跨语言对齐增强大模型—百聆

张绍磊

中国科学院 计算技术研究所

2024-01-04

中国科学院计算技术研究所
Institute of Computing Technology,Chinese Academy of Sciences

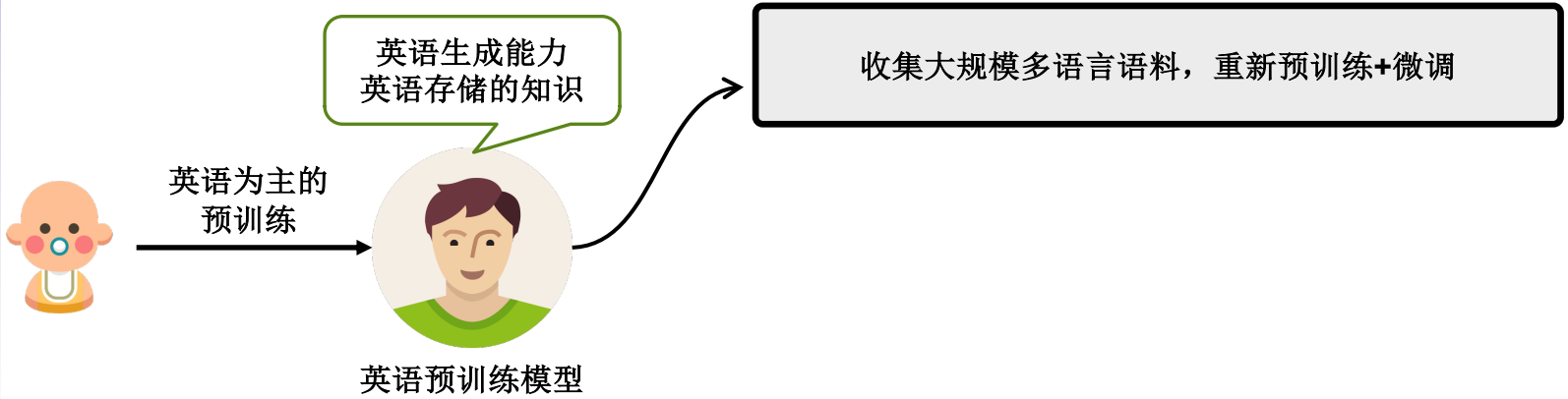
大部分LLM在英语语料上预训练

- 存在问题：LLaMA以英语为主，其他语言能力不强
- 英语为主的大模型 → 多语言大模型

大部分LLM在英语语料上预训练

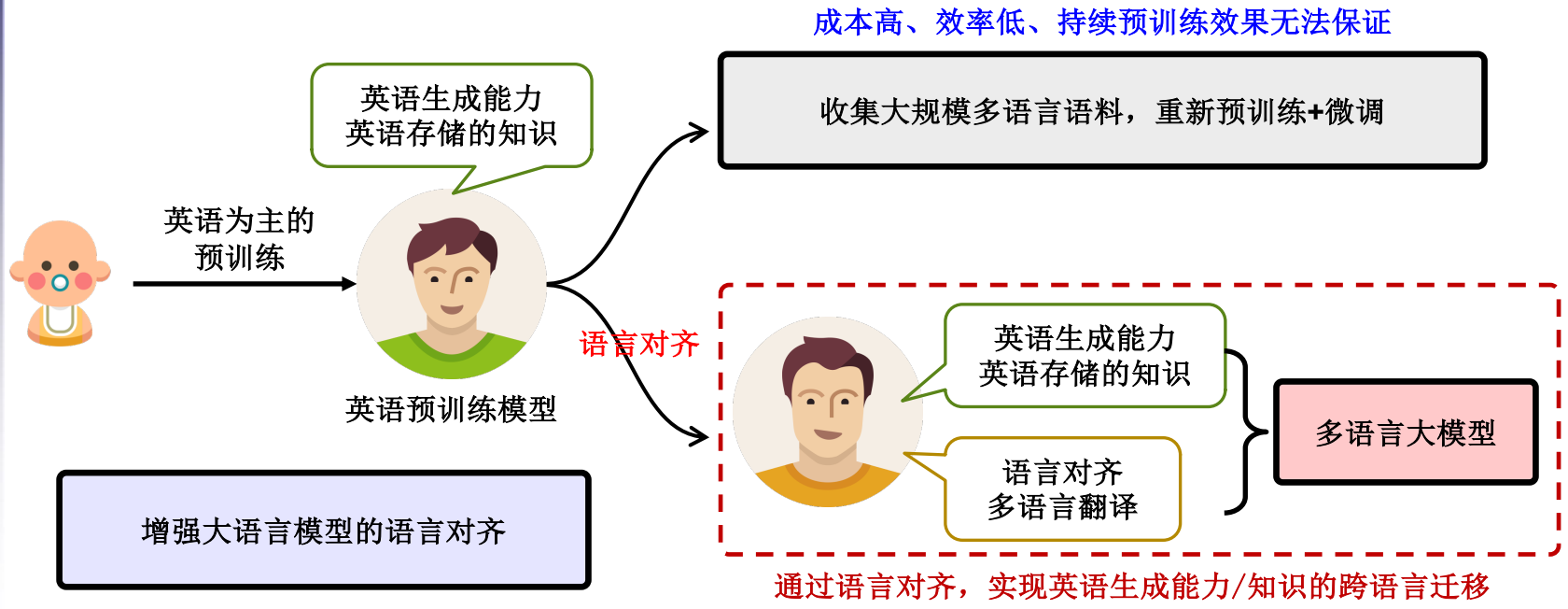
- 存在问题：LLaMA以英语为主，其他语言能力不强
- 英语为主的大模型 → 多语言大模型

成本高、效率低、持续预训练效果无法保证



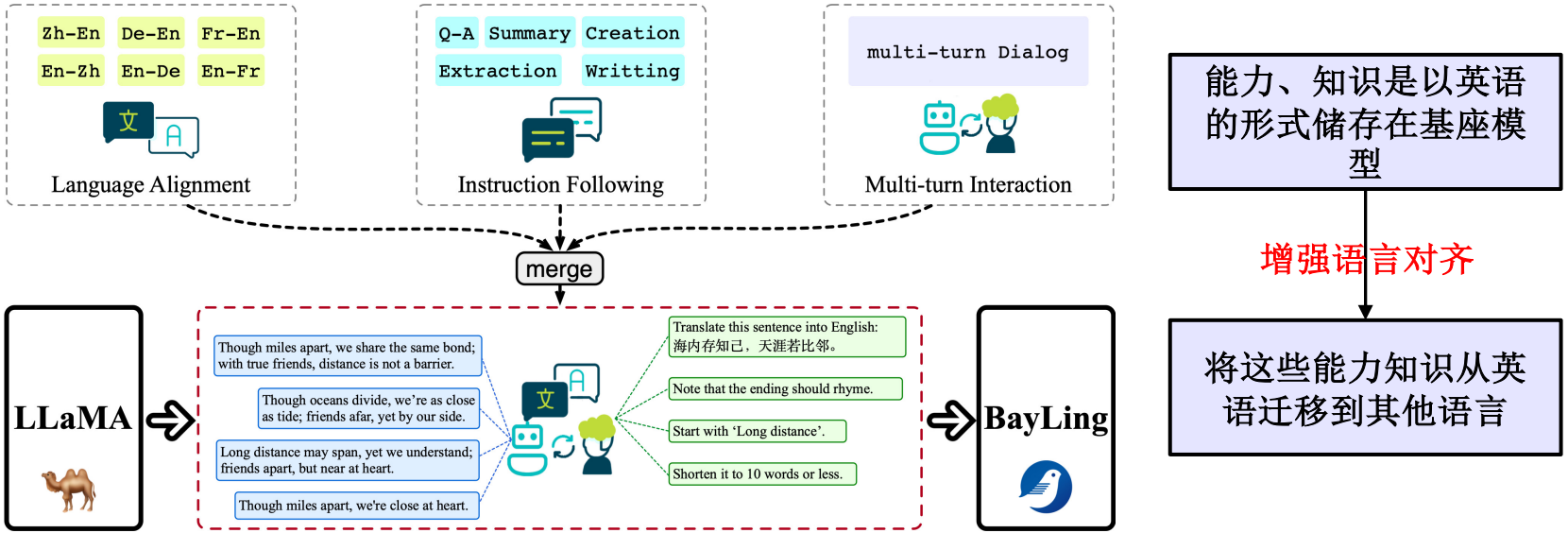
跨语言对齐 → LLM能力跨语言迁移

- 存在问题：LLaMA以英语为主，其他语言能力不强
- 英语为主的大模型 → 多语言大模型



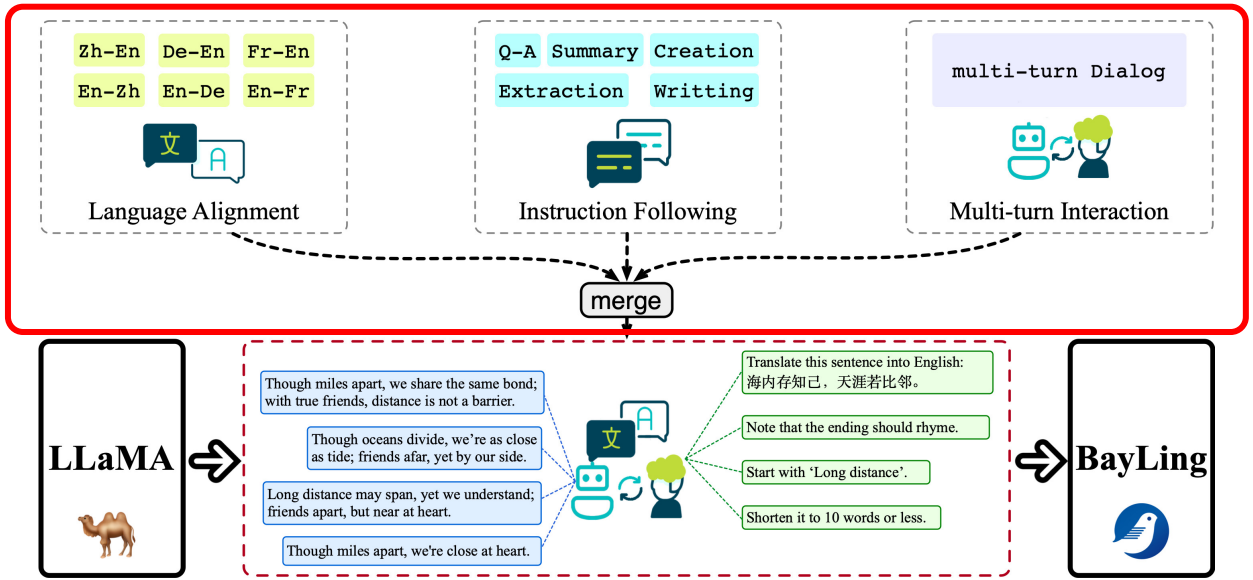
百聆大模型

- 通过交互式机器翻译来避免数据标注，同时提升语言生成和与人类对齐能力
  - 语言对齐：完成语言生成能力和与指令遵循能力从英语到其他语言的迁移。



百聆大模型

- 通过交互式机器翻译来避免数据标注，同时提升语言生成和与人类对齐能力
  - 语言对齐：完成语言生成能力和与指令遵循能力从英语到其他语言的迁移。
  - 复合任务：同时提升多语言、指令理解、多轮交互等多方面能力。



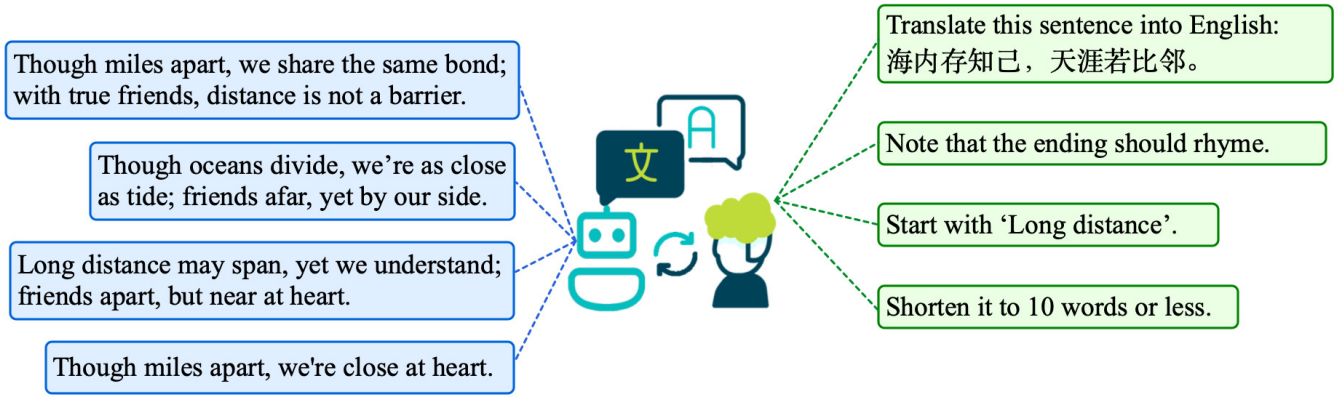
能力、知识是以英语的形式储存在基座模型

增强语言对齐

将这些能力知识从英语迁移到其他语言

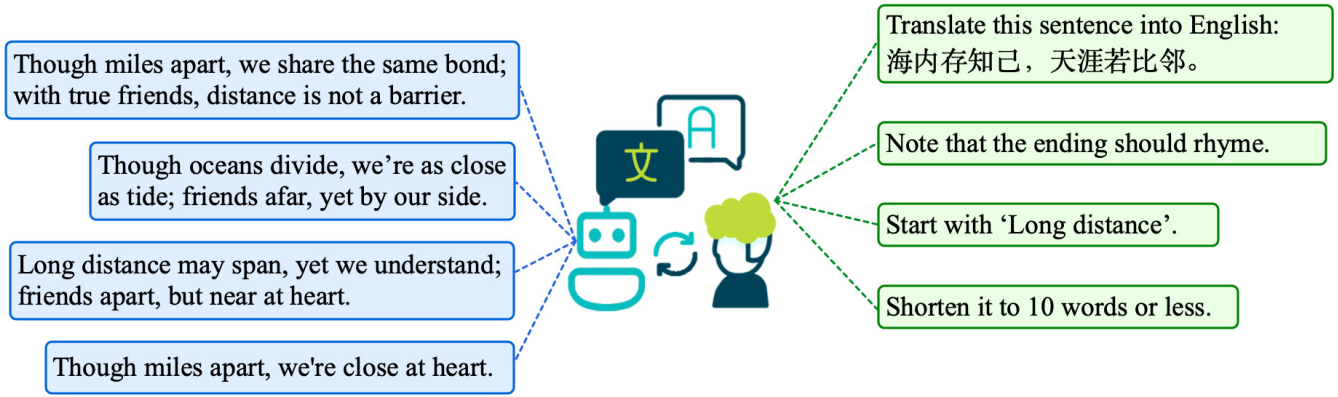
百聆大模型

- 通过交互式机器翻译来避免数据标注，同时提升语言生成和与人类对齐能力
  - 语言对齐：完成语言生成能力和与指令遵循能力从英语到其他语言的迁移。
  - 复合任务：同时提升多语言、指令理解、多轮交互等多方面能力。
  - 以人类为中心：交互式修改译文，增强了与人类意图对齐的能力。



百聆大模型

- 通过交互式机器翻译来避免数据标注，同时提升语言生成和与人类对齐能力
  - 语言对齐：完成语言生成能力和与指令遵循能力从英语到其他语言的迁移。
  - 复合任务：同时提升多语言、指令理解、多轮交互等多方面能力。
  - 以人类为中心：交互式修改译文，增强了与人类意图对齐的能力。
  - 数据质量高：**大部分为新闻语料**，质量高，几乎无毒性、偏见。



百聆大模型

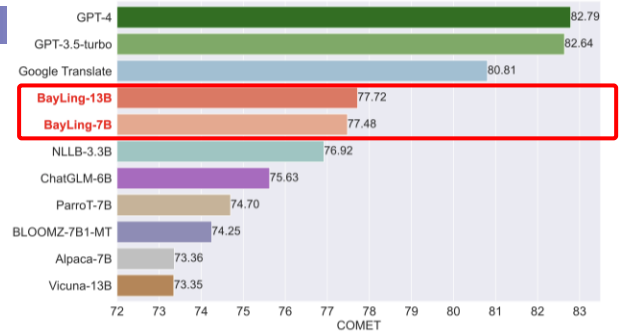
- 通过交互式机器翻译来避免数据标注，同时提升语言生成和与人类对齐能力
  - 语言对齐：完成语言生成能力和与指令遵循能力从英语到其他语言的迁移。
  - 复合任务：同时提升多语言、指令理解、多轮交互等多方面能力。
  - 以人类为中心：交互式修改译文，增强了与人类意图对齐的能力。
  - 数据质量高：**大部分为新闻语料**，质量高，几乎无毒性、偏见。

Source	Interactive	Languages		#Instances
Alpaca	Single-turn	English		52K
ShareGPT	Multi-turn	English-dominant		90K
		Instruction Languages	Translation Languages	
Interactive Translation	Multi-turn	English, Chinese	English, Chinese German, French	160K

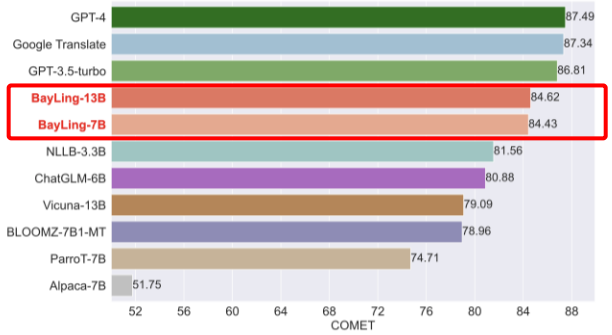
中英翻译结果

- 百聆-13B相比GPT-4
  - 取得 95% 翻译性能

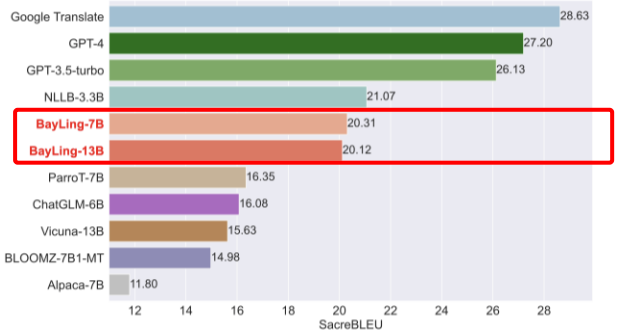
开源模型的最佳翻译性能



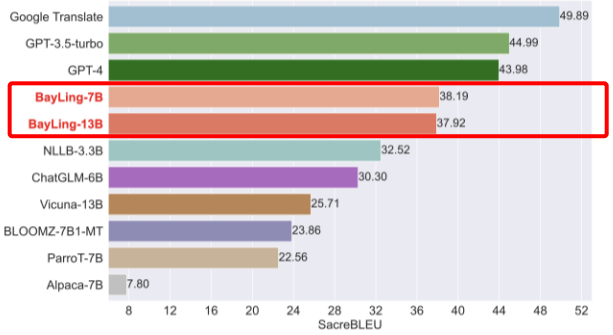
(a) COMET scores on Chinese-to-English translation



(b) COMET scores on English-to-Chinese translation



(c) BLEU scores on Chinese-to-English translation

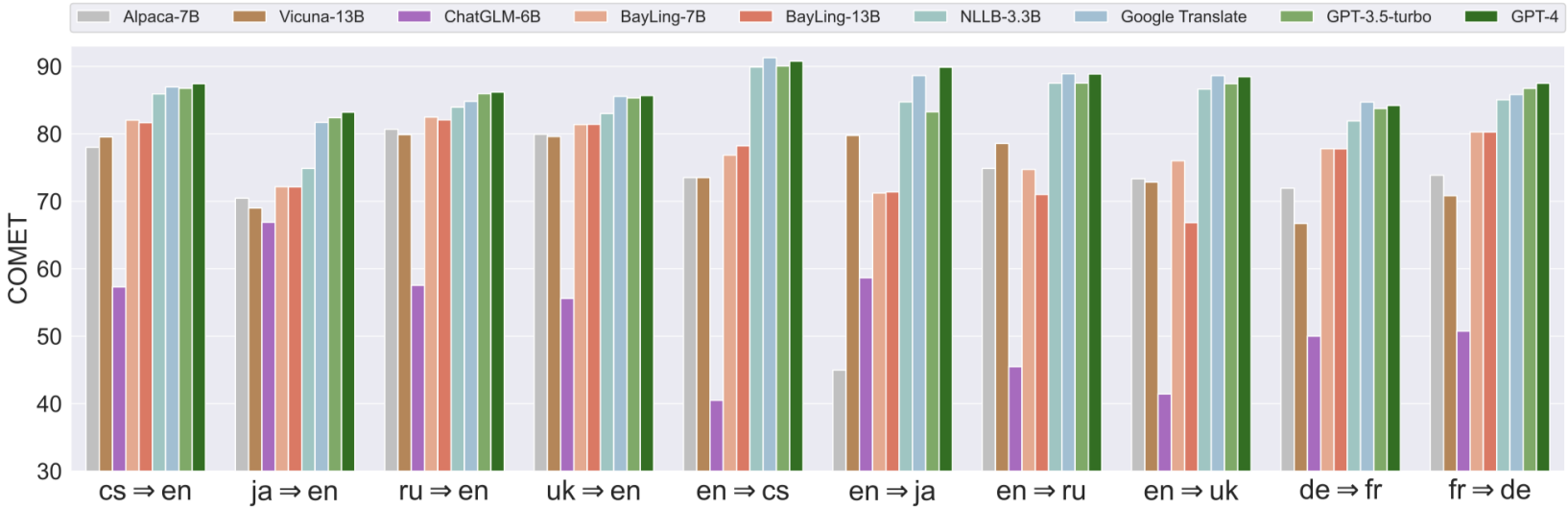


(d) BLEU scores on English-to-Chinese translation

图：WMT22 中英翻译测评

百聆Zero-shot翻译结果

- 在微调阶段未见过的语言上的翻译表现（零射翻译）

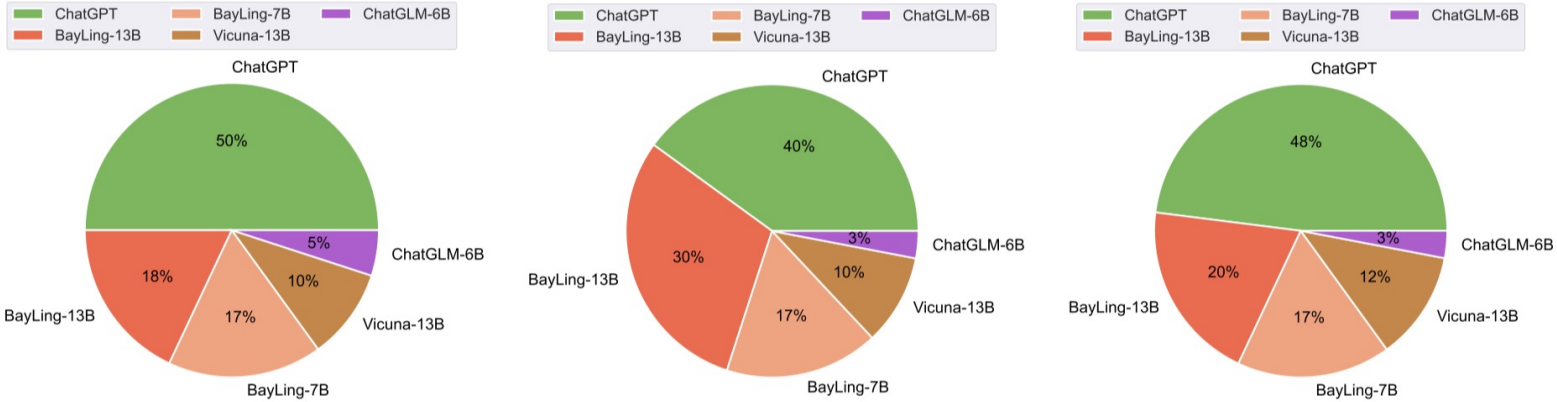


图： WMT22 多语言翻译（低资源语种）测评

利用**LLM**的泛化能力，将翻译能力从中英迁移到更多语种

交互式翻译人工评测胜率

- **英语专业标注员**同时和5个系统交互，选出表现最佳的一个
  - 百聆在翻译、指令遵循、多轮交互分别取得18%、30%、20%胜率
  - 人工评价仅落后于ChatGPT



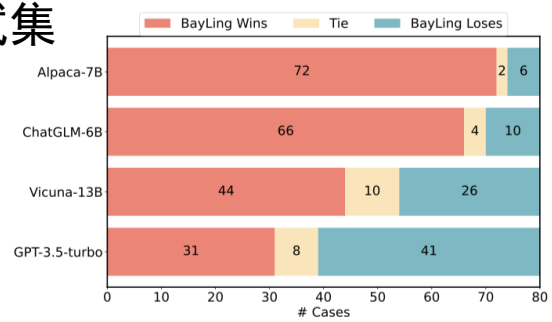
(a) Translation (b) Instruction following (c) Multi-turn interaction

图： 人工评价胜率

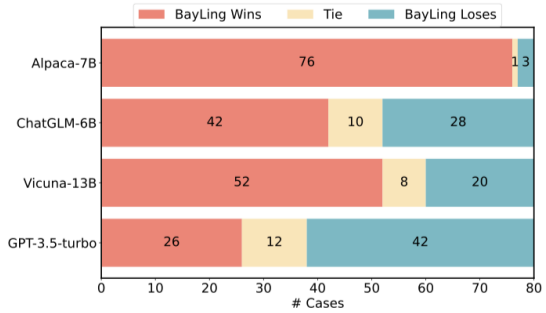
通用任务上的GPT-4评测结果

- 人工标注多轮中英指令测试集
  - Bayling-80
- 35% 优于 GPT-3.5-turbo
- 50% 不差于 GPT-3.5-turbo

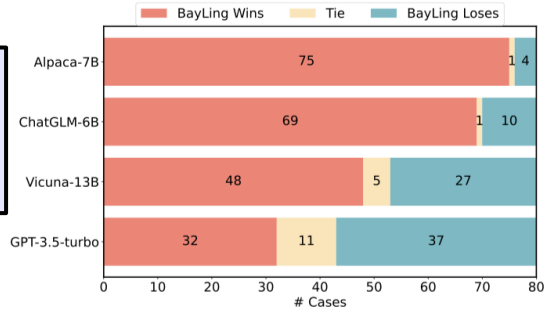
BayLing-13B表现出更强的
中文能力、多轮交互能力



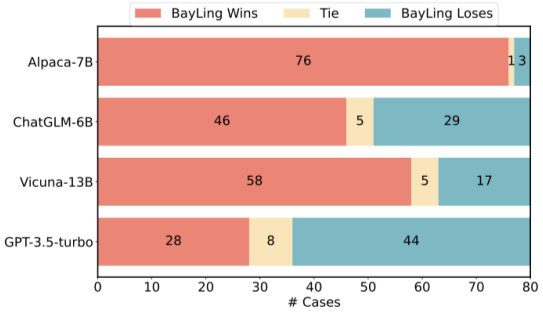
(a) Comparison on single-turn English instructions



(b) Comparison on single-turn Chinese instructions



(c) Comparison on multi-turn English instructions

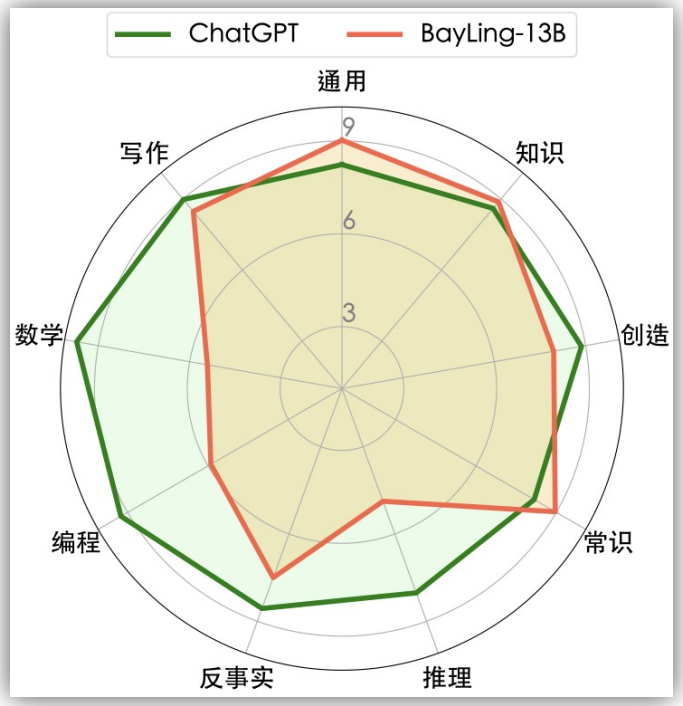


(d) Comparison on multi-turn Chinese instructions

图： 通用任务上GPT-4评价的胜/平/负

通用任务上与GPT-3.5-turbo能力比较

- 仅需**13B参数**，取得 GPT-3.5-turbo 89%的性能
  - **建议、写作、知识**上媲美 GPT-3.5-turbo
  - 数学、代码、推理弱于 GPT-3.5-turbo



表： 百聆相比**GPT-3.5-turbo**的得分

Instruction		GPT-3.5-turbo	BayLing-13B	Ratio
Single-turn	English	694.0	631.0	91%
	Chinese	687.0	592.0	86%
Multi-turn	English	700.5	643.0	92%
	Chinese	671.5	590.5	88%
Average		688.3	614.1	89%

AGIEval中文高考评测结果

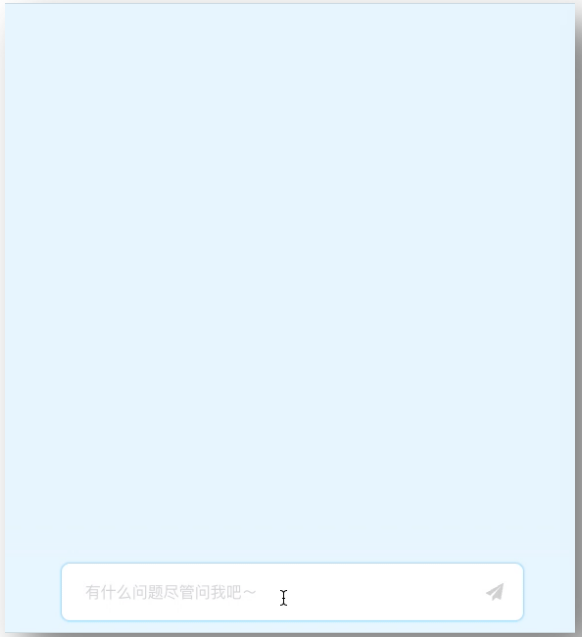
- 利用语言对齐将知识从英语迁移到中文
  - 基座模型中的知识以英语形式存储
  - 通过语言对齐将知识高效地迁移到中文，避免了用大量数据重新注入中文知识

大部分知识往往和语言无关 → 数学、物理、化学、生物提升更加明显

表： 百聆在高考测试集的得分

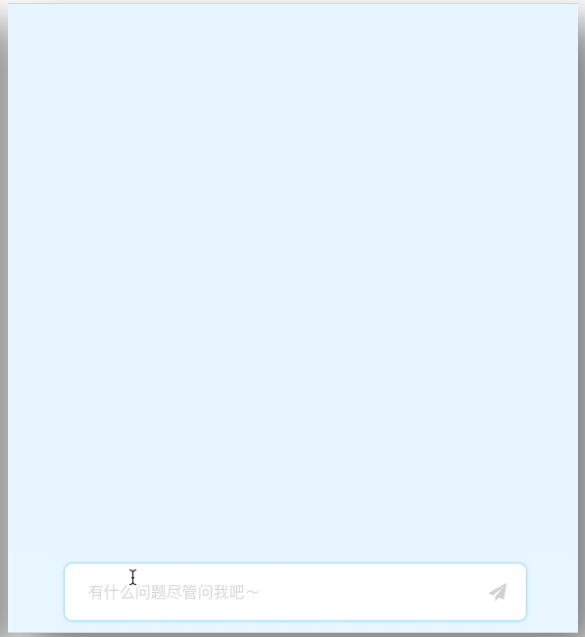
Systems	Avg.	GaoKao (%)								
		chinese	english	mathqa	physics	chemistry	biology	history	geography	mathcloze
GPT-3.5-turbo	43.87	42.68	86.27	30.48	21.00	44.44	46.19	59.57	63.32	0.85
BayLing -13B	32.13	29.27	69.28	29.34	21.50	36.71	30.00	34.04	38.19	0.85
BayLing-7B	28.20	27.64	55.56	26.78	24.50	29.95	29.05	33.19	27.14	0.00
ChatGLM-6B	31.83	31.71	52.29	26.50	16.00	27.54	28.10	54.04	47.74	2.54
Vicuna-13B	29.36	21.14	71.24	21.94	23.00	31.88	27.14	33.19	34.67	0.00
Alpaca-7B	20.03	24.80	36.27	17.95	6.00	20.77	20.95	24.68	27.14	1.69

诸多能力：丰富的知识



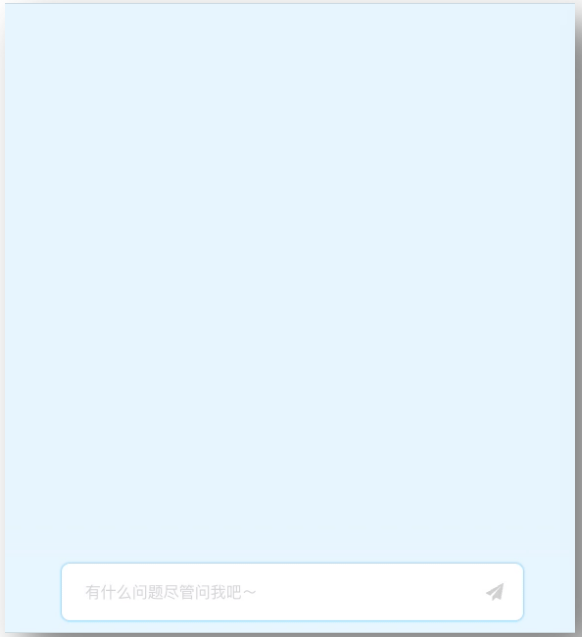
概念知识

专业知识



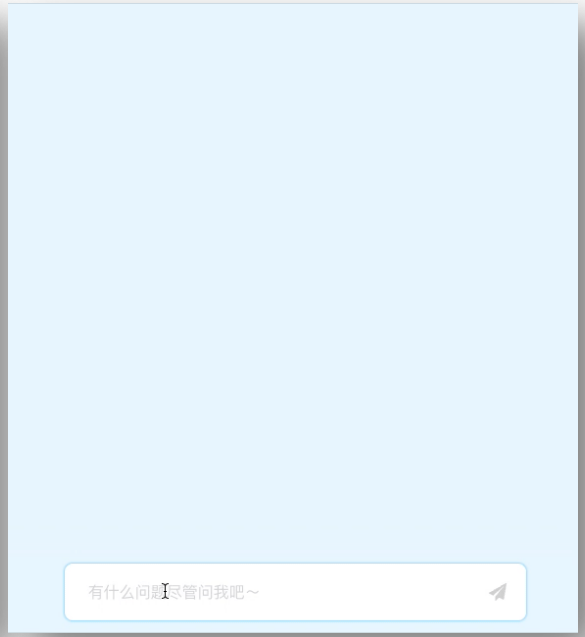
常识知识

诸多能力：多语言、跨语言能力



多语言翻译

跨语言任务



跨语言检索

百聆大模型

- **BayLing-13B**
  - 垂直领域（翻译任务）：取得开源的翻译大模型中的最佳性能
  - 通用领域：通用任务达到 GPT-3.5-turbo 89% 性能
    - 多语言、指令跟随、多轮对话

BayLing: Bridging Cross-lingual Alignment and Instruction Following through Interactive Translation for Large Language Models

Paper　　Demo　　Code & Model

中科院计算所 Institute of Computing Technology, Chinese Academy of Sciences

谢谢大家

张绍磊
中国科学院计算技术研究所
zhangshaolei20z@ict.ac.cn