Institute for Infocomm Research
I²R

SeaEval for Multilingual Foundation Models:
From Cross-Lingual Alignment to Cultural Reasoning

Nancy F. Chen

Fellow, Group Leader, Principal Investigator, Senior Principal Scientist
A*STAR

Keynote at ROCLING

20 October 2023

ARES PUBLIC

CREATING GROWTH, ENHANCING LIVES

A bit about myself

- IEEE SPS Distinguished Lecturer 2023, Program Chair for ICLR 2023, Singapore 100 Women in Tech 2021, ISCA Board, APSIPA Board of Governers
- 20+ yrs research experience in speech and language processing, machine learning
- Supervised 100+ students and staff
- Work experience at MIT Lincoln Lab, USA and A*STAR, Singapore
  - Technology translation: Government deployment, spin-off companies, IP licensing
  - Advisor & tech consultant to startups and multinationals

Nancy F. Chen

Team Members and Contributors of SeaEval

Wang Bin

Zhengyuan Liu

Huang Xin

Ding Yang

Fangkai Jiao

Siti Umairah

Nabilah

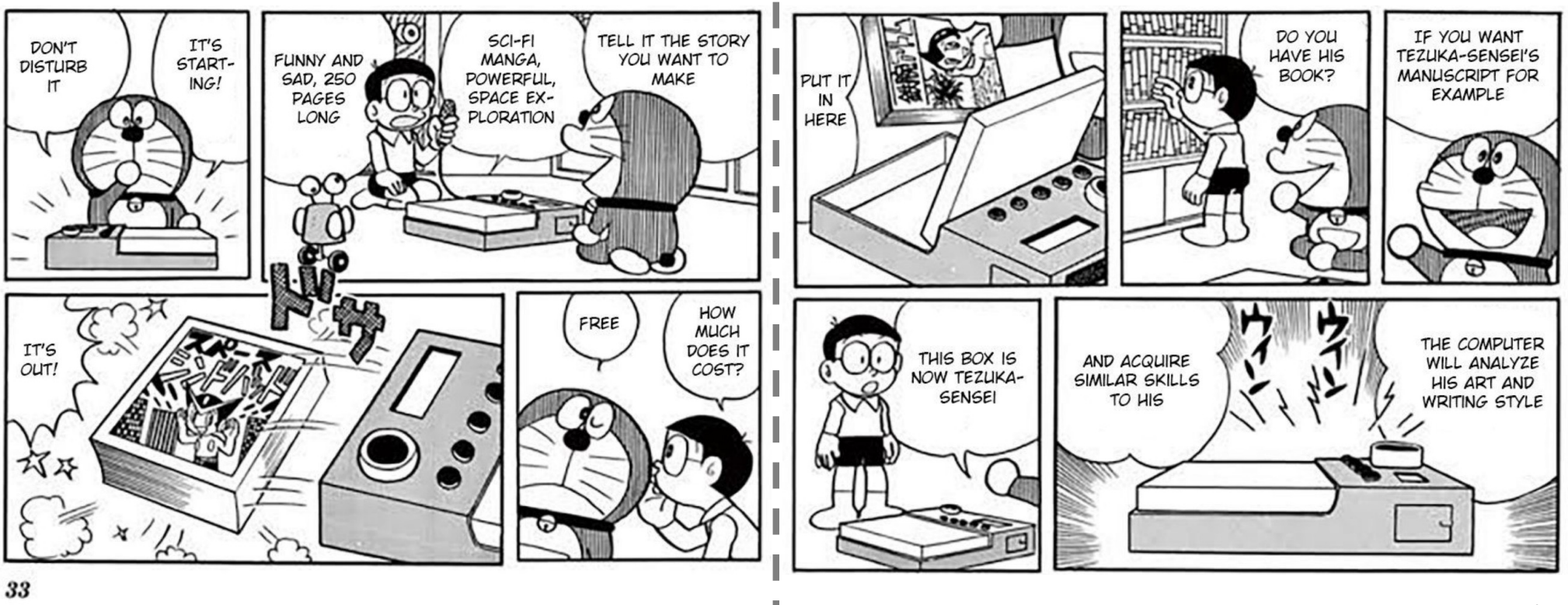
Maryam

Aw Ai Ti

Nancy Chen

Doreamon predicted Generative AI 44 years ago?!

[read from right to left, top to bottom]



What is Foundation Model?

Large Language Models (LLM): Foundation Model's First Success Story

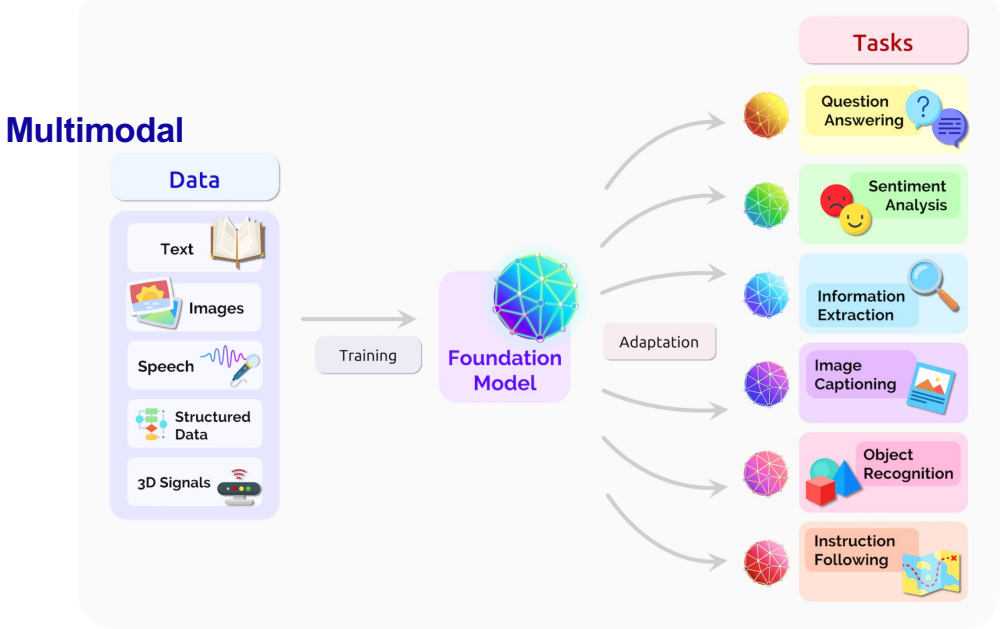


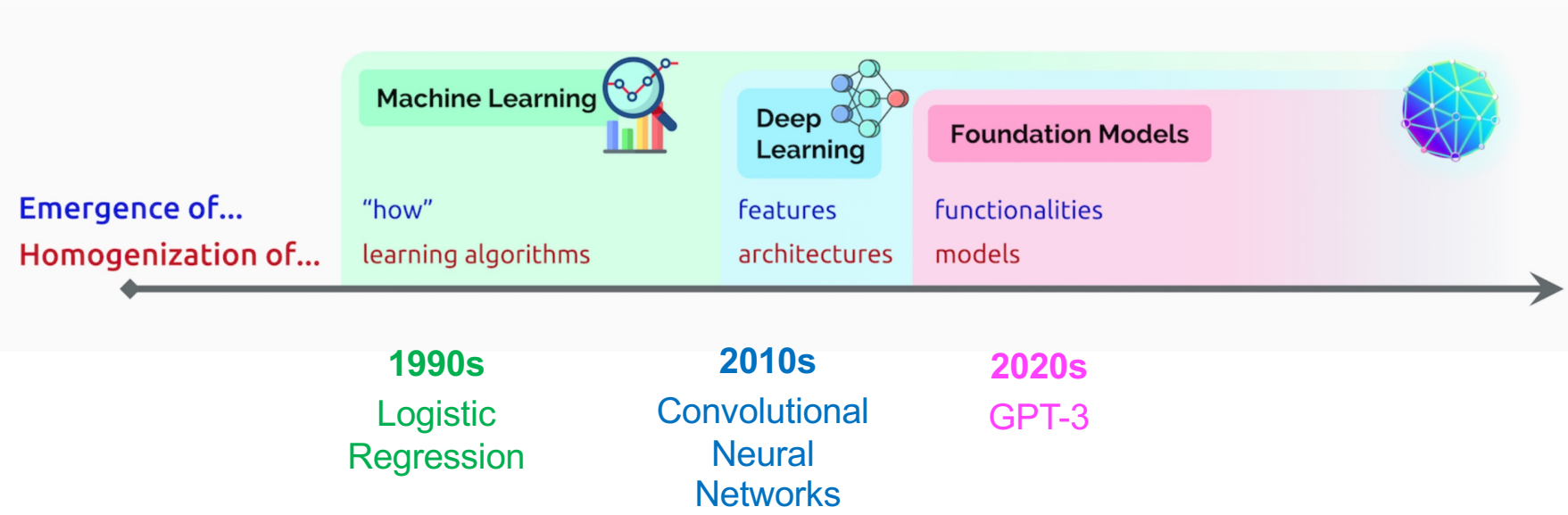
Fig. 2. A foundation model can centralize the information from all the data from various modalities. This one model can then be adapted to a wide range of downstream tasks.

A foundation model can be adapted to various task-specific models

Image adapted from
https://arxiv.org/pdf/2108.07258.pdf%20

The Story of AI for the Past 30 Years

Emergence & Homogenization enable ***Scale***

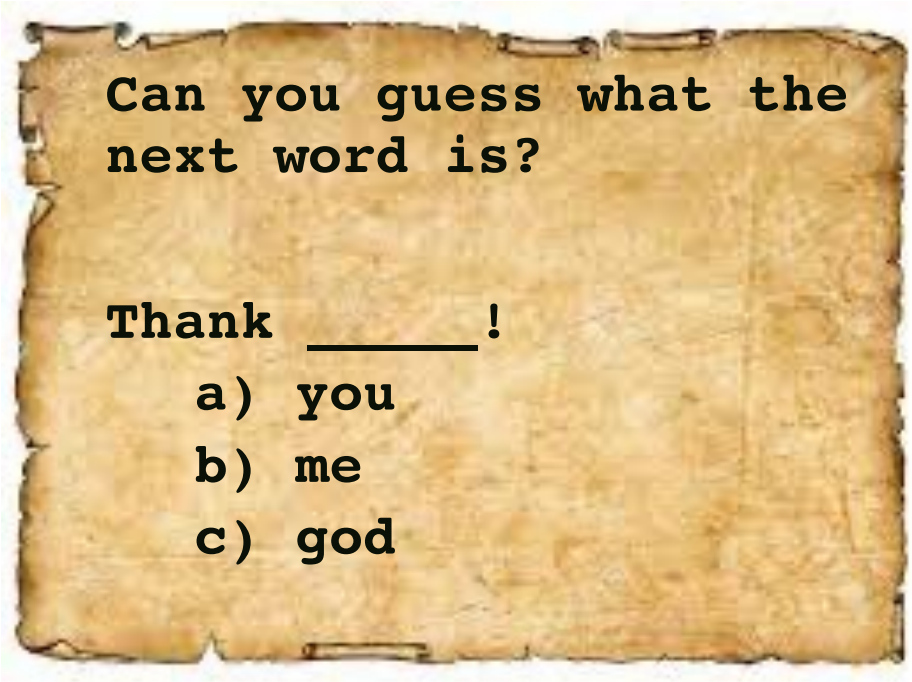


- **Emergence:** Behavior of a system is *implicitly induced* rather than explicitly constructed
- **Homogenization:** *Consolidation* of methodologies for building machine learning systems across a wide range of applications

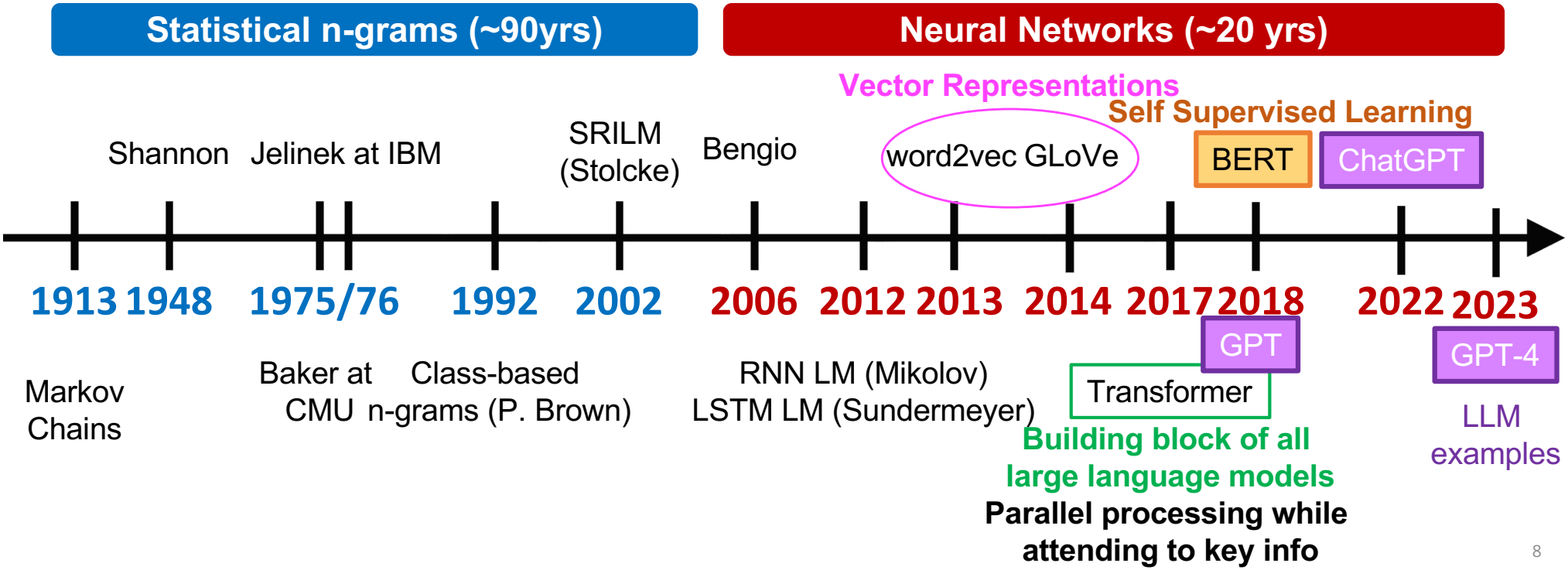
Image adapted from https://arxiv.org/pdf/2108.07258.pdf%20

What is a Language Model?

A classic word guessing game



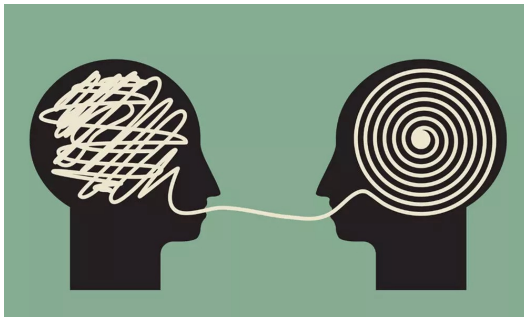
Language Models: A Historical Perspective



Capabilities of Multilingual Large Language Models

Can LLMs do multicultural reasoning?

Language



Reasoning

Culture

Humanity Runs on Coffee

Behind every successful woman is insane amounts of coffee

Can LLMs Understand Multicultural Practices?

Cultural Reasoning Example

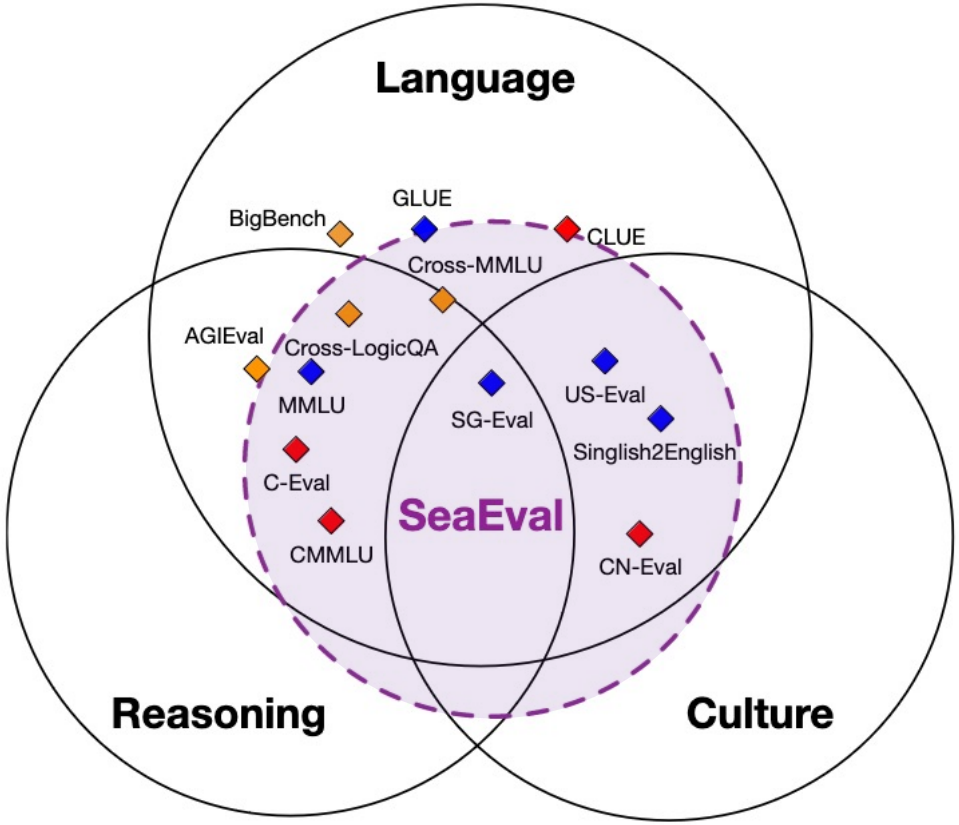
Question	*Which drink in Singapore has the highest calories?* *(A) Teh O* (B) *Teh Siew Dai* (C) *Kopi* (D) *Kopi C*
Multicultural Reasoning Steps	**Multilingual Understanding** (Hokkien) Teh = Tea (Cantonese) Siew Dai = Less Sweet/Sugar (Malay) Kopi = Coffee **Cultural/Personal Preferences** Teh = Tea + Condensed Milk + Sugar Teh O = Tea + Sugar Kopi = Coffee + Condensed Milk + Sugar Kopi C = Coffee + Evaporated Milk + Sugar **Reasoning with Dietary Knowledge** Condensed milk = Sweetened = Sugar was Added Sugar = Calories Pure Tea or Coffee = Almost No Calories
Answer	(C) Kopi

From Arabic roots: qahwah (قَهْوَة)

Black tea (without milk; O transliteration of 烏 in Hokkien)

C: transliteration of "diluted, thin" in Mandarin (稀/细)

SeaEval Benchmark



https://arxiv.org/abs/2309.04766

- SeaEval consists of **28 datasets**
  - **6 new datasets + Consistency Metric**
    - Cultural comprehension
    - Cross-lingual assessments
  - **5 languages:**
    - English
    - Chinese
    - Malay
    - Indonesian
    - Vietnamese
  - **4 task types:**
    - Cultural Understanding
    - Cross Lingual Consistency
    - Complex Reasoning
    - Standard NLP Tasks
  - **12,133 samples total**

Understanding the Boundaries of Multilingual LLMs

1. When instructions are paraphrased, do LLMs give the same answers?

2. For factual or scientific queries, will multilingual LLMs give consistent answers across languages?

3. Do LLMs (still) suffer from exposure bias (e.g., position bias, majority label bias)?

4. Can multilingual LLMs perform equally well on different languages?

Understanding the Boundaries of Multilingual LLMs

1. When instructions are paraphrased, do LLMs give the same answers?

2. For factual or scientific queries, will multilingual LLMs give consistent answers across languages?

3. Do LLMs (still) suffer from exposure bias (e.g., position bias, majority label bias)?

4. Can multilingual LLMs perform equally well on different languages?

Standard Accuracy Metric

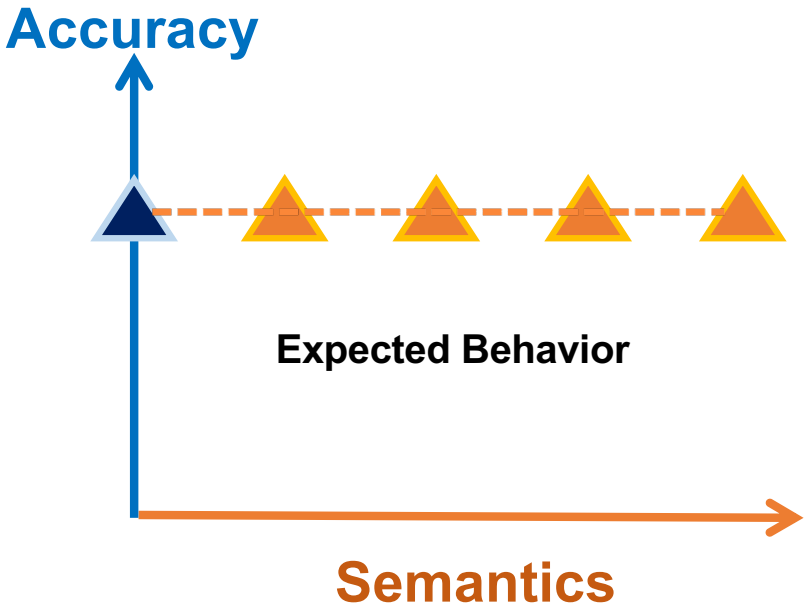
Standard Accuracy Metric

Better Performance

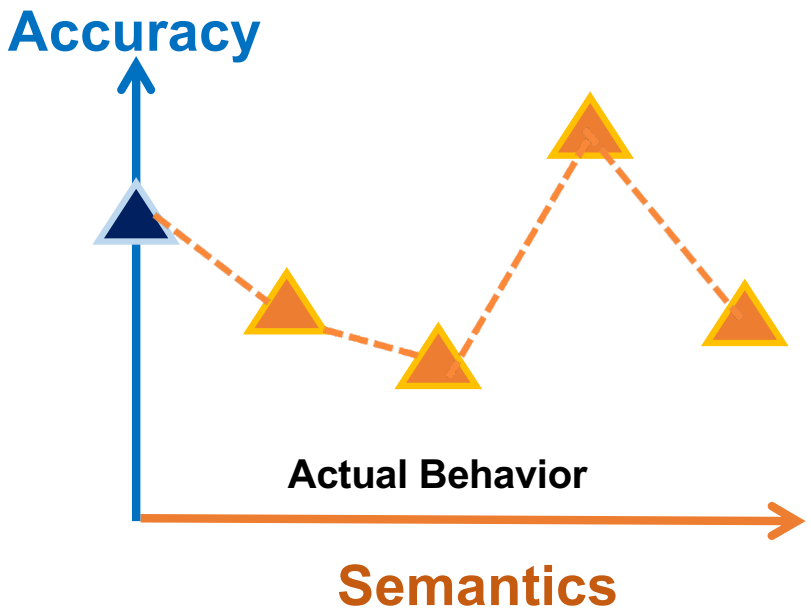
Poorer Performance

Assessing Performance by Extending to Semantics Dimension

What if I Paraphrase the Same Instructions Multiple Times?



What if I Paraphrase the Same Instructions Multiple Times?



LLM Performance is Sensitive to Variations in Paraphrased Instructions

Paraphrased Instructions Results in Varied Performance across

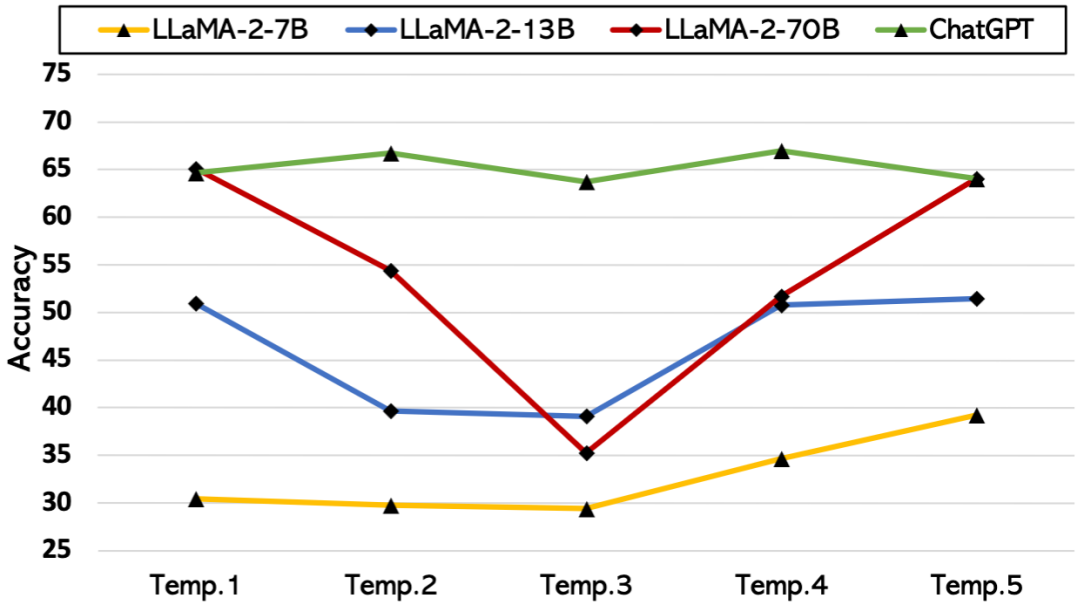


Figure 3: Performance on MMLU dataset with different instruction templates.

Paraphrasing Instructions Makes LLM Answer Differently

Qualitative Example

Please choose the correct answer from the options provided in the multi-choice question.

Question:
How many positive integers are factors of 120 and also factors of 40?

Choices:
(A) 1
(B) 3
(C) 4
(D) 8

Answer: (D) 8 

Please select the correct answer for the multi-choice question.

Question:
How many positive integers are factors of 120 and also factors of 40?

Choices:
(A) 1
(B) 3
(C) 4
(D) 8

Answer: (C) 4 

Examples from LLaMA-2-70B

Paraphrasing Instructions Makes LLM Answer Differently

Qualitative Example

Respond to the question by selecting the correct answer.

Question:
What are the tallest trees on Earth?

Choices:
(A) Sitka Spruce
(B) Giant Sequoia
(C) Coast Redwood
(D) Coast Douglas Fir

Answer: (C) Coast Redwood

Please answer the following multi-choice question by selecting the correct option.

Question:
What are the tallest trees on Earth?

Choices:
(A) Sitka Spruce
(B) Giant Sequoia
(C) Coast Redwood
(D) Coast Douglas Fir

Answer: (B) Giant Sequoia

Examples from LLaMA-2-70B

Understanding the Boundaries of Multilingual LLMs

1. When instructions are paraphrased, do LLMs give the same answers?

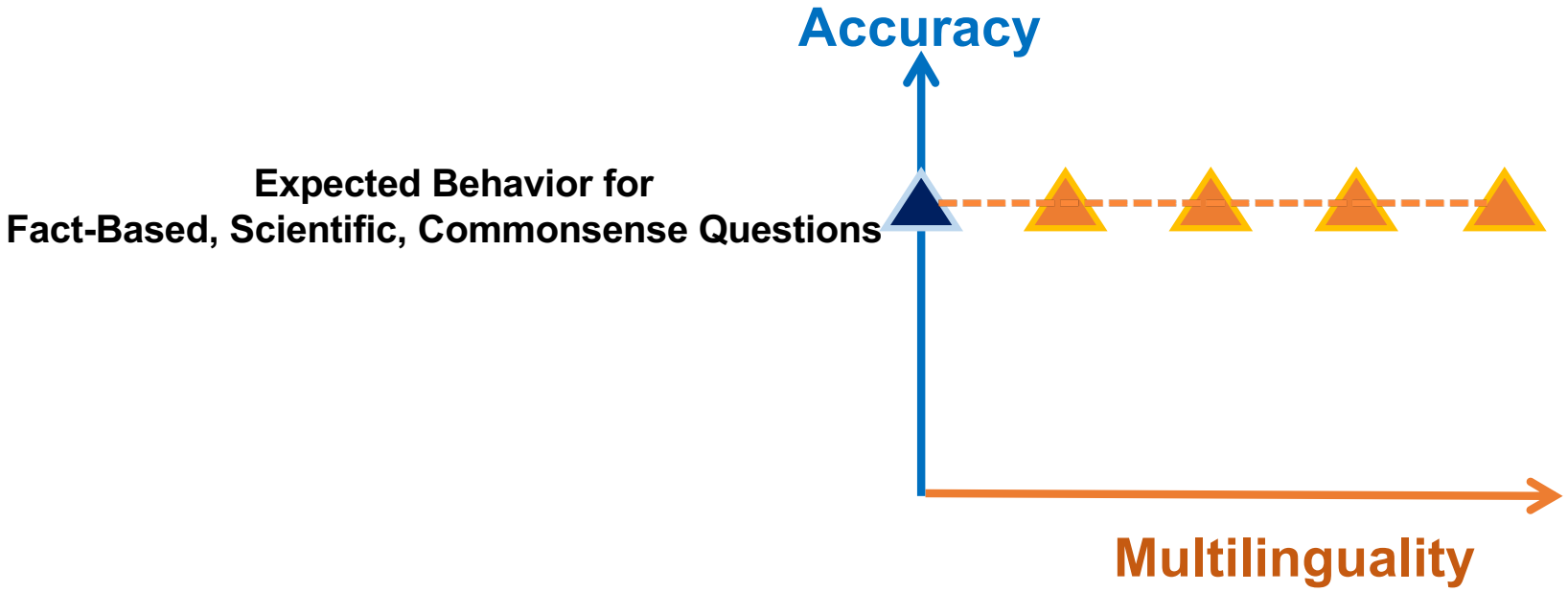
2. For factual or scientific queries, will multilingual LLMs give consistent answers across languages?

3. Do LLMs (still) suffer from exposure bias (e.g., position bias, majority label bias)?

4. Can multilingual LLMs perform equally well on different languages?

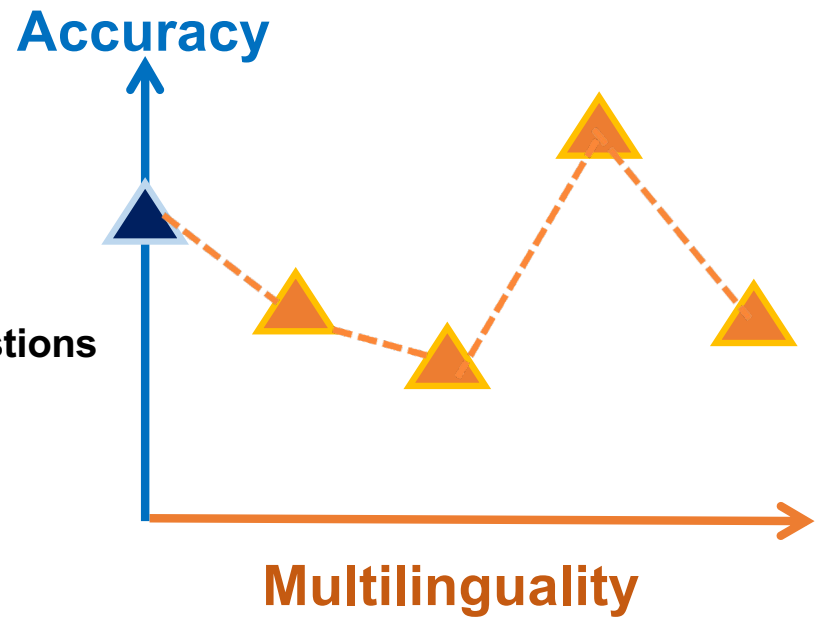
Assessing Performance by Extending to Multilinguality

What if I Ask the Same Question in Different Languages?



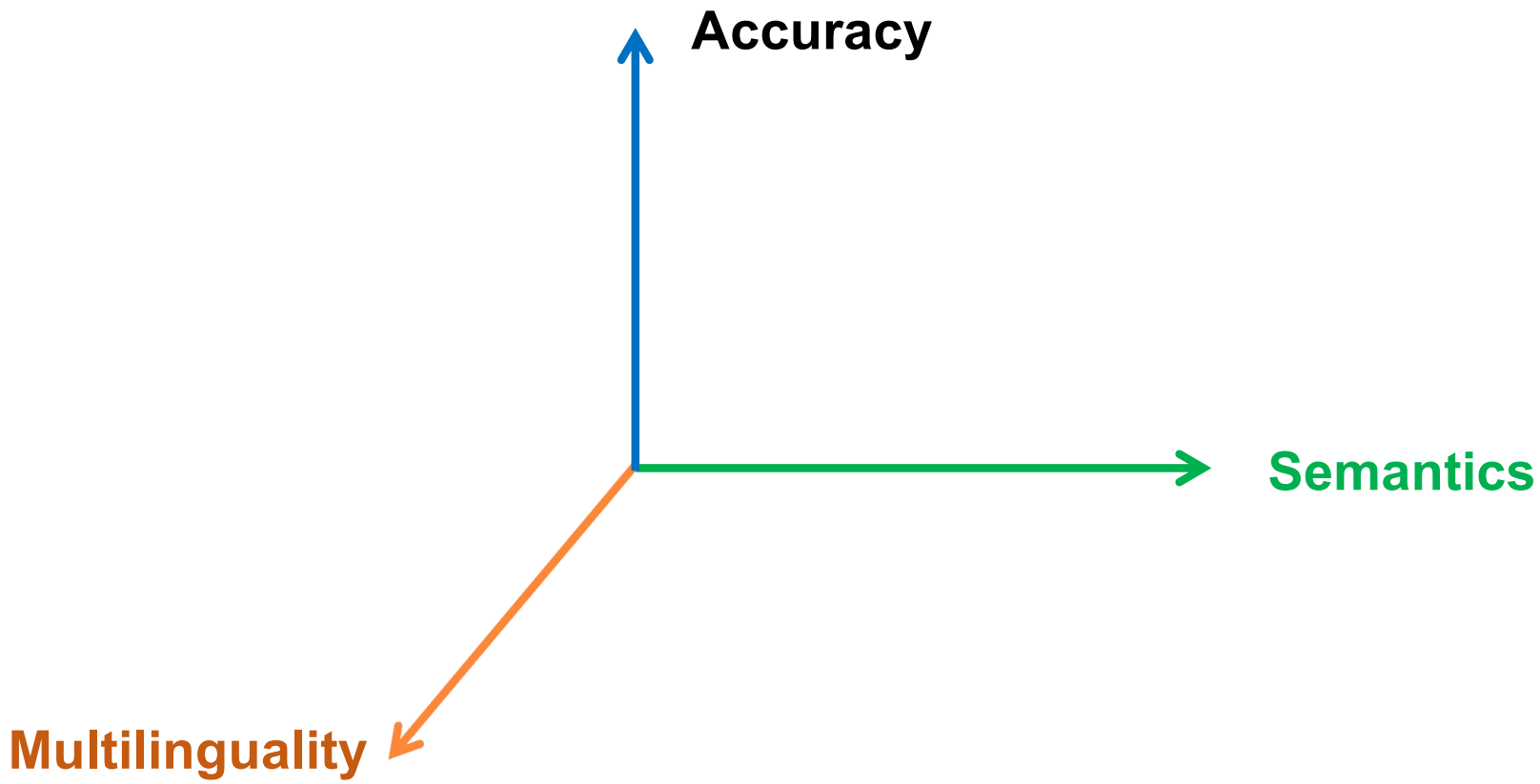
What if I Ask the Same Language in Different Languages?

Actual Behavior for
Fact-Based, Scientific, Commonsense Questions



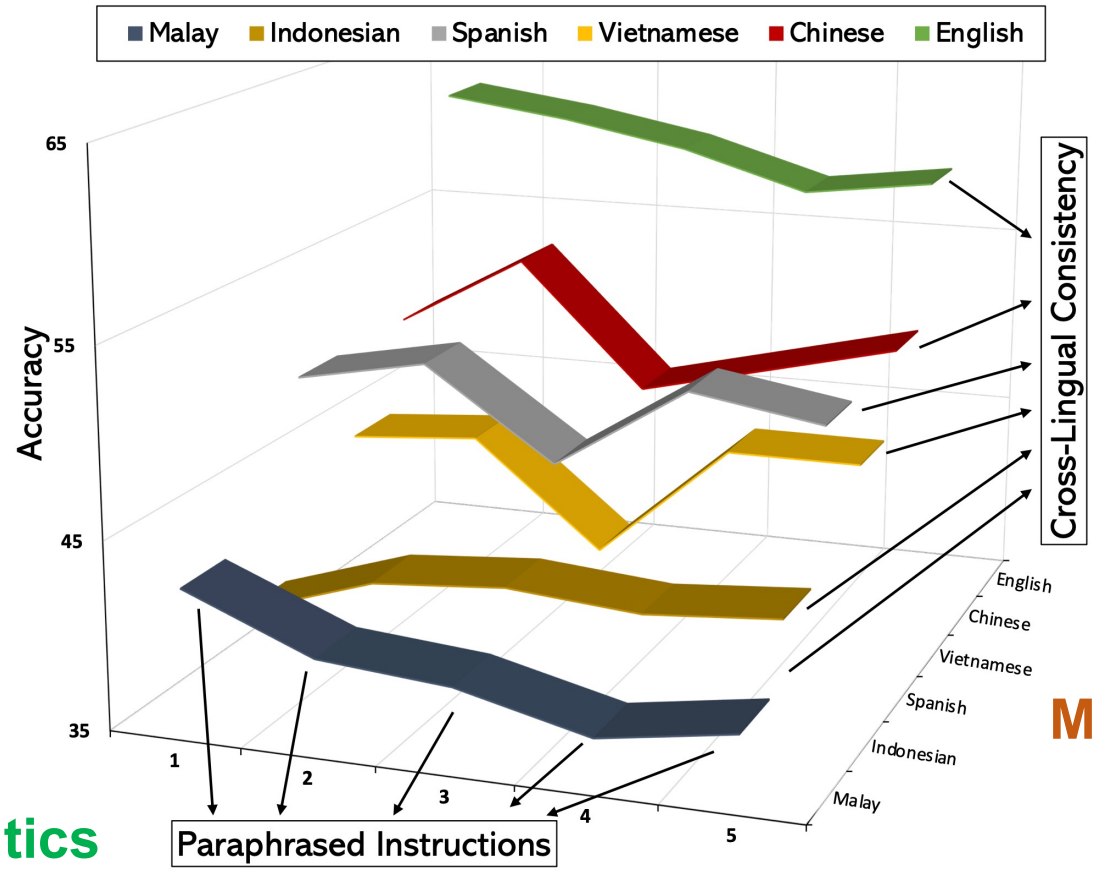
LLM Performance is Inconsistent Across Languages

Assessing Performance by Extending to Semantics & Multilinguality



Existing Metrics Overlook Variations in Semantics & Multilinguality

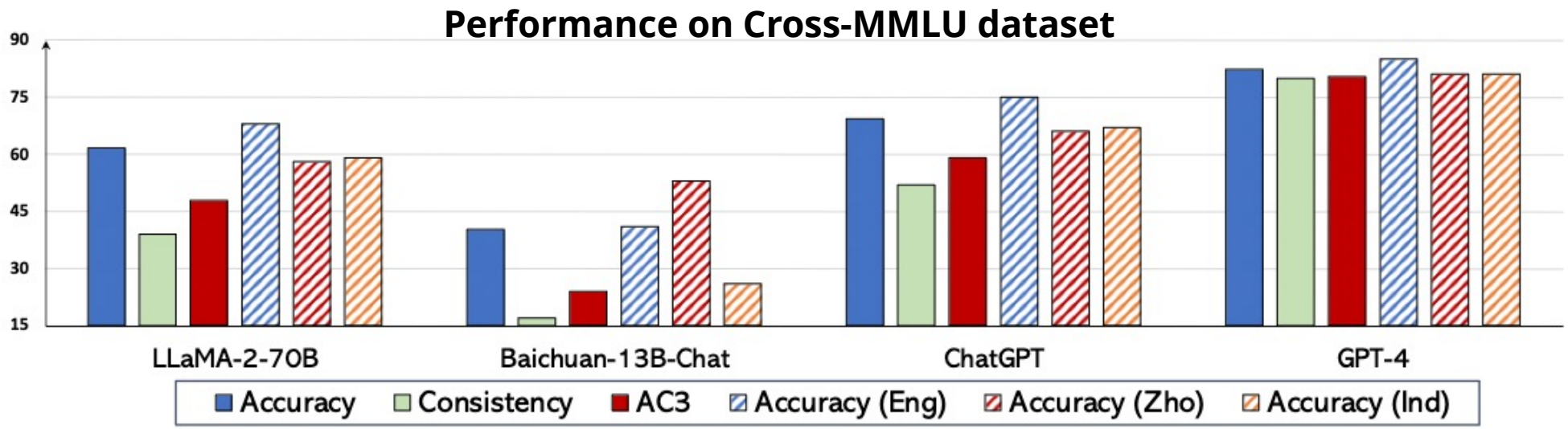
ChatGPT Performance on Cross-LogiQA



Semantics

Multilinguality

Cross-Lingual Inconsistency for Language Understanding



AC3: Harmonic Mean of Accuracy and Consistency

Cross-Lingual Inconsistency on Logical Reasoning

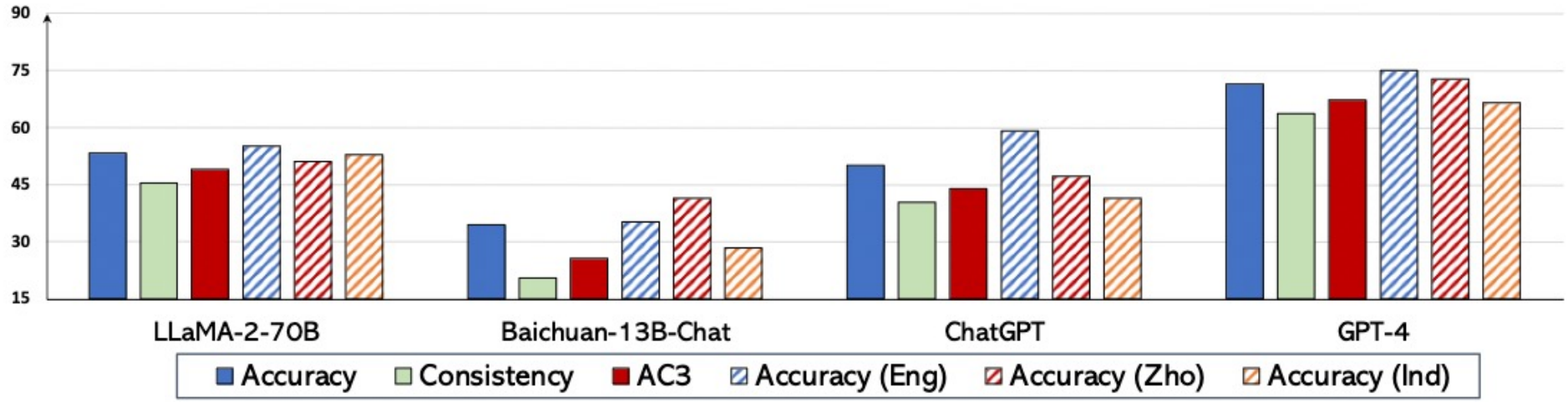


Figure 7: Detailed analysis on Cross-LogiQA dataset. The overall accuracy, consistency, and accuracy scores on three language portions are shown.

Quantifying Brittle Performance

A Spread of Inconsistence Performance

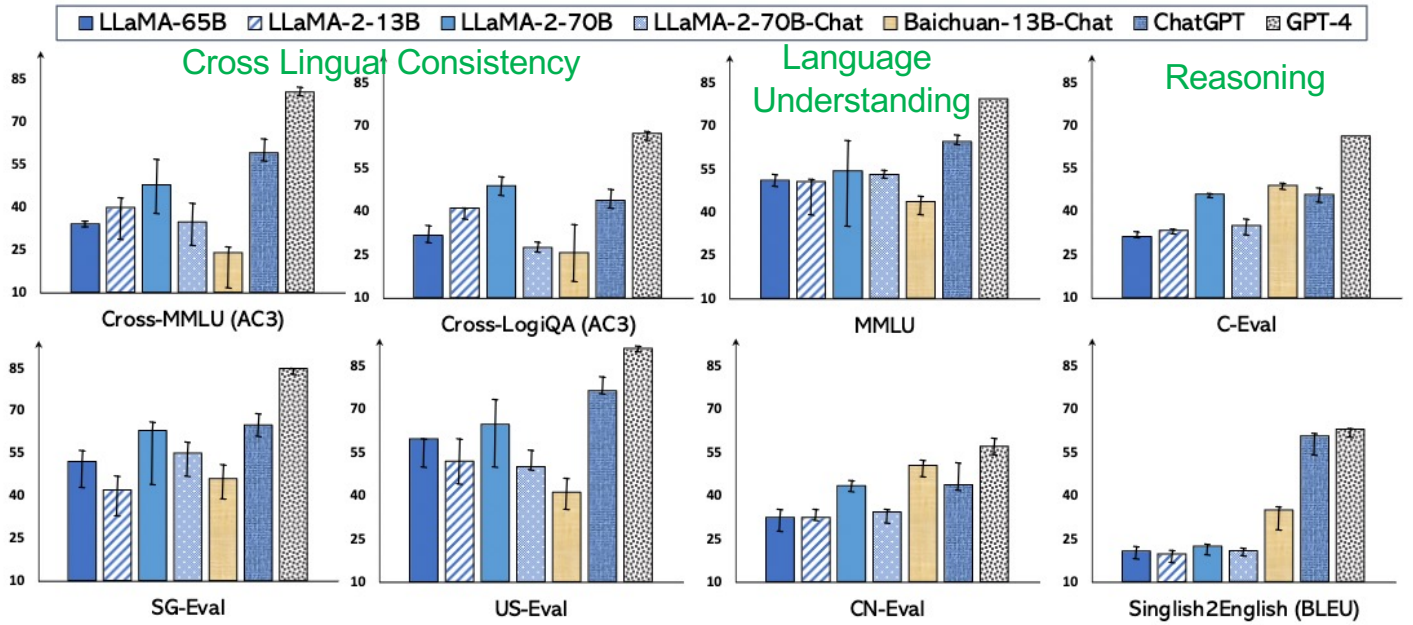


Figure 5: Evaluation results of representative LLMs on a subset of SeaEval tasks. AC3 score and BLEU are reported for Cross-MMLU and Singlish2English datasets, respectively, while Accuracy is shown for other datasets. The error bar represents the performance range arising from five different instruction templates.

Same Physics Question Asked in Different Languages

Non-English Models More Likely Wrong?

Language	English	Chinese	Indonesian	Spanish
Question	*Please choose the correct answer for the following question.* *When white light passes through a prism, the light that bends more than green is?* *(A) Red* *(B) Yellow* *(C) Blue* *(D) None of these*	回答下面问题，选择正确答案。 当白光通过棱镜时，比绿光弯曲更多的光是? (A) 红色的 (B) 黄色的 (C) 蓝色的 (D) 都不是	*Silakan pilih jawaban yang benar untuk pertanyaan berikut.* *Ketika cahaya putih melewati sebuat prisma, cahaya manakah yang memiliki sudut deviasi lebih besar daripada cahaya hijau?* *(A) Merah* *(B) Kuning* *(C) Biru* *(D) Tak ada satupun*	*Por favor elija la respuesta correcta para la siguiente pregunta.* *Cuando la luz blanca pasa a través de un prisma, la luz que se desvía más que la verde es* *(A) Rojo* *(B) Amarillo* *(C) Azul* *(D) Ninguna de las anteriores*
Answer	*The correct answer is:* *(C) Blue*	(A) 红色的 ***In English**: (A) Red*	*(D) Tak ada satupun* ***In English**: (D) None of them*	*La respuesta correcta es:* *(A) Rojo* ***In English**: (A) Red*
Correctness	✓	✗	✗	✗

ChatGPT Example from Cross-MMLU

Same Physics Question Asked in Different Languages

Inconsistent Results across Languages

Language	Korean	Malay	Turkish	German
Question	다음 문제에서 올바른 답을 선택해 주세요. 흰 빛이 프리즘을 통과할 때, 초록보다 더 많이 굴절되는 빛은 무엇입니까? (A) 빨강 (B) 노랑 (C) 파랑 (D) 이 중 어느 것도 아님	*Sila pilih jawapan yang betul untuk soalan berikut.* *Apabila cahaya putih melalui prisma, cahaya yang membengkok lebih daripada hijau ialah* *(A) Merah* *(B) Kuning* *(C) Biru* *(D) Tiada satu pun daripada ini*	*Lütfen aşağıdaki soru için doğru cevabı seçin.* *Beyaz ışık bir prizmadan geçtiğinde yeşilden daha fazla kırılan ışık* *(A) kırmızıdır* *(B) sarıdır* *(C) mavidir* *(D) Bunlardan hiçbiri*	*Bitte wählen Sie die richtige Antwort für die folgende Frage. Welches Licht wird stärker gebeugt als grünes Licht wenn weißes Licht durch ein Prisma fällt?* *(A) Rot* *(B) Gelb* *(C) Blau* *(D) Keine von diesen*
Answer	(C) 파랑 ***In English**: (C) Blue*	*(A) Merah* ***In English**: (A) Red*	*(C) mavidir* ***In English**: (C) Blue*	*(A) Rot* ***In English**: (A) Red*
Correctness	✓	✗	✓	✗

ChatGPT Example from Cross-MMLU

Same Physics Question Asked in Different Languages

Inconsistent Results across Languages

Language	Thai	French	Arabic
Question	โปรดเลือกคำตอบที่ถูกต้องสำหรับคำถามต่อไปนี้ เมื่อแสงสีขาวส่องผ่านทะลุปริซึม แสงที่หักเหมา -กกว่าสีเขียวคืองอมากกว่าสีเขียวจะเป็น (A) สีแดง (B) สีเหลือง (C) สีน้ำเงินน้ำ (D) ไม่มีข้อใด	*Veuillez choisir la bonne réponse à la question suivante.* *Lorsque la lumière blanche traverse un prisme, la lumière qui se courbe plus que la verte est* *(A) Rouge* *(B) Jaune* *(C) Bleue* *(D) Aucune d'eux*	اختر الإجابة الصحيحة للسؤال التالي. عندما يمر الضوء الأبيض عبر المنشور، الضوء الذي ينحني أكثر من الأخضر يكون *(A)* أحمر *(B)* أصفر *(C)* أزرق *(D)* لا شيء مما سبق
Answer	(D) ไม่มีข้อใด ***In English**: (D) None of these*	*(C) Bleue* ***In English**: (C) Blue*	*(C)* أزرق ***In English**: (C) Blue*
Correctness	✗	✓	✓

ChatGPT Example from Cross-MMLU

Same Physics Question Asked in Different Languages

Inconsistent Results across Languages

Language	Tamil	Portuguese	Vietnamese
Question	பின்வரும் கேள்விக்கான சரியான பதிலைத் தேர்ந்தெடுக்கவும். வெள்ளை ஒளி ஒரு ப்ரிஸம் வழியாக செல்லும் போது, பச்சை நிறத்தை விட அதிகமாக வளைக்கும் ஒளி *(A)* சிவப்பு *(B)* மஞ்சள் *(C)* நீலம் *(D)* இவற்றில் ஏதுமில்லை	*Por favor, escolha a resposta correta para a seguinte pergunta: Quando a luz branca passa por um prisma, qual das seguintes cores se dobra mais do que a cor verde?* *(A) Vermelho* *(B) Amarelo* *(C) Azul* *(D) Nenhuma destas*	Hãy chọn đáp án đúng cho câu hỏi sau. Khi ánh sáng trắng đi qua lăng kính thì ánh sáng lệch nhiều hơn ánh sáng xanh là (A) ánh sáng đỏ (B) ánh sáng vàng (C) ánh sáng xanh da trời (D) Không có cái nào trong số này
Answer	*(B)* மஞ்சள் ***In English**: (B) Yellow*	*(C) Azul* ***In English**: (C) Blue*	(C) ánh sáng xanh da trời ***In English**: (C) Blue*
Correctness	✗	✓	✓

ChatGPT Example from Cross-MMLU

Understanding the Boundaries of Multilinugal LLMs

1. When instructions are paraphrased, do LLMs give the same answers?

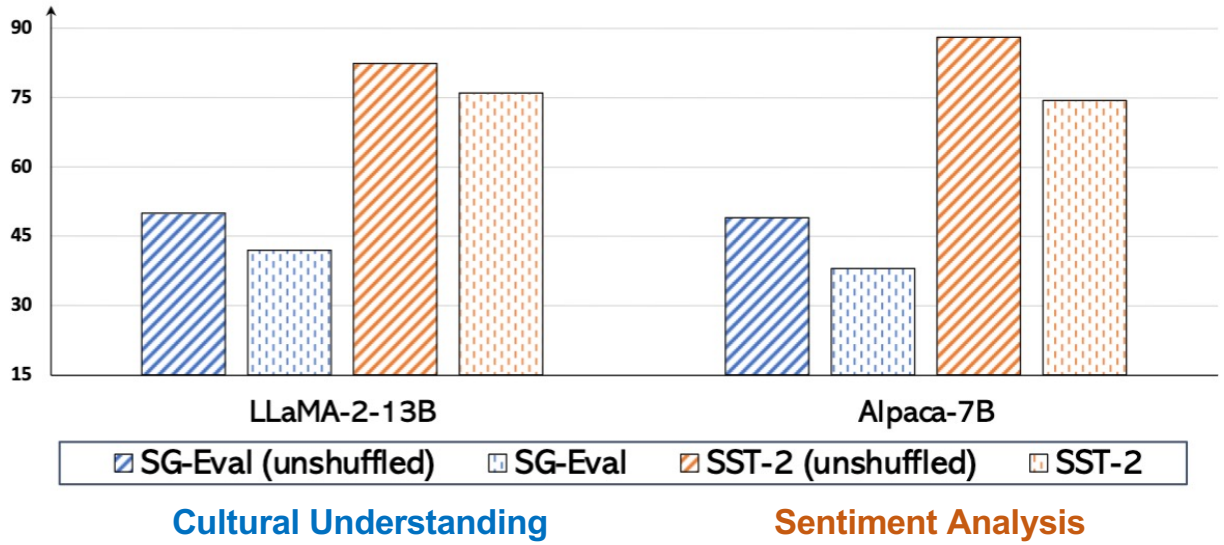
2. For factual or scientific queries, will multilingual LLMs give consistent answers across languages?

3. Do LLMs (still) suffer from exposure bias (e.g., position bias, majority label bias)?

4. Can multilingual LLMs perform equally well on different languages?

Exposure Bias in Classic NLP Tasks & Cultural Comprehension

When Labels in Test are Reshuffled, Performance Drops



Understanding the Boundaries of Multilingual LLMs

1. When instructions are paraphrased, do LLMs give the same answers?

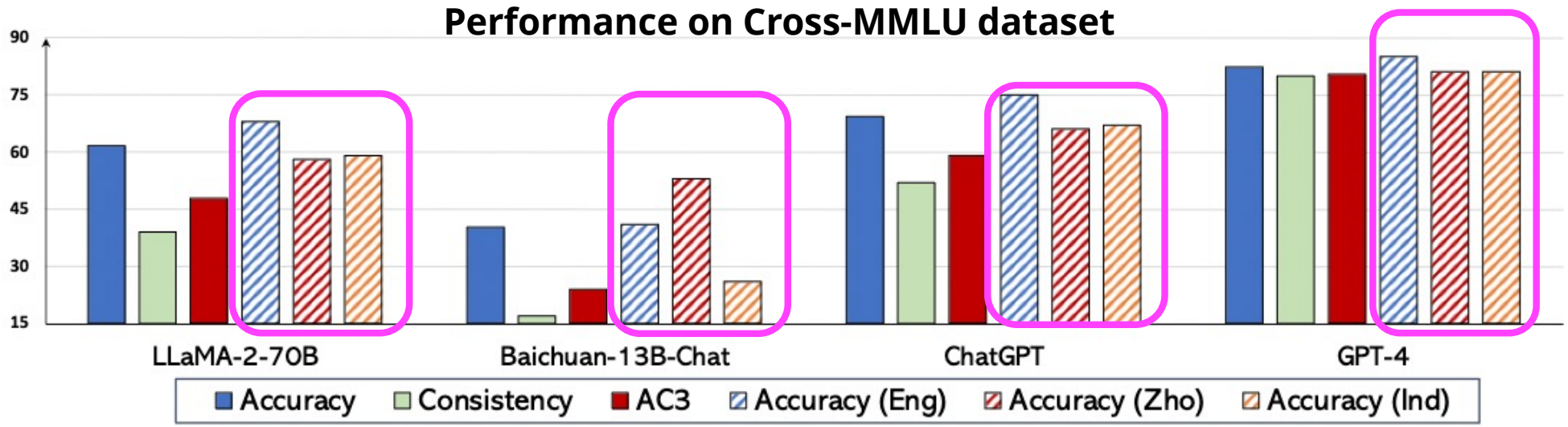
2. For factual or scientific queries, will multilingual LLMs give consistent answers across languages?

3. Do LLMs (still) suffer from exposure bias (e.g., position bias, majority label bias)?

4. Can multilingual LLMs perform equally well on different languages?

Models have not Attained Balanced Multilingual Capability

Language Understanding Cases



AC3: Harmonic Mean of Accuracy and Consistency

Models have not Attained Balanced Multilingual Capability

Logical Reasoning Cases

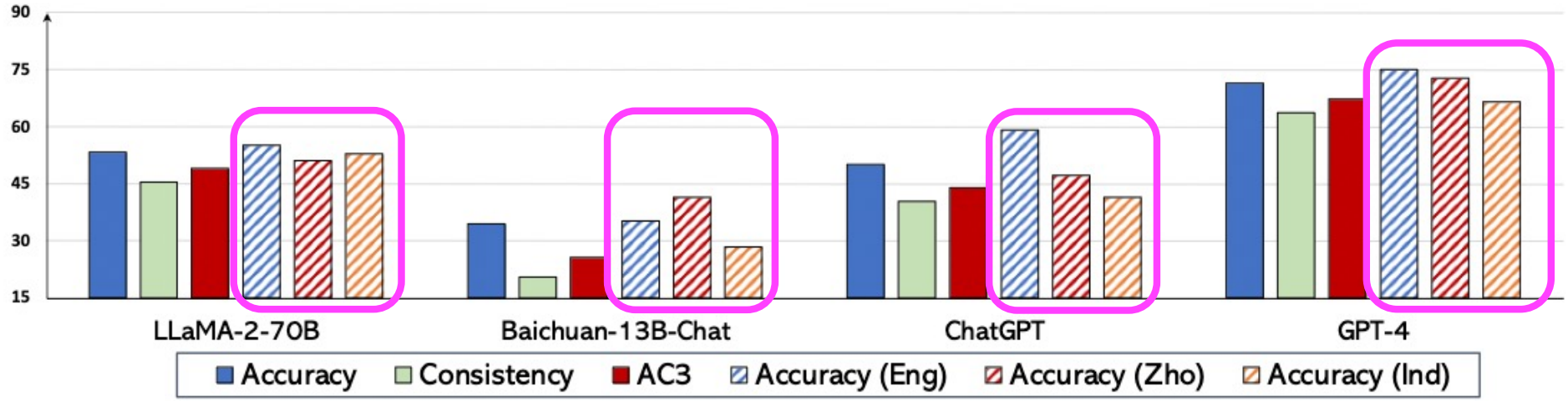


Figure 7: Detailed analysis on Cross-LogiQA dataset. The overall accuracy, consistency, and accuracy scores on three language portions are shown.

Catastrophic Forgetting in Multilingual Models after English Instruction Tuning

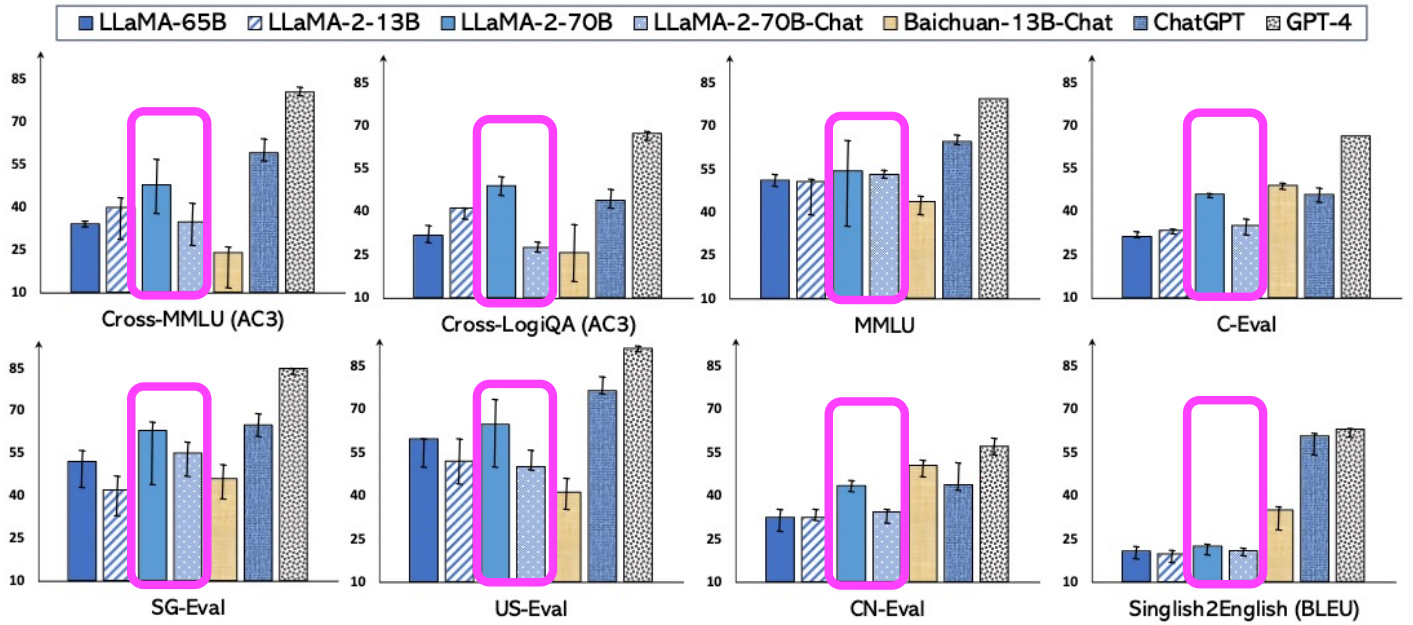


Figure 5: Evaluation results of representative LLMs on a subset of SeaEval tasks. AC3 score and BLEU are reported for Cross-MMLU and Singlish2English datasets, respectively, while Accuracy is shown for other datasets. The error bar represents the performance range arising from five different instruction templates.

Key Findings

1. When instructions are reworded, LLM often gives different answers

2. For factual or scientific queries, one would anticipate consistent answers across languages. However, many models fail to provide such consistency.

3. Many models still suffer from exposure bias (e.g., position bias, majority label bias)

4. Multilingually-trained models have not attained *balanced multilingual* capabilities

https://arxiv.org/abs/2309.04766

Conclusion

- Our endeavors underscore the need for more generalizable semantic representations and enhanced multilingual contextualization

- *SeaEval* can serve as a launchpad for in-depth investigations for multilingual and multicultural evaluations

https://arxiv.org/abs/2309.04766