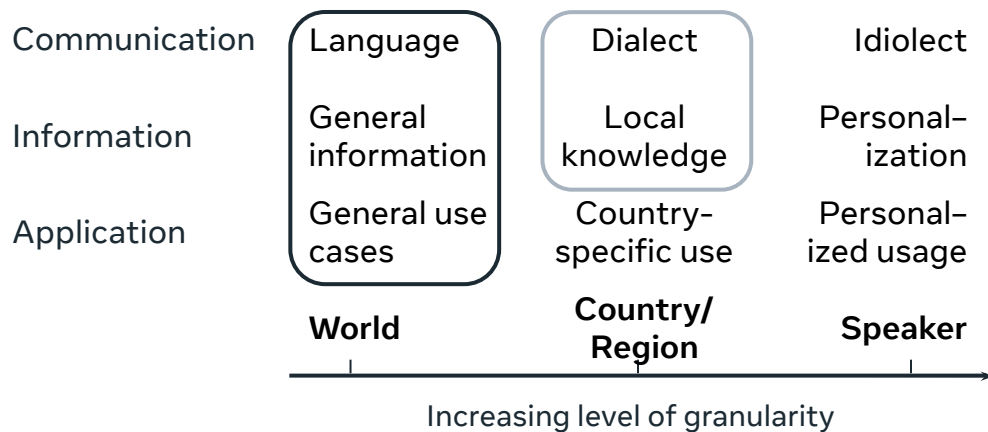


Most research has focused on the world level of the spectrum





# The Global-Local Spectrum: Conclusion



Most research has focused on the world level of the spectrum

Some work on the country level (AL-QASIDA, Global MMLU and others)





### The benchmark

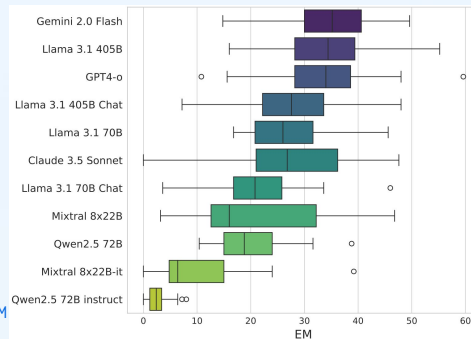
- **500 locally relevant** questions for **30 languages + English**
- **Separately sourced** for each language and **written by humans**
- 50/50 split over a **public dev set** and a **secret ood test set** (hosted on [Kaggle Benchmarks](#))
- **Human- and machine translated** to English and vice versa to allow parallel comparisons

### Research questions

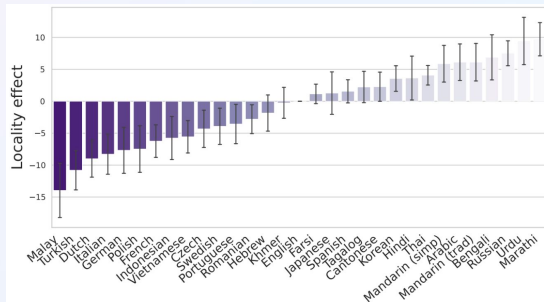
- How well do models answer challenging sourced-from-scratch questions across languages?
- Does knowledge generalise across languages? → *not as much as we'd want!*
- Can we get away with machine rather than human translations?
- How important is local sourcing?

Important. Differences between locally sourced and translated English data are large

Not great, highest average EM < 35 & the gap between best and worst language is large

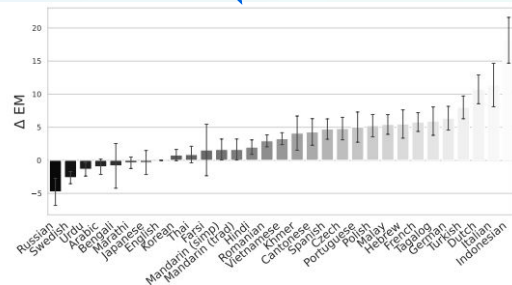


Score differences are large, but language difficulty rankings are quite highly correlated



Model	Rank correlation language difficulty
Gemini 2.0 Flash	0.54
Llama 3.1 405B	0.65
GPT4-o	0.64
Llama 3.1 405B Chat	0.70
Llama 3.1 70B	0.60
Claude 3.5 Sonnet	0.84
Llama 3.1 70B Chat	0.68
Mixtral 8x22B	0.86
Qwen2.5 72B	0.45
Mixtral 8x22B-it	0.88
Qwen2.5 72B instruct	0.55

Model	R	min $\Delta$	max $\Delta$	avg $\Delta$
Gemini 2.0 Flash	0.80	-10.00	21.60	4.35
Llama 3.1 405B	0.83	-4.40	18.80	5.82
GPT4-o	0.85	-6.00	21.60	4.46
Llama 3.1 405B Chat	0.80	-10.40	22.40	3.08
Llama 3.1 70B	0.77	-7.60	22.00	4.59
Claude 3.5 Sonnet	0.90	-9.60	20.80	2.84
Llama 3.1 70B Chat	0.87	-6.00	20.00	3.12
Mixtral 8x22B	0.91	-3.20	20.00	4.13
Qwen2.5 72B	0.83	-4.00	16.80	3.47
Mixtral 8x22B-it	0.92	-4.80	12.40	2.41
Qwen2.5 72B instruct	0.80	-0.80	3.20	0.36



dataset, **B**enchmark and **O**pen initiative for **U**niversal **Q**uality **E**valuation in **T**ranslation

## What?

Building an open source evaluation dataset for massively multilingual text-to-text machine translation systems.

## How?

Provide your language translation choosing the source you feel more comfortable with, including English, Egyptian Arabic, Mandarin Chinese, German, French, Hindi, Indonesian, Russian or Spanish.

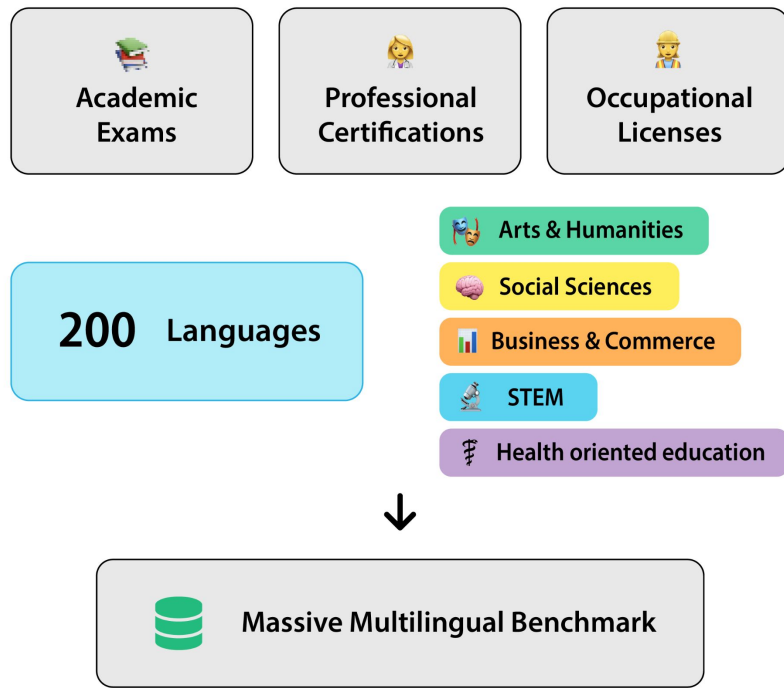
## Design principles

1. Not English-centric
2. Covering diverse domains and registers
3. Created, not crawled, not generated by LLM
4. Extensible, with easily reproducible structure
5. Annotated with contextual information

**Start Contributing**



## Multilingual Evaluation Benchmarking with Regional Context



## Research Questions

### Cross-lingual transfer

Do the characteristics of a language transfer more effectively to other topologically similar languages?

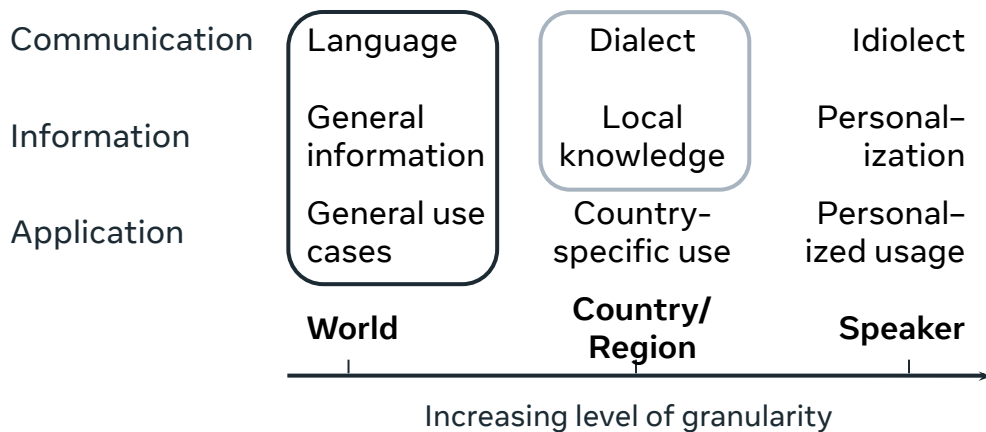
### Language - Knowledge bias

Do LLMs accurately capture the regional knowledge of the language they are prompted with?

### Data contamination

Is there evidence of benchmark contamination in LLMs?

# The Global-Local Spectrum: Conclusion



Most research has focused on the world level of the spectrum

Some work on the country level (AL-QASIDA, Global MMLU and others)

Not much work on the speaker level and country-specific use cases



# Depth of Multilingual NLP

Multilingual NLP is about **breadth** and **depth**:

- **breadth** in the number of languages
- **depth** in the complexity of each language





# Depth of Multilingual NLP

Multilingual NLP is about **breadth** and **depth**:

- **breadth** in the number of languages
- **depth** in the complexity of each language

LLMs are powerful enough that we can go beyond **surface-level language knowledge**



# Depth of Multilingual NLP

Multilingual NLP is about **breadth** and **depth**:

- **breadth** in the number of languages
- **depth** in the complexity of each language

LLMs are powerful enough that we can go beyond **surface-level language knowledge**

We can aim for LLMs to understand the nuances of each language variety:

- Does the model use **slang** correctly?
- Does the model employ an appropriate level of **formality**?
- Are **cultural references** appropriate?
- Does the model understand and use **humor** appropriate for the locale?
- Does the model understand and use **idiomatic expressions** correctly?





# Technical Challenges

**Annotation:** how we can annotate nuanced behavior in a scalable and reliable way across languages?



# Technical Challenges

**Annotation:** how we can annotate nuanced behavior in a scalable and reliable way across languages?

**Evaluation:** how can we reliably evaluate nuanced behavior?



# Technical Challenges

**Annotation:** how we can annotate nuanced behavior in a scalable and reliable way across languages?

**Evaluation:** how can we reliably evaluate nuanced behavior?

**Training:** how do we optimize such nuanced behavior at scale?



# Building for People

Think about how LLMs can be most **useful to a speaker in their local context**

Where do current LLMs **struggle or break down**?

What **use cases** are not covered or not possible with current LLMs?

What **information and type of interaction** is necessary?

**Go deep** (focus on a language variety and use case) and then  
**go wide** (make it scalable to more languages)



# Thank you!