

浅谈大语言模型时代的 科研选题

朱文昊
南京大学

提纲

- ◎ 第一部分（站在学生的角度）
 - ▶ 科研选题为何重要？怎样为选题做准备？
- ◎ 第二部分（站在机器翻译研究者的角度）
 - ▶ 在大语言模型时代，如何选择合适的科研课题？
- ◎ 第三部分（站在机器翻译研究者的角度）
 - ▶ 大语言模型时代的选题实践

第一部分

科研选题为何重要?
怎样为选题做准备?

科研过程概览

◎ 科研选题

- ▶ 选择大致研究方向



◎ 实验验证

- ▶ 确定具体解决方案



◎ 论文写作、宣传

- ▶ 包装、展示整体研究成果



图片来自：DALL·E 3

科研过程概览

◎ 科研选题

- ▶ 选择大致研究方向

选题内容决定方案可行性

◎ 实验验证

- ▶ 确定具体解决方案

◎ 论文写作、宣传

- ▶ 包装、展示整体研究成果



科研过程概览

◎ 科研选题

- ▶ 选择大致研究方向

◎ 实验验证

- ▶ 确定具体解决方案

◎ 论文写作、宣传

- ▶ 包装、展示整体研究成果



选题内容决定成果影响力

科研选题过程

◎ 了解前沿动态

▶ 大家在做什么?

◎ 解读发展趋势

▶ 大家关心什么?

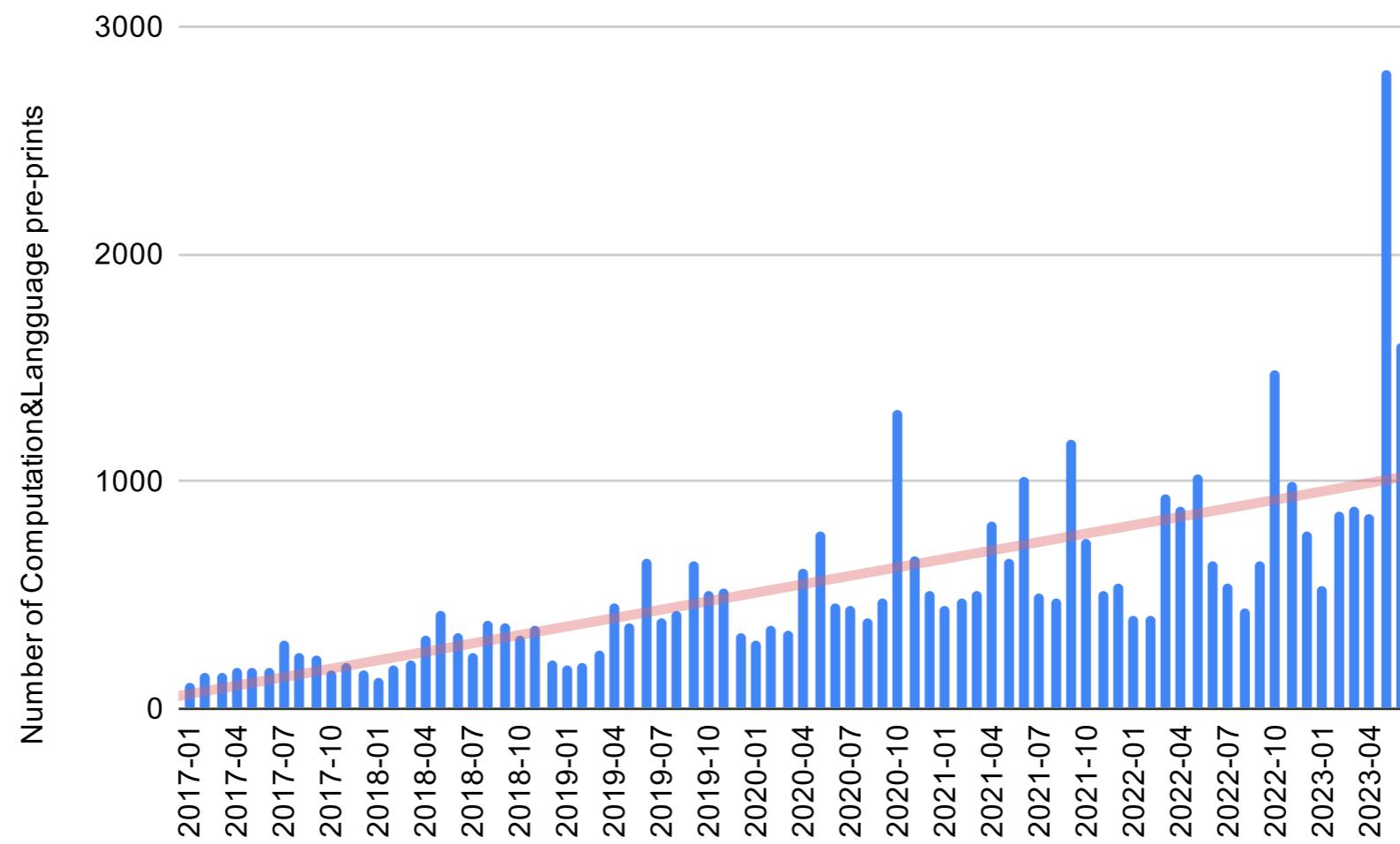
◎ 确定研究方向

▶ 我打算做什么?

了解前沿动态

◎ arXiv

- ▶ 浏览每日最新论文
- ▶ arXiv论文数目剧烈增长，浏览难度增加



图片来自：<https://jlibovicky.github.io/2023/06/30/Speeding-up-arXiv-browsing.html>

了解前沿动态

◎ 社交媒体

► 热点论文分享，节约阅读时间



Aran Komatsuzaki ✅ @arankomatsuzaki · Oct 18

VERA: Vector-Based Random Matrix Adaptation

...

Presents VeRA, which reduces the number of trainable parameters by 10x compared to LoRA, yet maintains the same performance

arxiv.org/abs/2310.11454

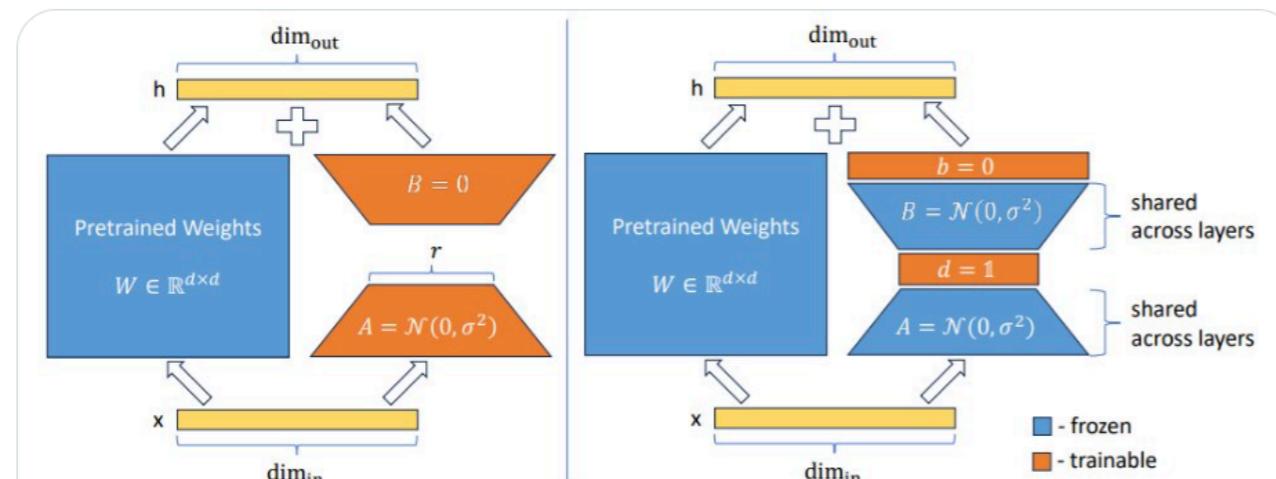


Figure 1: Schematic comparison of LoRA (left) and VeRA (right). LoRA updates the weights matrix W by training the low-rank matrices A and B , with intermediate rank r . In VeRA these matrices are frozen, shared across all layers, and adapted with trainable vectors d and b , substantially reducing the number of trainable parameters. In both cases, low-rank matrices and vectors can be merged into original weights matrix W , introducing no additional latency.

11

101

507

63.6K

↑

3:40 PM · Oct 18, 2023 · 15.7K Views



AK ✅
@_akhaliq

Subscribe

...

Hugging Face Daily papers email of 18 Oct 2023 is out

huggingface.co/papers

Daily Papers

by 🌎 AK

Here is my selection of papers for today (18 Oct):

- [VeRA: Vector-based Random Matrix Adaptation](#)
- [4K4D: Real-Time 4D View Synthesis at 4K Resolution](#)
- [LAMP: Learn A Motion Pattern for Few-Shot-Based Video Generation](#)
- [BitNet: Scaling 1-bit Transformers for Large Language Models](#)
- [EvalCrafter: Benchmarking and Evaluating Large Video Generation Models](#)
- [CrossCodeEval: A Diverse and Multilingual Benchmark for Cross-File Code Completion](#)
- [TEQ: Trainable Equivalent Transformation for Quantization of LLMs](#)
- [Approximating Two-Layer Feedforward Networks for Efficient Transformers](#)
- [Context-Aware Meta-Learning](#)
- [Set-of-Mark Prompting Unleashes Extraordinary Visual Grounding in GPT-4V](#)

Until next time,
AK

了解前沿动态

◎ 社交媒体

► 作者直接发布，评论区答疑解惑



Zeyuan Allen-Zhu
@ZeyuanAllenZhu

Part 3.2: Why do LLMs need Chain of Thoughts even for basic questions (e.g. was Biden born on an even day)? We show that LLMs cannot efficiently manipulate knowledge even if such knowledge is 100% extractable; + inverse knowledge search is just impossible.
arxiv.org/abs/2309.14402

Was Donald Trump born in an even day?

No, Donald Trump was born on June 14, 1946, which is an odd day.

Do you know what odd means?

Yes, I apologize for the oversight. You are correct. The number 14 is even. So, Donald Trump was born on an even day, June 14, 1946. Thank you for pointing that out.

Was Sabrina Zulma higher than Biden? Answer with a Yes or No.

Yes or no, was Donald Trump born earlier than Joe Biden?

Yes. X

Tell me why?

Donald Trump was born on June 14, 1946, while Joe Biden was born on November 20, 1942. Therefore, Joe Biden was born earlier than Donald Trump by almost four years.

AI at Meta and MBZUAI

1:33 AM · Sep 27, 2023 · 400.3K Views



Quanquan Gu @QuanquanGu · Oct 5

Intriguing findings! I've noticed that several examples, including the one in the picture, can be tackled by today's GPT-4. This prompts us to reconsider the definition of "reproducibility" in the era of #LLM. What applies to one version may not necessarily hold for another.

1

772

1

...



Zeyuan Allen-Zhu @ZeyuanAllenZhu · Oct 5

Agree. That's why we propose to focus on synthetic experiments and control the training set. As for GPT4 experiments, I suggest doing some larger-scale tests (instead of trying only Trump/Biden) as OpenAI may locally fix it at any time. I'm pasting below some of my run snapshots

Translate: What is character X in this (commonly-used) Chinese idiom?

Chinese Idiom Task

Prompt 1: 成语“辱不惊”的X是什么字?	GPT3.5 accuracy 9.4%,	GPT4 accuracy 17.6%
Prompt 2: 成语“宠X不惊”的X是什么字?	GPT3.5 accuracy 29.5%,	GPT4 accuracy 36.1%
Prompt 3: 成语“宠辱X惊”的X是什么字?	GPT3.5 accuracy 32.0%,	GPT4 accuracy 76.7%
Prompt 4: 成语“宠辱不X”的X是什么字?	GPT3.5 accuracy 56.7%,	GPT4 accuracy 90.6%

Translate: What is previous/next sentence of this (commonly-used) Chinese poem?

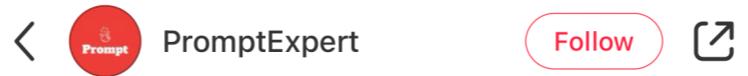
Inverse search: “西出阳关无故人”的上一句是什么? GPT3.5 accuracy 2.1%, GPT4 accuracy 7.3%
Forward search: “劝君更尽一杯酒”的下一句是什么? GPT3.5 accuracy 33.0%, GPT4 accuracy 66.5%

Chinese Poem Task

了解前沿动态

◎ 社交媒体

- ▶ 朋友圈、知乎、小红书也都是快速了解前沿研究动态的窗口
- ▶ 从多个窗口获取信息也可以查漏补缺



The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision)

Zhengyuan Yang*, Linjie Li*, Kevin Lin*, Jianfeng Wang*, Chung-Ching Lin*,
Zicheng Liu, Lijuan Wang*[#]
Microsoft Corporation

* Core Contributor ^ Project Lead

Abstract

Large multimodal models (LMMs) extend large language models (LLMs) with multi-sensory skills, such as visual understanding, to achieve stronger generic intelligence. In this paper, we analyze the latest model, GPT-4V(ision) [99–101, 1]¹, to deepen the understanding of LMMs. The analysis focuses on the intriguing tasks that GPT-4V can perform, containing test samples to probe the quality and genericity of GPT-4V’s capabilities, its supported inputs and working modes, and the effective ways to prompt the model. In our approach to exploring GPT-4V, we curate and organize a collection of carefully designed qualitative samples spanning a variety of domains and tasks. Observations from these samples demonstrate that GPT-4V’s unprecedented ability in processing arbitrarily interleaved multimodal inputs and the genericity of its capabilities together make GPT-4V a powerful multimodal generalist system. Furthermore, GPT-4V’s unique capability of understanding visual markers drawn on input images can give rise to new human-computer interaction methods such as visual referring prompting. We conclude the report with in-depth discussions on the emerging application scenarios and the future research directions for GPT-4V-based systems. We hope that this preliminary exploration will inspire future research on the next-generation multimodal task formulation, new ways to exploit and enhance LMMs to solve real-world problems, and gaining better understanding of multimodal foundation models. Finally, we acknowledge that the model under our study is solely the product of OpenAI’s innovative work, and they should be fully credited for its development. Please see the GPT-4V contributions paper [101] for the authorship and credit attribution: <https://cdn.openai.com/contributions/gpt-4v.pdf>.

关于 GPT-4V 你想知道的一切

这篇文章试图回答以下问题：

1. GPT-4V 支持哪些输入，有哪些使用模式？
2. GPT-4V 在各种领域和任务的能力怎么样？
3. 如何高效使用 GPT-4V？怎么写 prompt 更好？



VERA: VECTOR-BASED RANDOM MATRIX ADAPTATION

Dawid J. Kopiczko[#]
QUVA Lab
University of Amsterdam

Tijmen Blankevoort
Qualcomm AI Research¹

Yuki M. Asano
QUVA Lab
University of Amsterdam

ABSTRACT

Low-rank adaption (LoRA) is a popular method that reduces the number of trainable parameters when finetuning large language models, but still faces acute storage challenges when scaling to even larger models or deploying numerous per-user or per-task adapted models. In this work, we present Vector-based Random Matrix Adaptation (VeRA), which reduces the number of trainable parameters by 10x compared to LoRA, yet maintains the same performance. It achieves this by using a single pair of low-rank matrices shared across all layers and learning small scaling vectors instead. We demonstrate its effectiveness on the GLUE and E2E benchmarks, and show its application in instruction-following with just 1.4M parameters using the Llama2 7B model.

1 INTRODUCTION

In the era of increasingly large and complex language models, the challenge of efficient adaptation for specific tasks has become more important than ever. While these models provide powerful capabilities, their extensive memory requirements pose a significant bottleneck, particularly when adapting them for personalized use. Consider, for example, a cloud-based operating system assistant that continuously learns from and adapts to individual user behaviors and feedback. The need to store multiple checkpoints of finetuned models for each user rapidly escalates the required storage, even more so when multiple tasks come into play.

The situation is further exacerbated when we look at the state-of-the-art models like GPT-4 (OpenAI, 2023). Finetuning techniques like LoRA (Hu et al., 2022), while effective, still introduce considerable memory overhead. As an illustrative example, applying LoRA with a rank of 16 to the query and value layers of GPT-3 (Brown et al., 2020) would demand at least 288MB of memory, if stored in single-precision – at a million finetuned weights, e.g., one per user, that would amount to 275TB.

Given the recent proliferation of language models and their deployment in personalized assistants, edge devices, and similar applications, efficient adaptation methods are paramount. We believe there is untapped potential for even more efficient approaches. Previous work Aghajanyan et al. (2021) pointed out the low intrinsic dimensionality of pretrained models’ features. These studies reported numbers much lower than the trainable parameters used in LoRA, suggesting there is room for improvement.

In parallel to this, recent research has shown the surprising effectiveness of models utilizing random weights and projections (Peng et al., 2021; Ramanujan et al., 2020; Lu et al., 2022; Schrimpf et al., 2021; Franke et al., 2021). Such models serve as the basis of our proposed solution, Vector-based Random Matrix Adaptation (VeRA), which minimizes the number of trainable parameters introduced during finetuning by reparametrizing the weights matrices. Specifically, we employ “scaling vectors” to adapt a pair of frozen random matrices shared between layers. With this approach, many more versions of the model can reside in the limited memory of a single GPU.

In summary, our main contributions are as follows:

^{*}dj.kopiczko@gmail.com; [†]Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.
¹Datasets were solely downloaded and evaluated by the University of Amsterdam.

VeRA: LoRA of LoRAs, 比 LoRA 少 10 倍参数

论文：VeRA: Vector-based Random Matrix Adaptation

了解前沿动态

◎ 博客 (Anthropic's Blog, Yao Fu's Blog)

▶ 最新工作介绍，最新进展综述

The screenshot shows a list of three research posts from Anthropic's website:

- Collective Constitutional AI: Aligning a Language Model with Public Input** (Oct 17, 2023)
- Decomposing Language Models Into Understandable Components** (Oct 5, 2023)
- Towards Monosematicity: Decomposing Language Models With Dictionary Learning** (Oct 5, 2023)

内容来自：<https://www.anthropic.com/research>

[Sep 2023 | An Initial Exploration of Theoretical Support for Language Model Data Engineering. Part 1: Pretra...](#)

[Jun 2023 | A Stage Review of Instruction Tuning](#)

[May 2023 | Towards Complex Reasoning: the Polaris of Large Language Models](#)

[Dec 2022 | How does GPT Obtain its ability? Tracing Emergent Abilities of Language Models to their Sources](#)

[Nov 2022 | A Closer Look at Large Language Models Emergent Abilities](#)

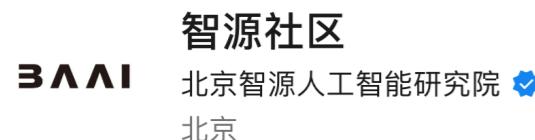
[Feb 2022 | Why S4 is Good at Long Sequence: Remembering a Sequence with Online Function Approximati...](#)

内容来自：<https://yaofu.notion.site/Yao-Fu-s-Blog-b536c3d6912149a395931f1e871370db>

了解前沿动态

◎ 公众号

- ▶ 最新访谈分享、最新工作介绍



继承学术出版严谨与系统，兼具新闻报道及时与多元；
为内行搭建思想交流媒介，以事实启迪公众对AI认知 >
584 篇原创内容
视频号: 智源社区
80 个朋友关注

消息 视频号 服务

作者精选

OpenAI科学家 Jason Wei专访：思维链灵感来源于冥想 | 智源独家

阅读 1.1万 赞 82 2个朋友读过



车万翔：ChatGPT时代，NLPer的危与机

阅读 1.1万 赞 60



复旦大学 数据智能与社会计算实验室 Lab of Data Intelligence and Social Computing (DISC) >
149 篇原创内容
47 个朋友关注

消息 服务

今天

大视觉语言模型基准数据集
ReForm-Eval：新瓶装旧酒，给旧...
阅读 1416 赞 6 20个朋友读过



10月 8 日

复旦大学数据智能与社会计算实验室 4篇论文被 EMNLP 2023 录用
阅读 3719 赞 11 17个朋友读过



南京大学 NLP 研究组 >
61 篇原创内容
166 个朋友关注

消息 服务

10月 12 日

通过语言对齐将大语言模型英语能力外推到非英语语言
阅读 737 赞 9 124个朋友读过



8月 14 日

【学术报告】大语言模型的多语言理解能力探究
阅读 650 赞 8 9个朋友读过



了解前沿动态

◎ 高校课程

- ▶ 内容每年更新，讲解内容更加结构化、系统化

Stanford XCS224U: Natural Language Understanding I Spring 2023

Stanford Online
50 videos

Taught by professor Christopher Potts, this professional Stanford Online course draws on theoretical concepts from linguistics, ...[More](#)

▶ Play all ⚡ Shuffle

Explanation methods for NLP models

Train/test evaluations cannot provide guarantees about behavior on new examples.

Probing methods illuminate internal representations but do not support causal inferences.

Attribution methods illuminate the causal dynamics of models but don't characterize their internal representations

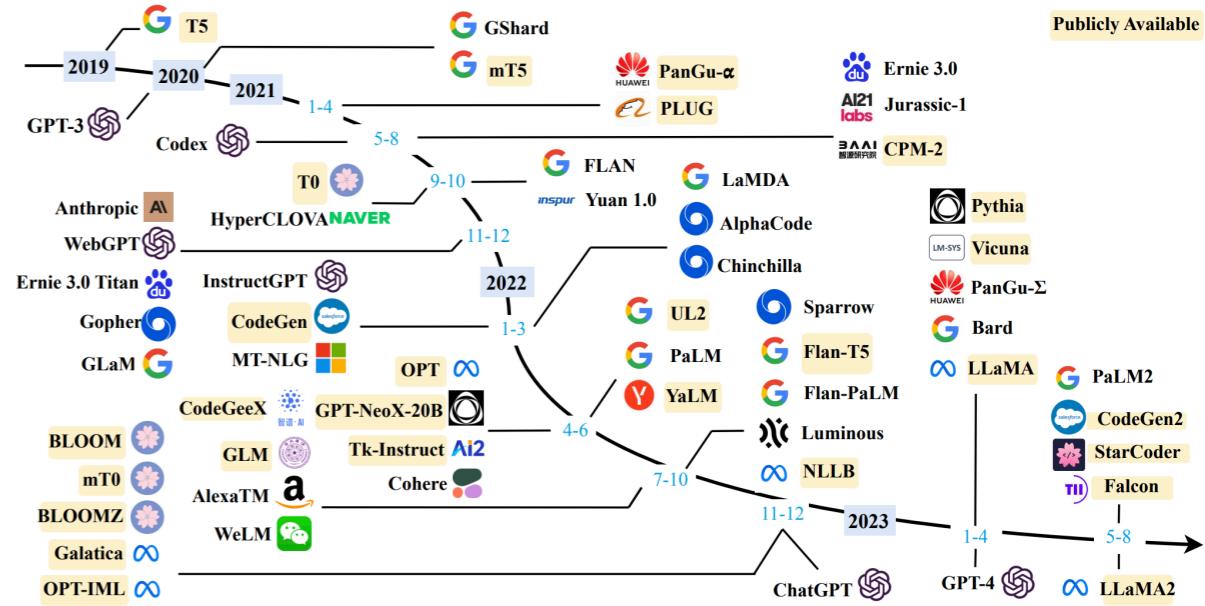
Active manipulations of model internal states provide causal insights and rich characterizations of those states.

Interchange intervention training: train models to conform to the structure of high-level symbolic models.

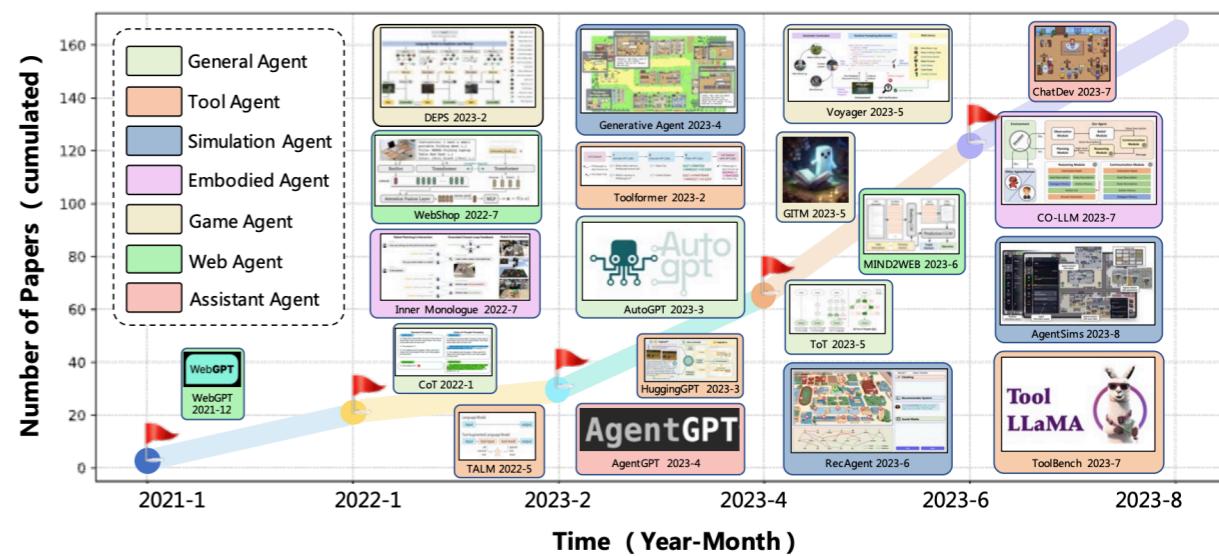
Stanford

解读发展趋势

- ◎ 对已有信息进行整理和总结
- ◎ 参考他人的整理和总结 (Survey)
 - 《A Survey of Large Language Models》
 - 《A Survey on In-context Learning》
 - 《Instruction Tuning for Large Language Models: A Survey》
 - 《A Survey on LLM-based Autonomous Agents》



图片来自：A Survey of Large Language Models



图片来自：A Survey on LLM-based Autonomous Agents

第二部分

在大语言模型时代，
如何选择合适的科研课题？

大语言模型对机器翻译产生的冲击

- ◎ 在大规模单语数据上训练的大语言模型具备很强的翻译能力
 - ▶ 训练数据改变 -> 学习范式改变
 - ▶ 模型架构改变 -> 交互模式改变

Model	<i>en ↔ zh newstest21</i>		<i>en ↔ de newstest21</i>	
Supervised baselines				
WMT'21 1st Place	70.0	66.6	76.9	76.9
WMT'21 2nd Place	69.7	66.3	76.3	76.7
WMT'21 3rd Place	69.7	65.8	76.0	76.4
Google Translate	69.5	65.0	76.4	75.7
Few-shot translation models				
PaLM	67.7	64.1	<u>75.9</u>	74.8
<i>Bilingual LMs (Beam)</i>	<u>62.6</u>	67.0	<u>74.9</u>	74.1
<i>Bilingual LMs (MBR)</i>	68.4	67.8	75.5	76.5
<i>Trilingual LM (Beam)</i>	65.3	65.3	<u>74.5</u>	74.4
<i>Trilingual LM (MBR)</i>	<u>68.9</u>	68.3	75.5	<u>76.8</u>

表格来自：Garcia et al. The unreasonable effectiveness of few-shot learning for machine translation

ZH 可以用英文翻译一下这首诗吗?
岱宗夫如何,齐鲁青未了,造化钟神秀,阴阳割分晓.
荡胸生层云,决眦入归鸟,会当凌绝顶,一览众山小.”

GPT-4 当然可以。这首诗是唐代诗人杜甫的《望岳》，以下是这首诗的英文翻译：
"How can Mount Tai be described?
The greenness of Qilu never fades.
Nature bestows it with divine beauty,
The distinction of Yin and Yang is clear at dawn.
A heaving chest creates layers of clouds,
Eyes squinting see birds returning.
When I reach its highest peak,
All the other mountains seem so small."

回答来自：GPT-4

拥抱大语言模型

◎ 基于大语言模型革新机器翻译范式

- 通过在上下文中提供额外信息，可以更好地控制翻译过程
- 模型的指令执行能力让人机交互过程更加便捷

Chain-of-Dictionary Prompting

"have" means "வேண்டும்" means "haben" means "avoir".

"4-month-old" means "4 மாத குழந்தை" means "4 Monate alt" means "4 mois".

"mice" means "எலிகள்" means "Maus" means "souris".

"non-diabetic" means "சர்க்கரை நோய்" means "nicht-diabetisch" means "non diabétique".

"used" means "பயன்படுத்தப்பட்ட" means "Gebrauch" means "utilisés".

"diabetic" means "சர்க்கரை நோயா" means "Diabetiker" means "diabétique".

"added." means "சேர்க்கப்பட்டது." means "- und hinzugef" means "ajoutée.". .

Translation Output

Translation from ChatGPT with higher quality:

"நாங்கள் இப்போது சர்க்கரை நோயற்ற 4 மாத வயது எலிகளைக் கொண்டு உள்ளோம், முன்னர் அவை சர்க்கரை நோயாக இருந்தன," அவர் சேர்ந்துள்ளார். "

Translated back to English using NLLB Translator for reader's convenience:

"We now have 4 month old diabetic rats, who were previously diabetic", he added.

图片来自：Lu et al. Chain-of-Dictionary Prompting Elicits Translation in Large Language Models

Below is a dialog consisting of instructions and responses. Write a response that completes the request.

Instruction:

Translate this sentence into English: 海内存知己，天涯若比邻。

Response:

Though miles apart, we share the same bond; with true friends, distance is not a barrier.

Instruction:

Note that the ending should rhyme.

Response:

Though oceans divide, we're as close as tide; friends afar, yet by our side.

Instruction:

Start with 'Long distance'.

Response:

Long distance may span, yet we understand; friends apart, but near at heart.

Instruction:

Shorten it to 10 words or less.

Response:

Though miles apart, we're close at heart.

...

Instruction of Lexically Constrained Translation Task

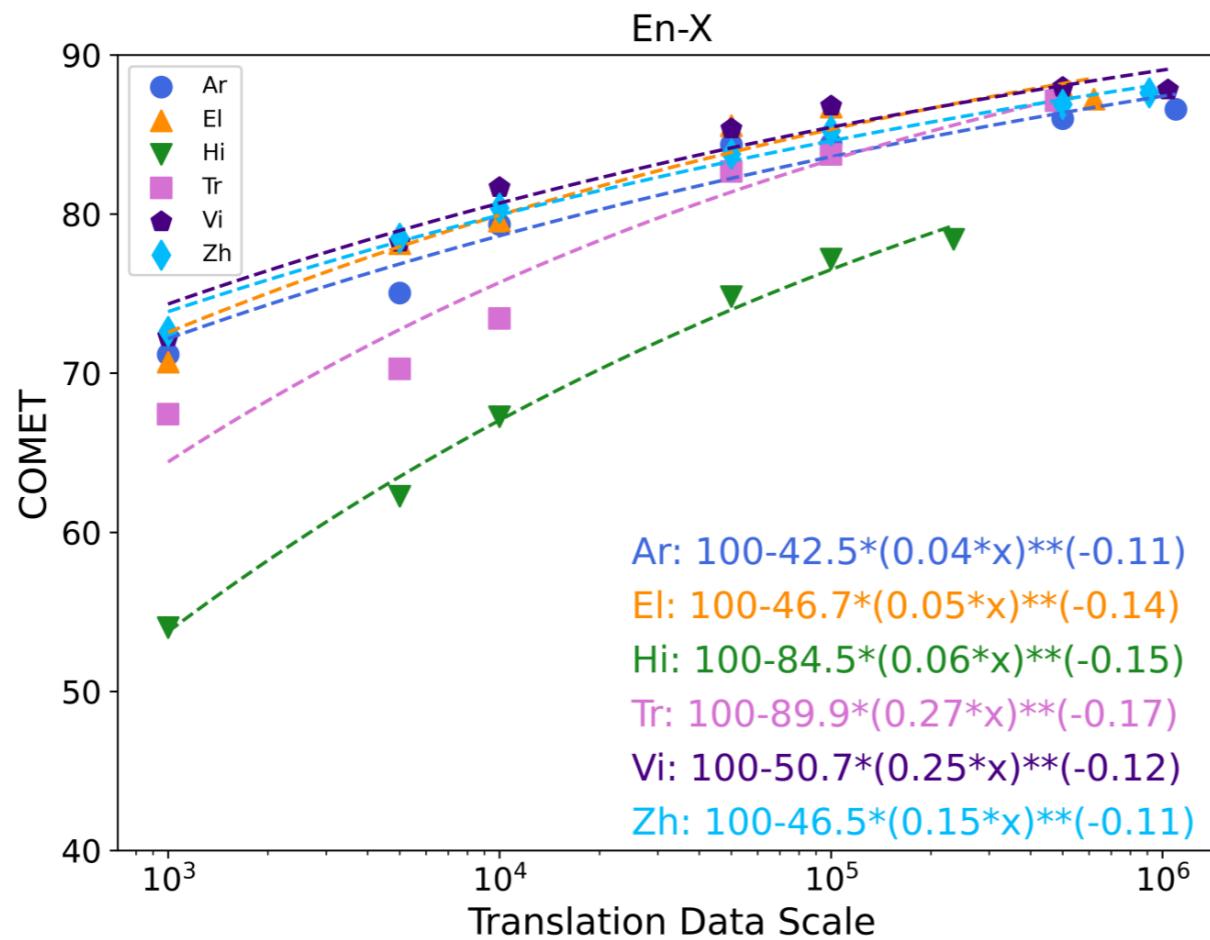
Provide the {target_language} translation of this sentence: {sentence}
Make sure to include these words: {constraint_1}, {constraint_2} ...

图片来自：Zhang et al. BayLing: Bridging Cross-lingual Alignment and Instruction Following through Interactive Translation for Large Language Models

拥抱大语言模型

◎ 基于大语言模型革新机器翻译范式

- 通过在上下文中提供额外信息，可以更好地控制翻译过程
- 模型的指令执行能力让人机交互过程更加便捷
- 如何利用可扩展的翻译数据，提升机器翻译性能上限



图片来自：Zhu et al. Extrapolating Large Language Models to Non-English by Aligning Languages

拥抱大语言模型

◎ 借助翻译任务的学习，激发模型的多语言能力

- ▶ 大语言模型的预训练语料以英文为主，在各语言上能力不平衡
- ▶ 能否以翻译任务为桥梁，将大模型的英语能力外推到非英语语言

ZH Please write a poem about Jinan, please rhyme.

 In Jinan where the springs do flow,
Beneath the sun's soft, golden glow.
By Yellow River's ancient side,
The city stands with timeless pride.

Baotu Spring, its waters clear,
Tales of old are whispered near.
Mountains guard on every side,
Jinan's beauty cannot hide.

Where willows sway and lotus bloom,
History whispers, dispelling gloom.
In every corner, stories blend,
Jinan's charm, it has no end.

ZH 请写一首关于济南的诗歌，要求押韵。

 济南城下泉声响，
百花洲边柳絮飘。
泉城夏日凉如水，
黄河之畔古风傲。

石门日照金山闪，
黑虎泉畔古迹悠。
四面山川护古城，
济南风情永不老。

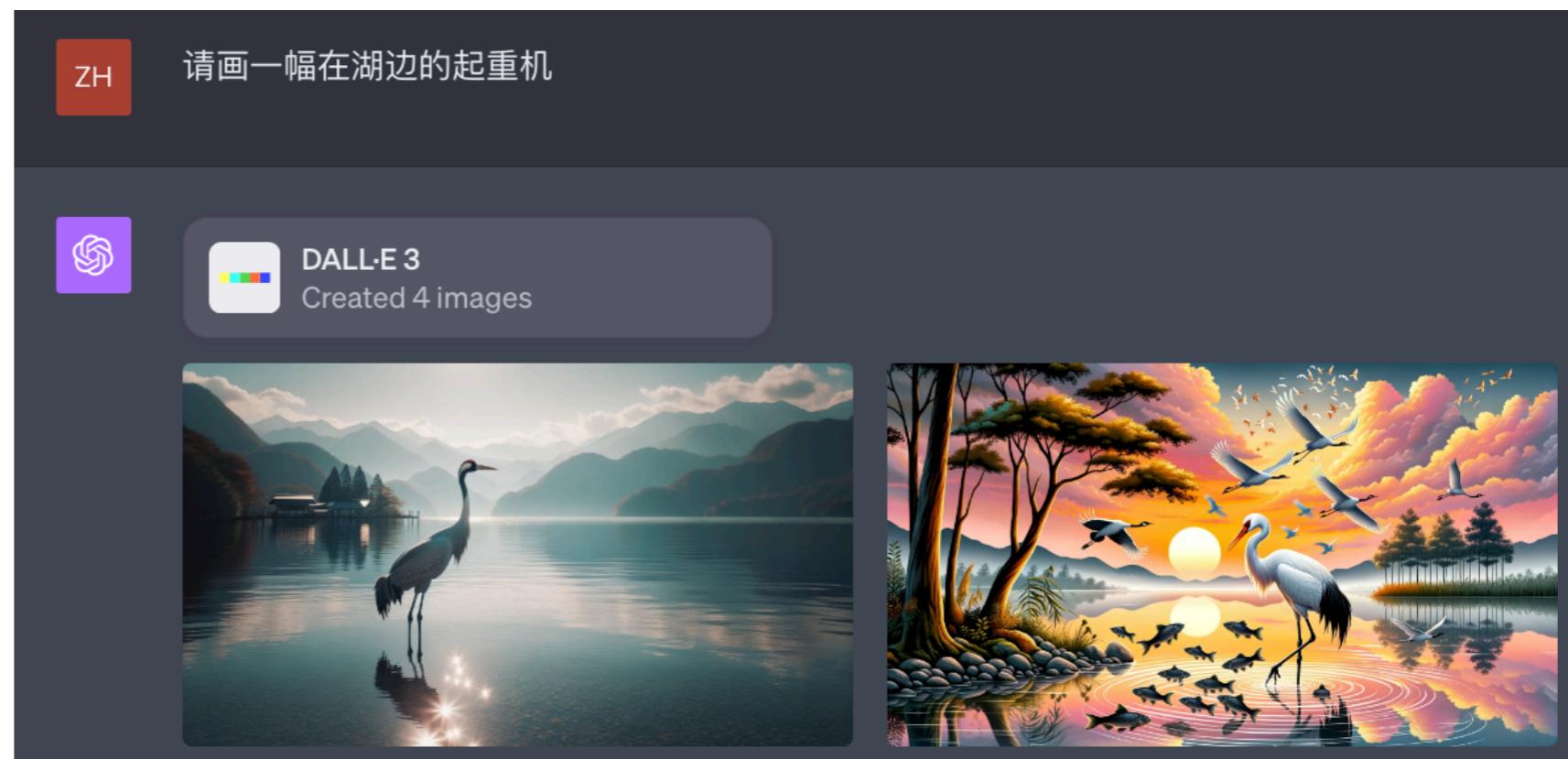
回答来自：GPT-4

拥抱大语言模型

◎ 借助翻译任务的学习，激发模型的多语言能力

- ▶ 大语言模型的预训练语料以英文为主，在各语言上能力不平衡
- ▶ 能否以翻译任务为桥梁，将大模型的英语能力外推到非英语语言
- ▶ 英语数据带来的偏见无处不在，能否借助语言对齐消除内在偏见

- 多轮对话能力
- 文生图能力

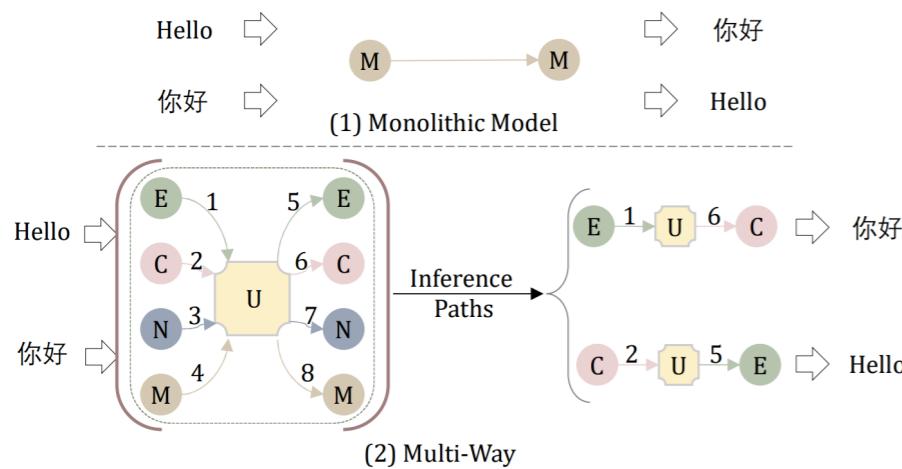


图片来自：DALL-E 3

坚守传统机器翻译

◎ 多语言机器翻译

- ▶ 消除语言障碍是机器翻译研究的终极目标
- ▶ 多语言机器翻译仍然面临重重挑战（参数竞争、高效解码、幻视）



图片来自：Yuan et al. Lego-MT: Learning Detachable Models for Massively Multilingual Machine Translation

MODEL	LOW RESOURCE		MID RESOURCE		HIGH RESOURCE	
	LP Fraction	Rate (%)	LP Fraction	Rate (%)	LP Fraction	Rate (%)
SMaLL100	2/7	0.213 _{0.00}	2/19	0.009 _{0.00}	1/5	0.017 _{0.00}
M2M (S)	5/7	0.261 _{0.08}	11/19	0.140 _{0.08}	0/5	0.000 _{0.00}
M2M (M)	3/7	0.083 _{0.00}	6/19	0.035 _{0.00}	0/5	0.000 _{0.00}
M2M (L)	4/7	0.296 _{0.08}	3/19	0.017 _{0.00}	0/5	0.000 _{0.00}
ChatGPT	4/7	0.059 _{0.08}	10/19	0.183 _{0.08}	0/5	0.000 _{0.00}

图片来自：Guerreiro et al. Hallucinations in Large Multilingual Translation Models.

The screenshot shows the SeamlessM4T interface with two main panels:

- AUTOMATIC SPEECH RECOGNITION:** Language: English. Text: "Recently, Google has launched its AI-powered note-taking app Notebook LM which promises new ways of taking and organizing notes."
- TEXT TRANSLATION:** Language: Chinese (simplified). Text: "最近,谷歌推出了AI驱动的笔录制应用程序Notebook LM,"
- SPEECH TRANSLATION:** Language: Chinese (simplified). Shows a speech-to-text transcription interface.

图片来自：SeamlessM4T

第三部分

大语言模型时代的选题实践

大语言模型时代的选题实践

- ◎ 选题一：大语言模型多语言翻译能力评估与分析

《Multilingual Machine Translation with Large Language Models:
Empirical Results and Analysis》

Wenhai Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang,
Lingpeng Kong, Jiajun Chen, Lei Li

- ◎ 选题二：通过语言对齐提升大语言模型的非英语能力

《Extrapolating Large Language Models to Non-English by Aligning
Languages》

Wenhai Zhu, Yunzhe Lv, Qingxiu Dong, Fei Yuan, Jingjing Xu, Shujian
Huang, Lingpeng Kong, Jiajun Chen, Lei Li

大语言模型多语言翻译能力评估与分析

◎ 评估对象

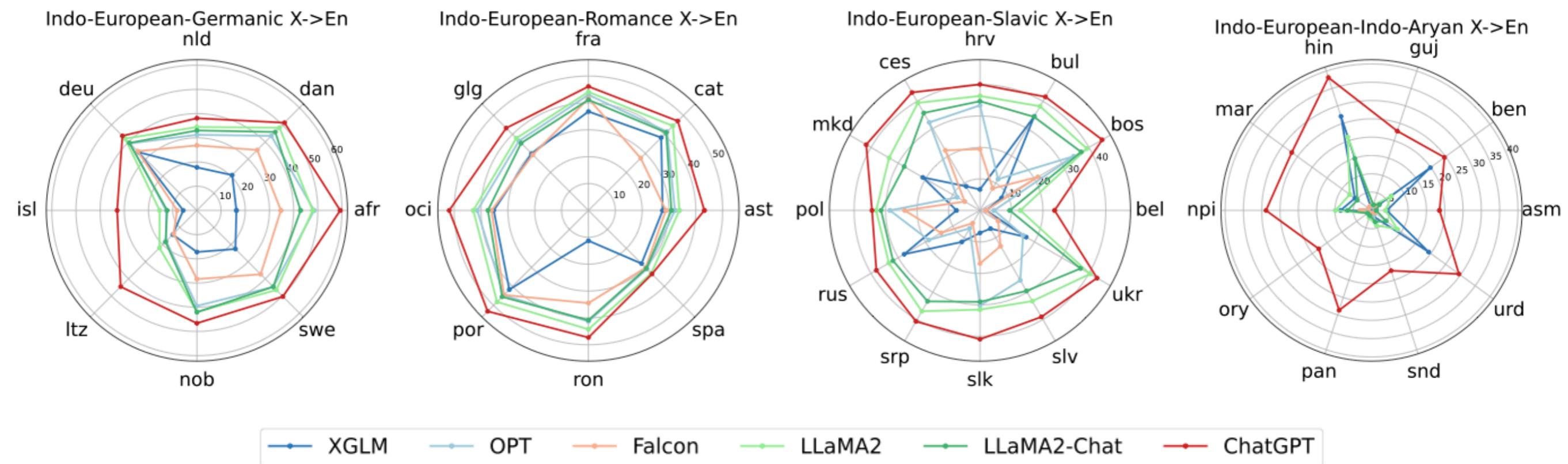
- ▶ 英语为主导的大模型：OPT-175B, LLaMA2-7B, Falcon-7B
- ▶ 覆盖多语言的大模型：XGLM-7B, BLOOMZ-7.1B, ChatGPT, GPT4
- ▶ 有监督基线模型：M2M-12B, NLLB-1.3B, Google Translate

◎ 评估过程

- ▶ 多语言翻译数据集：Flores-101
- ▶ 无泄漏数据集（新标注）
- ▶ 情景学习（避免改变模型参数）
- ▶ 102种语言，250个翻译方向

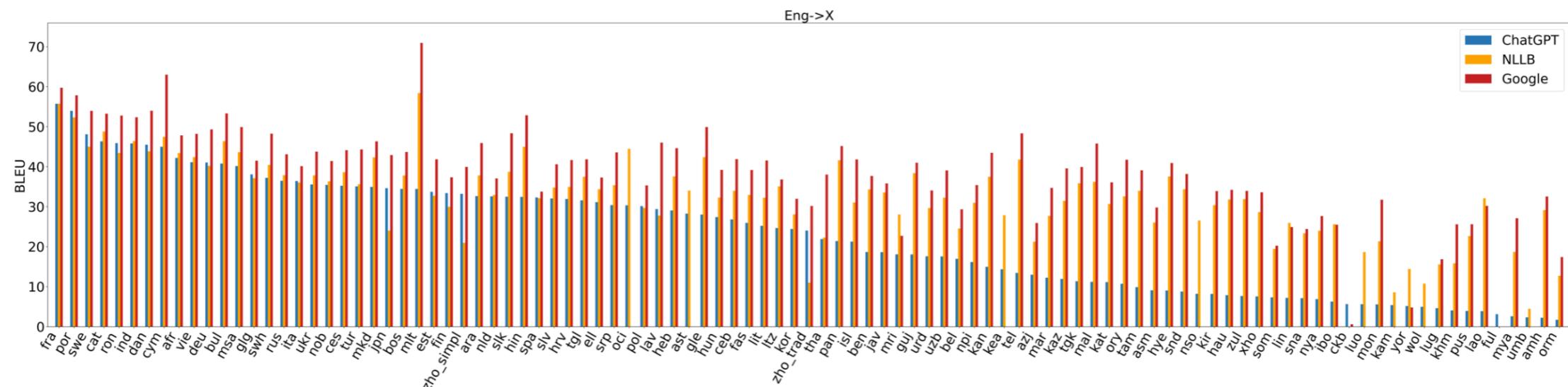
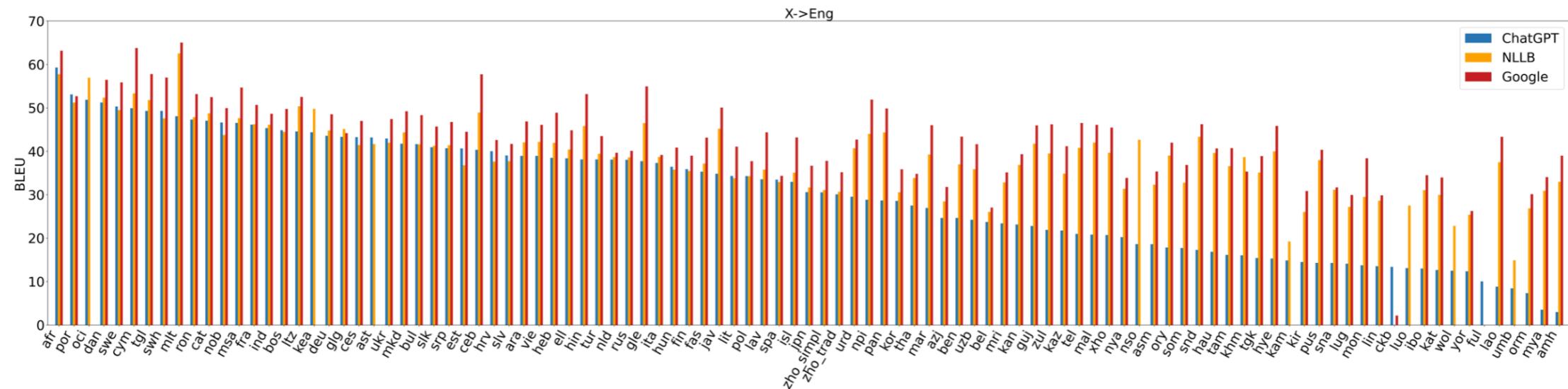
主要发现

- 在英语为主的大规模单语语料上训练后，大语言模型可以展现出不错的多语言翻译能力
- 其中ChatGPT/GPT4的多语言翻译能力是最强的



主要发现

- ChatGPT在16.67%的方向上超过了传统有监督模型NLLB
- 但是在低资源语言上，ChatGPT和传统机器翻译系统（NLLB、Google Translate）还有很大的差距。



主要发现

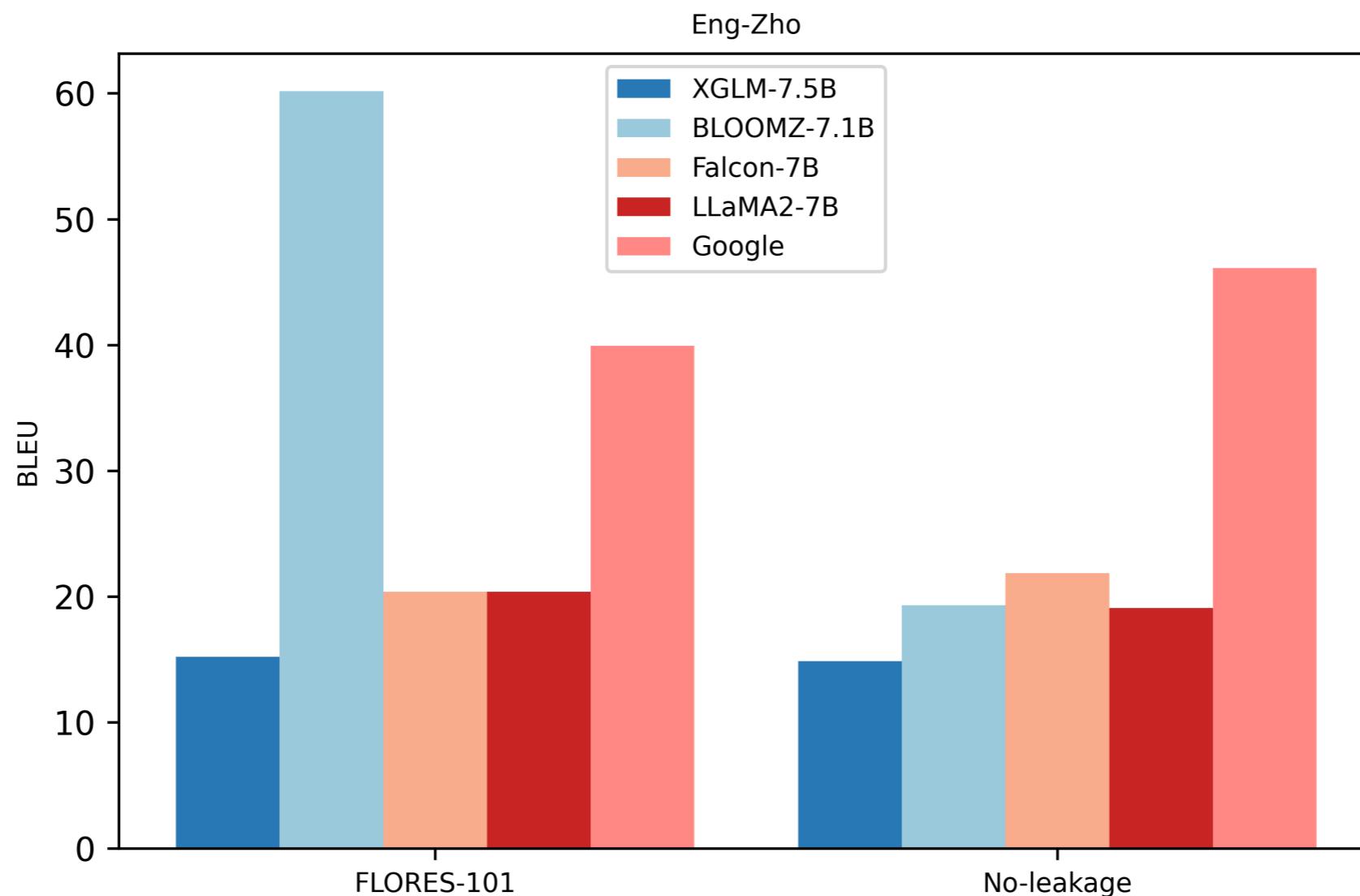
◎ 大模型内的语言不平衡现象

- ▶ 以英文为中心的翻译能力 > 以法文为中心的翻译能力 > 以中文为中心的翻译能力
- ▶ 在GPT4中，不平衡现象依然存在

Language	System	X-Eng	X-Fra	X-Zho	Eng-X	Fra-X	Zho-X
Deu	ChatGPT	43.56	41.63	26.83	41.02	33.75	26.25
	GPT4	47.04	44.56	29.36	44.69	34.79	27.39
Por	ChatGPT	53.09	46.09	27.43	53.95	42.29	28.29
	GPT4	52.81	46.79	27.43	55.89	44.19	33.28
Rus	ChatGPT	38.04	37.77	24.14	36.45	30.85	21.54
	GPT4	38.75	39.26	25.34	37.71	33.30	25.60
Hin	ChatGPT	38.15	33.75	20.79	32.44	23.85	18.05
	GPT4	45.88	38.60	26.15	35.30	28.32	21.17
Ita	ChatGPT	37.72	39.70	25.91	36.39	33.54	24.29
	GPT4	38.85	40.75	25.97	37.66	33.11	26.26
Ind	ChatGPT	45.33	40.27	27.35	45.80	35.43	25.98
	GPT4	47.54	43.28	27.71	47.97	35.94	26.62
Swh	ChatGPT	49.29	40.86	21.16	37.19	30.67	20.88
	GPT4	53.27	42.94	24.98	44.01	32.15	22.46
Ara	ChatGPT	38.94	38.46	24.85	32.64	26.22	18.82
	GPT4	43.29	39.58	27.52	36.91	27.13	20.12
Tur	ChatGPT	38.14	37.12	23.60	35.05	27.22	19.80
	GPT4	43.43	38.98	23.93	37.05	30.79	21.70
Tel	ChatGPT	20.97	20.07	12.43	13.40	8.69	5.42
	GPT4	35.02	30.96	20.90	20.67	15.61	11.84
Mya	ChatGPT	3.50	2.26	0.42	2.57	2.14	1.38
	GPT4	16.01	14.62	11.44	8.30	6.90	6.08
Fin	ChatGPT	35.90	36.31	25.63	33.38	26.43	20.93
	GPT4	40.17	38.85	24.85	35.33	30.69	23.53

数据泄漏问题

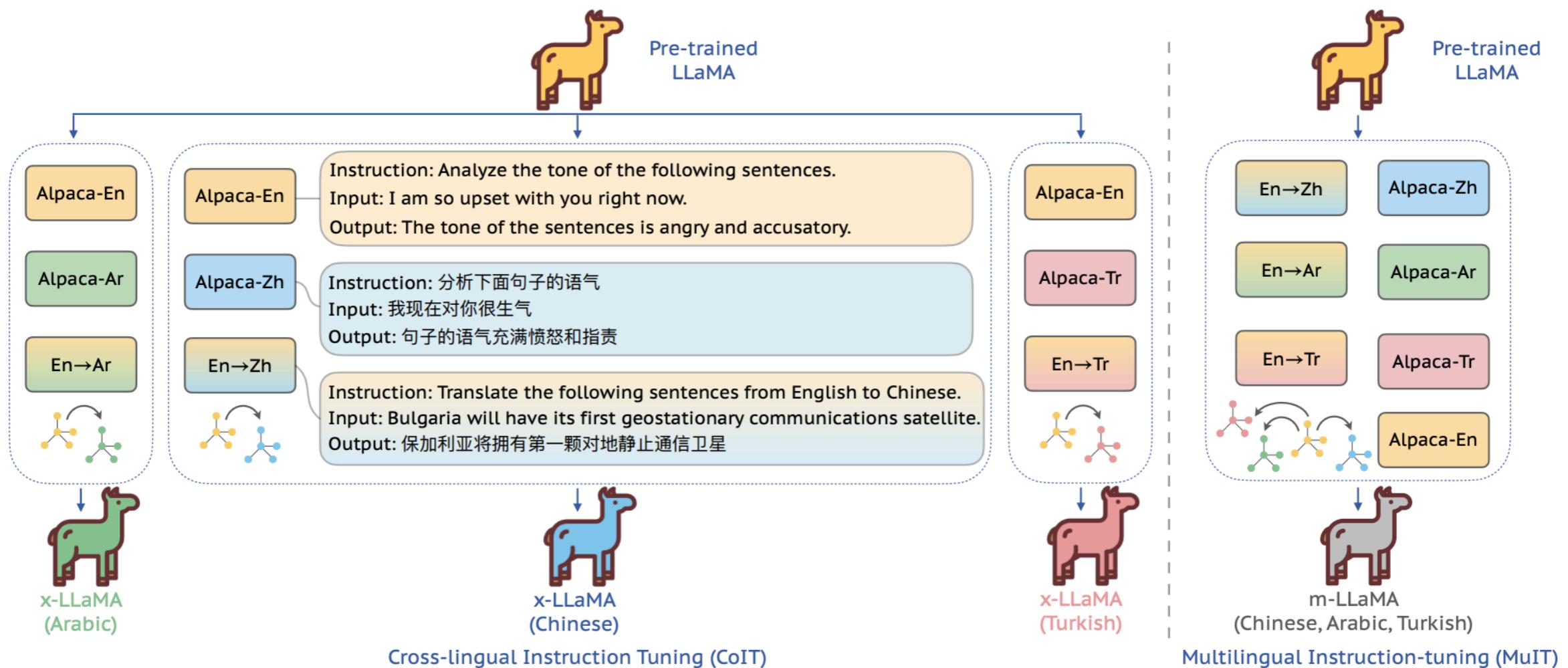
- ◎ BLOOMZ在Flores-101数据集和无泄漏数据集上的表现差别极大
- ◎ 大语言模型时代，使用开源数据评估模型能力存在风险



通过语言对齐提升大语言模型的非英语能力

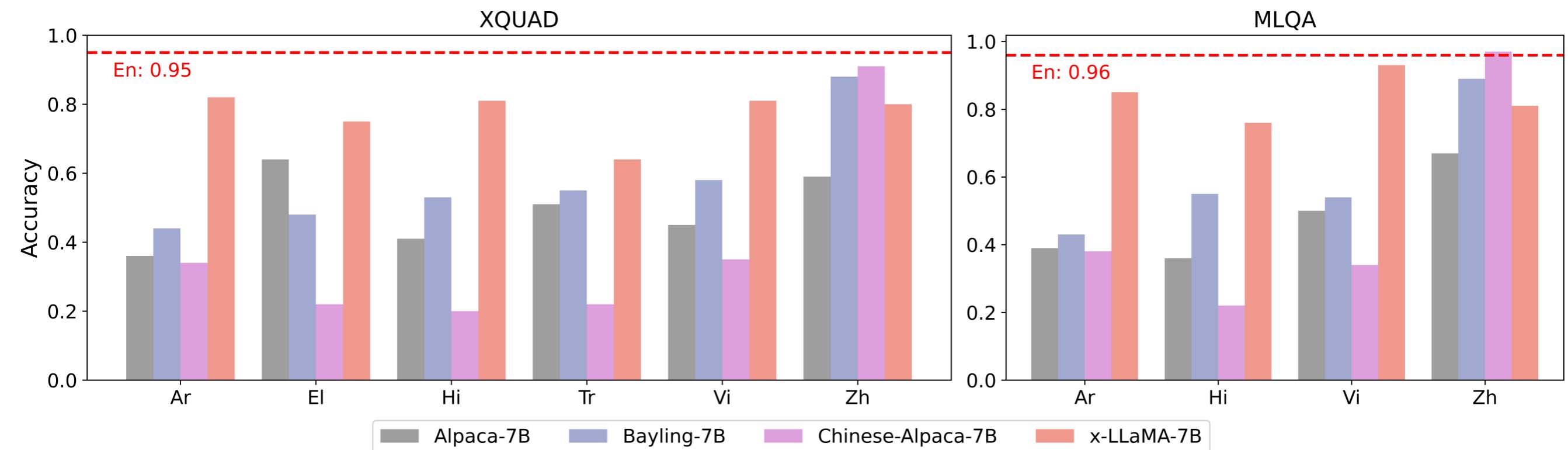
◎ 在指令微调时激发模型的非英语能力

- ▶ 多语言通用任务指令数据：执行多语言指令
- ▶ 翻译任务指令数据：将英语知识映射到非英语



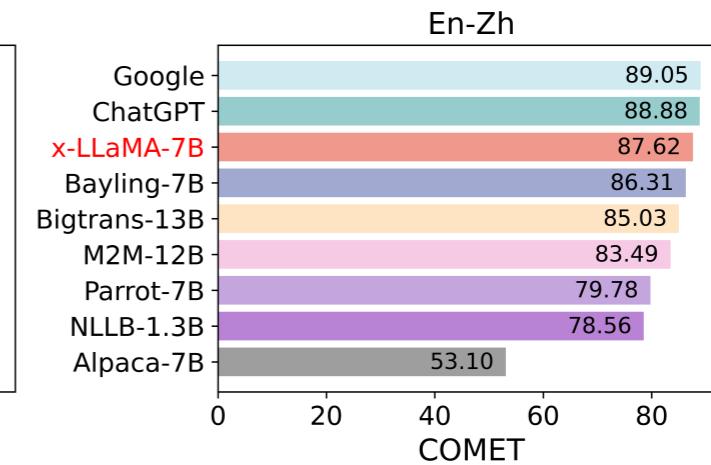
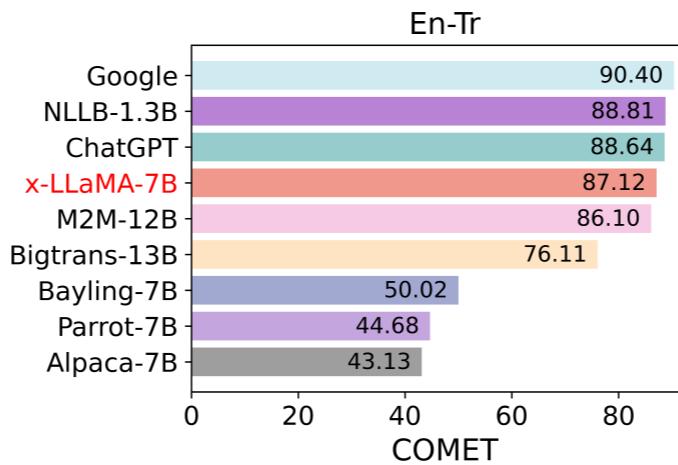
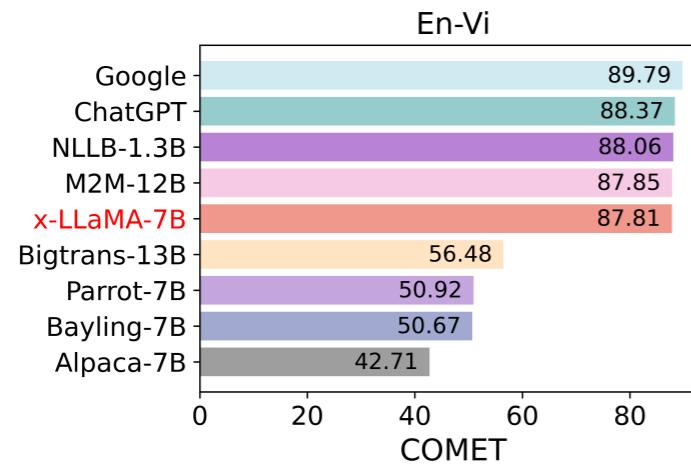
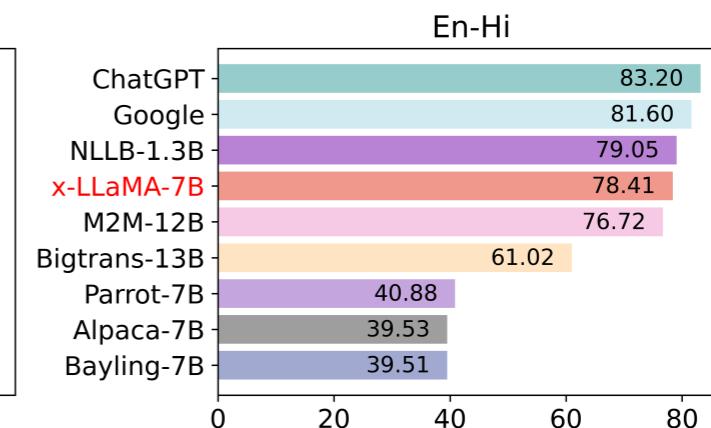
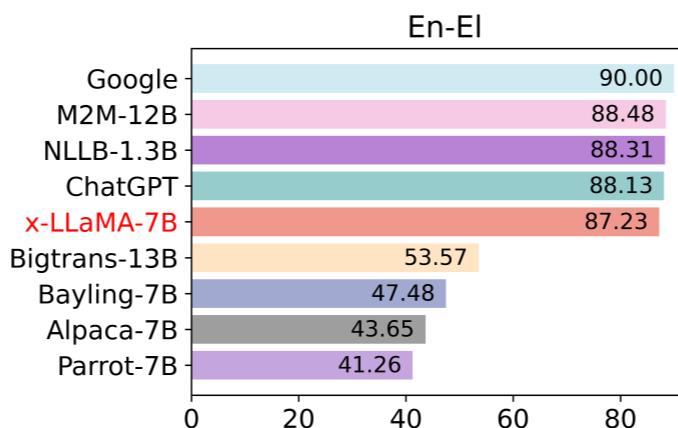
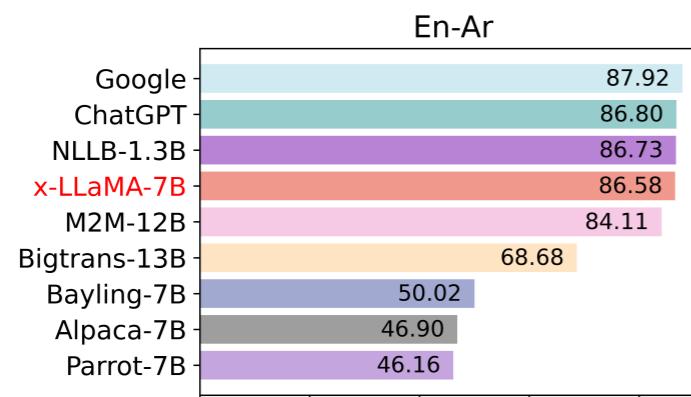
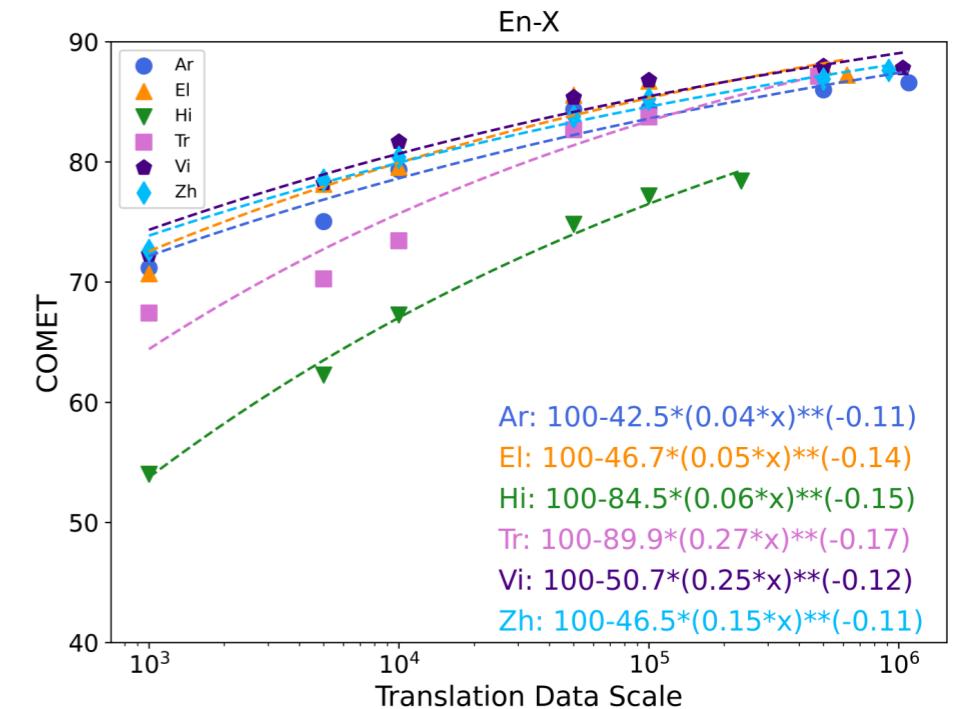
主要结果

- 相比于只用英语数据指令微调的Alpaca模型，x-LLaMA可以在非英语问答任务上取得大幅提升
- 使用开源平行语料，可以快速激发模型在多种非英语语言上的能力



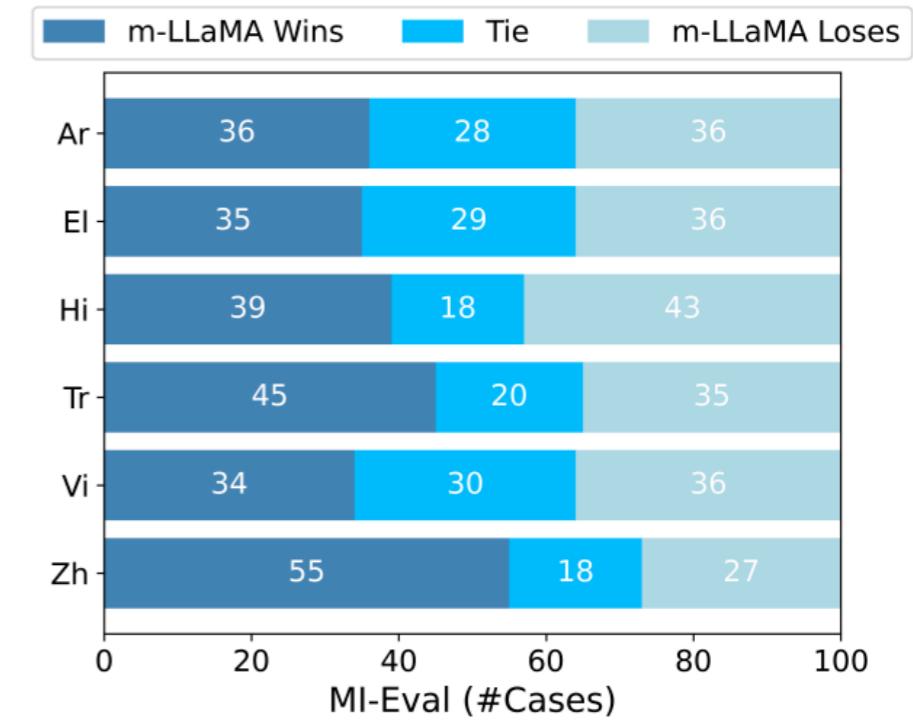
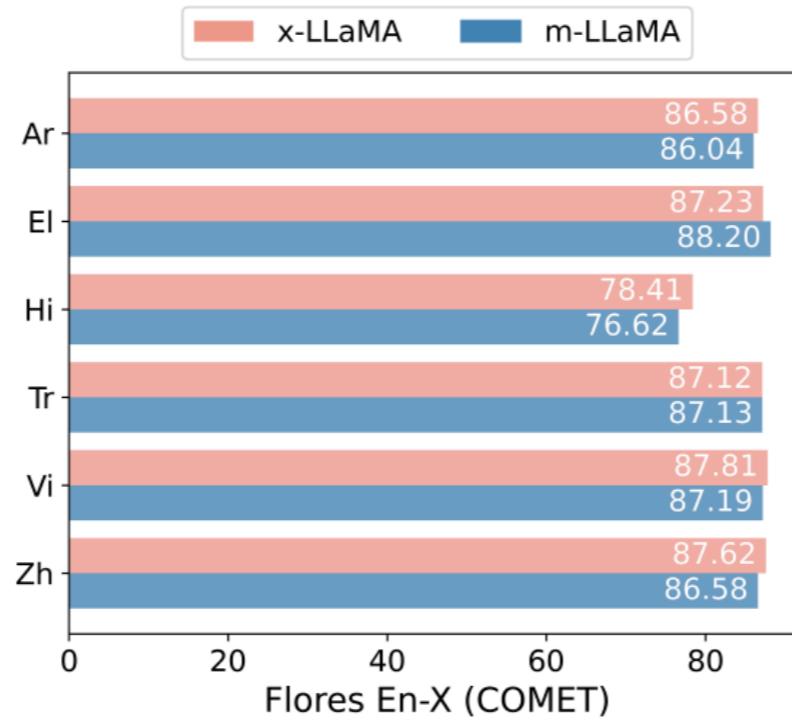
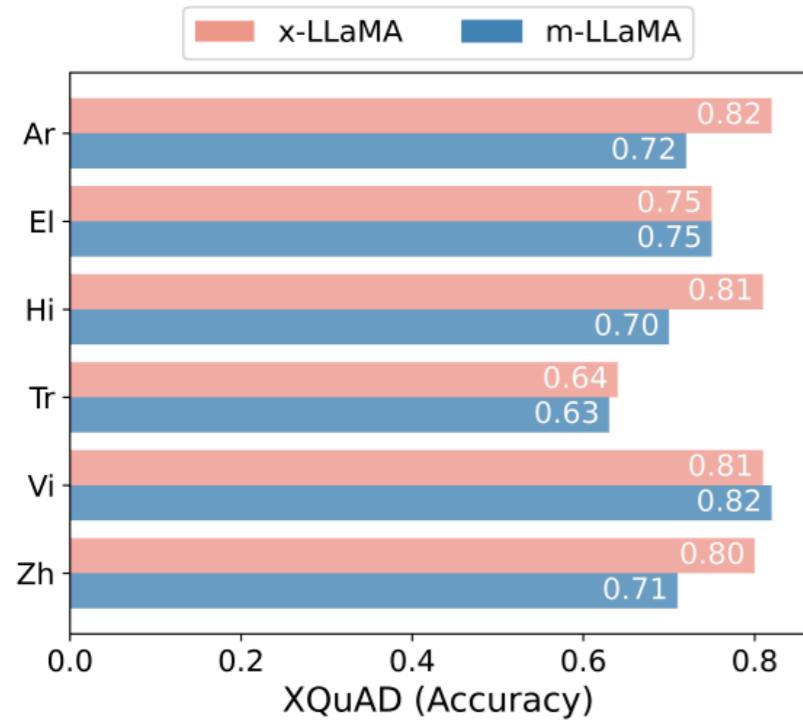
主要结果

- 使用可扩展的翻译数据对于提升模型非常重要
- x-LLaMA具备很好的翻译能力，在一半的方向上超过传统有监督模型M2M



主要结果

- 同时激发基座模型在多种非英语语言上的表现也是可能的
 - 执行多语言指令
 - 进行多语言机器翻译



总结

- ◎ 在整个科研过程中，科研选题至关重要
- ◎ 大语言模型为机器翻译研究带来了挑战和机遇
 - ▶ 翻译任务/数据可以在大模型时代产生更大的影响
- ◎ 我们的实践
 - ▶ 大语言模型多语言翻译能力评估与分析
<https://arxiv.org/pdf/2304.04675.pdf>
 - ▶ 通过语言对齐提升大语言模型的非英语能力
<https://arxiv.org/pdf/2308.04948.pdf>

谢谢！