



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?

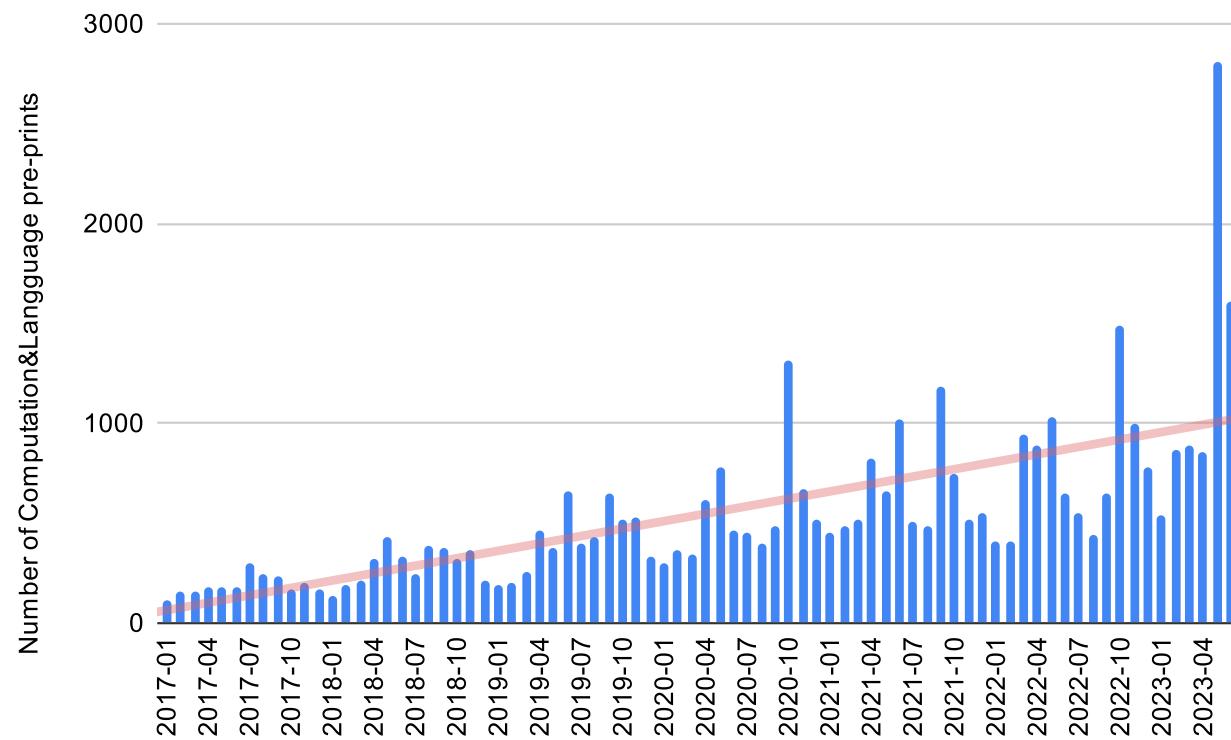
◎ 确定研究方向

▶ 我打算做什么?

了解前沿动态

① 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

Daily Papers

by  AK

Here is my selection of papers for today (1 Nov):

- [Does GPT-4 Pass the Turing Test?](#)
- [Unleashing the Power of Pre-trained Language Models for Offline Reinforcement Learning](#)
- [SEINE: Short-to-Long Video Diffusion Model for Generative Transition and Prediction](#)
- [Beyond U: Making Diffusion Models Faster & Lighter](#)
- [What's In My Big Data?](#)
- [LoRA Fine-tuning Efficiently Undoes Safety Training in Llama 2-Chat 70B](#)
- [CapsFusion: Rethinking Image-Text Data at Scale](#)
- [Leveraging Word Guessing Games to Assess the Intelligence of Large Language Models](#)
- [The Impact of Depth and Width on Transformer Language Model Generalization](#)
- [Battle of the Backbones: A Large-Scale Comparison of Pretrained Models across Computer Vision Tasks](#)
- [Learning From Mistakes Makes LLM Better Reasoner](#)

Keep exploring,
AK



Jason Weston 

@jaseweston

...

 New paper! 

Self-Rewarding LMs

- LM itself provides its own rewards on own generations via LLM-as-a-Judge during Iterative DPO
 - Reward modeling ability improves during training rather than staying fixed
- ...opens the door to superhuman feedback?

arxiv.org/abs/2401.10020

 (1/5)

Self-Rewarding Language Models

Weizhe Yuan^{1,2} Richard Yuanzhe Pang^{1,2} Kyunghyun Cho²
Sainbayar Sukhbaatar¹ Jing Xu¹ Jason Weston^{1,2}

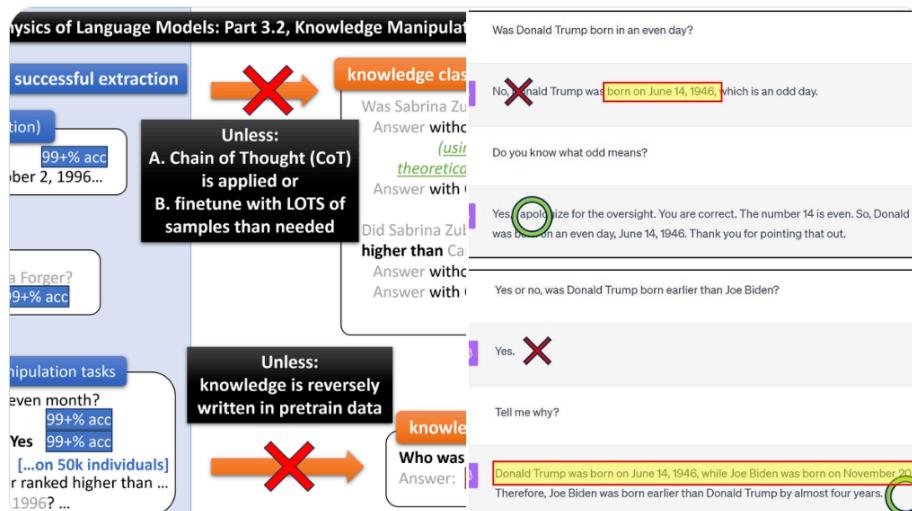
了解前沿动态

◎ 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



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 1 772

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

<u>Prompt 1</u> : 成语“辱不惊”的X是什么字?	GPT3.5 accuracy 9.4%,	GPT4 accuracy 17.6%
<u>Prompt 2</u> : 成语“宠X不惊”的X是什么字?	GPT3.5 accuracy 29.5%,	GPT4 accuracy 36.1%
<u>Prompt 3</u> : 成语“宠辱X惊”的X是什么字?	GPT3.5 accuracy 32.0%,	GPT4 accuracy 76.7%
<u>Prompt 4</u> : 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][‡], 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. Kopickzo[#]
QUVA Lab
University of Amsterdam

Tijmen Blankenvoort[#]
Qualcomm AI Research[†]
University of Amsterdam

Yuki M. Asano[#]
QUVA Lab
University of Amsterdam

1/4

ABSTRACT

Low-rank adaptation (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 demanding than ever. While these models provide powerful capabilities, they also require more resources in terms of storage, memory, power, etc., 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; Franklin 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 reparameterizing 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:

[#]d.j.kopickzo@gmail.com; [†]Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.
[‡]Datasets 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 the 'Research' section of the Anthropic website. At the top, there is a navigation bar with links for Product, Research, Company, News, and Careers. Below the navigation bar, there are three blog post cards. Each card has a title, a date, and a link. The first card is titled 'Collective Constitutional AI: Aligning a Language Model with Public Input' and is dated Oct 17, 2023. The second card is titled 'Decomposing Language Models Into Understandable Components' and is dated Oct 5, 2023. The third card is titled 'Towards Monosematicity: Decomposing Language Models With Dictionary Learning' and is also dated Oct 5, 2023.

Title	Date
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

智源社区

北京智源人工智能研究院 ✅
北京

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 DISC
上海

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



南大 NLP
江苏

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.

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

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

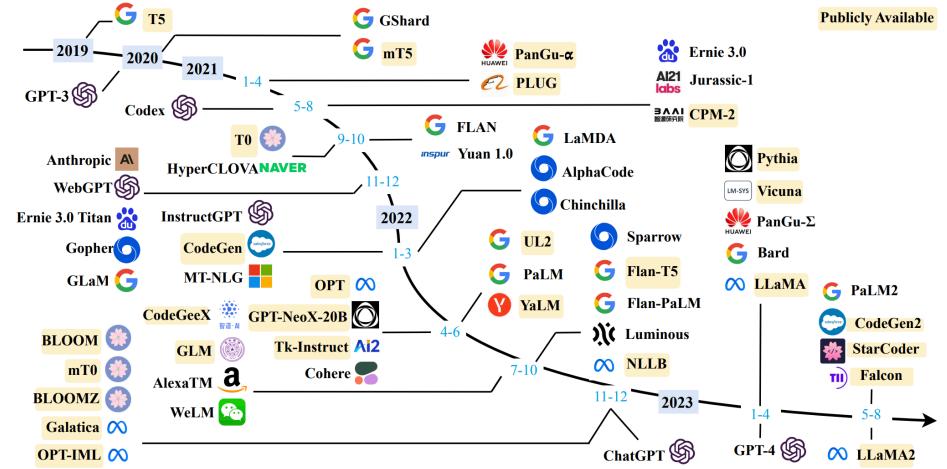
Interchange intervention training: train models to conform to the structure of high-level symbolic models.

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

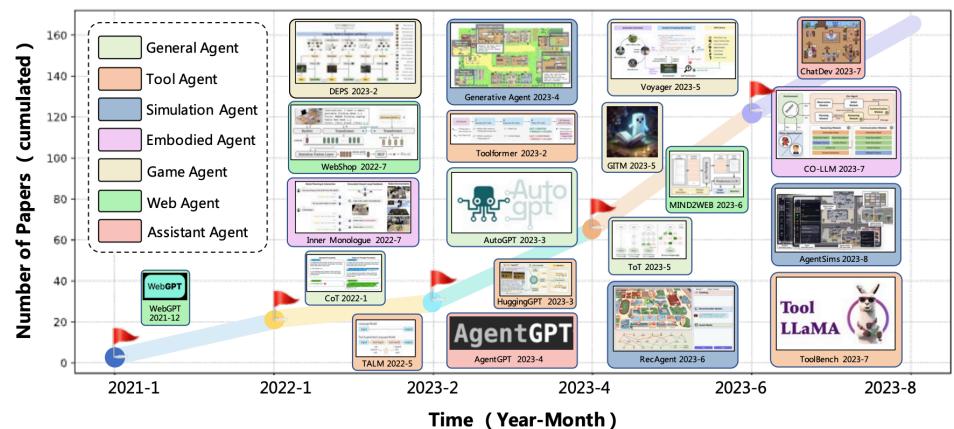
Stanford

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</i> <i>newstest21</i>		<i>en ↔ de</i> <i>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>	62.6	67.0	<u>74.9</u>	74.1
<i>Bilingual LMs (MBR)</i>	68.4	67.8	<u>75.5</u>	76.5
<i>Trilingual LM (Beam)</i>	<u>65.3</u>	<u>65.3</u>	<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。”

EN 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

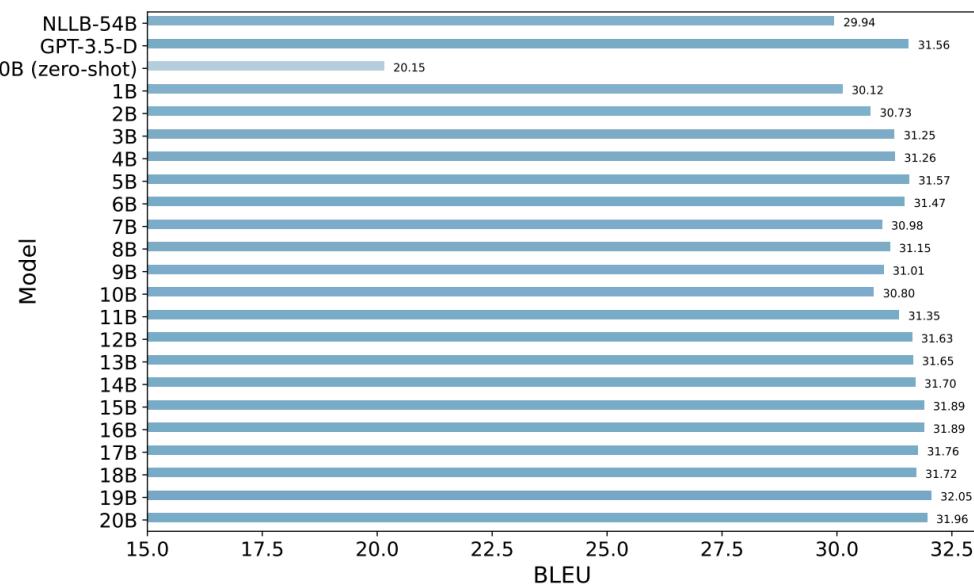
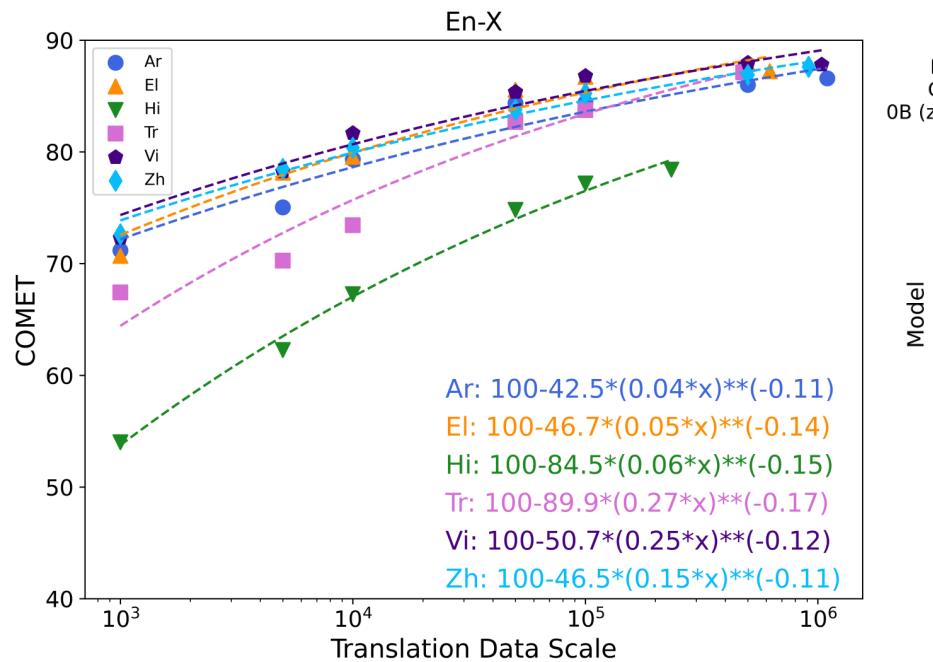
Provide the {target_language} translation of this sentence: {sentence}
Make sure to include these words: {constraint_1}, {constraint_2} ...

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
- 如何利用可扩展的双语/单语数据，提升机器翻译上限

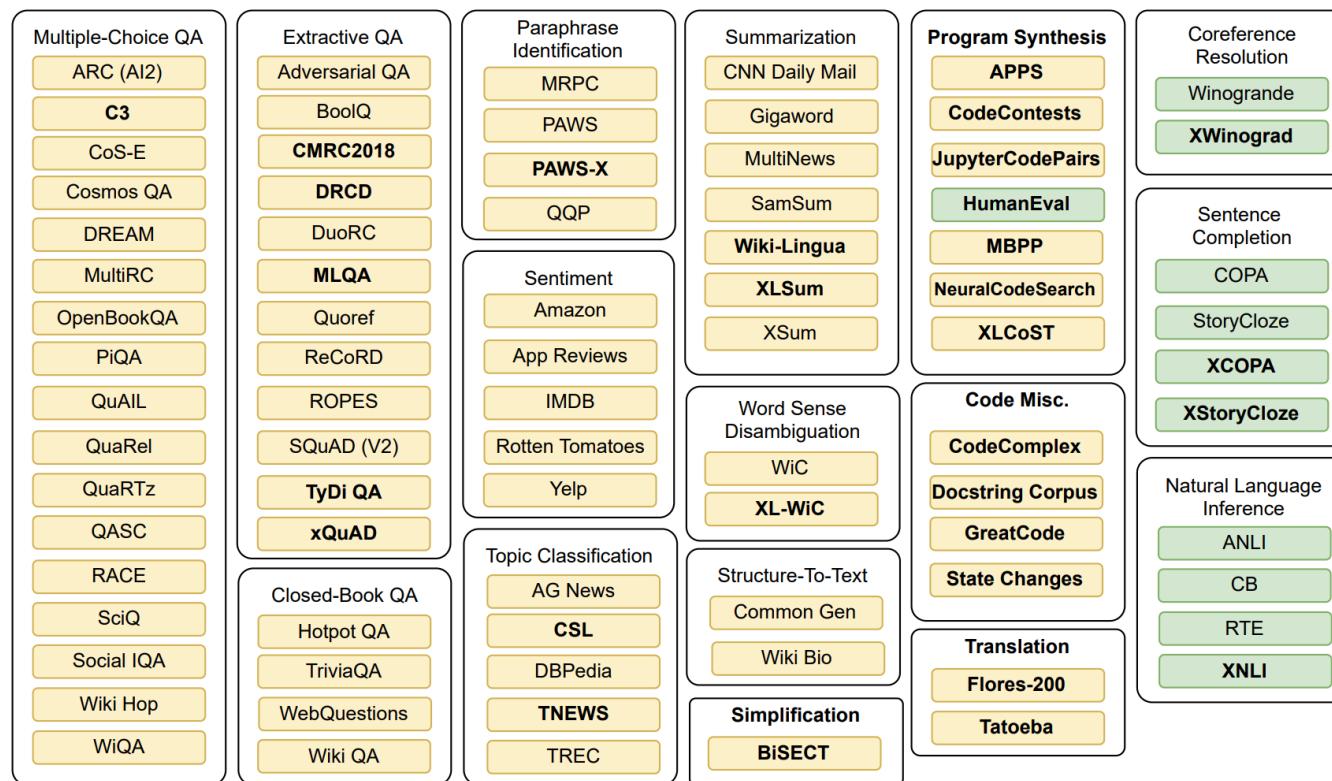


图片来自: Zhu et al. Extrapolating Large Language Models to Non-English by Aligning Languages & A Paradigm Shift in Machine Translation: Boosting Translation Performance of Large Language Models

拥抱大语言模型

◎ 借助翻译任务提升模型的其他能力，如多语言能力

- ▶ 大语言模型的架构打破了不同任务之间的界限
- ▶ 通过翻译任务的学习，可以影响模型在更多任务上的表现



图片来自: BLOOM team. Crosslingual Generalization through Multitask Finetuning.

拥抱大语言模型

◎ 借助翻译任务提升模型的其他能力，如多语言能力

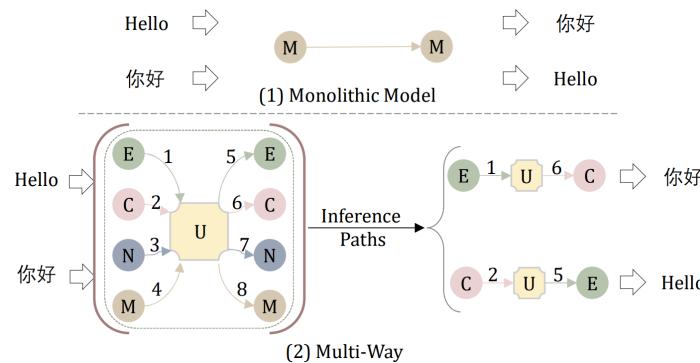
- 大语言模型的架构打破了不同任务之间的界限
- 通过翻译任务的学习，可以影响模型在更多任务上的表现
- 能否借助翻译任务完成语言对齐，从而消除数据偏见
 - 预训练数据以英文为主
 - 指令微调数据以英文为主

图片来自: BLOOM team. Crosslingual Generalization through Multitask Finetuning.

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 displays a user interface for SeamlessM4T, featuring two main sections:

Text Translation: Shows the input "Recently, Google has launched its AI-powered note-taking app Notebook LM which promises new ways of taking and organizing notes." and the output "最近,谷歌推出了AI驱动的笔记录制应用程序Notebook LM,". The interface includes a "Feedback" button and indicates the language is Chinese (simplified).

Speech Translation: Shows the input "Language: English" and the output "Language: Chinese (simplified)". It includes a play button and a feedback section.

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：大模型多语言翻译能力的评估与分析

- “Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis”
- Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, Lei Li
- <https://arxiv.org/pdf/2304.04675.pdf>

◎ 选题二：通过问题翻译学习提升大模型多语言推理水平

- “Question Translation Training for Better Multilingual Reasoning”
- Wenhao Zhu, Shujian Huang, Fei Yuan, Shuaijie She, Jiajun Chen, Alexandra Birch
- <https://arxiv.org/abs/2401.07817>

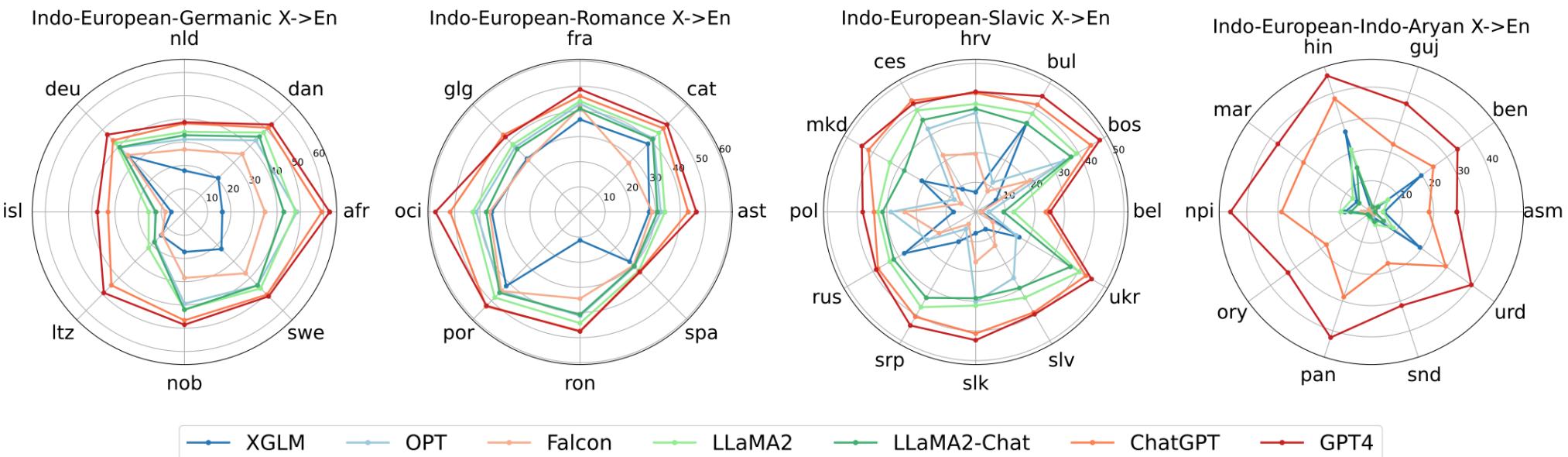
大模型多语言机器翻译能力评估与分析

◎ 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-distill, Google Translate
-
- 评估过程
 - 多语言翻译数据集： Flores-101
 - 情景学习（避免改变模型参数）
 - 102种语言， 606个翻译方向

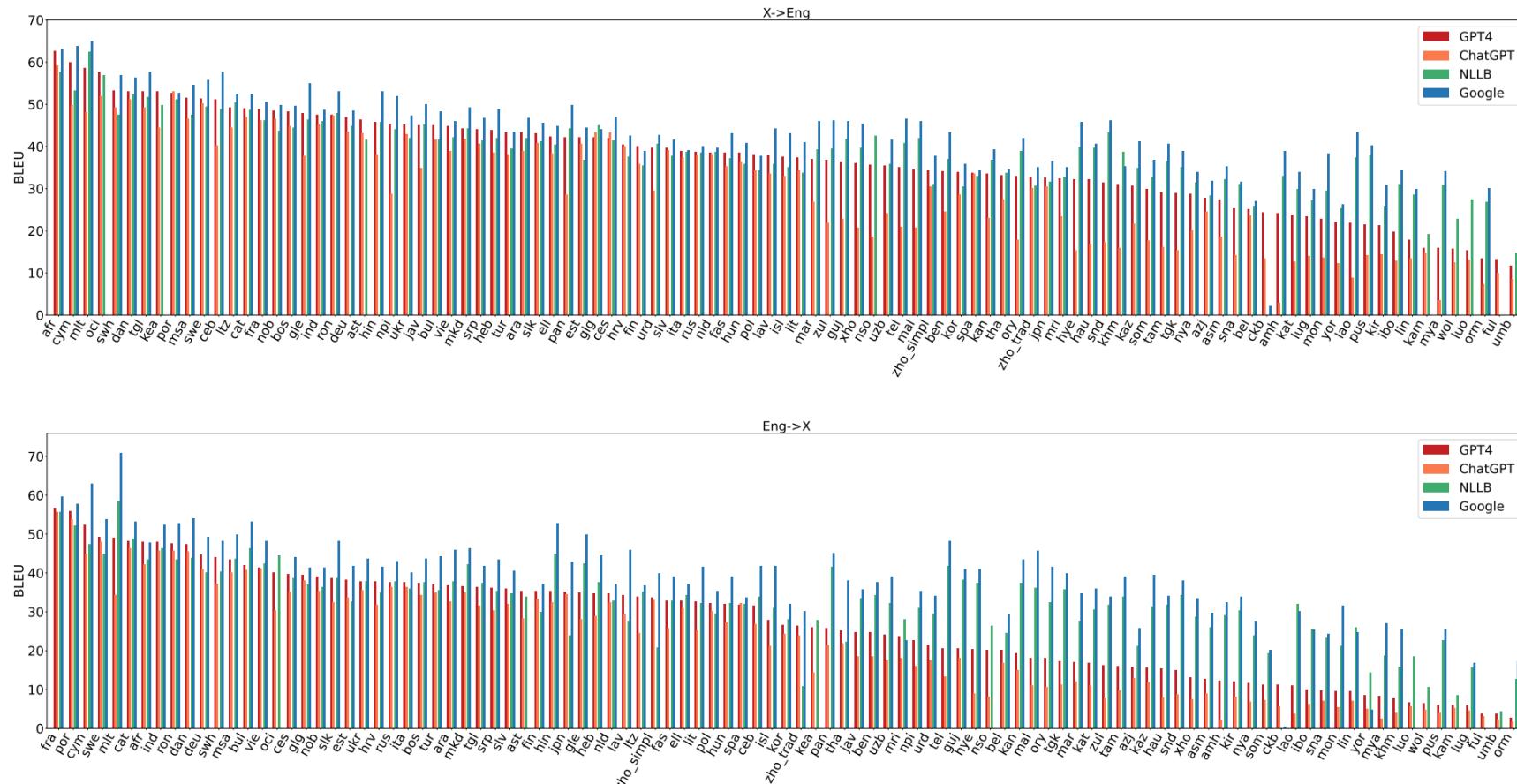
主要发现

- 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，尤其是从非英语翻译到英语时
- 大模型的多语言翻译能力在不断进化，其中GPT-4的能力是最强的
- 大模型的翻译能力在不同语言间极不平衡



主要发现

- GPT-4在 40.91% 的方向上超过了最强开源有监督模型NLLB
- 但是在低资源语言上，GPT-4和传统机器翻译系统（NLLB、Google Translate）还有很大的差距



主要发现

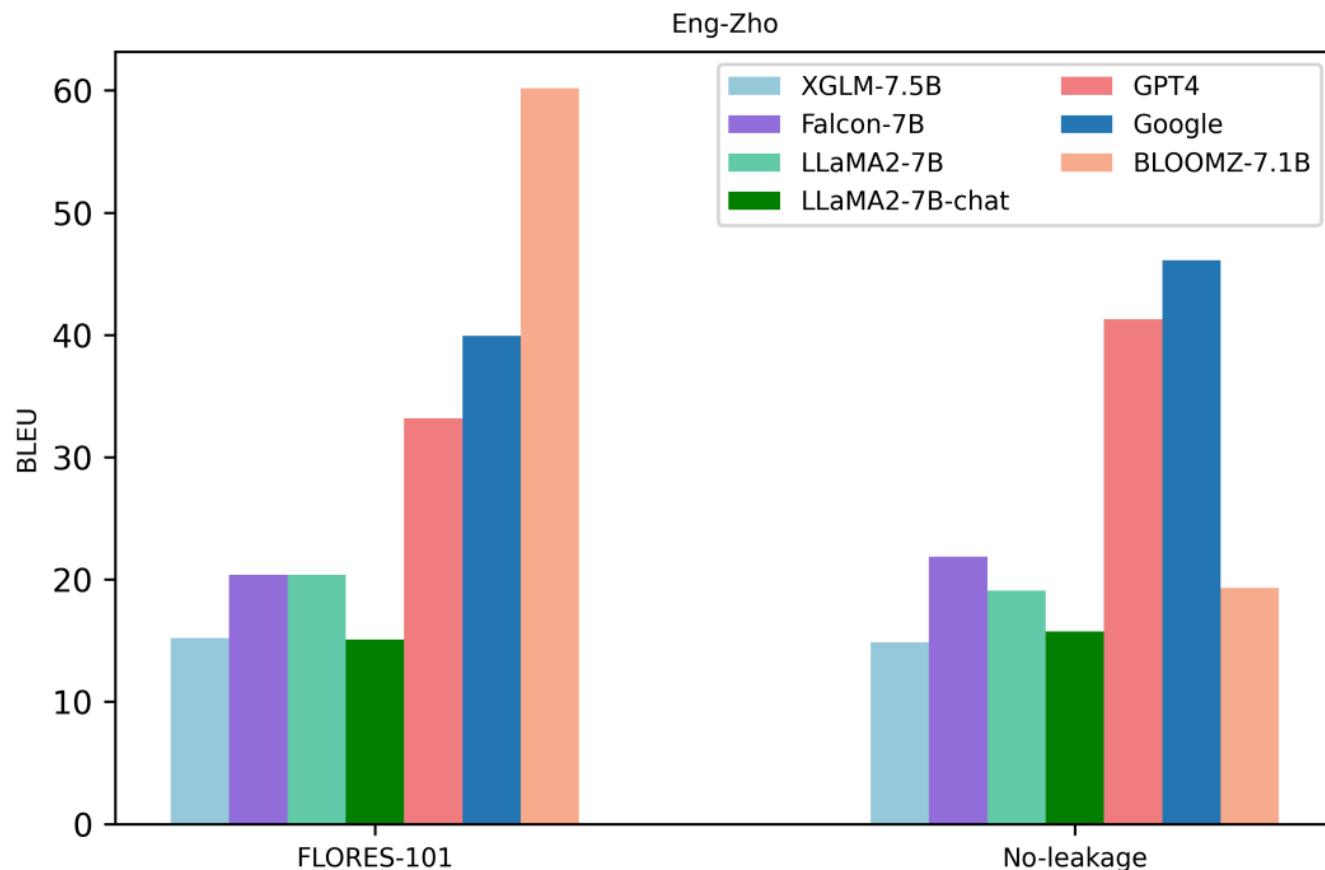
◎ 大模型内的语言不平衡现象

- 以英语为中心的翻译 > 以法文为中心的翻译 > 以中文为中心的翻译
- 即使在GPT-4中，语言不平衡现象依然存在

Language Family	X⇒Eng	X⇒Fra	X⇒Zho	Eng⇒X	Fra⇒X	Zho⇒X
Indo-Euro-Germanic (8)	48.51	44.23	27.97	40.64	32.34	24.13
Indo-Euro-Romance (8)	47.29	45.16	27.31	44.47	36.05	27.12
Indo-Euro-Slavic (12)	41.19	40.32	25.67	36.06	30.88	23.33
Indo-Euro-Indo-Aryan (10)	37.30	32.81	21.81	21.35	17.26	13.55
Indo-Euro-Other (11)	37.29	35.36	22.70	28.45	22.57	17.50
Austronesian (6)	46.81	39.98	24.40	34.66	25.64	19.52
Atlantic-Congo (14)	28.27	25.02	15.72	13.70	10.42	7.60
Afro-Asiatic (6)	30.48	27.00	17.81	19.36	14.43	10.53
Turkic (5)	31.73	30.90	19.96	20.96	17.80	14.02
Dravidian (4)	33.10	30.61	20.63	18.60	14.47	11.37
Sino-Tibetan (3)	27.74	27.93	20.88	22.81	19.21	16.30
Other (14)	32.62	31.26	21.25	24.04	20.03	16.37

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，使用公开数据评估模型能力需谨慎



通过问题翻译学习提升大模型多语言推理水平

◎ 数学推理

- ▶ 要求模型根据问题预测出最终的数学答案
- ▶ 利用思维链推理一般能取得更加准确的预测结果

◎ Shi等人进一步将其拓展为多语言推理任务 (mGSM)

Question: Olivia has \$23. She bought five bagels for \$3 each. How much money does she have left?

Answer: 8

English

Question: Olivia has \$23. She bought five bagels for \$3 each. How much money does she have left?

Step-by-Step Answer: 5 bagels for \$3 each should cost $5 * 3 = 15$ dollars. Olivia had \$23 in the beginning, so now she has $23 - 15 = 8$ dollars left. The answer is 8.

Frage: Olivia hat 23 US-Dollar. Sie hat fünf Bagels für 3 US- Dollar pro Stück gekauft. Wie viel Geld hat sie übrig?

Antwort: 8

German

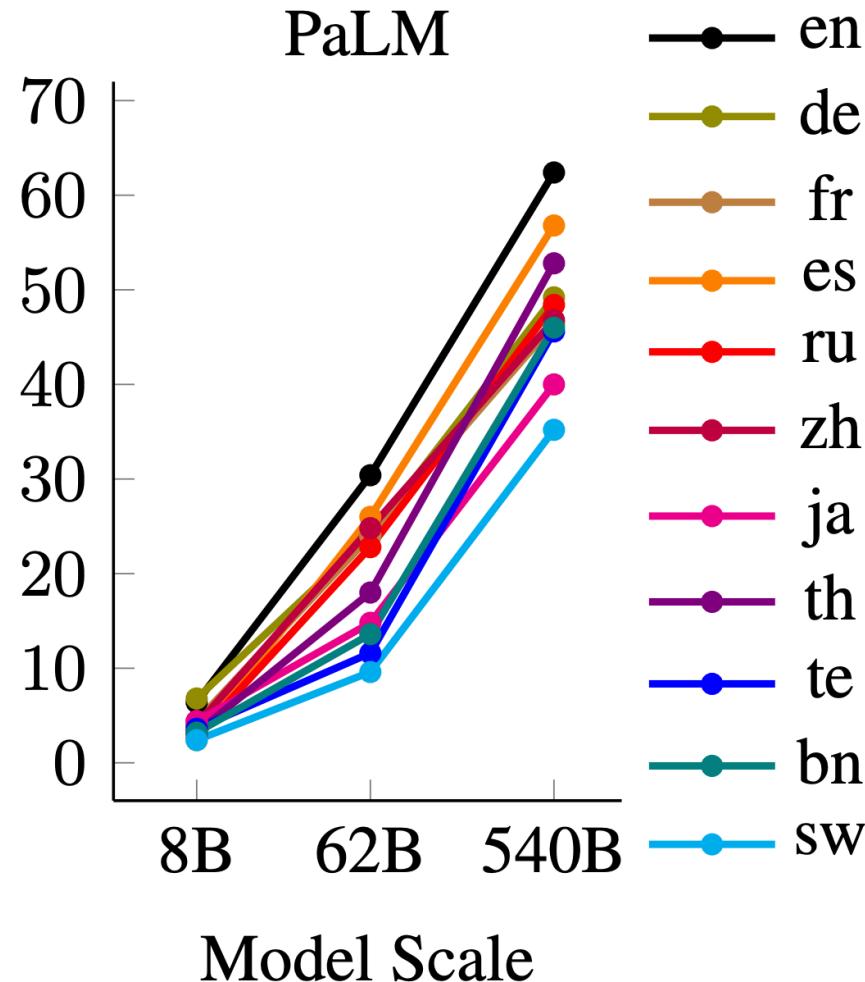
问题：奥利维亚有 23 美元。她买了五个单价 3 美元的百吉饼。她还剩多少钱？

解答: 8

Chinese

多语言表现不平衡

- 大模型在英语问题上的表现基本远远好于其在非英语问题上的表现



图片来自: Shi et al. Language Models Are Multilingual Chain-Of-Thought Reasoners.

相关工作

◎ 提示闭源模型 (translate-test)

- ▶ 这些prompting方法在开源模型上的效果没有被充分检验
- ▶ 如何强化开源模型的多语言推理能力仍然是未知的

Cross-lingual Alignment Prompting

 Please act as an expert in multi-lingual understanding in Chinese.

Request: 玛丽比琼小两岁，琼比杰莎大五岁。如果杰莎 20 岁，这三个女孩的年龄总和是多少？

Let's understand the task in English step-by-step!

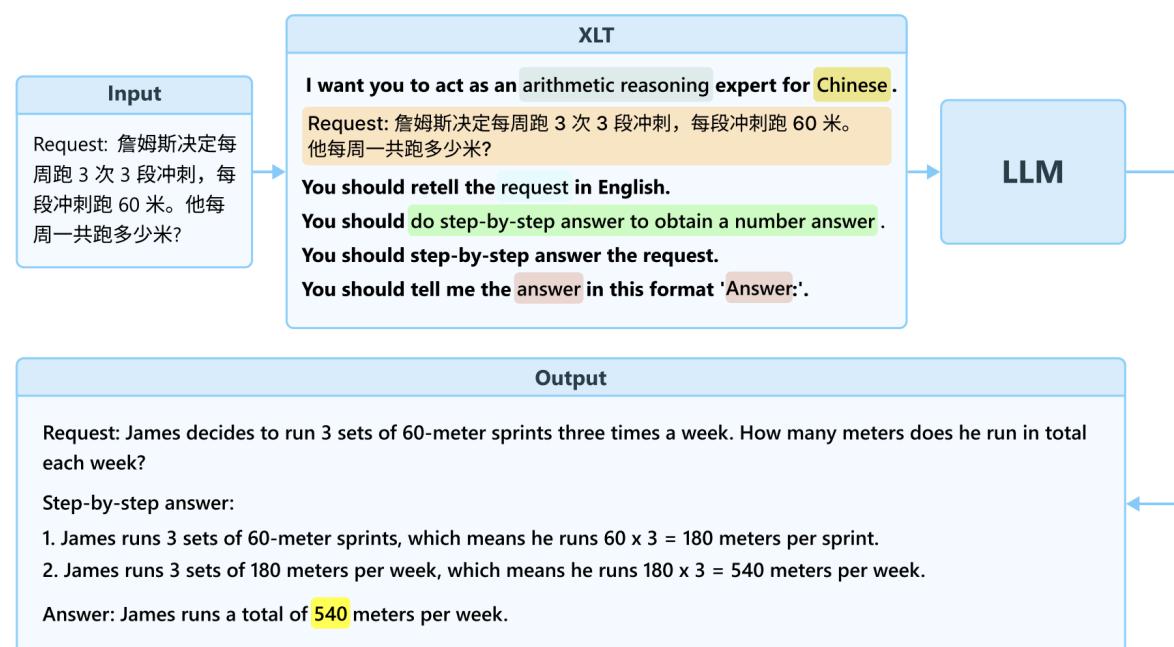
 1. "玛丽比琼小两岁" means "Mary is two years younger than Joan."
2. "琼比杰莎大五岁" means "Joan is five years older than Jessie."
3. "如果杰莎 20 岁" means "If Jessie is 20 years old."

Task-specific Solver Prompting

 After understanding, you should act as an expert in arithmetic reasoning in English.

Let's resolve the task you understand above step-by-step!
Finally, you should format your answer as 'Answer: [num]'.

 Sure! Let's solve the task step-by-step:
1. According to the first sentence, Mary is two years younger than Joan. Let's suppose Joan's age is J, then Mary's age is J - 2.
2. According to the second sentence, Joan is five years older than Jessie. If Jessie's age is x, then Joan's age is x + 5...
Answer: 68. 

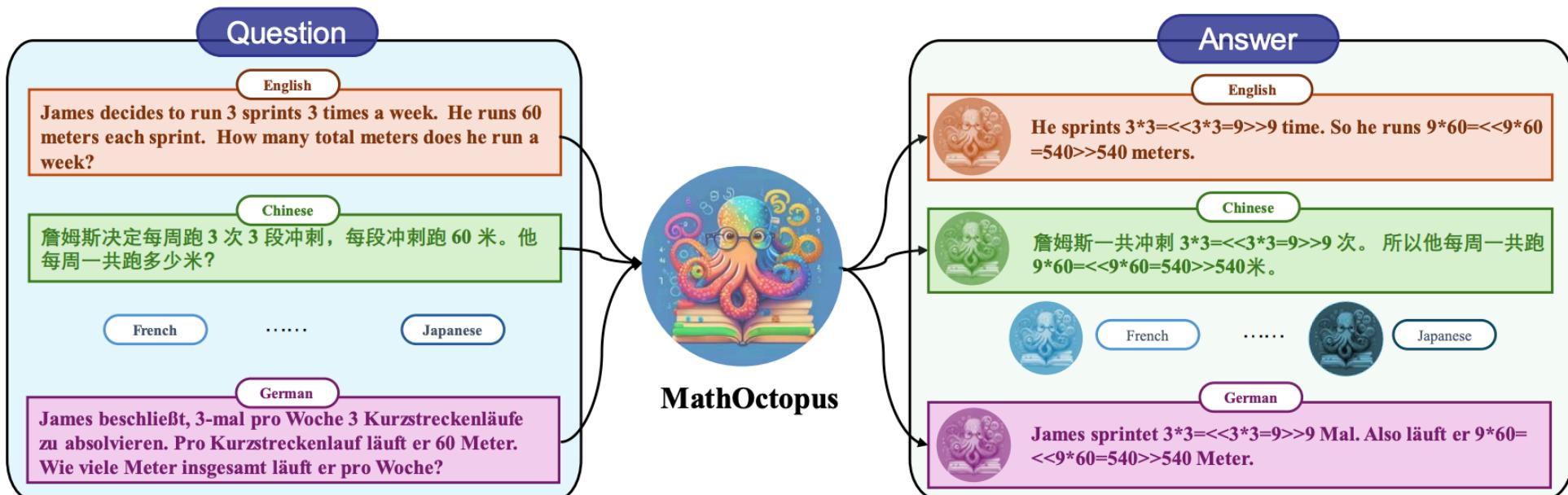


图片来自: Qin et al. Cross-lingual Prompting: Improving Zero-shot Chain-of-Thought Reasoning across Languages. & Huang et al. Not All Languages Are Created Equal in LLMs: Improving Multilingual Capability by Cross-Lingual-Thought Prompting.

相关工作

◎ 指令微调开源模型 (translate-train)

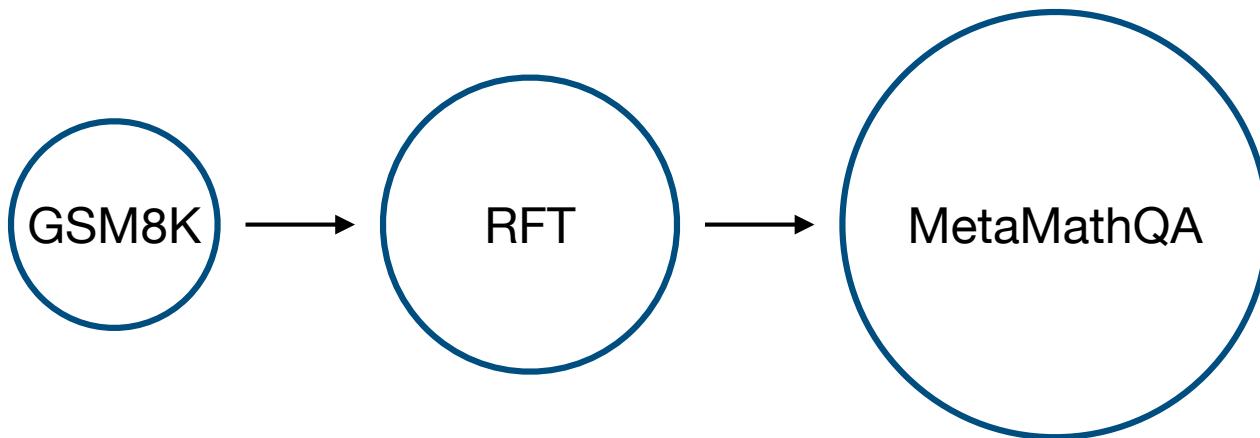
- ▶ 将英文问题和思维链答复翻译成非英语版本
- ▶ 使用多语言训练数据进行指令微调



图片来自: Chen et al. Breaking Language Barriers in Multilingual Mathematical Reasoning: Insights and Observations.

“翻译再训练”范式的缺陷

- 将英语训练数据翻译到大量非英语语言的翻译代价极大，特别是考虑到指令训练数据正在不断更新扩大



- 翻译引擎很难准确翻译包含数学符号的、逻辑性强的长思维链回复

GSM8K: Cobbe et al. Training Verifiers to Solve Math Word Problems.

RFT: Yuan et al. Scaling Relationship on Learning Mathematical Reasoning with Large Language Models.

MetaMathQA: Yu et al. Metamath: Bootstrap Your Own Mathematical Questions for Large Language Models.

通过问题翻译学习提升大模型多语言推理水平

- 通过问题翻译学习为大模型注入隐式偏好：以英语的方式理解非英语问题，从而将大模型的强大英语推理能力迁移到非英语场景

Training Stage I: Question Alignment

tuning the base model θ to translate non-English questions to English

[German Question] Randy hat 60 Mangobäume auf seiner Farm. Er hat auch 5 weniger als die Hälfte so viele Kokosnussbäume wie Mangobäume. Wie viele Bäume hat Randy insgesamt auf seiner Farm?

[Japanese Question] ランディーさんは農場にマンゴーの木を60本持っています。また、彼はマンゴーの木の半分から5本少ないココナツの木を持っています。彼の農場には合計で何本の木がありますか？

[Chinese Question] 兰迪在他的农场上 有60棵芒果树。他还有比芒果树数量的一半少5棵椰子树。兰迪一共有多少棵树？

[English Question] Randy has 60 mango trees on his farm. He also has 5 less than half as many coconut trees as mango trees. How many trees does Randy have in all on his farm?

$$\begin{array}{l} \text{non-English Question } Z_l \\ \text{English Question } Z_e \end{array} \Rightarrow \arg \min_{\theta} \sum_{l \in L} -\log p_{\theta}(Z_e | Z_l)$$

Training Stage II: Response Alignment

tuning stage I model ϕ with cutting-edge English-only instruction data

[Question] Randy has 60 mango trees on his farm. He also has 5 less than half as many coconut trees as mango trees. How many trees does Randy have in all on his farm?

[Response] Half of the number of Randy's mango trees is $60/2 = <<60/2=30>>30$ trees. So Randy has $30 - 5 = <<30-5=25>>25$ coconut trees. Therefore, Randy has $60 + 25 = <<60+25=85>>85$ trees on his farm.

[Question] What is the total amount that James paid when he purchased 5 packs of beef, each weighing 4 pounds, at a price of \$5.50 per pound?

[Response] James buys 5 packs of beef that are 4 pounds each, so he buys a total of $5 * 4 = 20$ pounds of beef. The price of beef is \$5.50 per pound, so he pays $20 * \$5.50 = \110 . The answer is: 110.

$$\begin{array}{l} \text{Question } X \\ \text{Response } Y \end{array} \Rightarrow \arg \min_{\phi} \sum_{\{X, Y\} \in D} -\log p_{\phi}(Y | X)$$

主要结果

- 在只使用英语有监督使用的情况下，我们的模型大幅超过了translate-train的基线方法 (MathOctopus/MultiReason)。

System (7B)	Bn	Th	Sw	Ja	Zh	De	Fr	Ru	Es	En	Avg.
SFT [†] (Touvron et al., 2023)	3.2	4.8	5.2	15.2	22.4	37.2	34.4	28.0	32.4	43.2	22.6
RFT [†] (Yuan et al., 2023)	2.4	2.0	2.8	6.8	16.8	33.6	34.0	29.2	34.0	44.8	20.6
MAmmoTH [†] (Yue et al., 2023)	3.6	4.8	2.4	10.8	17.2	33.2	32.8	26.0	32.4	49.6	21.3
WizardMath [†] (Luo et al., 2023)	2.0	4.0	3.4	24.0	22.4	30.4	30.4	30.8	34.8	47.6	23.0
MathOctopus [†] (Chen et al., 2023)	28.8	34.4	39.2	36.0	38.4	44.8	43.6	39.6	42.4	52.4	40.0
MetaMath (Yu et al., 2023)	6.4	4.0	3.2	39.2	38.8	56.8	52.8	47.2	58.0	63.2	37.0
MultiReason	26.8	36.0	36.8	33.2	42.4	42.8	40.8	42.4	42.8	47.2	39.1
MonoReason	7.6	5.6	5.2	34.0	45.2	54.0	56.8	51.6	58.8	65.5	38.4
QAlign + MonoReason (Ours)	32.4	39.6	40.4	44.0	48.4	54.8	56.8	52.4	59.6	68.0	49.6
System (13B)	Bn	Th	Sw	Ja	Zh	De	Fr	Ru	Es	En	Avg.
SFT [†] (Touvron et al., 2023)	6.0	6.8	7.6	25.2	32.8	42.8	40.8	39.2	45.2	50.4	29.7
RFT [†] (Yuan et al., 2023)	3.2	4.4	3.6	26.4	33.6	38.4	44.8	41.6	46.8	52.0	29.5
MAmmoTH [†] (Yue et al., 2023)	3.6	5.2	1.6	19.2	31.2	45.6	39.6	36.8	50.0	56.4	28.9
WizardMath [†] (Luo et al., 2023)	6.4	5.6	5.6	22.0	28.0	40.4	42.0	34.4	45.6	52.8	28.3
MathOctopus [†] (Chen et al., 2023)	35.2	46.8	42.8	43.2	48.8	44.4	48.4	47.6	48.0	53.2	45.8
MetaMath (Yu et al., 2023)	11.6	6.4	7.6	42.8	49.2	64.8	65.2	63.6	65.2	67.2	44.4
MultiReason	37.6	42.2	44.0	43.2	53.6	47.6	54.0	48.0	54.8	56.4	48.1
MonoReason	12.4	11.2	6.4	42.0	46.0	64.0	62.4	61.6	64.8	68.4	43.9
QAlign+ MonoReason (Ours)	38.4	49.6	46.0	52.4	59.2	62.0	62.4	64.4	67.2	69.2	57.1

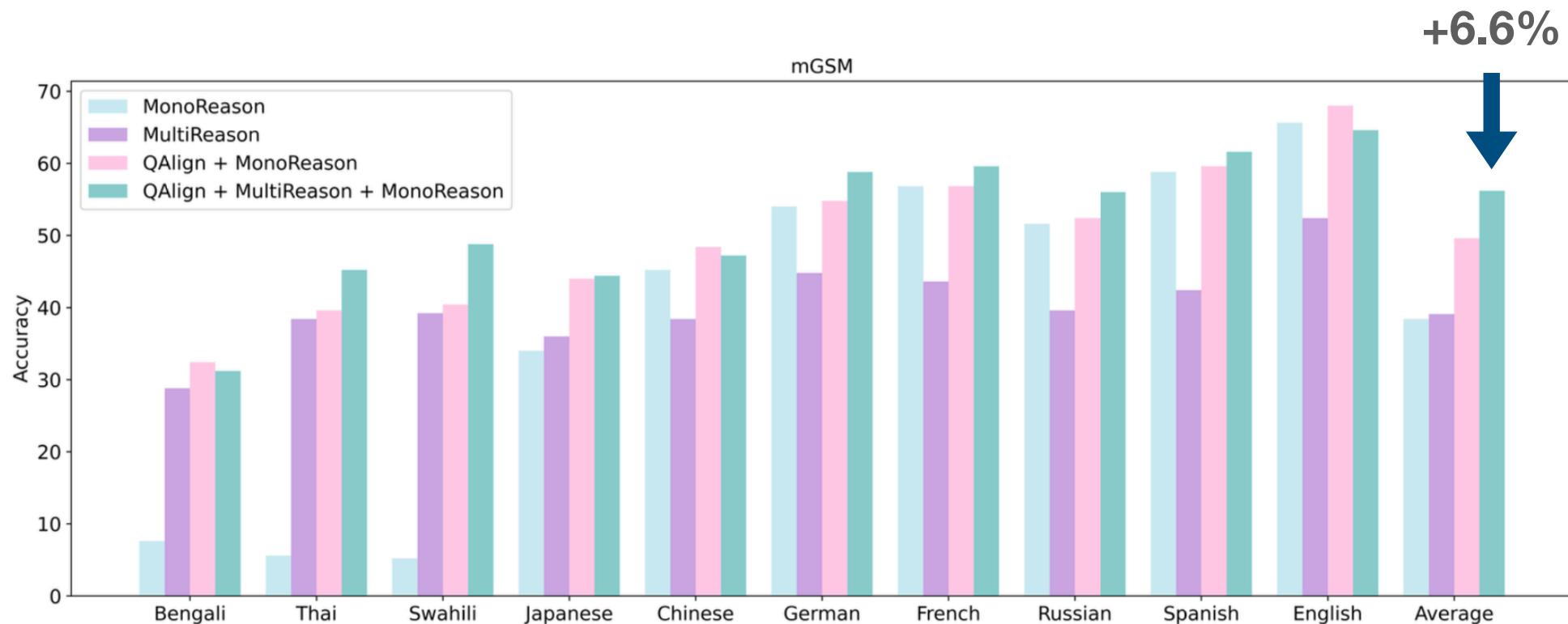
主要结果

◎ 在域外测试集上，我们的模型也展现出了更好的鲁棒性。

System (7B)	Bn	Th	Sw	Ja	Zh	De	Fr	Ru	Es	En	Avg.
SFT [†] (Touvron et al., 2023)	11.5	18.2	17.2	31.6	35.2	39.0	39.1	39.1	39.2	38.8	30.9
RFT [†] (Yuan et al., 2023)	7.7	16.9	14.9	33.9	34.9	40.8	41.5	39.5	42.5	42.7	31.3
MAmmoTH [†] (Yue et al., 2023)	4.3	6.3	4.2	26.7	26.8	39.6	39.9	33.7	42.9	45.1	26.3
WizardMath [†] (Luo et al., 2023)	16.1	17.0	10.3	37.9	36.3	39.2	37.7	37.4	44.8	48.5	32.5
MathOctopus [†] (Chen et al., 2023)	31.8	39.3	43.4	41.1	42.6	48.4	50.6	46.9	49.4	50.7	44.1
MetaMath (Yu et al., 2023)	14.2	17.8	16.5	53.2	53.1	61.4	60.7	58.9	61.2	65.5	46.3
MultiReason	27.6	36.5	42.4	40.9	43.2	44.3	46.7	42.3	45.5	48.0	41.3
MonoReason	15.0	17.1	15.4	51.9	54.4	60.9	62.2	59.3	63.3	65.5	46.2
QAlign + MonoReason (Ours)	41.7	47.7	54.8	58.0	55.7	62.8	63.2	61.1	63.3	65.3	57.2
System (13B)	Bn	Th	Sw	Ja	Zh	De	Fr	Ru	Es	En	Avg.
SFT [†] (Touvron et al., 2023)	13.9	23.4	19.8	41.8	43.3	46.2	47.8	47.8	46.1	50.9	38.1
RFT [†] (Yuan et al., 2023)	12.2	24.8	19.4	42.4	42.3	45.1	45.2	46.5	45.6	47.1	37.1
MAmmoTH [†] (Yue et al., 2023)	5.0	13.7	12.9	42.2	47.7	52.3	53.8	50.7	53.9	53.4	38.6
WizardMath [†] (Luo et al., 2023)	13.7	16.3	12.5	29.5	37.0	48.7	49.4	43.8	49.4	56.3	35.7
MathOctopus [†] (Chen et al., 2023)	35.2	41.2	46.8	39.2	52.0	47.2	48.0	45.6	53.2	56.4	46.5
MetaMath (Yu et al., 2023)	14.6	15.7	17.4	57.0	56.6	67.3	64.7	63.7	65.9	67.7	49.1
MultiReason	35.0	41.3	44.6	49.9	48.1	53.3	53.2	51.6	52.5	54.5	48.4
MonoReason	20.6	20.5	19.1	57.0	58.8	68.4	68.1	67.5	68.9	68.9	51.8
QAlign + MonoReason (Ours)	49.2	55.5	55.2	64.3	63.8	69.5	68.1	66.4	66.4	67.6	62.6

主要结果

- ◎ 在我们的框架中融入多语言有监督数据还可以进一步提升模型的多语言推理表现



主要结果

- 我们还对语言对齐阶段的实现进行了充分的探索，发现翻译任务的数据内容、数据领域和翻译方向都对语言对齐效果有着巨大影响。
- 虽然在不同的实现下，模型的英语能力十分近似，但是只有使用X-En的问题翻译任务才能将模型的英语推理能力迁移到非英语。

Data	Direction	MGSM		MSVAMP	
		Non-En	En	Non-En	En
<i>Question</i>	X→En	47.6	68.0	56.5	65.3
<i>Question</i>	En→X	36.2	68.0	48.3	64.4
<i>Response</i>	X→En	46.4	67.2	52.1	64.9
<i>Response</i>	En→X	42.8	68.0	49.0	63.9
<i>Flores-101</i>	X→En	36.3	68.0	46.8	65.4

Question (English)	Gold Answer
There are 27 unicorns left in the world. One third of them are in the Scottish Highlands. Two thirds of the Scottish unicorns are female. How many female Scottish unicorns are there?	6
Chain-of-Thought Response	Prediction
<p>[MonoReason]</p> <p>One third of the unicorns are in the Scottish Highlands, so there are $27/3 = 9$ unicorns in the Highlands.</p> <p>Two thirds of the Scottish unicorns are female, so there are $9 * 2/3 = 6$ female Scottish unicorns.</p> <p>The answer is: 6</p>	6 ✓
<p>[QAlign + MonoReason]</p> <p>One third of the unicorns are in the Scottish Highlands, so there are $27/3 = 9$ unicorns in the Highlands.</p> <p>Two thirds of the unicorns in the Highlands are female, so there are $9 * 2/3 = 6$ female unicorns in the Highlands.</p> <p>Therefore, there are 6 female Scottish unicorns.</p> <p>The answer is: 6</p>	6 ✓
Question (Chinese)	Gold Answer
<p>世界上剩下27只独角兽，其中三分之一在苏格兰高地。</p> <p>苏格兰高地的独角兽的三分之二是雌性。苏格兰雌性独角兽有多少只？</p>	6
Chain-of-Thought Response	Prediction
<p>[MonoReason]</p> <p>If three-fourths of the remaining 27 one-horned animals are in the Scottish Highlands, then there are $27 * 3/4 = 21$ one-horned animals in the Scottish Highlands.</p> <p>If two-thirds of the one-horned animals in the Scottish Highlands are female, then there are $21 * 2/3 = 14$ female one-horned animals in the Scottish Highlands.</p> <p>Therefore, there are 14 female one-horned animals in the Scottish Highlands.</p> <p>The answer is: 14</p>	14 X
<p>[QAlign + MonoReason]</p> <p>If there are 27 remaining unicorns in the world, and one-third of them are in the Scottish Highlands, then there are $27/3 = 9$ unicorns in the Scottish Highlands.</p> <p>If two-thirds of the unicorns in the Scottish Highlands are female, then there are $9 * 2/3 = 6$ female unicorns in the Scottish Highlands.</p> <p>Therefore, there are 6 female unicorns in the Scottish Highlands.</p> <p>The answer is: 6</p>	6 ✓

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://github.com/NJUNLP/MMT-LLM>

- 通过问题翻译学习提升大模型多语言推理水平

- <https://github.com/NJUNLP/QAlign>

- 通过语言对齐提升大语言模型的非英语能力

- <https://github.com/NJUNLP/x-LLM>

相关代码、数据均已开源
欢迎大家参考和讨论！

谢谢！