新疆大学计算机科学与技术学院
(网络空间安全学院)

面向资源匮乏语言的大语言模型：从机器翻译到多模态翻译及未来探索

米尔阿迪力江·麦麦提

MNLP2026, 西宁, 青海
20260726

Outline

- 大语言模型时代的低资源语言处理
- 面向低资源语言的大语言模型机器翻译探索
- 多模态大语言模型驱动的低资源翻译探索
- 面向低资源语言大模型的未来探索
- 总结

大语言模型时代的低资源语言处理

The Three Paths to Achieving AI

实现人工智能的三条路径

LLM = 通过数据学习 + 受人脑启发 + 输入经验知识

1 从数据中学习

通过数据驱动，ML方法模仿人类智能

- 利用大量数据训练模型
- 发现数据中的模式和规律
- 通过机器学习/深度学习方法逼近人类智能行为

2 受人脑启发

解明人脑机制，基于相同原理实现人类智能

- 研究人脑的结构与工作机制
- 借鉴神经元、突触等生物原理
- 构建类脑模型，实现类人智能

3 输入经验知识

将知识通过规则等方式教给计算机，进行符号处理

- 将人类经验知识形式化
- 通过规则、逻辑、知识图谱等方式输入计算机
- 基于符号推理和逻辑处理实现智能

LLM

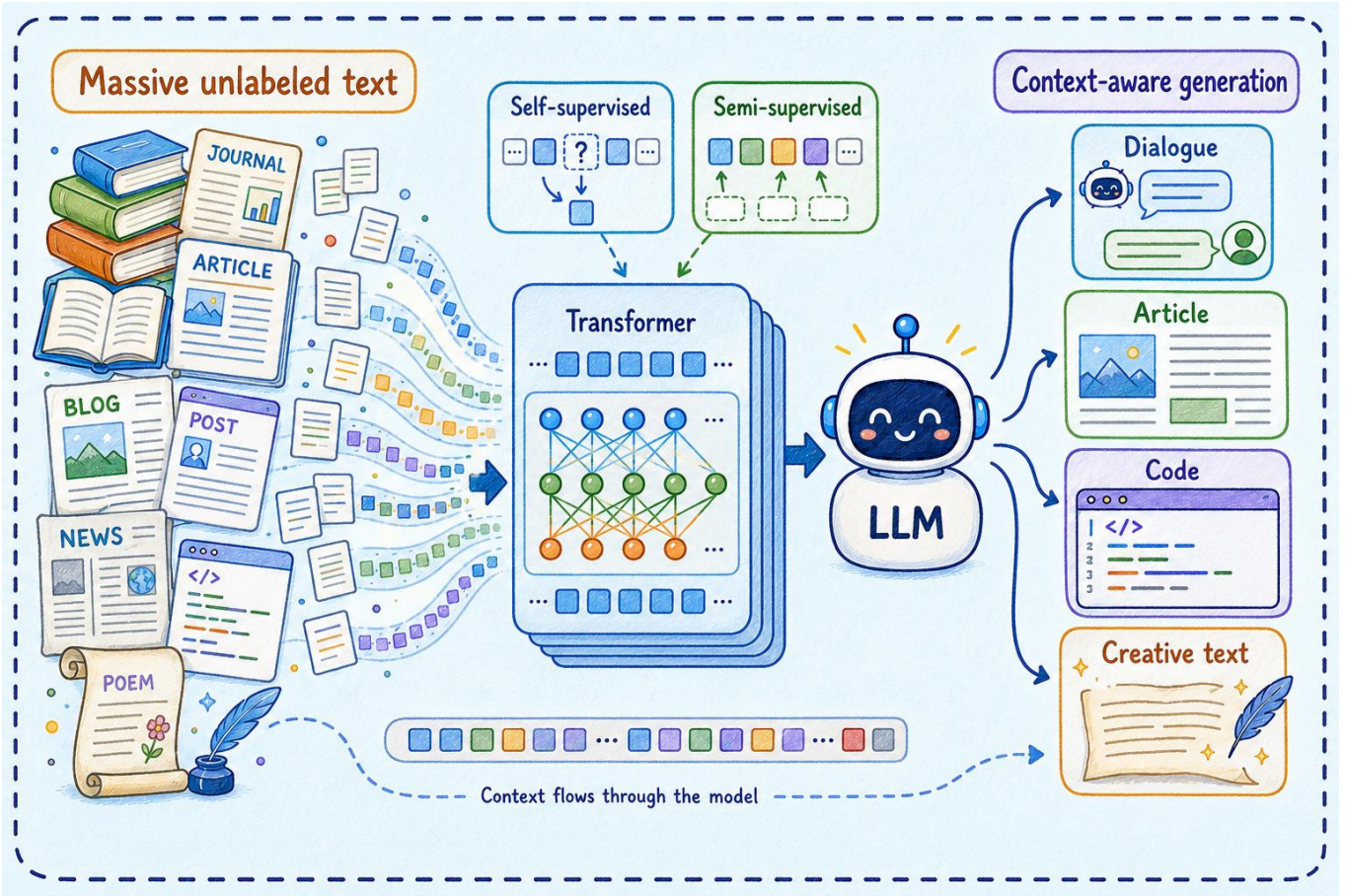
融合三条路径的优势，迈向通用人工智能

- 从数据中学习：提供泛化能力
- 受人脑启发：提供高效结构与机制
- 输入经验知识：提供可靠知识与可控性

Actually, the LLM is ….

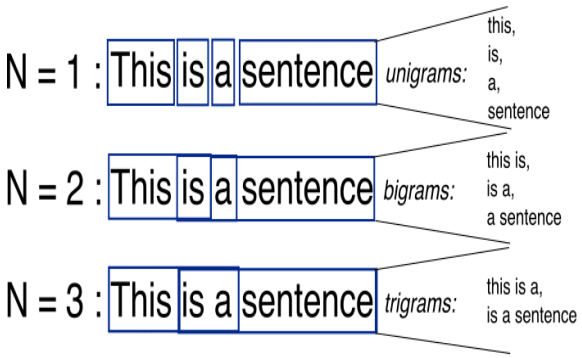
大规模语言模型（LLM）是基于**Transformer** 架构的模型，它们通过**自监督**或**半监督学习**在大量未标记的文本上进行广泛训练。这些模型旨在生成听起来自然且与上下文相关的文本，适用于各种风格和格式。

Large-scale Pre-trained Language Model

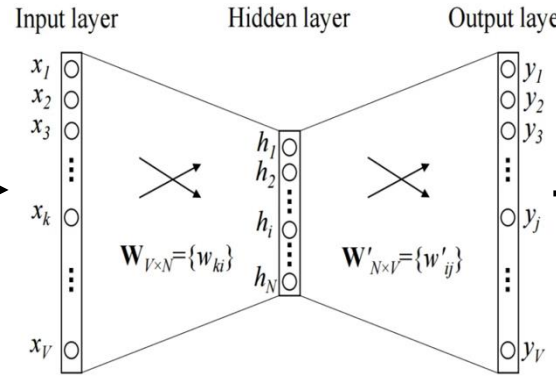


N-gram LM ➡ LLM

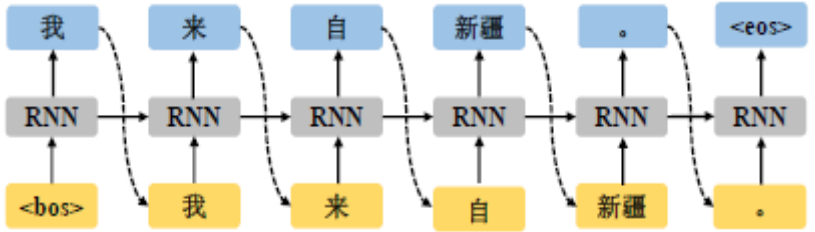
Language Models



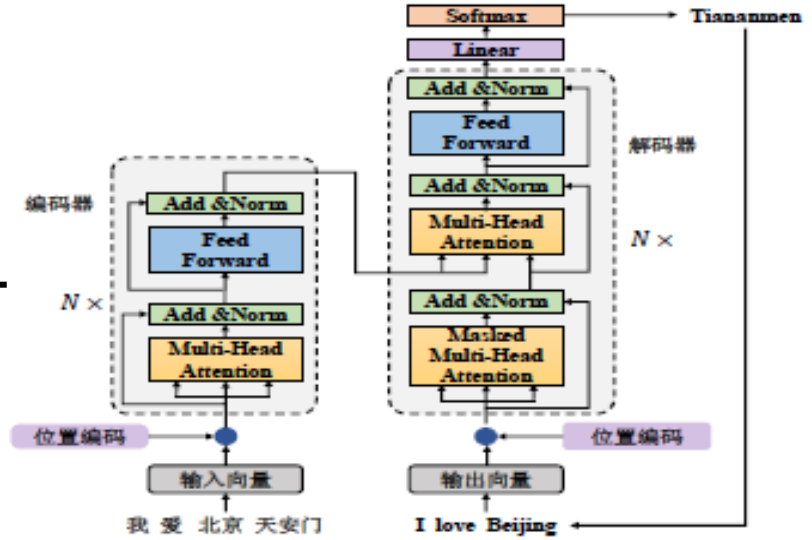
N-gram LM



Word2Vec



RNNLM

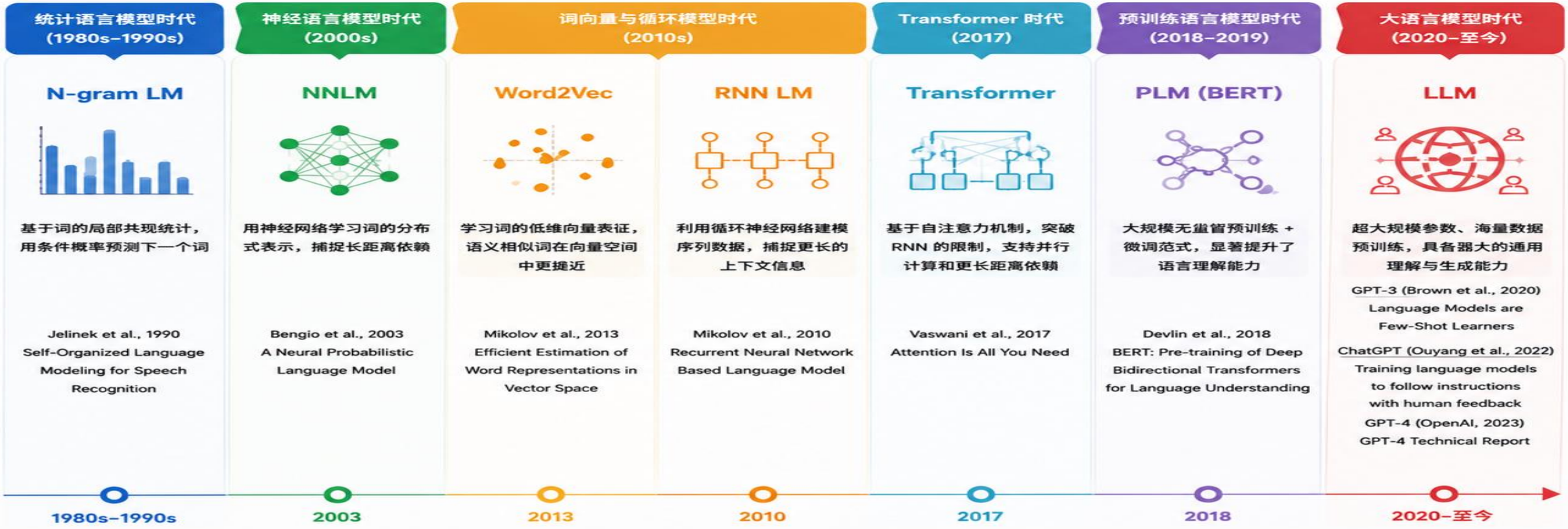


Transformer

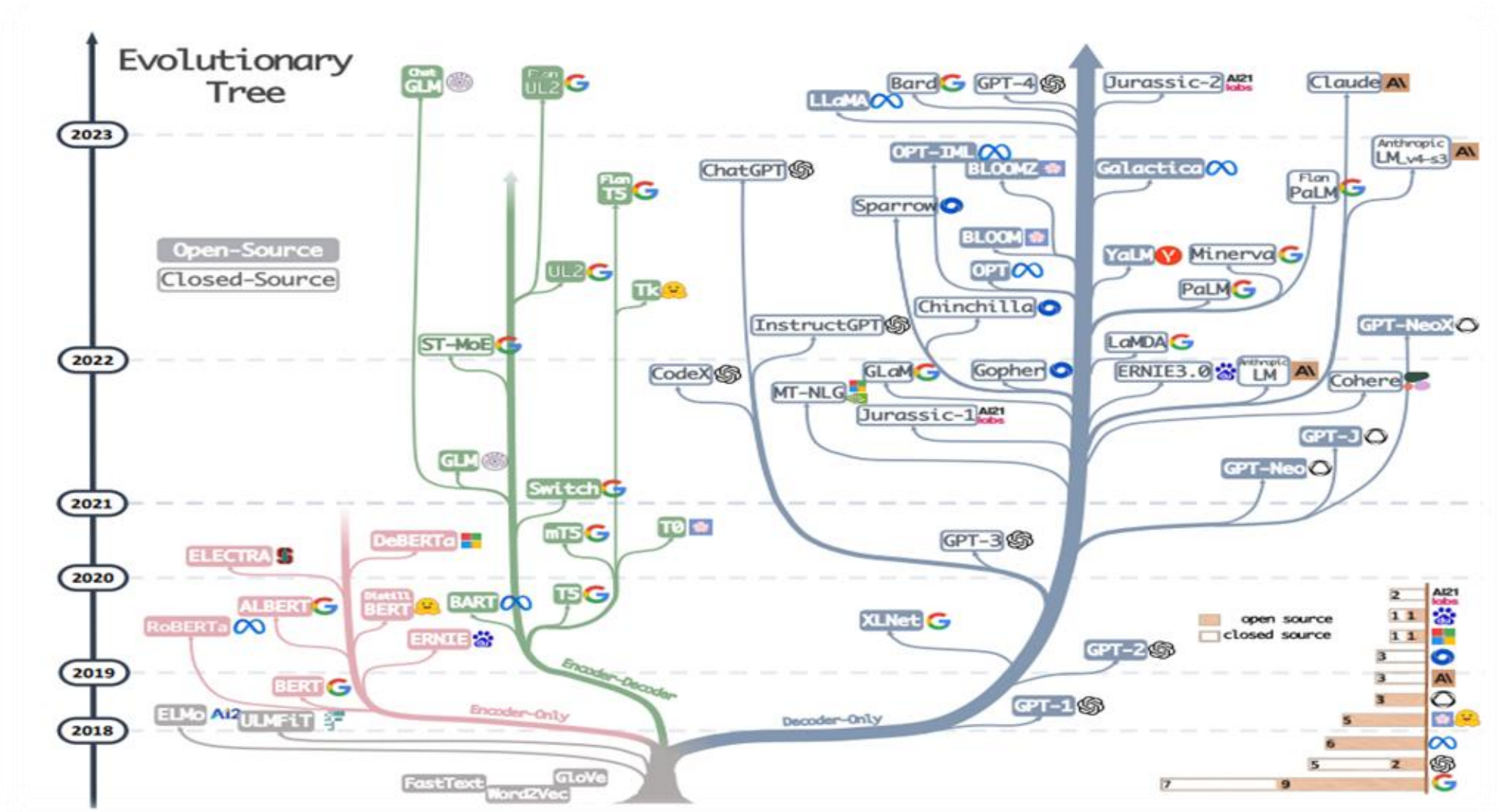
PLM

LLM

N-gram LM ➡ LLM


统计语言模型时代 (1980s–1990s)
神经语言模型时代 (2000s)
词向量与循环模型时代 (2010s)
Transformer 时代 (2017)
预训练语言模型时代 (2018–2019)
大语言模型时代 (2020–至今)
N-gram LM
基于词的局部共现统计，用条件概率预测下一个词
Jelinek et al., 1990
Self-Organized Language Modeling for Speech Recognition
NNLM
用神经网络学习词的分布式表示，捕捉长距离依赖
Bengio et al., 2003
A Neural Probabilistic Language Model
Word2Vec
学习词的低维向量表征，语义相似词在向量空间中更接近
Mikolov et al., 2013
Efficient Estimation of Word Representations in Vector Space
RNN LM
利用循环神经网络建模序列数据，捕捉更长的上下文信息
Mikolov et al., 2010
Recurrent Neural Network Based Language Model
Transformer
基于自注意力机制，突破 RNN 的限制，支持并行计算和更长距离依赖
Vaswani et al., 2017
Attention Is All You Need
PLM (BERT)
大规模无监督预训练 + 微调范式，显著提升了语言理解能力
Devlin et al., 2018
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
LLM
超大规模参数、海量数据预训练，具备强大的通用理解与生成能力
GPT-3 (Brown et al., 2020)
Language Models are Few-Shot Learners
ChatGPT (Ouyang et al., 2022)
Training language models to follow instructions with human feedback
GPT-4 (OpenAI, 2023)
GPT-4 Technical Report
1980s–1990s
2003
2013
2010
2017
2018
2020–至今
局部统计
只能建模短距离依赖
分布式表示
开始捕捉语义信息
序列建模
处理更长上下文
注意力机制
并行计算，长距离依赖
理解能力
深层语义理解
通用智能
生成、推理、对话等多种能力

The Evolutionary of LLM

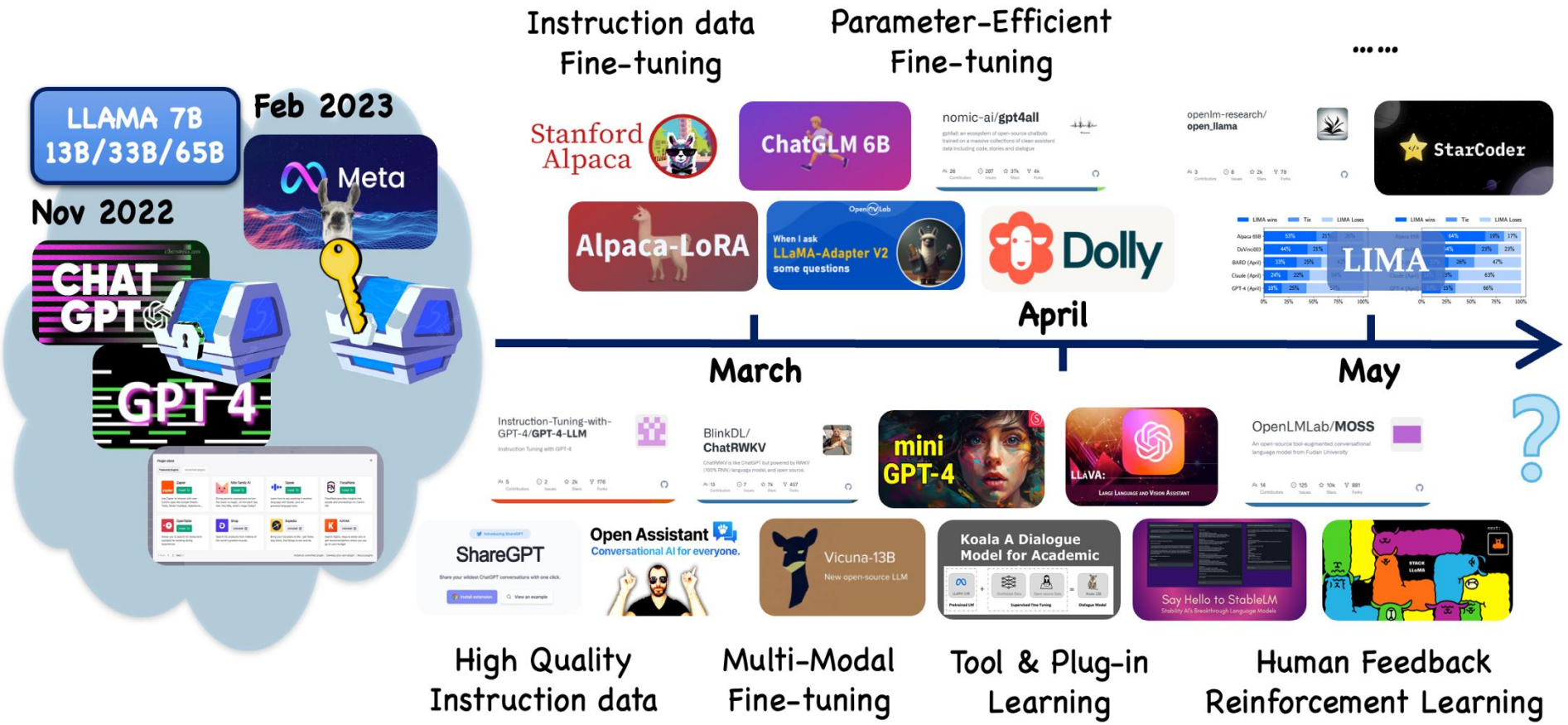


(Yang et al., 2023)

The Evolutionary of LLM

Basic Fact - foundational models are rapidly evolving

A growing number of instruction-tuned models have emerged



(Ding et al., CCL2023)

The 4 Core Capabilities of LLMs

1 语言理解
Language Understanding

理解低资源语言的语义、意图和上下文含义

乌尔都语 (Urdu)

کل کے اجلاس میں شرکت کرنے والے تمام افراد کا شکریہ.

براہ کرم آئندہ ہفتے کی تیاری کیجیے کریں.

▶ **理解**：感谢参会者，并提醒下周做好准备。

乌兹别克语 (Uzbek)

Ertangi yig'ilishda ishtirok etgan barcha odamlarga rahmat. Iltimos, kelgusi hafta uchun tayyorgarlik ko'ring.

▶ **理解**：感谢参会者，并提醒下周做好准备。

2 指令遵循
Instruction Following

理解并严格遵循用户指令，完成指定任务

用户指令（乌尔都语）

مندرجہ ذیل جملے کو انگریزی میں ترجمہ کریں

میں آج بازار گیا کیونکہ مجھے کچھ سامان خریدنا تھا۔

模型输出 (English)

I went to the market today because I had to buy some groceries.

用户指令（乌兹别克语）

Quyidagi jumlani xitoy tiliga tarjima qiling:
Men bugun bozorga bordim, chunki menga bir oz oziq-ovqat kerak edi.

模型输出（中文）

我今天去了市场，因为我需要买一些食品。

The 4 Core Capabilities of LLMs

3 上下文学习
In-context Laarrning

通过少量示例学习新任务，具备强大的泛化能力

示例（乌尔都语 → 英语 翻译）

مثال 1: a: میں گھر جا رہا ہوں۔ → I am going home.

مثال 2: a: وہ وہ کتاب پڑھ رہی ہے۔ → She is reading a book.

مثال 3. a: میں کل سفر کرنا ہے۔ → We have to travel tomorrow.

.....

新句子（乌尔都语） → وہ اپنے دوست سے ملنے پارک گیا۔

模型输出（English） → He went to the park to meet his friend.

示例（乌兹别克语 → 中文 翻译）

1. Men ovqat yedim. → 我吃了饭。
2. U maktabga bordi. → 他去上学了。
3. Biz film ko'rdik. → 我们看了电影。

.....

新句子（乌兹别克语） → U do'kon bilan suhbatlashdi.

模型输出（中文） → 他和店主聊了天。

4 推理能力
Reasoning

进行逻辑推理、因果分析和复杂问题解决

问题（乌尔都语）

اگر ایک شخص کے پاس 3 سیب ہیں اور وہ اپنے دوست کو 2 سیب دے دیتا۔ ہے تو اس کے پاس کتنے سیب باقی رہیں گے؟ براہ کرم قدم قدم بہ قدم حل کریں۔

模型推理过程 (Urdu)

1. اس شخص کے پاس ابتدا میں 3 سیب ہیں
2. اس نے اپنے دوست کو 2 سیب دیئے۔ 2.
3. سیب بچے = 3 - 2 = 1

جواب اس کے پاس 1 سیب باقی رہیں گے۔

问题（乌兹别克语）

Agar bir kitobning narxi 15,000 so'm bo'lsa va men 3 ta kitob olsam, 20,000 so'm pul bilan yetarli bo'ladimi? Qadam-baqadam yeching.

模型推理过程 (Uzbek)

1. Bitta kitob narxi = 15,000 so'm
2. 3 ta kitob narxi = 15,000 × 3 = 45,000 so'm
3. Mening pulim = 20,000 so'm
4. 20,000 < 45,000, shuning uchun yetarli emas.

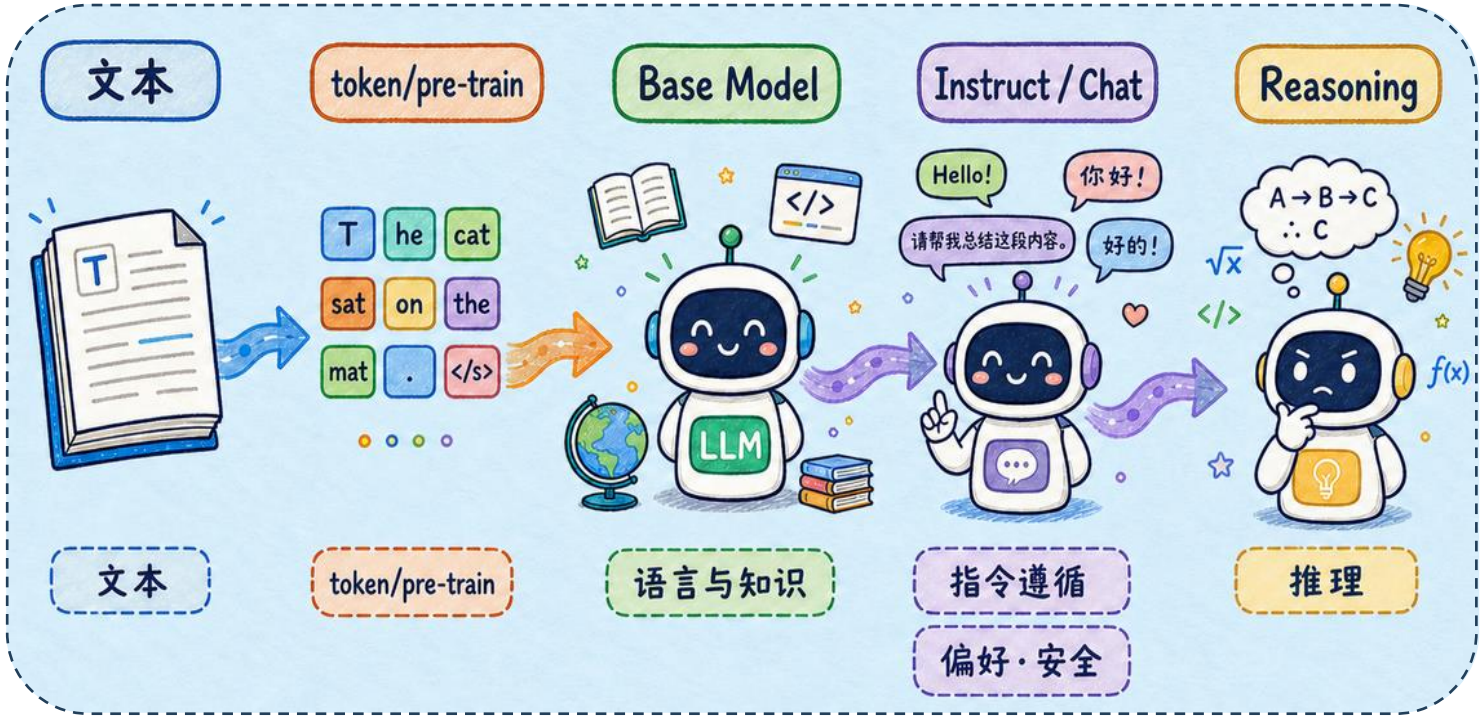
Javob: Yo'q, yetarli bo'lmaydi.

LLM Capabilities Across Different Training Stages

每个阶段解决不同问题

文本先被编码为Token；预训练学习语言与知识；后训练再塑造指令遵循、偏好、推理和安全行为

- 预训练回答“模型会不会”；
- 后训练回答“模型能否按要求做好”；
- 评测回答“模型是否达到使用门槛”；



主线：Tokenizer决定模型看到什么，预训练决定模型知道什么，后训练决定模型如何回答

Tokenizer: Converting Natural Language into Discrete Tokens

词表并不是人工列出的全部单词，而是从代表性语料中学习得到的一组字符、子词与特殊Token

准备语料 → 学习词表 → 冻结使用

01 规范化与预切分

统一编码和空白，再按语言规则得到可继续合并的初始片段。

02 BPE或Unigram训练

反复合并高频片段或选择最优子词集合，形成固定词表与**ID映射**。

03 编解码验证

检查压缩率、多语种与代码覆盖；预训练和推理必须使用同一Tokenizer。

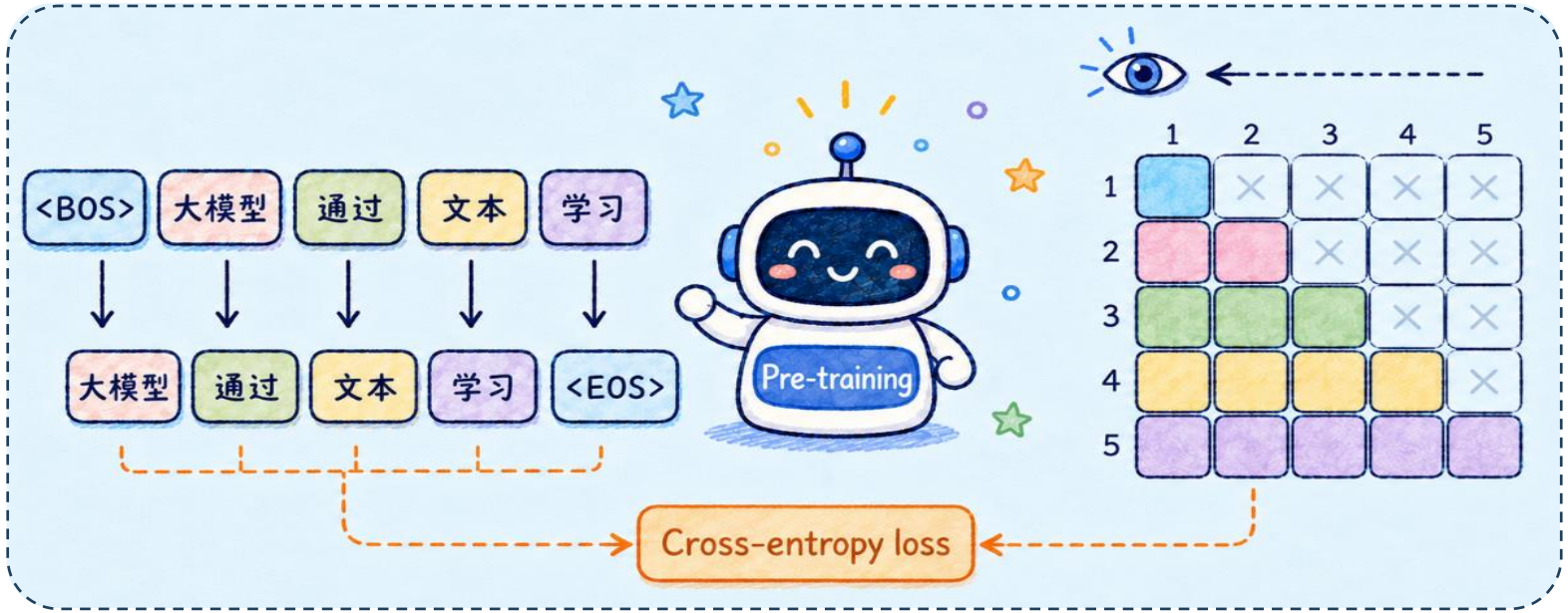
Tokenizer 改变序列长度与词义边界：切得过碎会增加计算量，覆盖不足会损害多语种与代码能力

Pre-training: Autoregressive Next-Token Prediction

训练目标

把文本右移一位就得到监督信号

给定$x_1 \ldots x_t$，模型学习提高真实下一个Token x_{t+1}的概率；所有位置的预测误差共同形成交叉熵Loss。



输入

一段Token序列
只看当前位置左侧

目标

最小化交叉熵
提高真Token概率

产物

Base Model
会续写、含知识

自监督的关键：训练材料既是输入，也是标签；规模化文本因此可以直接转化为学习信号

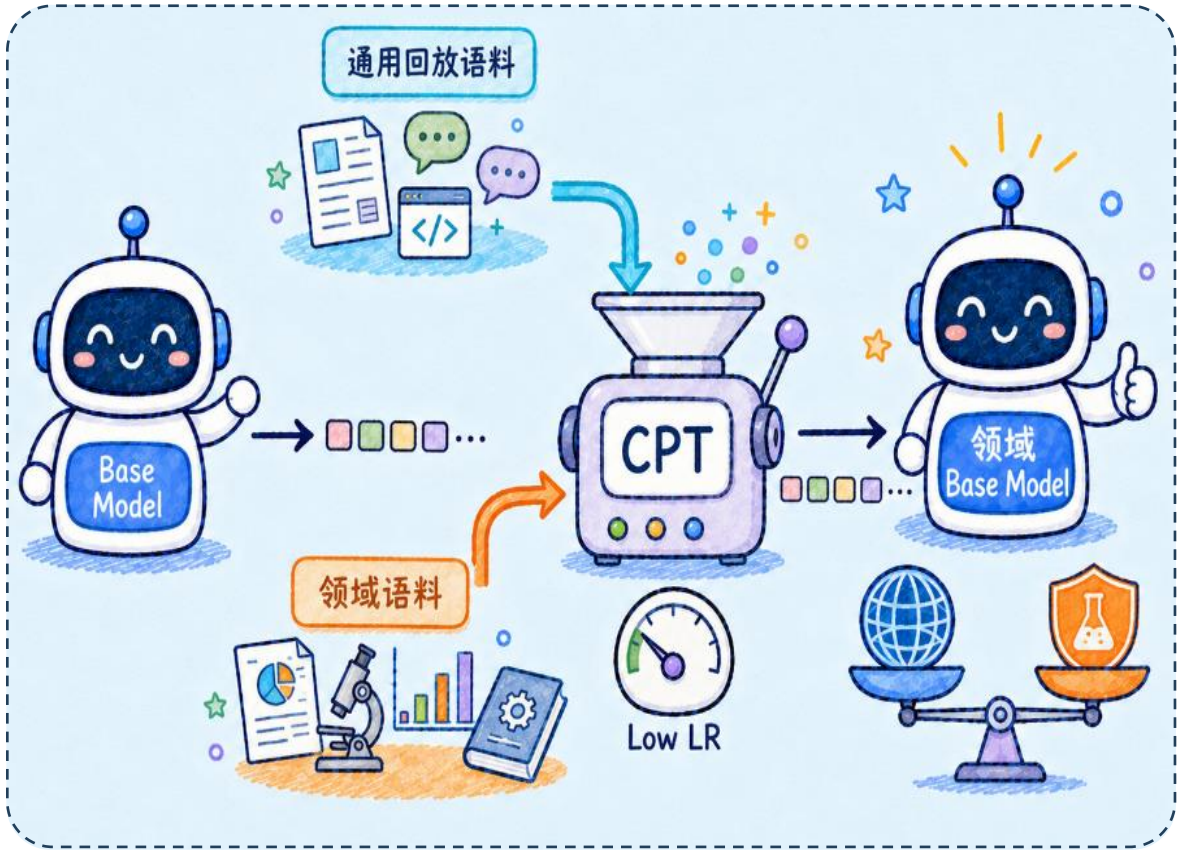
Continued Pre-training: Enhancing Domain && Language-Specific Capabilities

持续预训练 (CPT)

在通用Base Model上继续使用语言建模目标，可补充领域、语言、时效知识或更长上下文能力

保持目标函数，只改变训练数据分布

仍然做下一Token预测；混合领域、多语种、代码或长文档，使Base Model向目标能力继续移动。



数据重配	低率续训	回归验证
提高领域比例 保留通用回放	更低学习率 控制Token数	评估领域增益 监测通用遗忘

Base Model 擅长补全文本，却未必稳定遵循指令；要变成可用助手，还必须进入后训练

SFT: Learning Instructions, Dialogues, and Desired Responses

训练形式仍是下一Token预测，但数据变成“系统提示—用户指令—助手回答”，Loss通常只计算目标回答部分

构造示例 → **模板化训练** → **得到Instruct**

01 收集与筛选

人工编写、专家数据或经过审核的合成回答；覆盖问答、代码、推理、工具和多轮对话。

02 格式化并计算Loss

套用统一Chat Template，通常遮罩System与User部分，只监督Assistant目标回答。

03 训练与回归

全参微调 (FFT) 或参数高效微调 (PEFT)，比如Prefix tuning，LoRA等。

SFT 提供“应该怎样回答”的正向示范，但只看标准答案，尚未充分学习多个可行回答之间的优劣

Post-training: From Next-Token Prediction to Helpful AI Assistance

后训练是基于**预训练模型**，采用指令微调和强化学习等技术使模型能够遵循指令，并生成符合人类偏好的内容，同时具备推理能力

SFT指令微调

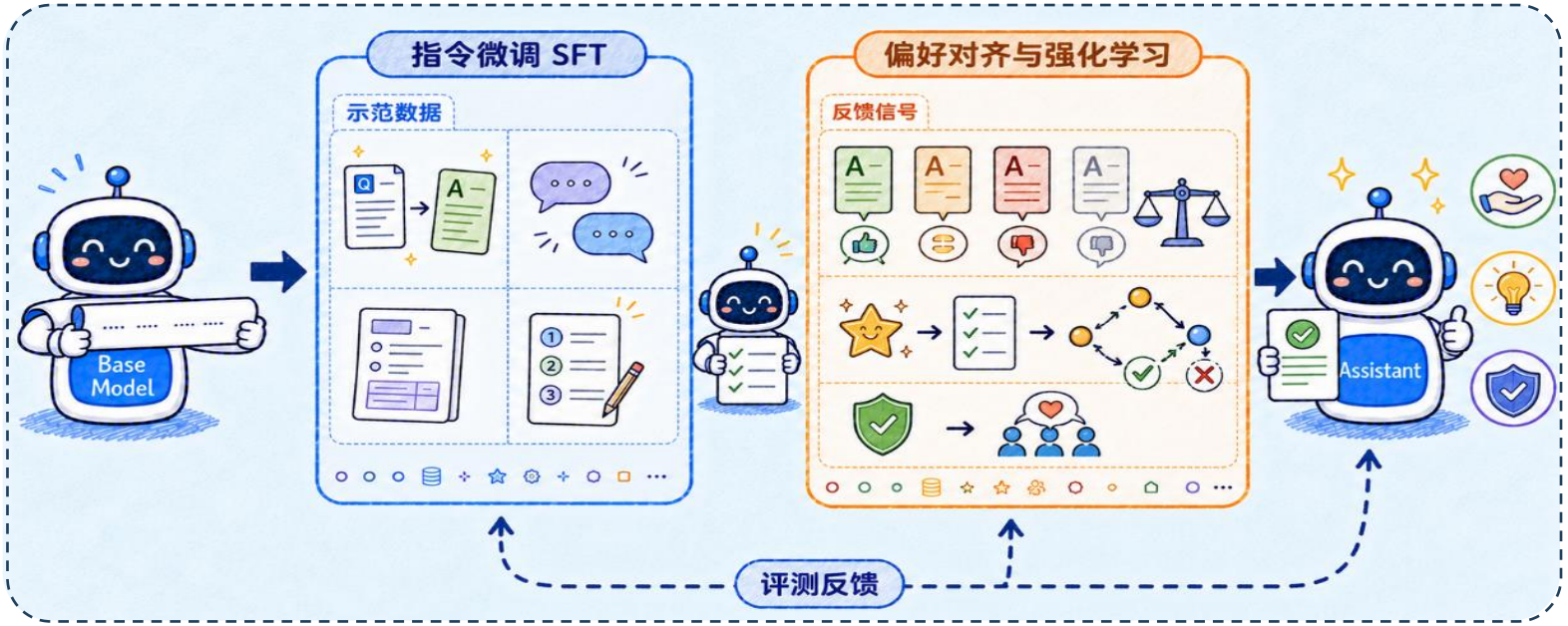
学习指令—**回答**
对话与格式规范

偏好对齐

比较多个答案
DPO或RLHF

推理强化

多次采样与验证
RLVR / GRPO



后训练的共同目标：改变回答分布，让高质量、有帮助且安全的回答更容易被模型生成

RL: Enhancing Reasoning and Preference Alignment after SFT

模型通过可验证奖励或偏好奖励获得反馈，持续提高高质量回答的生成概率。

生成回答

模型对同一问题采样一个或多个候选答案及推理过程

计算奖励

规则验证、奖励模型或偏好比较对答案与行为质量进行评分。

策略更新

提高高奖励回答的概率，抑制低质量行为

推理能力强化

数学、代码、逻辑与复杂问题求解
典型路线： RLVR
常用算法： GRPO、PPO等

行为偏好对齐

有用性、指令遵循、安全性与回答风格
典型路线： RLHF、RLAIF
常用方法： PPO等强化学习算法，或DPO等直接偏好优化方法

强化学习通过可验证奖励提升推理能力，通过偏好奖励优化有用性与安全性

RL for MT: Learning Translation Reasoning through Relative Advantage Optimization

ExTrans 将 DeepSeek-R1 的译文作为示例，并用 DeepSeek-v3 比较策略译文与示例译文，生成RL奖励

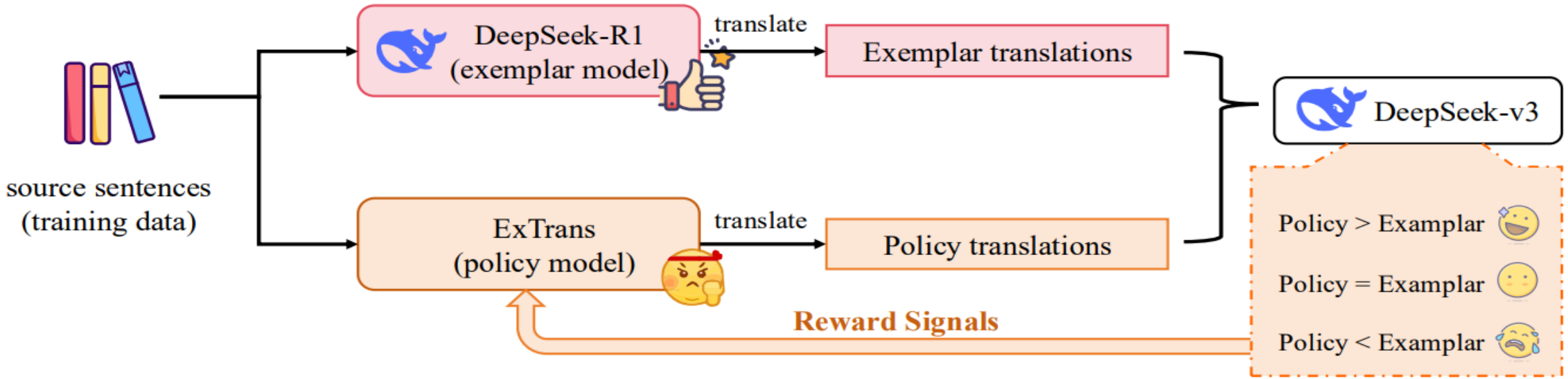
让翻译模型与更强示例“同题比较”

策略译文优于示例译文时，获得更高奖励；奖励还结合格式正确性、思维过程质量与 CometKiwi 分数。

1 示例译文
DeepSeek-R1 生成exemplar translation

2 策略译文
ExTrans policy 生成思考+译文

3 GRPO更新
DeepSeek-v3 比较质量 优势转化为奖励



Wang, J et al. ExTrans: Multilingual Deep Reasoning Translation via Exemplar-Enhanced Reinforcement Learning. NAACL2025

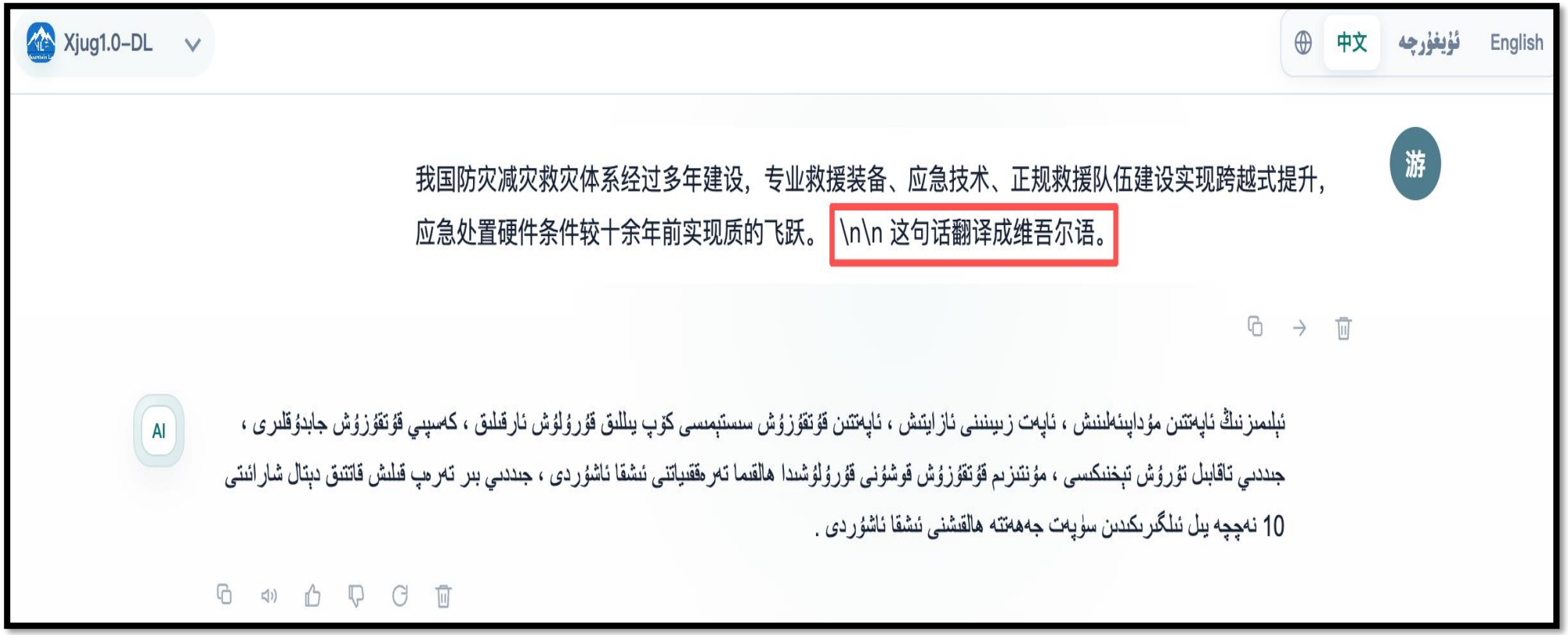
TianChi LLM


新对话
对话记录
新对话
07/26 01:30
游客用户
未登录状态，可注册或登录后保存个人信息。
Xjug1.0-DL
中文
ئۇيغۇرچە
English
NLP
Mountain Lab
你好，有什么可以帮你?
我是你的智能助手，可以回答问题、协助写作与翻译。
给维语智能助手发送消息
深度思考

TianChi LLM

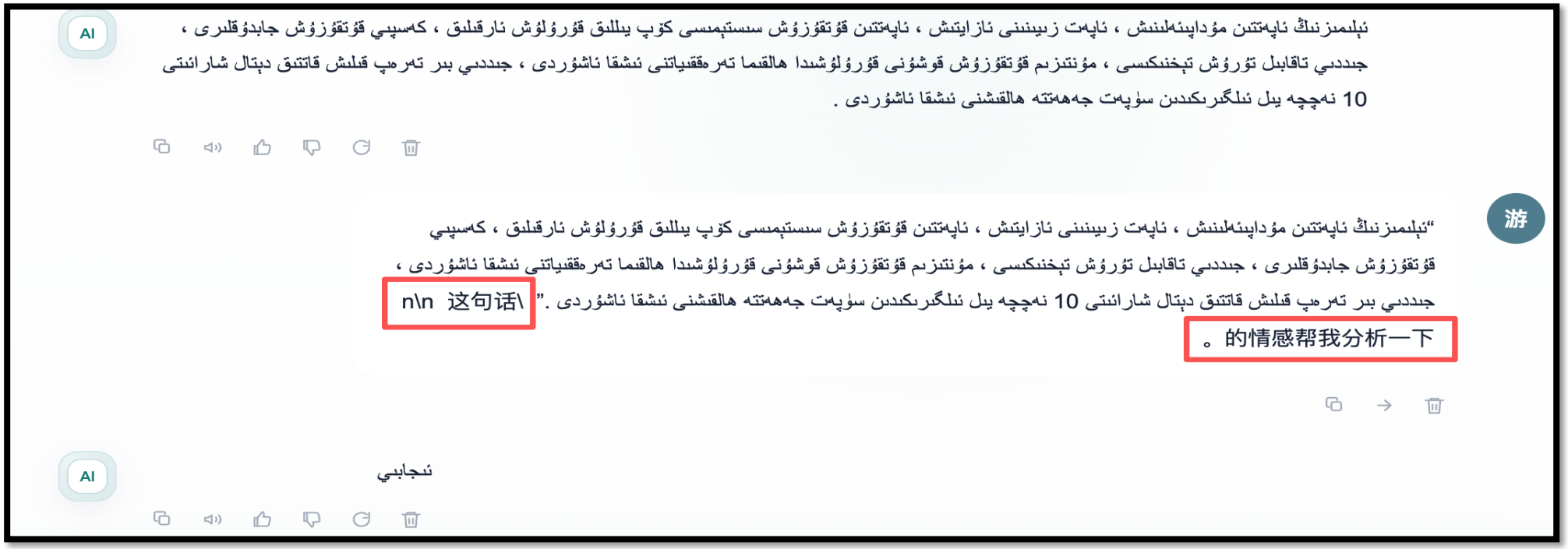

中文
ئۇيغۇرچە
English
Xjug1.0-DL
يېڭى سۆھبەت
سۆھبەت تارىخى
يېڭى سۆھبەت
01:30 07/26
NLP
Mountain Lab
ياخشىمۇسىز، سىزگە قانداق ياردەم بېرەلەيمەن؟
مەن سىزنىڭ ئەقلىي ياردەمچىڭىز، سوئاللارغا جاۋاب بېرىش، يېزىش ۋە تەرجىمىگە ياردەملىشىمەن.
ئۇيغۇرچە ئەقلىي ياردەمچىگە ئۇچۇر يېزىڭ
چوڭقۇر تەپەككۇر
مېھمان
تىزىمغا كىرمىگەن.
تىزىملىتىپ ياكى كىرىپ
شەخسىي
ئۇچۇرلىرىڭىزنى
ساقلىيالايسىز.

MT Task

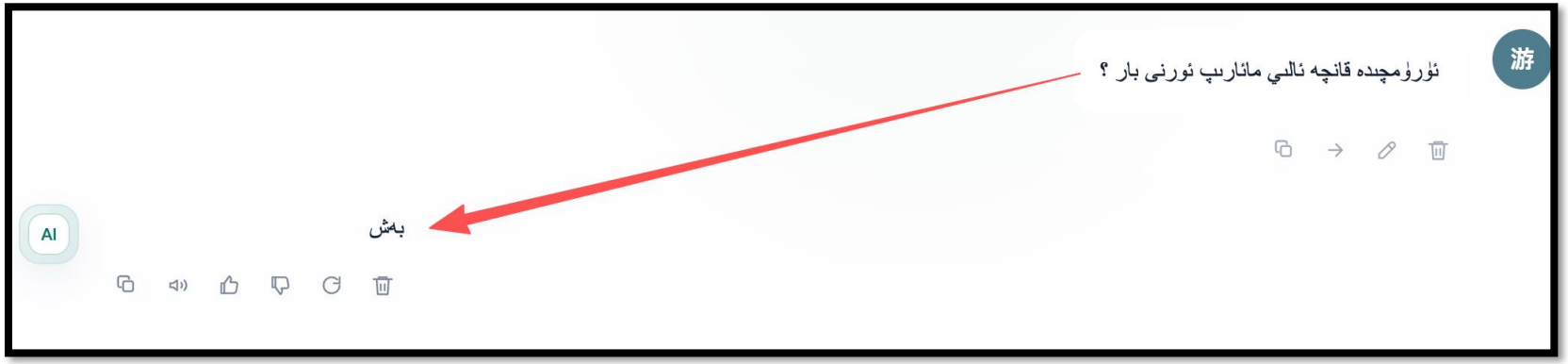

Xjug1.0-DL
中文
ئۇيغۇرچە
English
游
我国防灾减灾救灾体系经过多年建设，专业救援装备、应急技术、正规救援队伍建设实现跨越式提升，
应急处置硬件条件较十余年前实现质的飞跃。 \n\n 这句话翻译成维吾尔语。
AI
ئېلىمىزنىڭ ئاپەتتىن مۇداپىئەلىنىش ، ئاپەت زىيىنىنى ئازايتىش ، ئاپەتتىن قۇتقۇزۇش سىستېمىسى كۆپ يىللىق قۇرۇلۇش ئارقىلىق ، كەسىپي قۇتقۇزۇش جابدۇقلىرى ،
جىددي تاقابىل تۇرۇش تېخنىكىسى ، مۇنتىزىم قۇتقۇزۇش قوشۇنى قۇرۇلۇشىدا ھالقىما تەرەققىياتنى ئىشقا ئاشۇردى ، جىددي بىر تەرەپ قىلىش قاتتىق دېتال شارائىتى
10 نەچچە يىل ئىلگىرىكىدىن سۈپەت جەھەتتە ھالقىشنى ئىشقا ئاشۇردى .

TianChi LLM

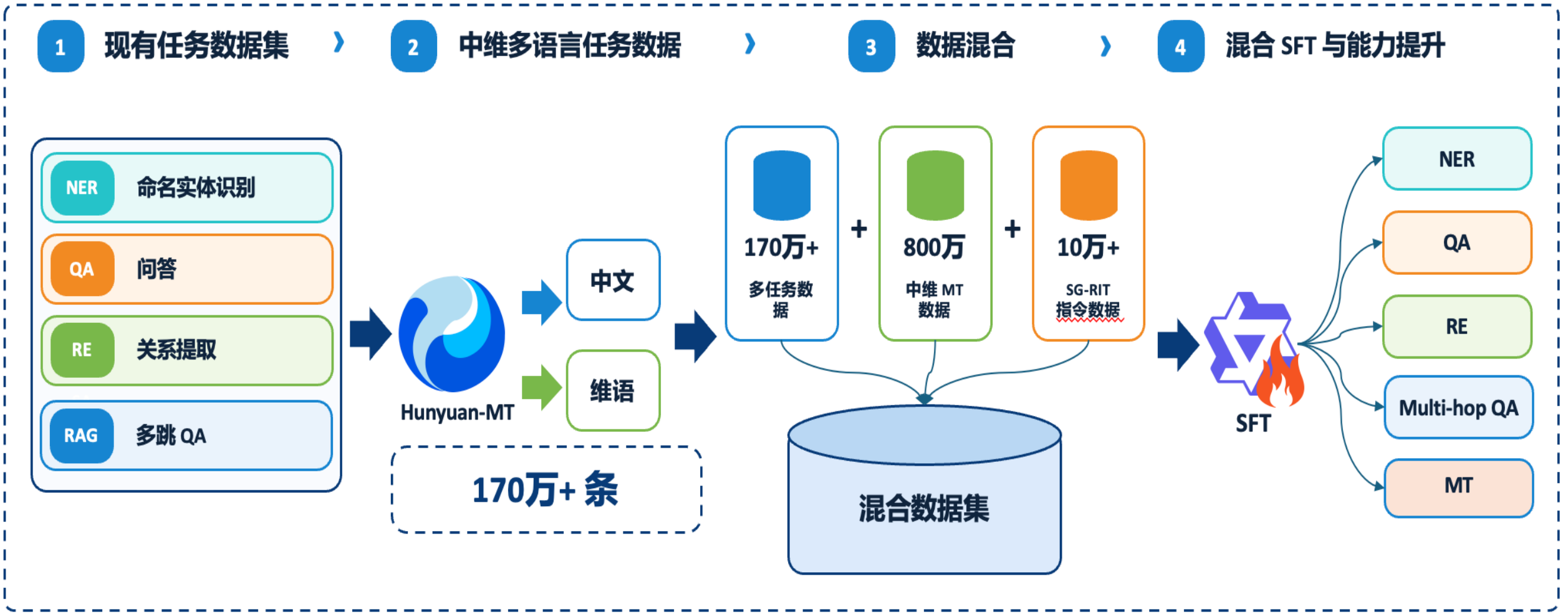
SA Task



QA Task



Multi-task Instruction Tuning: Enhancing LLM Capabilities for Low-Resource Languages



目标：以高质量混合语料提升中文、维吾尔语等多语言场景的关键任务能力

Evaluation Results

机器翻译

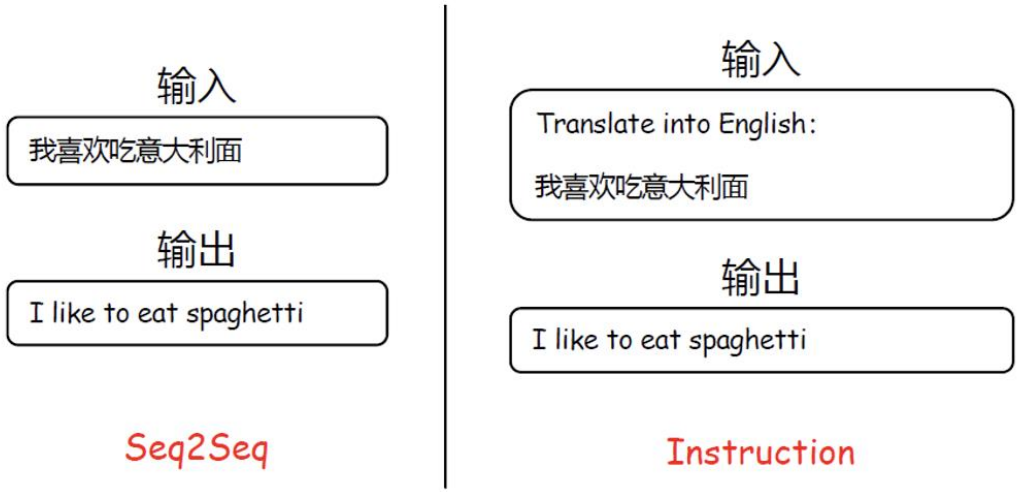
模型	中文 → 维语					
	BLEU	ChrF++	COMET	XCOMET	BLEURT	Avg.
Qwen3-32B	5.34	23.36	55.56	27.67	27.76	27.94
Ours	**24.02**	**43.56**	**81.77**	**68.96**	**31.29**	**49.92**
Hunyuan-MT-7B	19.85	41.03	82.72	70.67	30.59	48.97
Ours	**27.17**	**42.83**	**83.41**	**72.58**	**33.86**	**51.97**

问答、多跳问答、命名实体识别、关系提取

语言	模型	任务			
		QA	Multi-hop QA	NER	RE
中文	Qwen3-32B	23.65	3.98	4.00	5.78
	Ours	**62.43**	**73.20**	**87.31**	**42.64**
维语	Qwen3-32B	10.91	2.15	6.80	10.47
	Ours	**51.12**	**68.08**	**64.11**	**21.13**

Machine Translation with LLM

- General Model v.s. Specific Task（Translation）
  - Specifying model behavior through instruction.
  - In-context Learning (ICL)

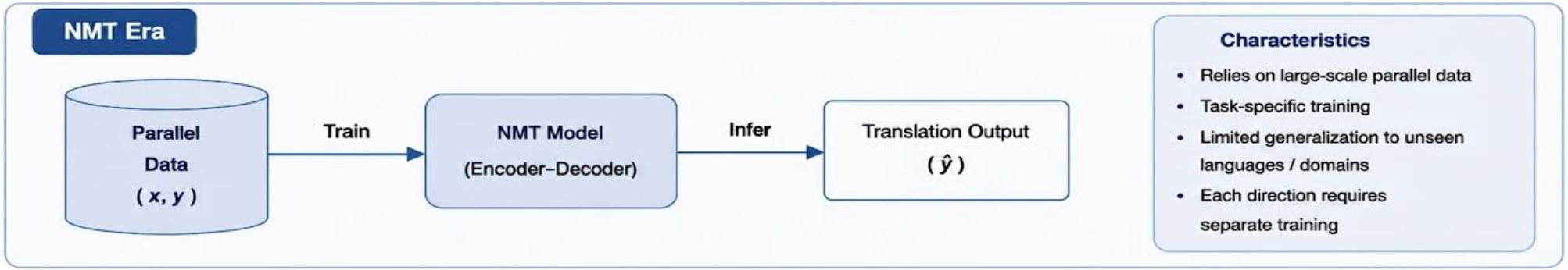


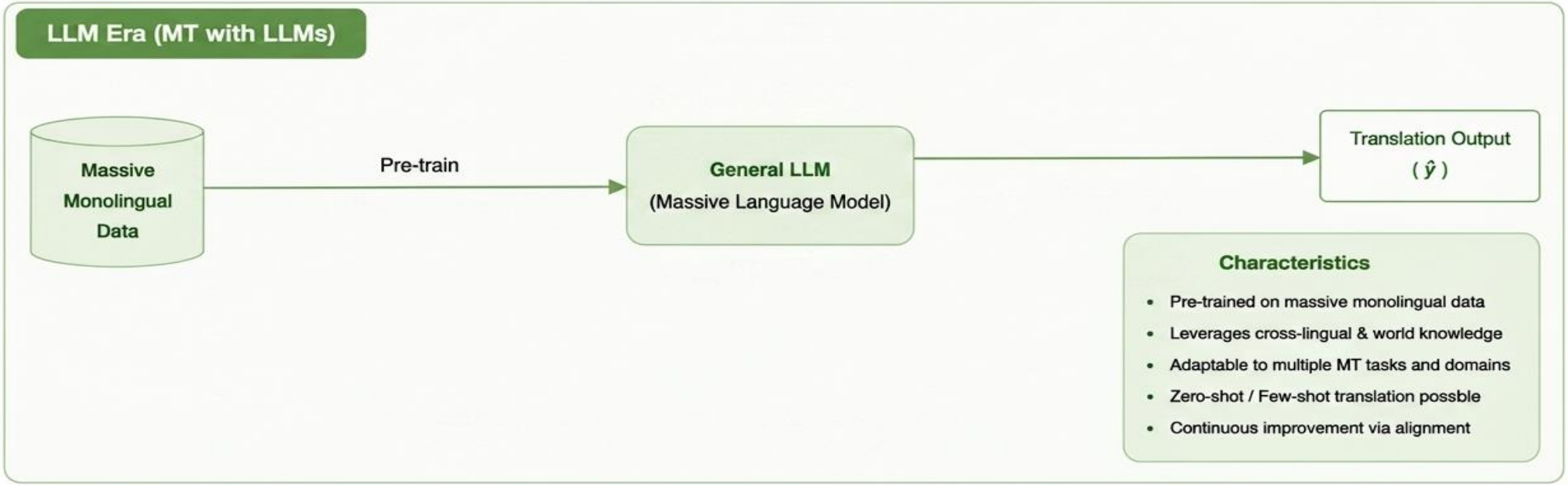
Acquiring the ability to directly **follow instructions** through instruction learning.

A **small number of examples** were obtained during **pre-training**.

Prof. Huang CCL 2023

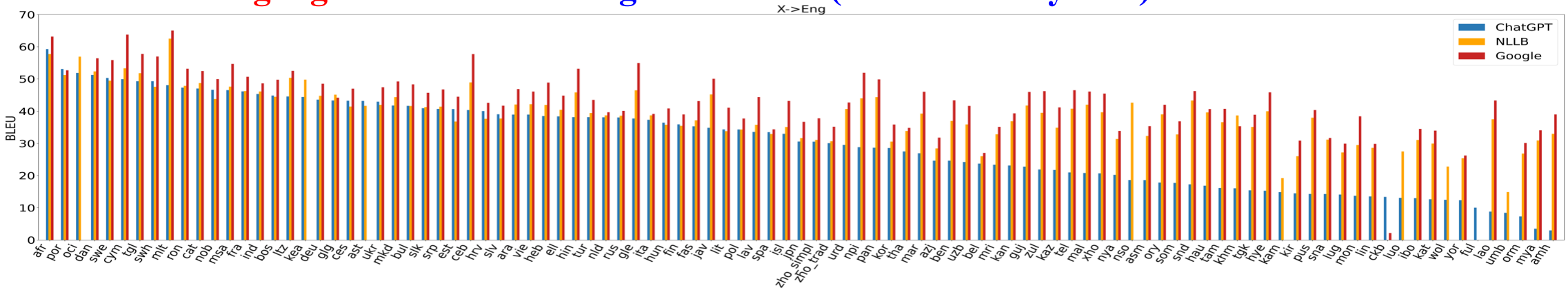
Paradigm Shift in Machine Translation: From NMT to LLM-era

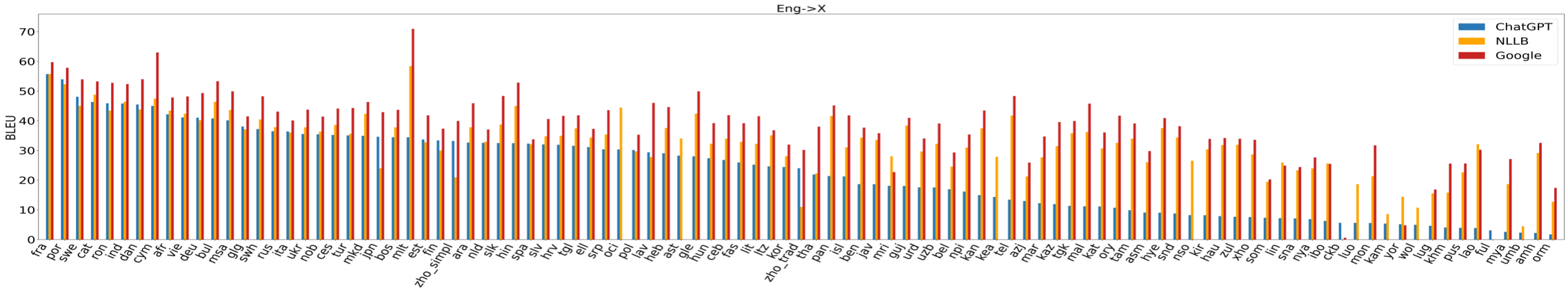




LLM-based MT v.s. NMT

- **LLMs still lag behind strong supervised baselines:** ChatGPT (16.67%) v.s. NLLB (83.33%)
- **Low-resource languages: far behind Google Translate (commercial system)**





Zhu W, et al. Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis, arXiv:2304.04675

Prof. Huang CCL 2023

Pros of LLM-based MT

- **语境理解能力**
  - 提升翻译的准确性和流畅程度
- **交互式翻译功能**
  - 用户能够为翻译提供实时反馈或指导
- **多功能文本处理**
  - 能够翻译正式文本、口语和俚语
- **文档级翻译功能**
  - 在长文本和复杂文档中保持一致性和语境信息
- **多语言功能**
  - 能同时进行多种语言的翻译

Context

Prof. Liu CCMT 2023

Pros of LLM-based MT

Interactive Translation

AD

Please translate it into English:
大型语言模型已被证明在使用少量样本学习的情况下，能在各种自然语言任务中表现出色。为了进一步了解规模对少量样本学习的影响，我们训练了一个具有5400亿参数、密集激活的Transformer语言模型，我们称之为Pathways语言模型。

The translation of the provided text into English is:

"Large language models have been shown to perform well in various natural language tasks with limited sample learning. To further understand the impact of scale on limited sample learning, we trained a Transformer language model with 540 billion parameters and dense activations, which we call the Pathways language model."

Versatile Text Processing

AD

Please translate "远水救不了近火" into English.

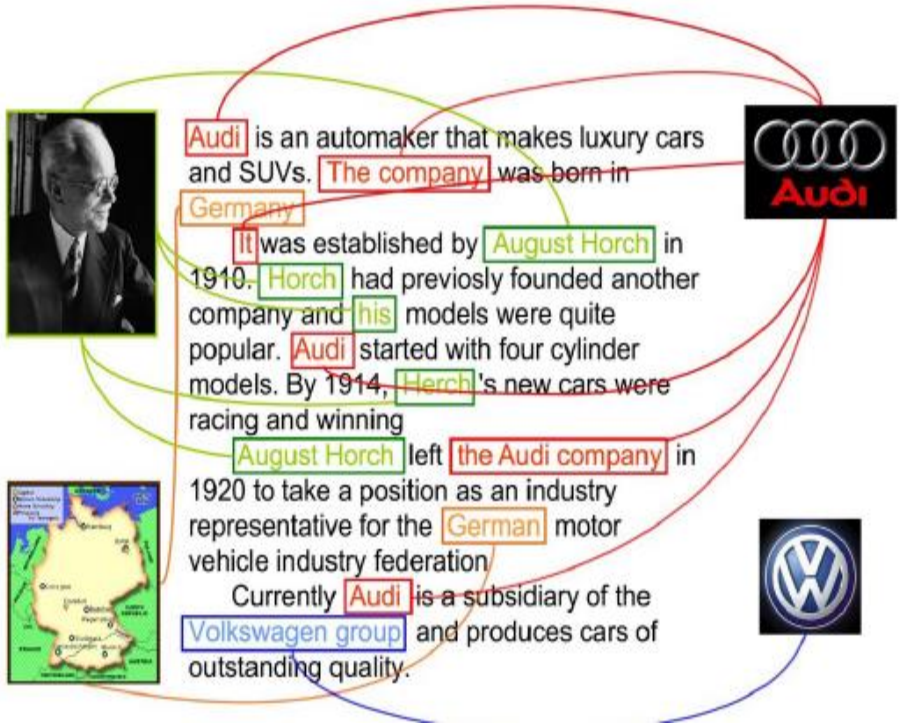
The Chinese phrase "远水救不了近火" literally translates to "distant water cannot put out a nearby fire." This idiom is used to express the idea that help that is too far away is of no use in an emergency or urgent situation. In other words, when you're facing an immediate problem, solutions that are too far off in time or space won't be of any help.

Prof. Liu CCMT 2023

Pros of LLM-based MT

Document-Level Translation

Translate this document from English to Chinese:



奥迪是一家生产豪华汽车和SUV的汽车制造商。该公司起源于德国。它由奥古斯特·霍希（August Horch）于1910年创立。霍希此前曾创立过另一家公司，其车型相当受欢迎。奥迪起初生产四缸车型。到1914年，霍希的新车开始参加比赛并获胜。

奥古斯特·霍希于1920年离开奥迪公司，担任德国机动车行业联合会的行业代表。

目前，奥迪是大众集团的子公司，生产质量卓越的汽车。

Wang et al. Document-Level Machine Translation with Large Language Models. EMNLP 2023.

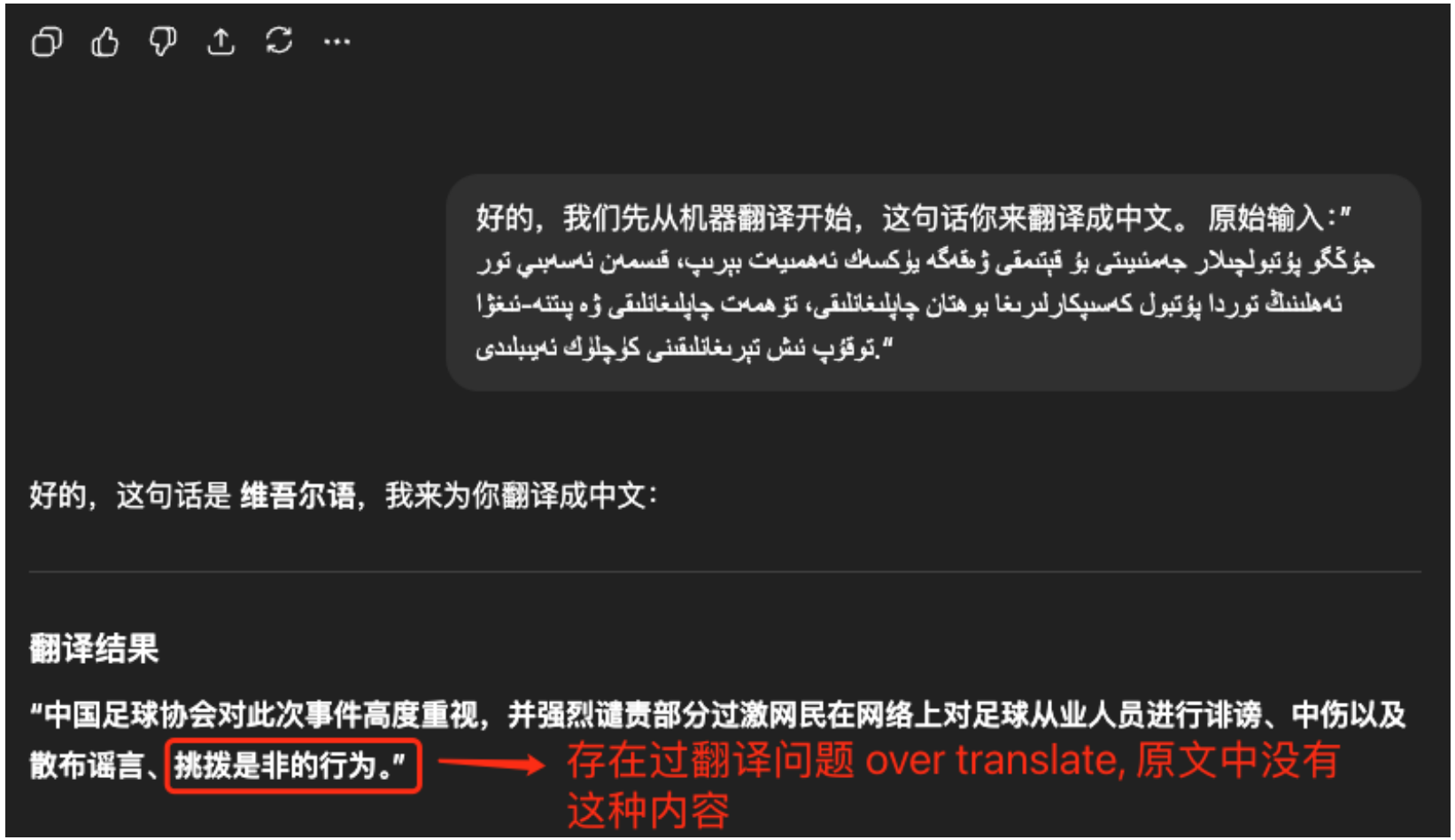
Prof. Liu CCMT 2023

Cons of LLM-based MT

1. Domain Adaptation
   - Limited domain knowledge transfer
   - Catastrophic forgetting
2. Low-Resource Languages
   - Limited performance on rare language pairs
   - Insufficient linguistic resources
3. Evaluation
   - Traditional metrics cannot fully measure translation quality
4. Efficiency
   - High cost for fine-tuning and deployment

Challenges of LLM-based MT on LRLs

Original Input: “ جۇڭگو پۇتبولچىلار جەمئىيىتى بۇ قېتىمقى ۋەقەگە يۈكسەك ئەھمىيەت بېرىپ، قىسمەن ئەسەبىي تور ئەھلىنىڭ توردا پۇتبول كەسپكارلىرىغا بوھتان چاپلىغانلىقى، تۆھمەت چاپلىغانلىقى ۋە پىتنە-ئىغۋا توقۇپ ئىش تېرىغانلىقىنى كۈچلۈك ئەيىبلىدى.”



Ug→Zh translation sample

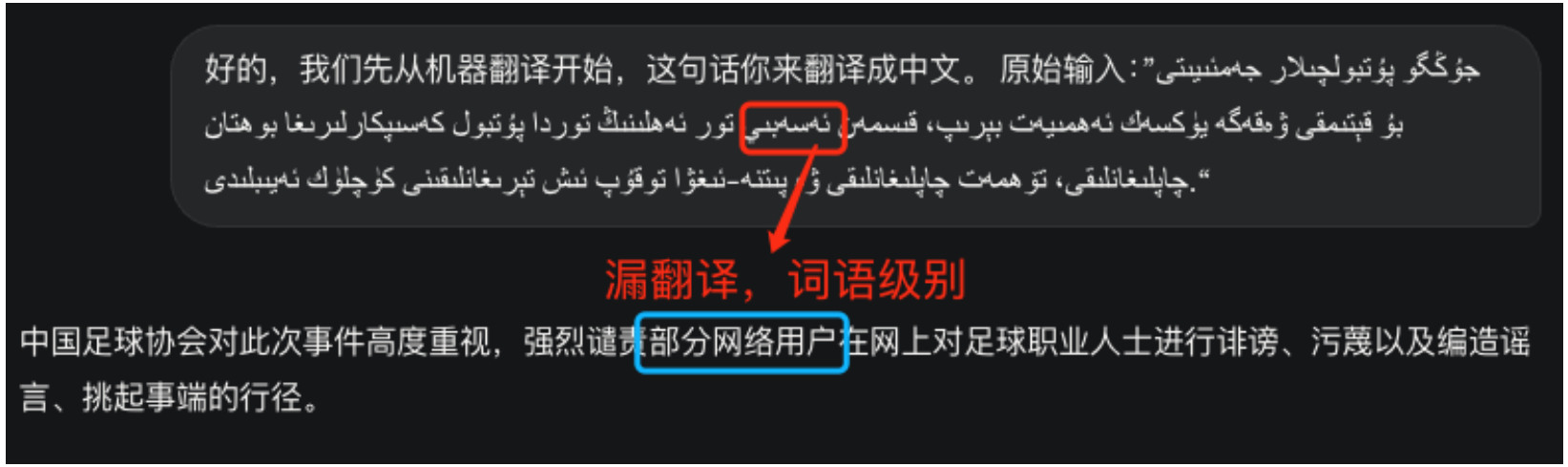
Challenges of LLM-based MT on LRLs

Original Input： “ جۇڭگو پۇتبولچىلار جەمئىيىتى بۇ قېتىمقى ۋەقەگە يۈكسەك ئەھمىيەت بېرىپ، قىسمەن ئەسەبىي تور ئەھلىنىڭ توردا پۇتبول كەسىپكارلىرىغا بوھتان چاپلىغانلىقى، تۆھمەت چاپلىغانلىقى ۋە پىتنە-ئىغۋا توقۇپ ئىش تېرىغانلىقىنى كۆچلۈك ئەيىبلىدى.”

Ug→Zh translation sample

Challenges of LLM-based MT on LRLs

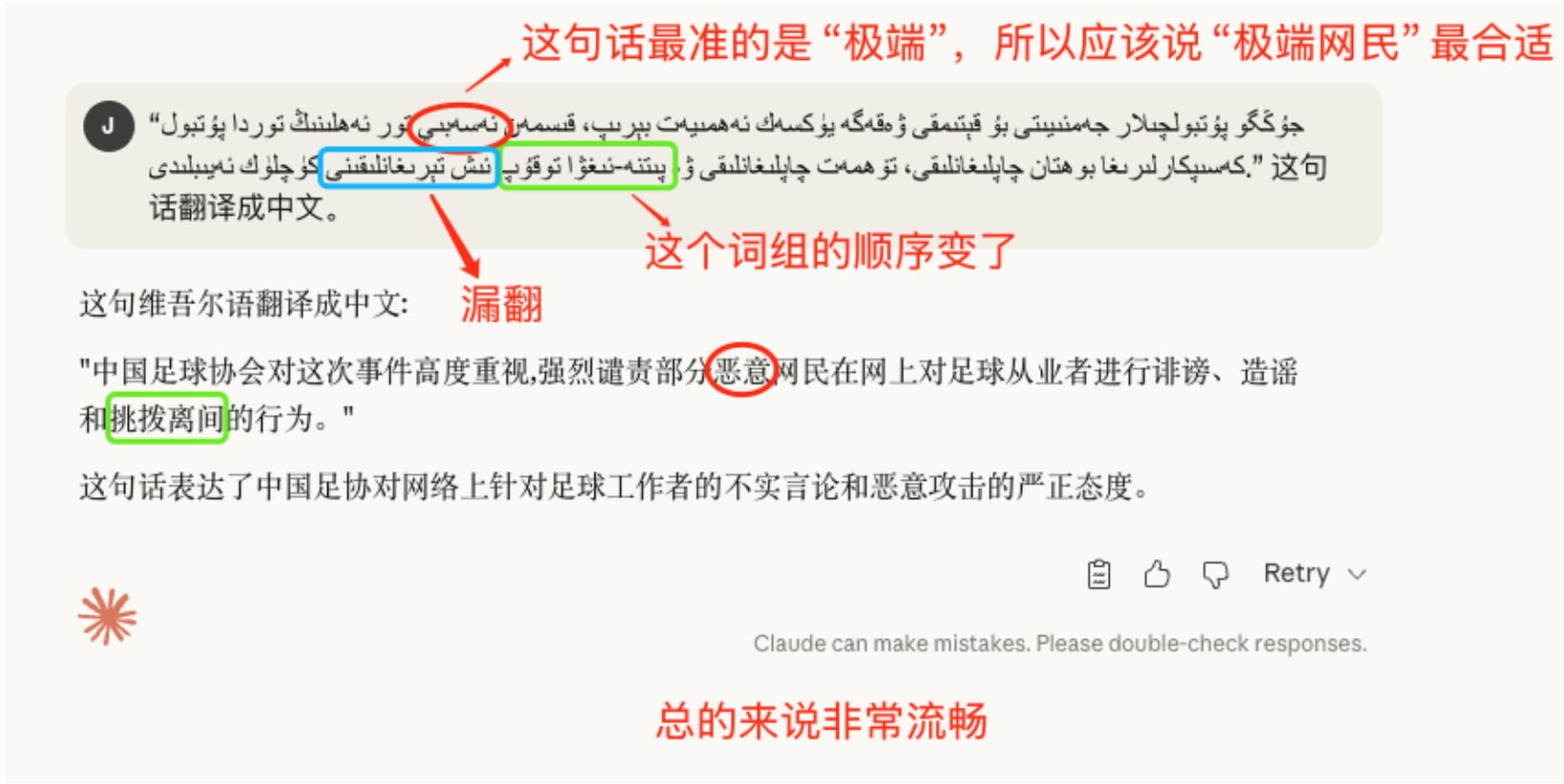
Original Input: “ جۇڭگو پۇتبولچىلار جەمئىيىتى بۇ قېتىمقى ۋەقەگە يۈكسەك ئەھمىيەت بېرىپ، قىسمەن ئەسەبىي تور ئەھلىنىڭ توردا پۇتبول كەسىپكارلىرىغا بوھتان چاپلىغانلىقى، تۆھمەت چاپلىغانلىقى ۋە پىتنە-ئىغۋا توقۇپ ئىش تېرىغانلىقىنى كۈچلۈك ئەيىبلىدى.”



Ug→Zh translation sample

Challenges of LLM-based MT on LRLs

Original Input: “ جۇڭگو پۇتبولچىلار جەمئىيىتى بۇ قېتىمقى ۋەقەگە يۈكسەك ئەھمىيەت بېرىپ، قىسمەن ئەسەبىي تور ئەھلىنىڭ توردا پۇتبول كەسىپكارلىرىغا بوھتان چاپلىغانلىقى، تۇھمەت چاپلىغانلىقى ۋە پىتنە-ئىغۋا توقۇپ ئىش تېرىغانلىقىنى كۈچلۈك ئەيىبلىدى.”



Ug→Zh translation sample

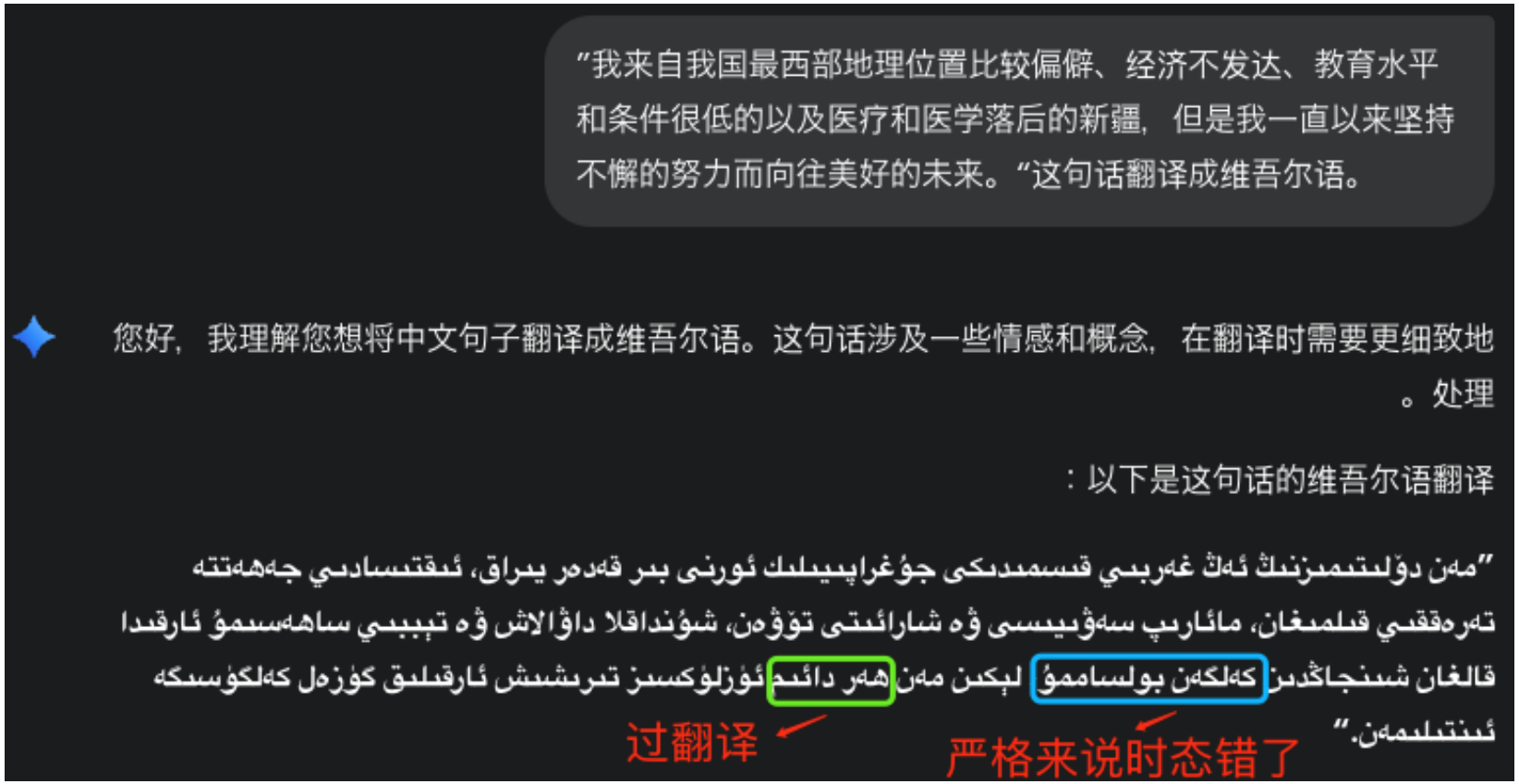
Challenges of LLM-based MT on LRLs

Original Input： “我来自我国最西部地理位置比较偏僻、经济不发达、教育水平和条件很低的以及医疗和医学落后的新疆，但是我一直以来坚持不懈的努力而向往美好的未来。”

Zh →Ug translation sample

Challenges of LLM-based MT on LRLs

Original Input： “我来自我国最西部地理位置比较偏僻、经济不发达、教育水平和条件很低的以及医疗和医学落后的新疆，但是我一直以来坚持不懈的努力而向往美好的未来。”



Zh →Ug translation sample

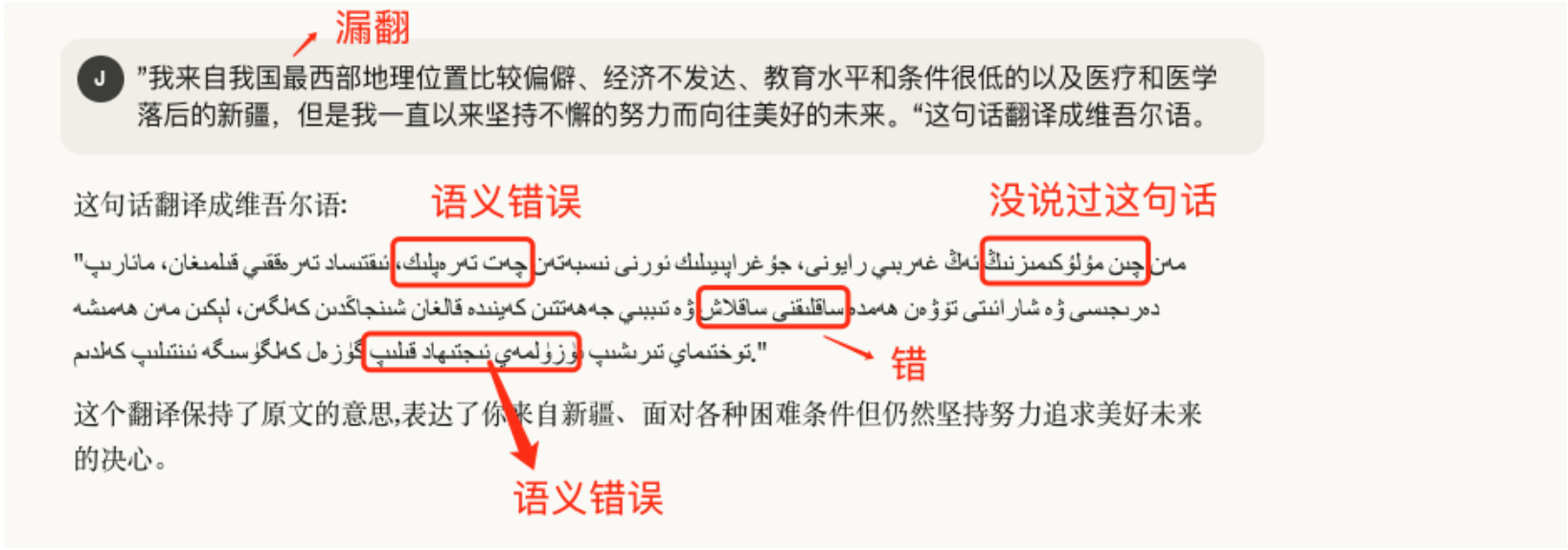
Challenges of LLM-based MT on LRLs

Original Input： “我来自我国最西部地理位置比较偏僻、经济不发达、教育水平和条件很低的以及医疗和医学落后的新疆，但是我一直以来坚持不懈的努力而向往美好的未来。”

Zh →Ug translation sample

Challenges of LLM-based MT on LRLs

Original Input： “我来自我国最西部地理位置比较偏僻、经济不发达、教育水平和条件很低的以及医疗和医学落后的新疆，但是我一直以来坚持不懈的努力而向往美好的未来。”



Zh →Ug translation sample

MT Performance Comparison

Zh →Ug translation sample




Ug→Zh translation sample



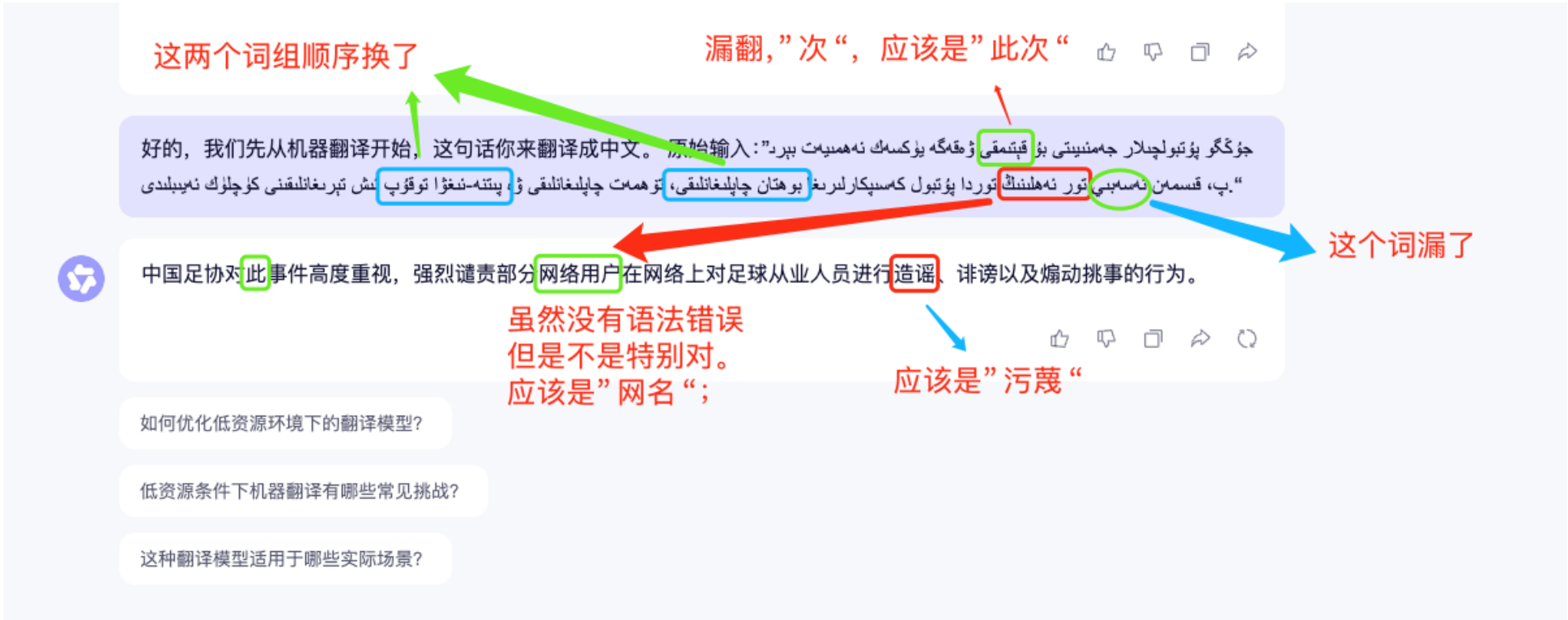

Challenges of LLM-based MT on LRLs

Original Input： “جۇڭگو پۇتبولچىلار جەمئىيتى بۇ قېتىمقى ۋەقەگە يۈكسەك ئەھمىيەت بېرىپ، قىسمەن ئەسەبىي تور ئەھلىنىڭ توردا پۇتبول كەسىپكارلىرىغا بوھتان

Ug→Zh translation sample

Challenges of LLM-based MT on LRLs

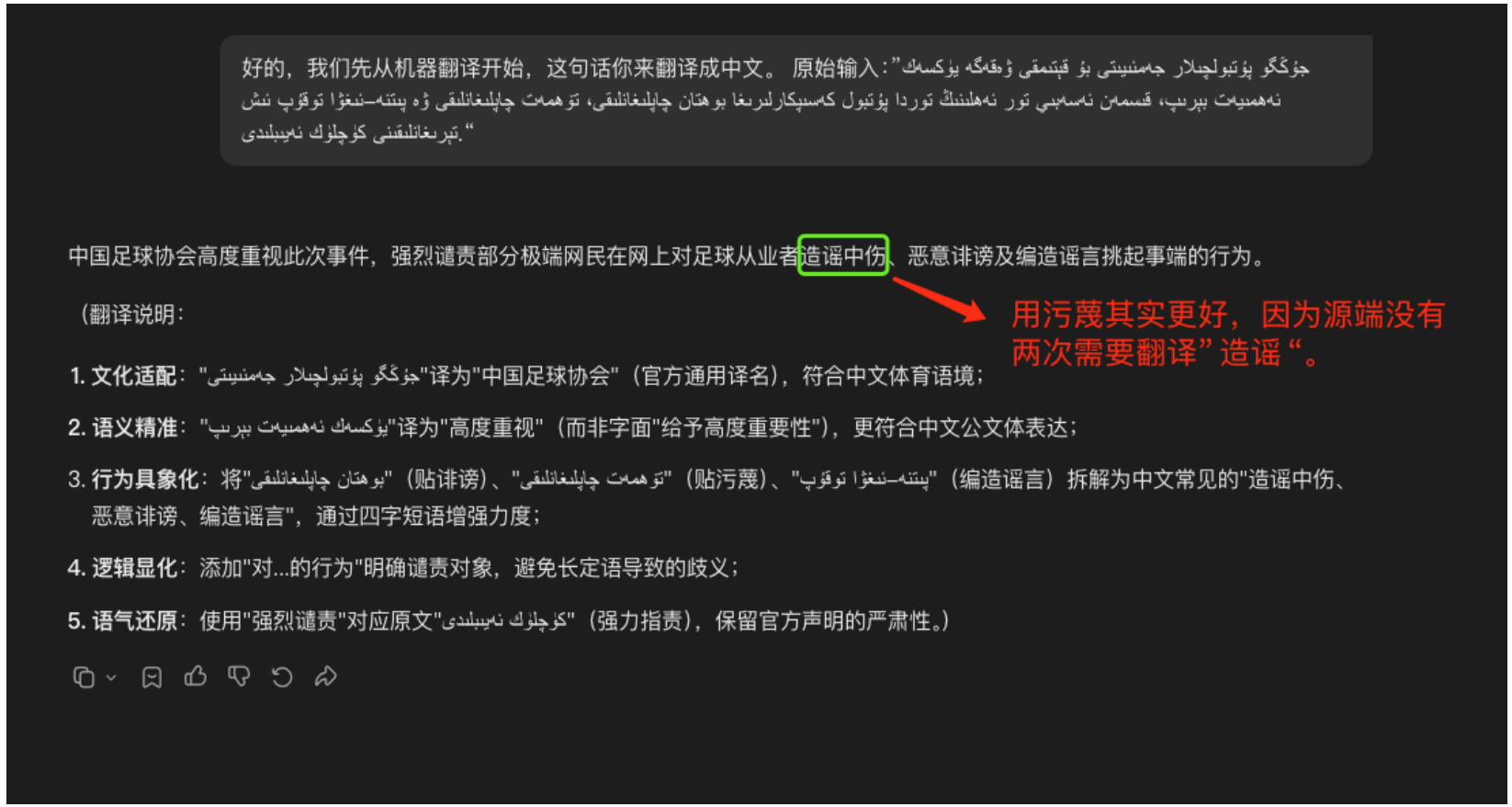
Original Input: “ جۇڭگو پۇتبولچىلار جەمئىيىتى بۇ قېتىمقى ۋەقەگە يۈكسەك ئەھمىيەت بېرىپ، قىسمەن ئەسەبىي تور ئەھلىنىڭ توردا پۇتبول كەسىپكارلىرىغا بوھتان چاپلىغانلىقى، تۆھمەت چاپلىغانلىقى ۋە پىتنە-ئىغۋا توقۇپ ئىش تېرىغانلىقىنى كۈچلۈك ئەيىبلىدى.”



Ug→Zh translation sample

Challenges of LLM-based MT on LRLs

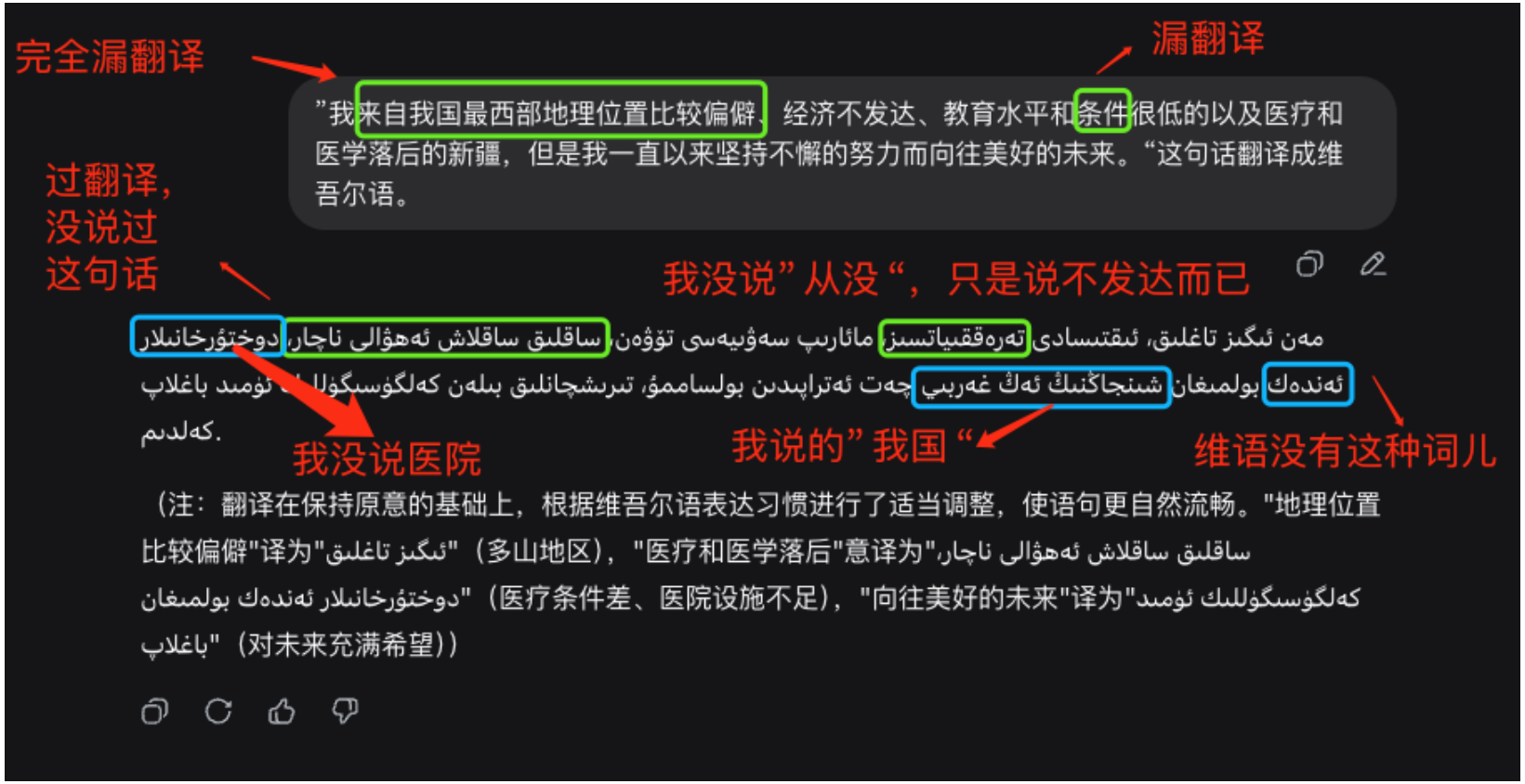
Original Input： “ جۇڭگو پۇتبولچىلار جەمئىيىتى بۇ قېتىمقى ۋەقەگە يۈكسەك ئەھمىيەت بېرىپ، قىسمەن ئەسەبىي تور ئەھلىنىڭ توردا پۇتبول كەسىپكارلىرىغا بوھتان چاپلىغانلىقى، تۆھمەت چاپلىغانلىقى ۋە پىتنە-ئىغۋا توقۇپ ئىش تېرىغانلىقىنى كۈچلۈك ئەيىبلىدى.”



Ug→Zh translation sample

Challenges of LLM-based MT on LRLs

Original Input： “我来自我国最西部地理位置比较偏僻、经济不发达、教育水平和条件很低的以及医疗和医学落后的新疆，但是我一直以来坚持不懈的努力而向往美好的未来。”



Zh →Ug translation sample

Challenges of LLM-based MT on LRLs

Original Input： “我来自我国最西部地理位置比较偏僻、经济不发达、教育水平和条件很低的以及医疗和医学落后的新疆，但是我一直以来坚持不懈的努力而向往美好的未来。”

Zh →Ug translation sample

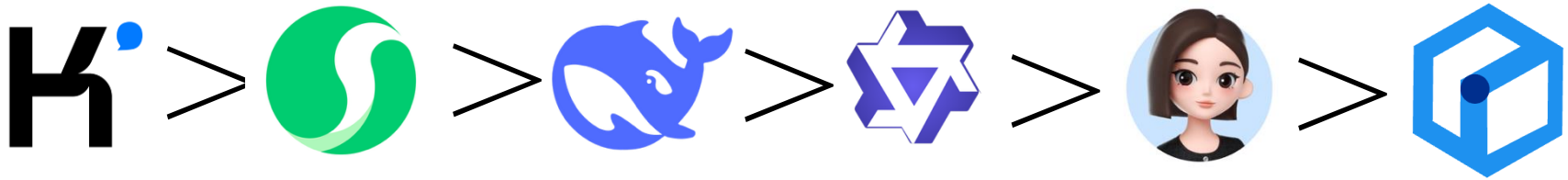
Challenges of LLM-based MT on LRLs

Original Input： “我来自我国最西部地理位置比较偏僻、经济不发达、教育水平和条件很低的以及医疗和医学落后的新疆，但是我一直以来坚持不懈的努力而向往美好的未来。”

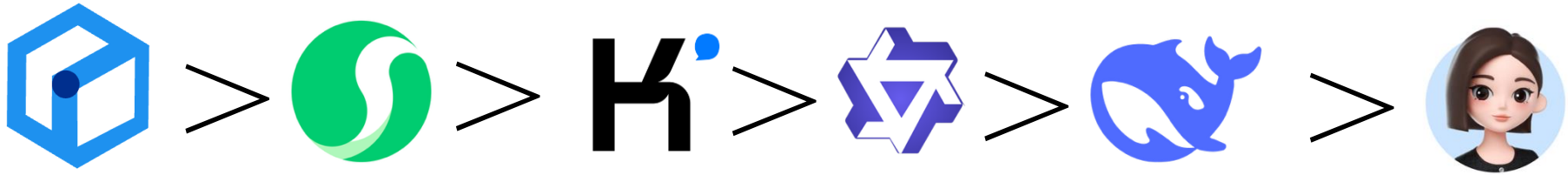
Zh →Ug translation sample

MT Performance Comparison

Ug→Zh translation sample



Zh →Ug translation sample



Main Challenges of LLM-based MT on LRLs

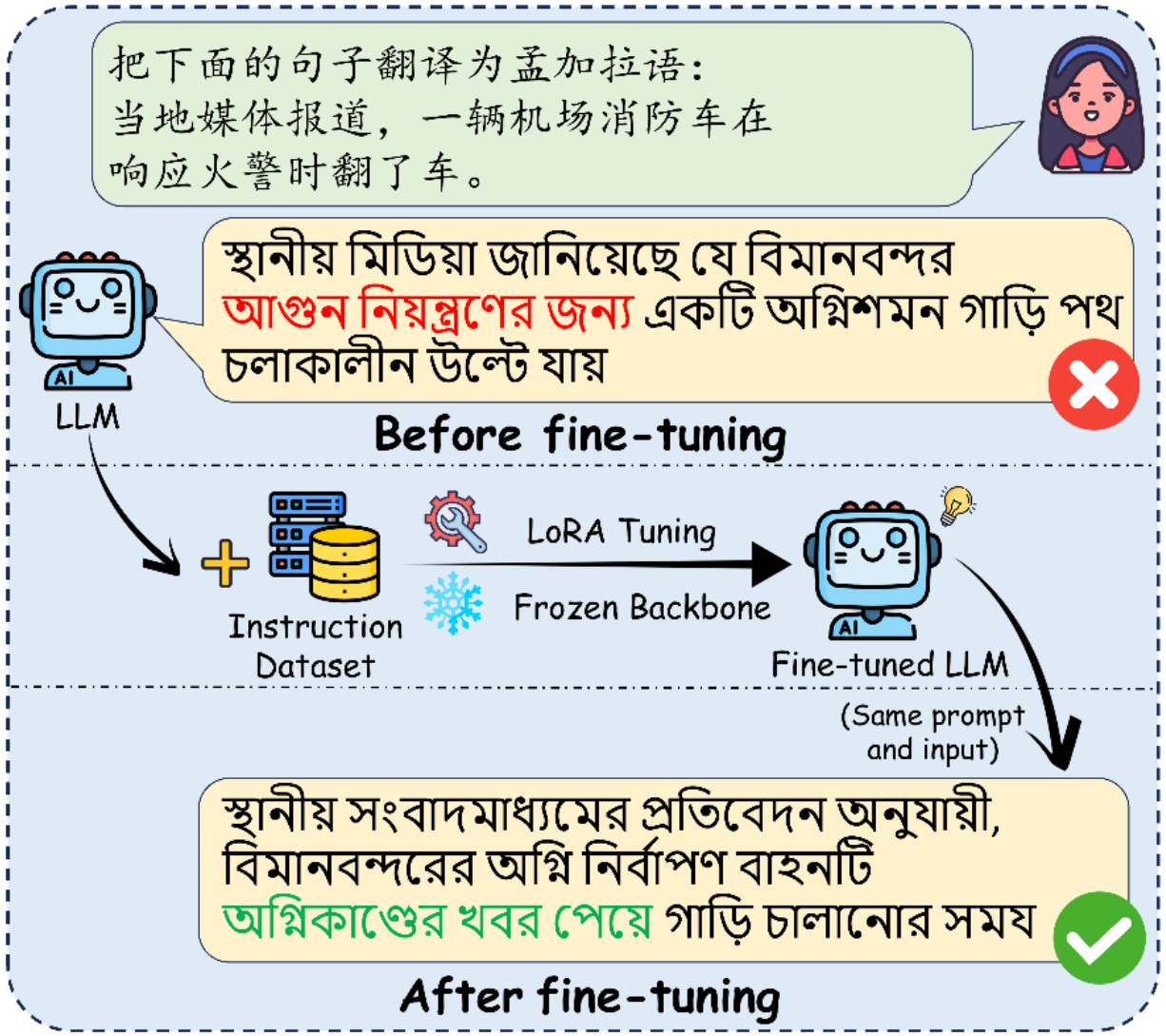
- **Omission** – Missing content or phrases
- **Over-translation** – Unnecessary or redundant content
- **Semantic Errors** – Incorrect meaning or misinterpretation
- **Syntactic Errors** – Grammatical or structural issues
- **Word-level Errors** – Inaccurate lexical choices
- ……

Outline

- 大语言模型时代的低资源语言处理
- 面向低资源语言的大语言模型机器翻译探索
- 多模态大语言模型驱动的低资源翻译探索
- 面向低资源语言大模型的未来探索
- 总结

Strategies for Text-only LLM-based MT

Constructing LRLs Instruction



Low-Resource Bottleneck: LLM performance constrained by scarce, low-quality instruction datasets

Proposed Solution: Refined Instruction Tuning — an automated pipeline for high-quality Chinese→X instruction corpora

COLING2027 under review

Constructing LRLs Instruction

- Experiment
  - Dataset
    - Constructed high-quality instruction datasets for 8 low-resource Chinese → X directions (with approximately 5k instances per direction), covering Uyghur (ug), Tibetan (bo), Persian (fa), Hebrew (he), Urdu (ur), Bengali (bn), Vietnamese (vi), and Indonesian (id).

Region	Language	Size	AverageLength
China	zh→ug	5k	98.93
	zh→ur	5k	106.36
Middle East	zh→fa	5k	102.74
	zh→he	5k	99.09
South Asia	zh→ur	5k	104.03
	zh→bn	5k	97.08
Southeast Asia	zh→vi	5k	97.04
	zh→id	5k	97.27

COLING2027 under review

Constructing LRLs Instruction

- Experiment
  - Main experiment
    - Main experimental results on the FLOWERS+, IWSLT, and CCMatrix test sets, with scores being the average of four evaluation metrics: Chrf++, COMET, XCOMET-XL, and BLEURT.

Method	**FLORES+** ($zh \rightarrow xx$)								
	fa	he	vi	id	bn	ur	ug	bo	Avg.
Madlad (Kudugunta et al., 2023)	60.31	61.74	70.58	75.08	56.85	54.10	42.99	40.65	57.79
Direct (Yang et al., 2025)	64.25	56.79	74.30	72.11	55.53	52.99	35.52	40.76	56.53
COD (Lu et al., 2024)	60.70	60.76	72.62	77.10	60.75	53.48	38.03	**49.35**	59.10
MAPS (He et al., 2024)	66.15	63.09	74.95	78.44	63.21	55.28	37.15	48.22	60.81
CompTrans (Zebaze et al., 2025)	62.92	58.51	73.59	77.17	53.10	52.10	34.37	46.58	57.29
TEaR (Feng et al., 2025)	65.82	**63.37**	75.07	78.54	**65.26**	55.80	38.65	45.40	61.00
Ours	**66.29**†	62.29	**75.57**†	**79.07**†	57.45	**58.11**†	**49.66**†	45.84	**61.79**

Table 2: Comparison results on the FLORES+ benchmark ($zh \rightarrow xx$). Scores in individual language columns represent the macro-average of four evaluation metrics. The 'Avg.' column denotes the mean score across all eight evaluated target languages. '†' indicates that our method achieves statistically significant improvements over the SOTA baseline across all individual metrics with $p < 0.01$.

COLING2027 under review

Constructing LRLs Instruction

- Human Evaluation
  - Adequacy
  - Fluency
  - Faithfulness
  - 1-5

Lan.	Eval.	Fluen.	Adeq.	Faith.
fa	E^1_{fa}	4.96	4.87	4.87
	E^2_{fa}	4.30	4.47	4.56
ug	E^1_{ug}	4.97	4.95	4.99
	E^2_{ug}	4.99	4.98	4.89
bo	E^1_{bo}	3.77	3.69	3.68
vi	E^1_{vi}	4.53	4.91	4.94
bn	E^1_{bn}	4.90	4.87	4.68
ur	E^1_{ur}	4.18	4.27	4.41
Avg.		**4.58**	**4.63**	**4.63**

Table 6: Human evaluation results for the constructed instruction dataset. Column headers are abbreviated as follows: Lan. (Language), Eval. (Evaluator IDs), Fluen. (Fluency), Adeq. (Adequacy), and Faith. (Faithfulness).

COLING2027 under review

Multi-Source Knowledge

- Motivation

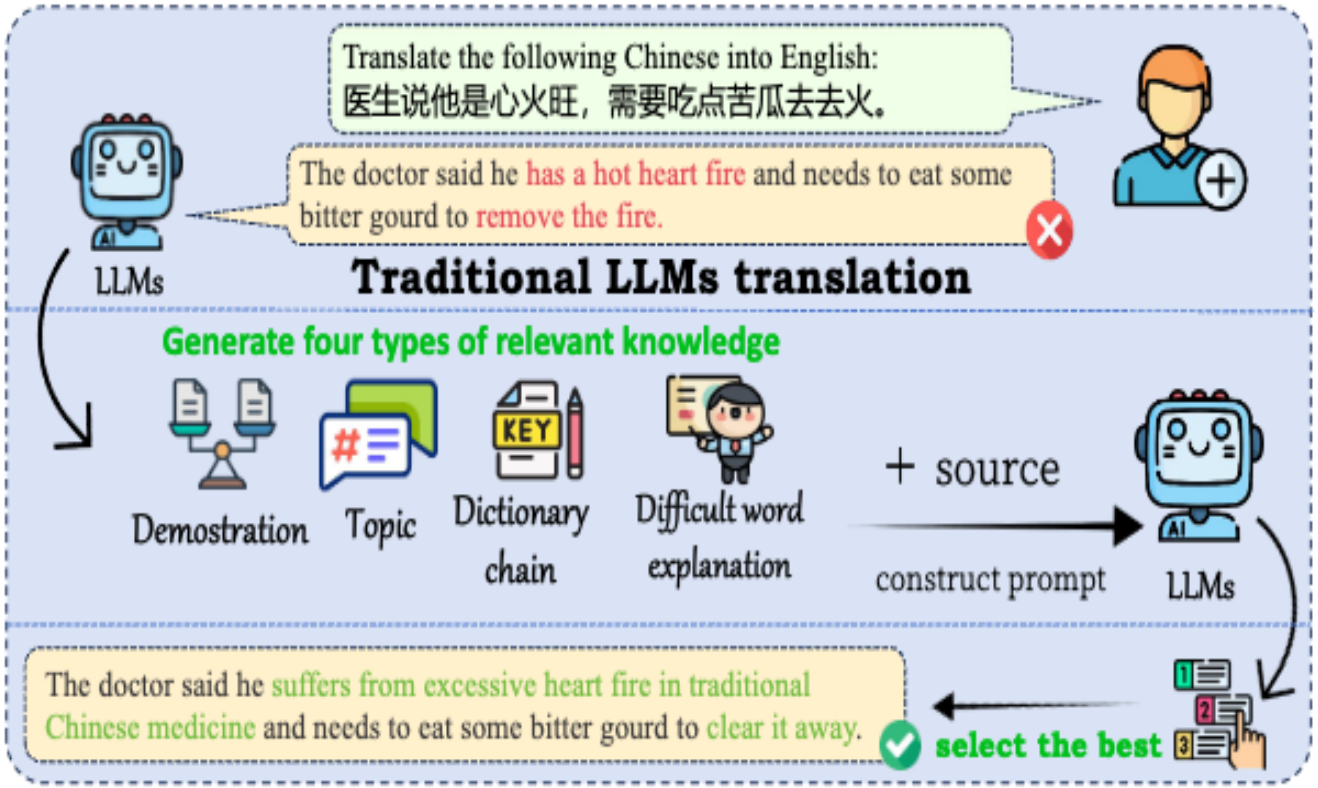


Figure 1: Multi-source reasoning for low-resource languages translation

- Inherent Limitations: LLMs struggle in low-resource translation via **Direct Prompting** due to a lack of explicit linguistic reasoning.
- Information Deficiency: Existing CoT methods suffer from coarse granularity, failing to systematically integrate the deep, **multi-source knowledge** required for precise translation.

CCMT2026 under review

Multi-Source Knowledge

- Dataset & Evaluation Metrics & Models
  - Dataset
    - The experiments utilize the official FLORES-200 benchmark alongside custom-curated test sets from IWSLT and CCMatrix, which were manually cleaned and filtered to ensure high-quality evaluation across diverse low-resource contexts.

Dataset	Language Pairs	Size (per pair)
FLORES-200	En ↔ {Fa, Ur, Lo, Uz}	1,012
IWSLT	Zh → {Fa, He, Id, Vi}	1,000
CCMatrix	Zh → {Fa, He, Id, Bn, Vi}	1,000

Table 1: Statistics of Evaluation Datasets

CCMT2026 under review

- Experiment
  - Main experiment

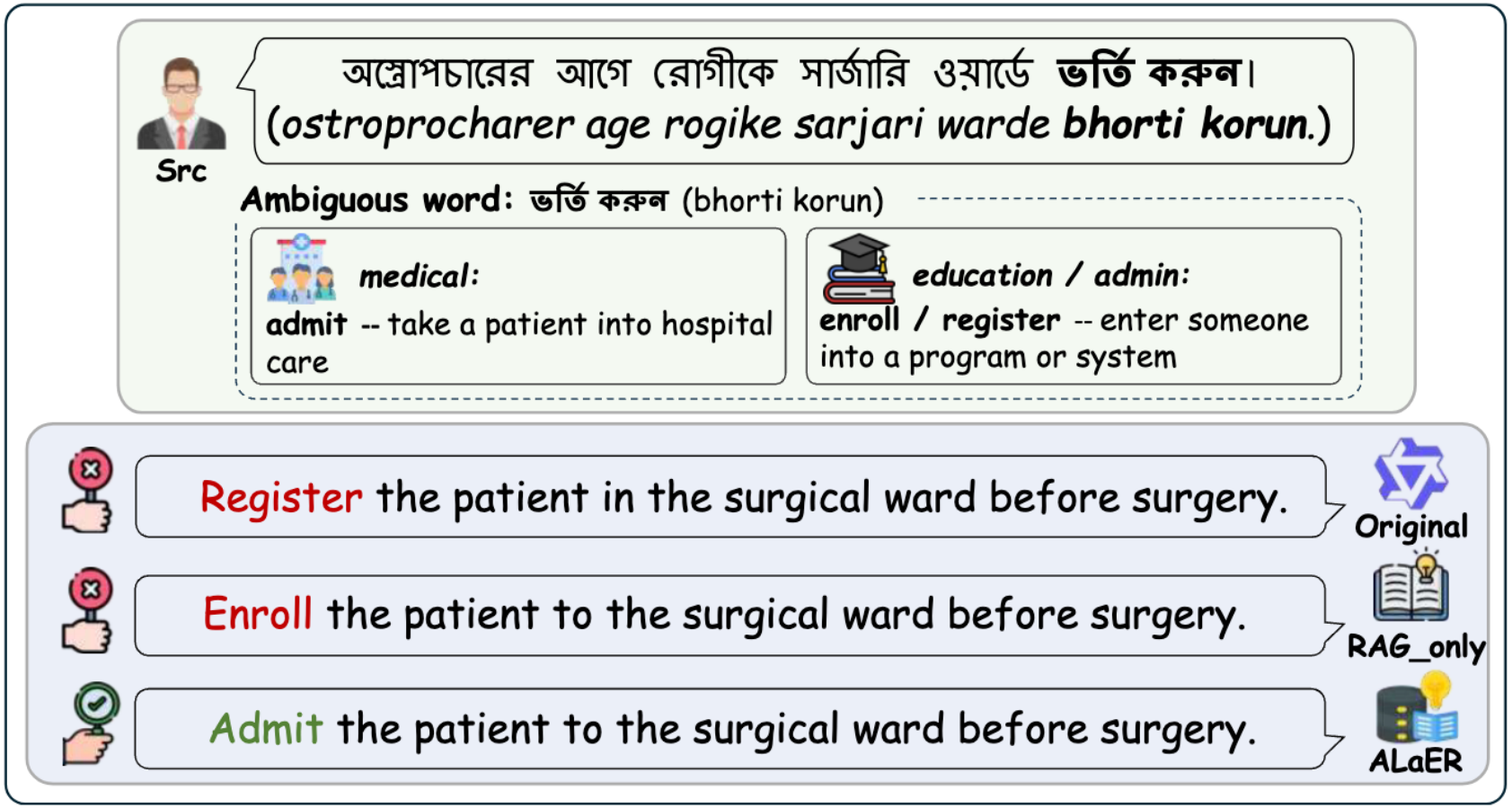
Table 2: Results on the IWSLT dataset (**Chinese** → **XX**) of Qwen3-32B evaluated by XCOMET-XL.

Method	**Zh-Fa**	**Zh-He**	**Zh-Id**	**Zh-Vi**	**Avg.**
Madlad (Kudugunta et al., 2023)[27]	70.04	71.74	81.68	76.53	75.00
Direct (Yang et al., 2025)[28]	74.88	71.24	84.87	81.62	78.15
MAPS (He et al., 2024)[4]	<u>78.03</u>	<u>75.94</u>	83.23	**86.46**	<u>80.91</u>
COD (Lu et al., 2024)[8]	73.56	72.58	79.44	83.91	77.37
TEaR (Feng et al., 2025)[29]	75.72	72.64	<u>84.87</u>	81.11	78.59
CompTra (Zebaze et al., 2025)[30]	74.33	72.29	81.11	<u>84.54</u>	78.07
Ours	**78.73**†	**76.56**‡	**85.66**‡	83.16	**81.03**

CCMT2026 under review

Dual-RAG: Word Sense Disambiguation

- Motivation



- Low-resource languages lack sufficient high-quality bilingual data, making traditional RAG hard to apply.
- LLMs are vulnerable to lexical ambiguity in **cross-domain** translation, which may cause semantic hallucinations.
- Existing RAG systems rely on **coarse sentence-** or document-level retrieval and cannot precisely disambiguate high-risk ambiguous words.

COLING2027 under review

Dual-RAG: Word Sense Disambiguation

- Dataset
  - We evaluate Domain-aware RAG on three benchmarks: X-Bench, WMT, and IWSLT. As shown in the table, the X-Bench is a multi-domain benchmark that includes six Languages, such as Bengali (bn), Hungarian (hu), Urdu (ur), Persian (fa), Malay (ms), Indonesian (id), and **seven domains**. For each Language, we sample from public OPUS corpora.

Lan.	OpenSubtitles	TED2020	QED	Tanzil	wikimedia	WikiMatrix	Europarl	WMT-News	TEP	Size
hu	✓	✓	✓			✓	✓	✓		1.2K
ms	✓	✓	✓	✓	✓					1.1K
ur	✓	✓	✓	✓	✓					0.8K
bn	✓	✓	✓	✓		✓				0.9K
fa		✓	✓	✓		✓			✓	0.9K
id	✓	✓	✓	✓		✓				1.2K

Table 1: Source corpora and domain coverage of X-Bench across language subsets. Checkmarks indicate whether a source corpus is included in each language subset; detailed domain links are provided in the Appendix (see Table 11).

COLING2027 under review

Dual-RAG: Word Sense Disambiguation

- Experiment Results

Method	WMT		IWSLT		
	bn	ne	bn	uz	vi
Direct	78.11	80.36	71.13	61.32	76.86
RAG_only	77.74	79.16	73.24	63.50	77.12
Threshold-RAG	78.08	79.77	73.11	62.87	77.05
Filter-RAG	77.92	79.81	73.14	62.87	77.05
ICES	78.12	78.61	71.11	61.17	76.83
CoD	78.12	80.21	71.11	61.17	76.83
CompTra	77.98	80.05	71.12	61.19	76.88
Ours	**78.30***	**80.37**	**74.34***	**64.47***	**77.79***

Table 5: Results on WMT and IWSLT. "$\star$" denotes significant improvement over the strongest non-ours baseline ($p < 0.01$). More detailed results are given in Appendix Table 32 and Table 33.

Item	Content
Source (fi)	Kirja täytyy saada ensin valmiiksi ja sitten tehdään **käsikirjoitus**.
Reference	Well, I've gotta write the book first, John. Then, you know, they get somebody to write the **screenplay**.
Ambiguity	**käsikirjoitus**
Sense contrast	**screenplay / script / manuscript**
Direct	The book has to be completed first, and then a **manuscript** will be made. ✗
RAG_only	The book has to be completed first, and then the **manuscript** is made. ✗
Ours	The book has to be completed first, and then a **screenplay** will be written. ✓

Table 9: A MuCoW case illustrating lexical ambiguity in *käsikirjoitus*. Ours resolves the intended sense, while Direct and RAG_only choose the wrong literal sense.

COLING2027 under review

Retrieve, Refine, and Translate: LLM-Based Translation for Low-Resource Languages

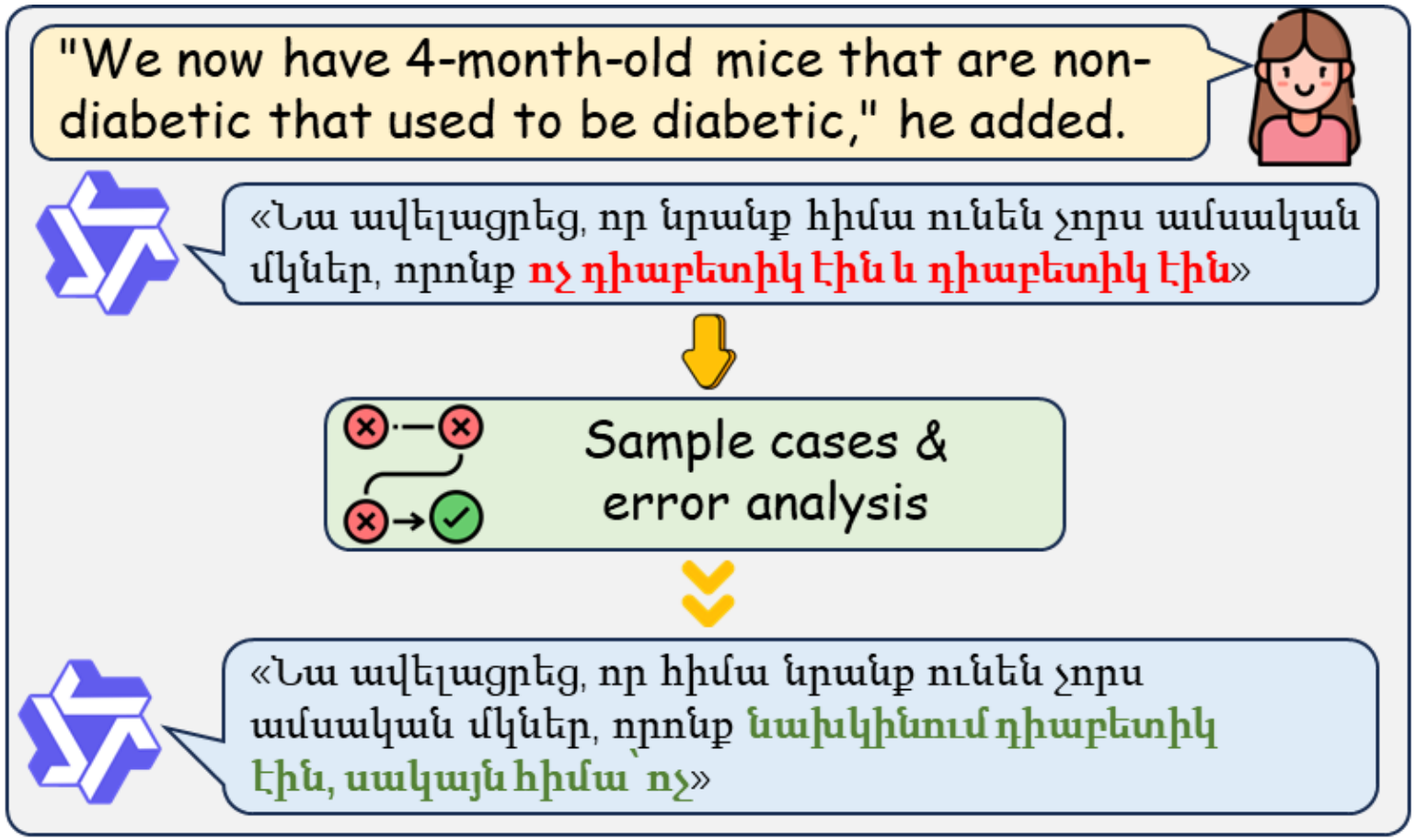
Shibo Zhang[1,2,3,4], Mieradilijiang Maimaiti[1,2,3,4,*], Zhengyi Guo[1,2,3,4], Dezhi Wang[1,2,3,4], Wu Le[5], Zhuofei Xie[5], Jiawei Chen[5], Wushouer Silamu[1,2,3,4],

[1]School of Computer Science and Technology, Xinjiang University, Urumqi, China
[2]Xinjiang Laboratory of Multi-Language Information Technology, Xinjiang University, Urumqi, China
[3]Xinjiang Multilingual Information Technology Research Center, Urumqi, China
[4]International Research Laboratory of Silk Road Multilingual Cognitive Computing, Urumqi, China
[5]Integrated Laboratory for Space, Air, and Ground Systems
Email: {zhangshibo, guozhengyi, kyriezhizhi}@stu.xju.edu.cn, {miradeljan51, wushour}@xju.edu.cn, alaia@richnetworks.hk, zhx34@pitt.edu, syscjw@zohomail.cn

Zhang et al., IJCNN 2026

RAG and Self-Refined Knowledge

- Motivation



- Although LLMs demonstrate promising translation and **self-refinement capabilities**, their performance remains severely constrained in low-resource scenarios.
- Leveraging external knowledge via **RAG** assists LLM translation, but **relying** solely **on** contextual parallel examples and single-round generation is insufficient to resolve diverse errors.

Zhang et al., IJCNN 2026

RAG and Self-Refined Knowledge

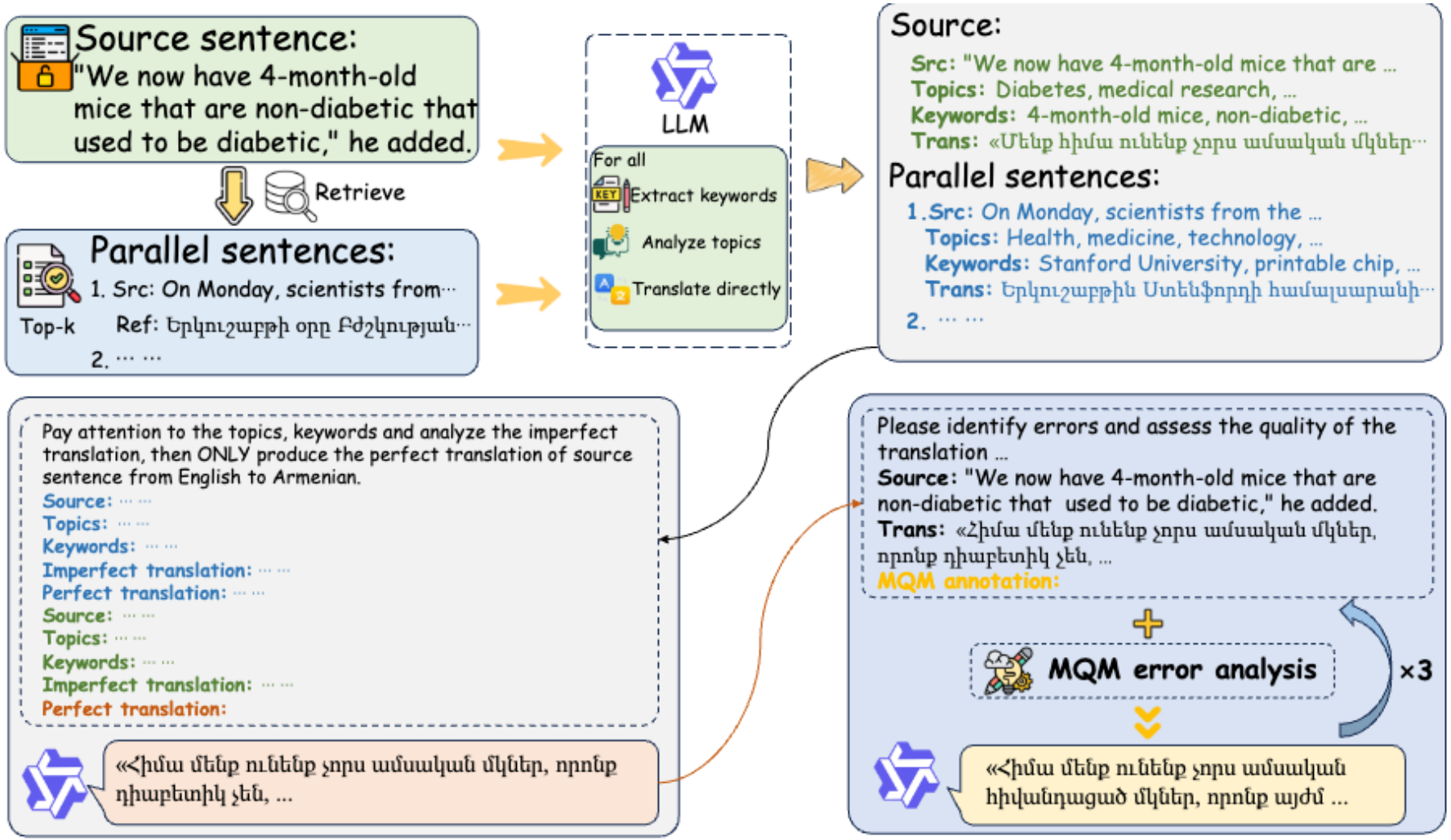


Fig. 2. Overview of the proposed framework. It integrates Implicit Refinement using retrieved knowledge and semantic constraints with Explicit Refinement via iterative ($T = 3$) MQM-based self-correction.

Zhang et al., IJCNN 2026

RAG and Self-Refined Knowledge

- Experiment
  - Main experiment
    - Main experimental results on the FLORES-200, NTREX-128, and TICO-19 with XCOMET-XL and BLEURT-20 metrics.

TABLE III
RESULTS ON THE FLORES-200 DATASET (ENGLISH → XX) EVALUATED BY XCOMET AND BLEURT

Methods	Armenian		Azerbaijani		Hebrew		Lao	
	XCOMET	**BLEURT**	**XCOMET**	**BLEURT**	**XCOMET**	**BLEURT**	**XCOMET**	**BLEURT**
0-shot	68.57	72.56	64.92	62.45	70.72	67.34	49.49	62.06
Vanilla RAG	71.47	74.90	68.63	64.31	71.50	68.06	55.55	67.35
COD	70.83	73.57	66.64	63.04	70.44	66.99	52.22	63.80
MAPS	75.53	76.53	71.31	65.20	76.33	71.27	57.05	**68.00**
TEaR	71.19	73.66	67.66	63.29	73.42	68.94	51.36	63.61
CompTra	61.11	62.71	67.32	63.32	70.74	67.53	49.99	52.54
Ours	**76.56**	**76.87**	**72.15**	**65.58**	**78.57**	**72.07**	**57.65**	67.64

Methods	Khmer		Tamil		Urdu		Bengali	
	XCOMET	**BLEURT**	**XCOMET**	**BLEURT**	**XCOMET**	**BLEURT**	**XCOMET**	**BLEURT**
0-shot	50.98	57.35	53.01	74.83	66.84	56.38	67.33	73.98
vanilla rag	55.28	60.51	55.10	76.77	68.15	56.66	68.48	74.87
COD	51.24	57.49	53.25	74.60	63.65	55.81	66.22	74.01
MAPS	57.11	61.87	**57.87**	77.51	71.04	**57.56**	71.31	75.81
TEaR	52.93	58.88	55.23	76.42	67.86	56.65	68.44	74.32
CompTra	44.95	44.42	42.03	59.77	64.45	56.23	54.68	63.02
Ours	**57.74**	**61.97**	57.38	**77.75**	**71.52**	56.95	**71.45**	**75.94**

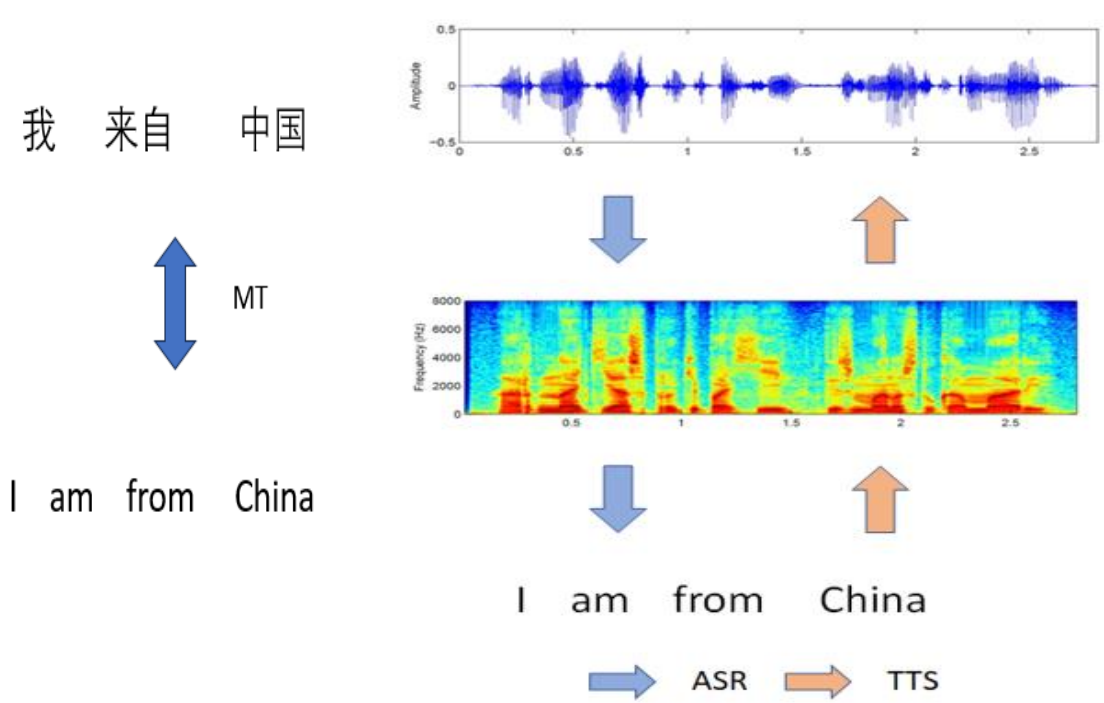
Zhang et al., IJCNN 2026

Outline

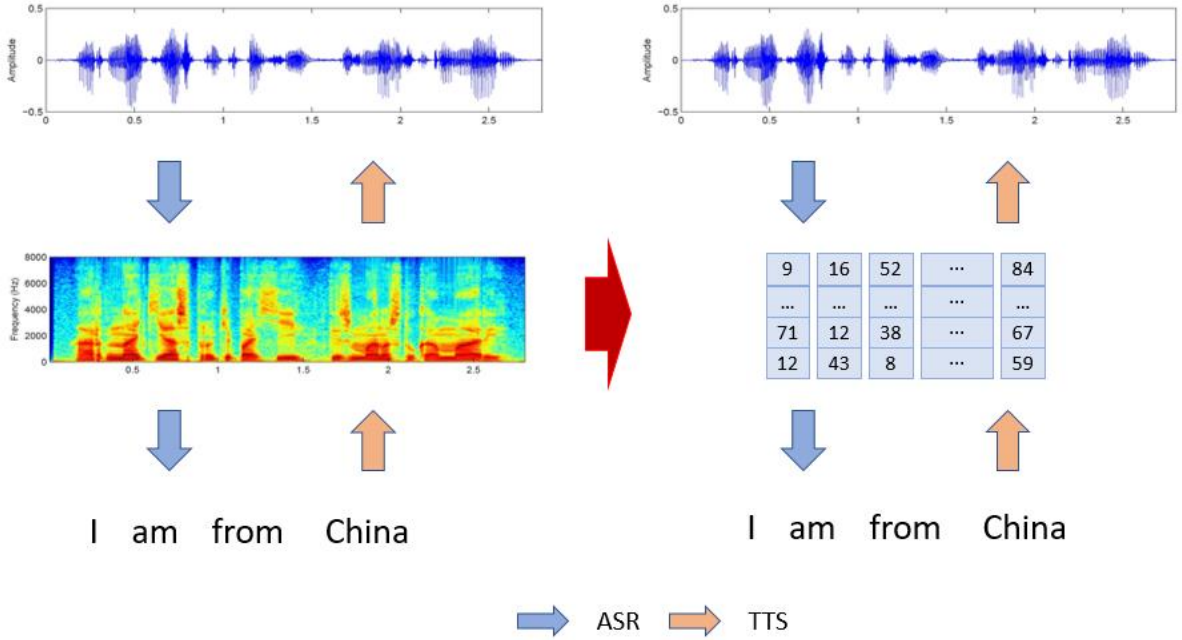
- 大语言模型时代的低资源语言处理
- 面向低资源语言的大语言模型机器翻译探索
- 多模态大语言模型驱动的低资源翻译探索
- 面向低资源语言大模型的未来探索
- 总结

Findings on Multi-modal LLM-based MT

Multi-modal LLM-based MT --- Speech



Speech Processing vs NLP



From Continuous Signals to Discrete Tokens

Prof. Liu CCMT2023

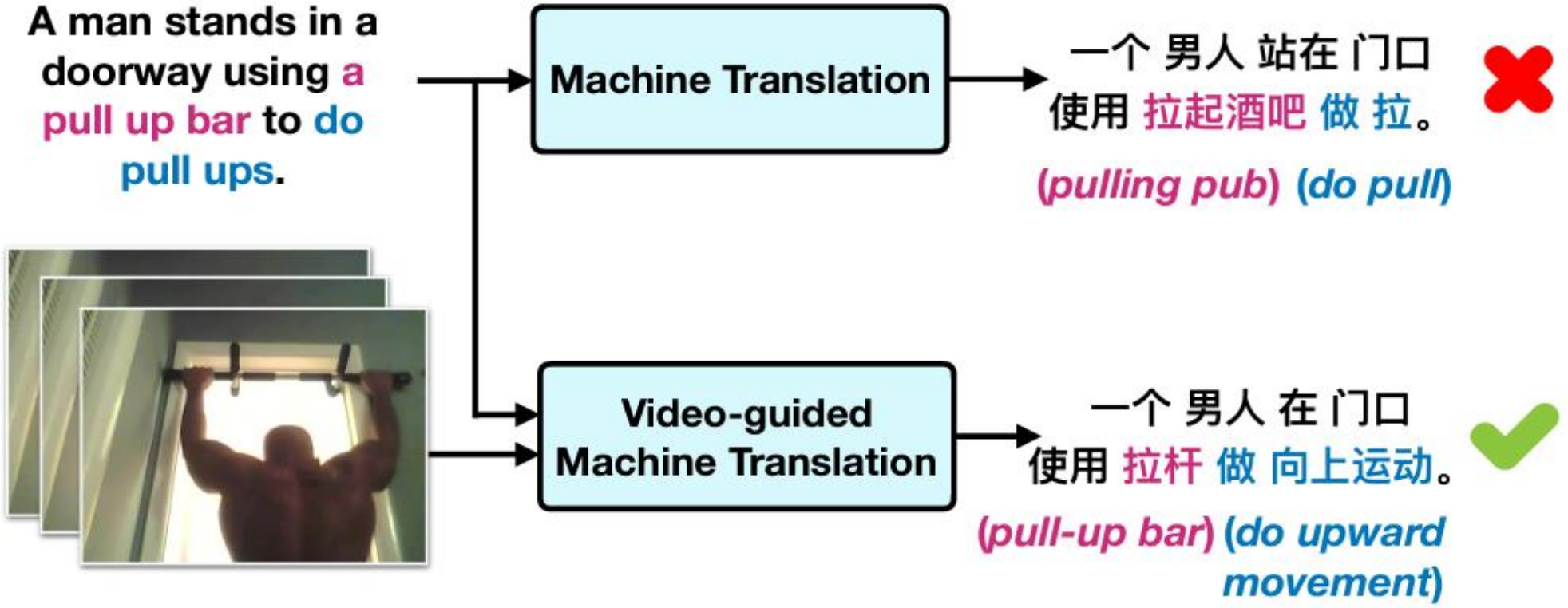
Multi-modal LLM-based MT --- Image

Image Captioning Model

There is a teacher teaching his students in the classroom.

有一个老师在教室里教他的学生。

Multi-modal LLM-based MT --- Video



Prof. Wang CCMT2023

Vision-to-Text: Benchmarking Multimodal LLMs on Extremely Low-Resource Languages

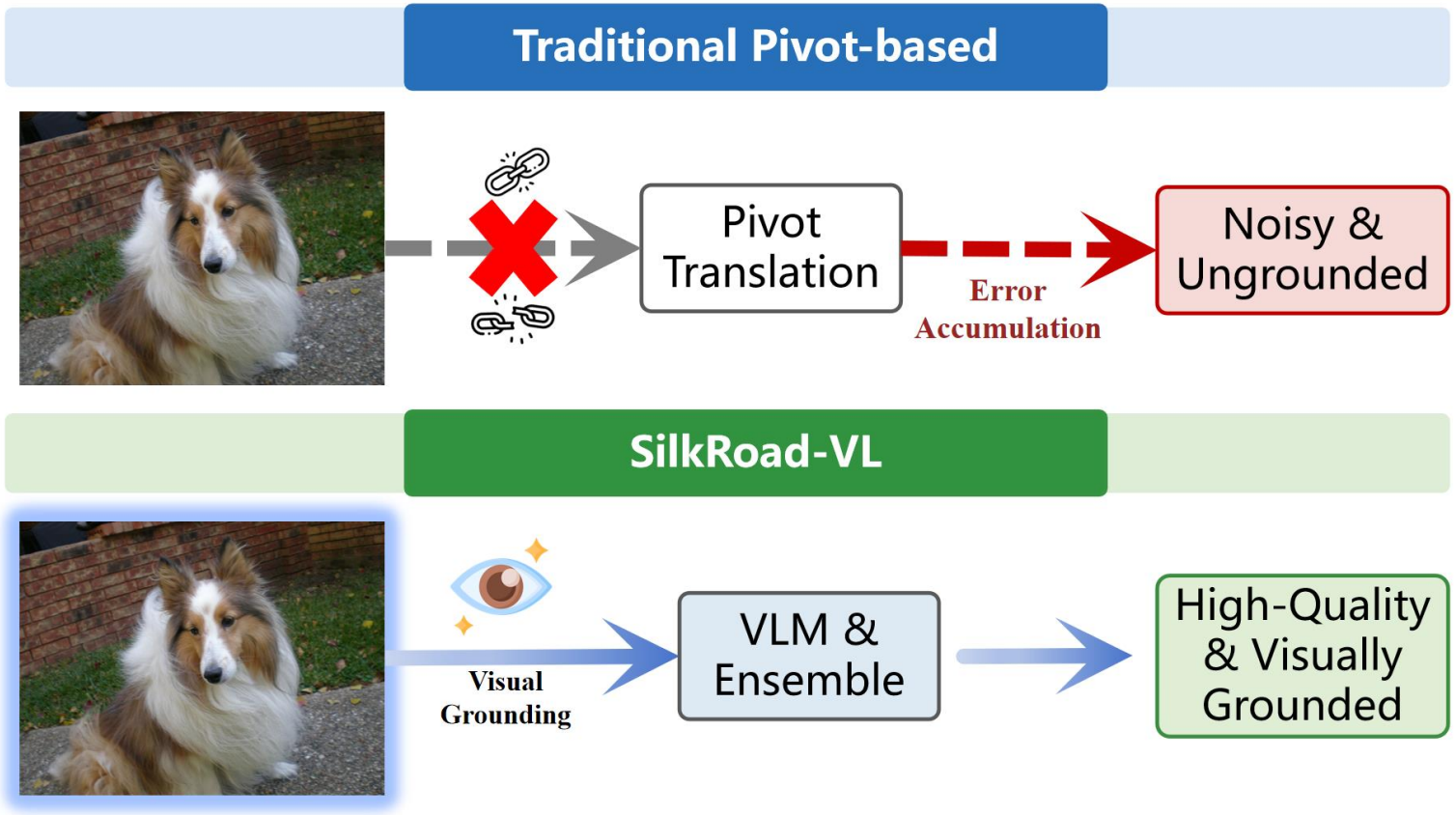
Shuoshuo Hou[1,2,3,4], Mieradilijiang Maimaiti[1,2,3,4,*], Zhexin Li[1,2,3,4], Jiaxin Wang[1,2,3,4], Ahmad Hassan[1,2,3,4], Kaishaer Jiapaer[1], Nilufar Abdurakhmonova[5], Roza Urinbayeva[5], Madina Mansurova[6], Shormakova Assem[6], Gulnar Murat[7], Le Wu[8], and Wushouer Silamu[1,2,3,4]

[1]School of Computer Science and Technology, Xinjiang University, Urumqi, China
[2]Xinjiang Laboratory of Multi-Language Information Technology, Xinjiang University, Urumqi, China
[3]Xinjiang Multilingual Information Technology Research Center, Urumqi, China
[4]Joint International Research Laboratory of Silk Road Multilingual Cognitive Computing, Urumqi, China
[5]Department of Computational Linguistics, National University of Uzbekistan, Tashkent, Uzbekistan
[6]Faculty of Information Technology and Artificial Intelligence, Al-Farabi Kazakh National University, Almaty, Kazakhstan
[7]Kapshagay Bidai Onimderi, Kapshagay, Kazakhstan
[8]Integrated Laboratory for Space, Air, and Ground Systems, Changji, China
Email: {yumoya, lizhexin, 107552503749, k20241401814_}@stu.xju.edu.cn,
{miradeljan51, wushour}@xju.edu.cn, {n.abduraxmonova, orinboyeva_r}@nuu.uz,
ahmadhassan2067@gmail.com, madina.mansurova@kaznu.edu.kz,
Shormakova.asem1@kaznu.edu.kz, DrGulnarMurat@hotmail.com, alaia@richnetworks.hk

Hou et al., IJCNN 2026

Constructing Multi-modal Datasets for LRLs

- Motivation



- Traditional pivot translation accumulates errors and lacks visual grounding, resulting in noisy and less accurate translations.
- We introduce visual information to directly support translation, improving translation quality and image-text consistency.

Hou et al., IJCNN 2026

Constructing Multi-modal Datasets for LRLs

- Experiment
  - Dataset statistics

Table 9: Human evaluation results on SilkRoad-VL.

Language	Fluency	Adequacy	Visual Relevance
ug	4.61	4.70	4.73
uz	4.10	4.03	4.74
kk	4.33	4.33	4.67
ur	4.80	4.19	4.69
Average	**4.46**	**4.31**	**4.71**

Table 2: Statistics of the SilkRoad-VL dataset, grouped by region.

Language	Total	Short		Long	
		Count	Length	Count	Length
Central Asia					
kk	26.1k	13.0k	14.8	13.2k	31.6
ky	23.2k	12.8k	15.4	10.3k	33.0
tg	2.4k	1.8k	17.0	0.6k	40.0
ug	8.7k	3.7k	15.7	5.0k	34.7
uz	14.9k	8.2k	15.2	6.7k	33.0
South Asia					
bn	34.4k	16.0k	16.0	18.4k	37.8
ur	9.6k	5.1k	23.3	4.5k	56.6
Middle East					
fa	36.3k	18.1k	20.7	18.1k	51.3
he	30.8k	16.1k	14.7	14.7k	31.9
Southeast Asia					
id	38.6k	19.0k	17.3	19.7k	38.8
vi	37.3k	18.5k	23.1	18.8k	55.8
Total / Average	**262.3k**	**132.3k**	**17.6**	**130.1k**	**41.6**

Hou et al., IJCNN 2026

Constructing Multi-modal Datasets for LRLs

- Experiment

Table 4: Results on the SilkRoad-VL test set for languages outside Central Asia using BERTScore (BS), COMET-Kiwi (CK), and CLIPScore (CS). Underlined values denote the best zero-shot baseline, and bold values denote the best overall result.

Model	South Asia						Middle East						Southeast Asia					
	bn			ur			he			fa			id			vi		
	BS	CK	CS	BS	CK	CS	BS	CK	CS	BS	CK	CS	BS	CK	CS	BS	CK	CS
Qwen3-VL	92.22	77.01	16.81	90.61	67.83	16.55	93.83	74.66	18.48	94.33	79.67	16.78	97.03	83.55	22.95	95.89	83.51	20.47
Qwen2.5-VL	89.72	63.69	16.89	87.46	54.11	17.34	91.88	63.58	18.60	93.04	74.65	16.08	95.43	80.41	22.70	95.06	82.33	20.46
LLaVA-v1.6	75.07	31.27	**18.46**	79.29	29.81	16.64	80.84	27.46	18.34	82.42	29.07	**17.23**	91.38	56.01	23.29	88.94	47.41	21.03
LLaVA-1.5	77.95	34.08	17.19	82.26	32.22	17.38	83.34	30.93	18.25	84.71	34.41	17.18	93.50	64.35	**23.89**	92.53	61.65	**21.09**
InternVL3	87.23	53.61	16.72	84.62	40.28	**17.42**	90.64	55.21	18.30	89.95	53.17	16.97	95.51	65.30	22.52	95.42	71.51	20.63
Qwen3-Text	92.67	78.24	16.83	91.68	69.88	17.32	94.27	74.30	18.55	94.90	79.59	16.67	97.22	83.96	23.18	96.36	**83.79**	20.49
LLaVA-v1.6+FT	95.09	73.80	16.74	95.16	66.60	17.27	97.61	71.81	18.39	96.99	71.58	16.71	98.32	71.61	22.83	97.69	72.72	20.64
Ours	**95.43**	**82.22**	16.71	**95.52**	**79.58**	17.23	**97.76**	**82.96**	**18.65**	**97.29**	**83.70**	17.04	**98.48**	**84.23**	22.88	**97.87**	83.55	20.52

Hou et al., IJCNN 2026

RAG + Multi-Modal Fusion

- Motivation
  - Directly injecting visual information into multimodal translation is often unstable and may introduce noise.
  - In low-resource settings, text-only retrieval is **weak** at **disambiguating** short or templated sentences.
  - Therefore, we propose using Visual Neighbor Memory as **auxiliary evidence** for disambiguation rather than letting vision dominate generation.

NAACL2027 under review

RAG + Multi-Modal Fusion

- Experiment

Method	Short-Caption Retrieval				Low-Resource Transfer			Distant LR
	VG test				LoResMT			UMMT
	bn	**hi**	**ml**	**ha**	**ha**	**kr**	**ti**	**uy**
Pure-text								
Direct (Yang et al. 2025)	57.79	61.73	51.65	32.91	29.63	19.58	25.35	29.63
CoD (Lu et al. 2024)	57.36	60.68	51.72	33.19	29.93	19.74	26.27	29.69
CompTra (Zebaze, Sagot, and Bawden 2025)	57.37	60.70	51.69	33.12	29.92	24.65	26.09	29.69
MAPS (He et al. 2024)	58.85	62.27	55.20	34.91	31.33	22.39	27.63	33.97
Visual prompting								
DeDiT-style Prompting (Liu et al. 2025)	55.01	61.41	47.49	32.72	27.04	20.85	18.17	18.16
Direct multimodal LLM								
Qwen3-VL-8B-Instruct (Bai et al. 2025)	55.60	57.27	46.09	32.28	28.26	19.54	17.11	22.65
Qwen3-VL-32B-Instruct (Bai et al. 2025)	59.20	60.77	50.76	32.86	26.63	18.11	19.17	27.97
InternVL3.5-38B-Instruct (Wang et al. 2025)	55.94	62.11	48.33	31.41	26.63	18.63	16.84	18.45
Our variants								
Text-only	62.16	62.50	58.30	43.96	33.19	24.05	27.48	37.72
Fusion ($\alpha = 0.9$)	62.11	62.64	58.11	43.81	33.20	24.03	27.26	37.66
VNM (ours)	**63.36**	**63.81**	**59.75**	**44.32**	**34.07**	**25.13**	**28.21**	**39.92**

Table 1: Main results on the core English→low-resource benchmarks (Avg only). Best results are in bold, and second-best results are underlined. Full metrics are reported in Tables S8–S11 of the Supplementary Material.

NAACL2027 under review

RAG + Multi-Modal Fusion

- Case study

Image

Source Sentence
men playing baseball

Reference
बेसबॉल खेल रहे पुरुष *(men playing baseball.)*

Direct Output
क्रिकेट खेलते हुए पुरुष *(men playing cricket.)*

Text-RAG Output
महिलाएं बेसबॉल खेल रही हैं। *(women playing baseball.)*

Image-RAG Output
महिलाएं बेसबॉल खेल रही हैं। *(women playing baseball.)*

VNM-RAG Output
बेसबॉल खेल रहे पुरुष ✓ *(men playing baseball.)*

Hindi

Image

Source Sentence
A man is grilling out in his backyard.

Reference
بىر ئەر ئۆزىنىڭ ئارقا ھويلىسىدا كاۋاپ پىشۇرىۋاتىدۇ. *(a man is grilling in his backyard.)*

Direct Output
ئەتەكى ئۆيەكىنىڭ ئالدىدا چاچىلا. *(incorrect translation.)*

Text-RAG Output
ئەتە ئۆيىنىڭ ئىچىدە چاچىلاپ تۇرۇ. *(incorrect translation.)*

Image-RAG Output
ئەتە ئۆيىنىڭ ئىچىدە چاچىلاپ تۇرۇ. *(incorrect translation.)*

VNM-RAG Output
بىر ئەر ئارقا ھويلىدا كاۋاپ پىشۇرۇۋاتىدۇ. *(a man is grilling in the backyard.)* ✓

Uyghur

NAACL2027 under review

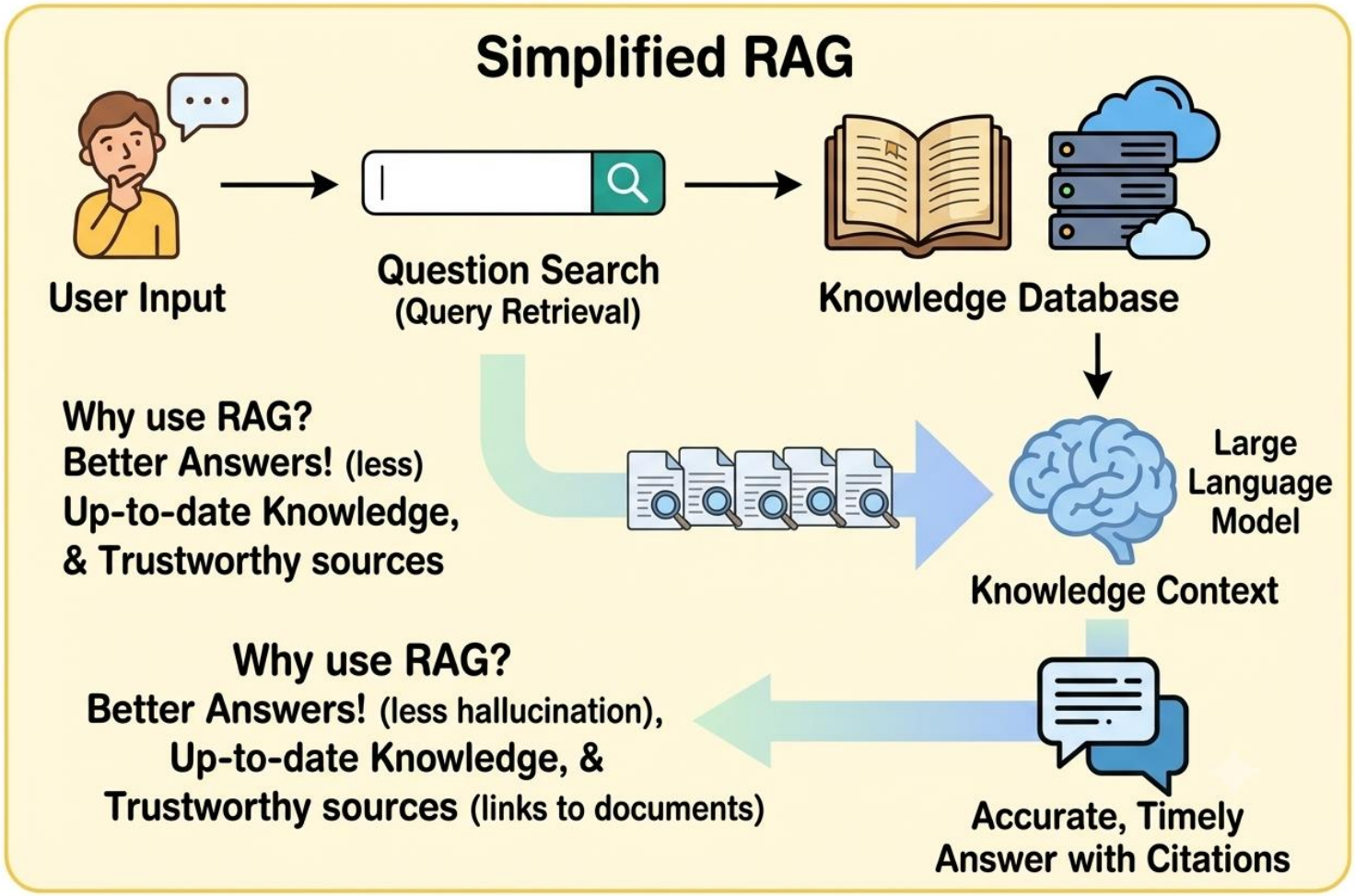
Outline

- 大语言模型时代的低资源语言处理
- 面向低资源语言的大语言模型机器翻译探索
- 多模态大语言模型驱动的低资源翻译探索
- 面向低资源语言大模型的未来探索
- 总结

Investigations on RAGs

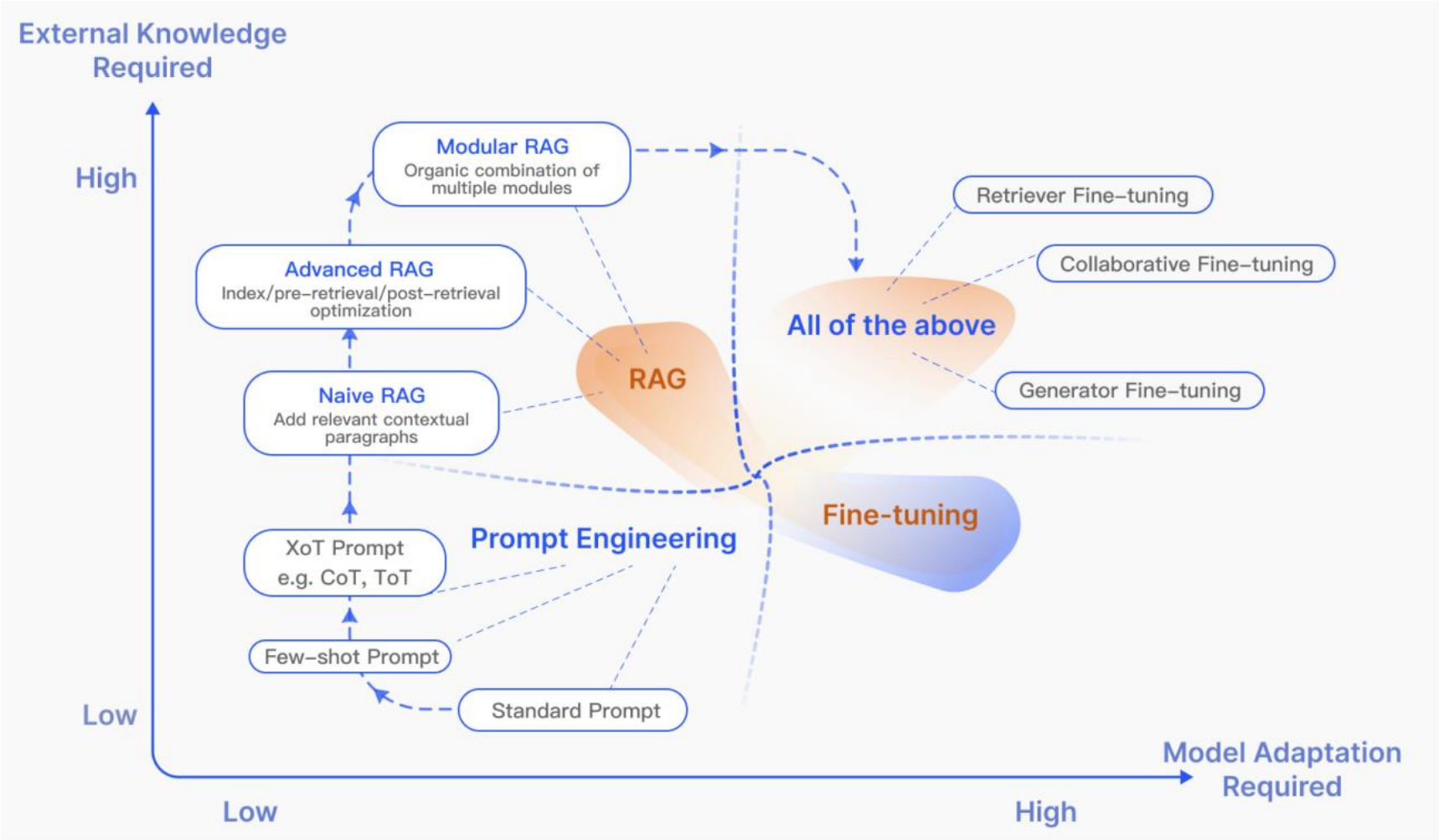
What is RAG

- RAG（Retrieval-Augmented Generation，检索增强生成）是一种将外部知识检索与大语言模型生成相结合的技术，通过先检索相关文档，再基于检索结果生成答案。



RAGs

- RAG与其它提升LLM的方法相比，对高质量的外部知识库具有较高的要求，因为检索到的内容直接参与LLM的思考和回复。



Retrieval-Augmented Generation for Large Language Models: A Survey，Arxiv 2024, Gao et al.

Why RAG

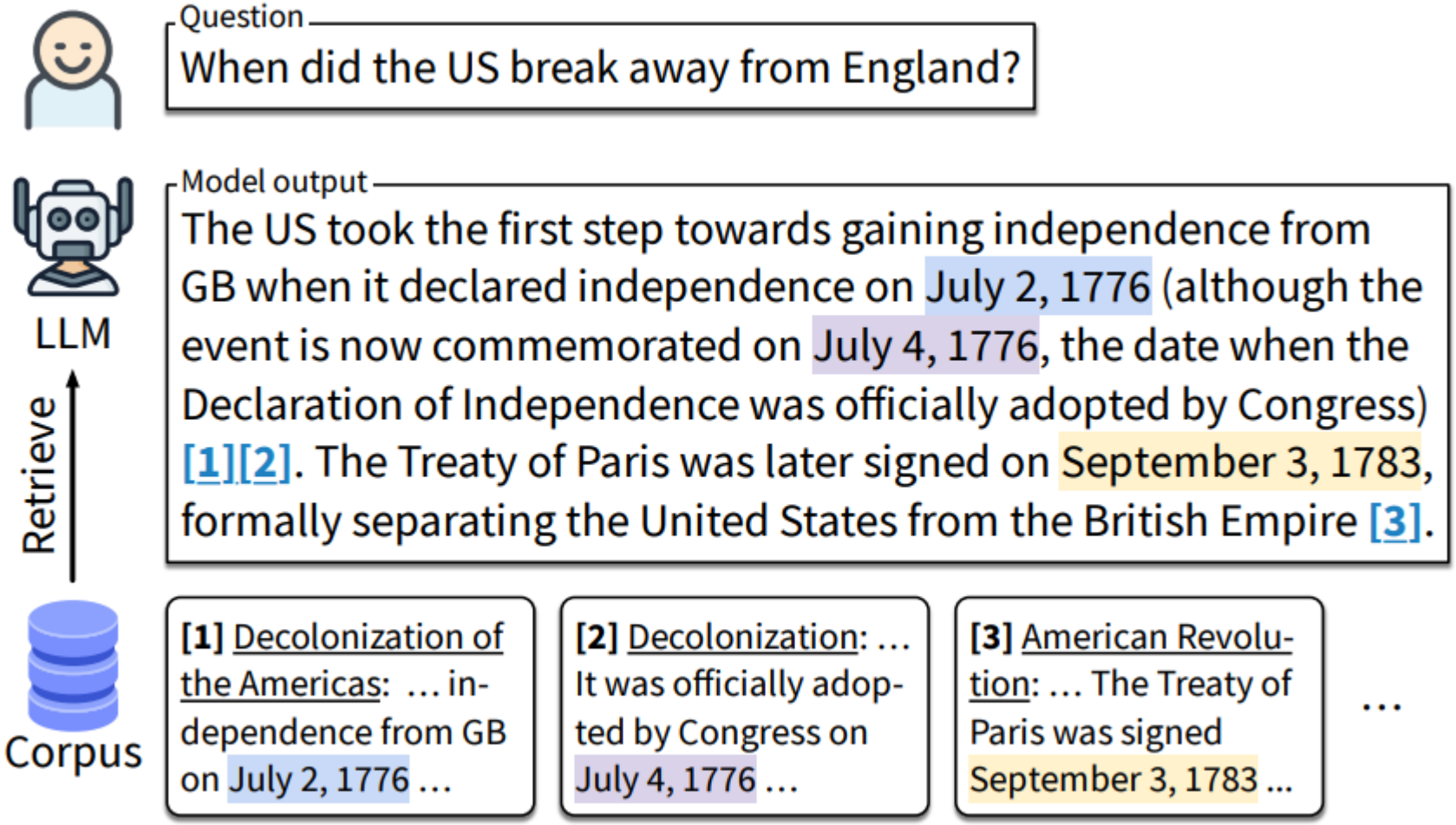
- LLM的知识过时问题：无法掌握快速变化的信息

Type	Question	Answer (as of this writing)
never-changing	*Has Virginia Woolf's novel about the Ramsay family entered the public domain in the United States?*	***Yes**, Virginia Woolf's 1927 novel To the Lighthouse entered the public domain in 2023.*
never-changing	*What breed of dog was Queen Elizabeth II of England famous for keeping?*	***Pembroke Welsh Corgi** dogs.*
slow-changing	*How many vehicle models does Tesla offer?*	*Tesla offers **six** vehicle models: Model S, Model X, Model 3, Model Y, Tesla Semi, and Cybertruck.*
slow-changing	*Which team holds the record for largest deficit overcome to win an NFL game?*	*The record for the largest NFL comeback is held by the **Minnesota Vikings**.*
fast-changing	*Which game won the Spiel des Jahres award most recently?*	***Dorfromantik** won the 2023 Spiel des Jahres.*
fast-changing	*What is Brad Pitt's most recent movie as an actor*	*Brad Pitt is credited as Keith in **IF**.*
false-premise	*What was the text of Donald Trump's first tweet in 2022, made after his unbanning from Twitter by Elon Musk?*	*He **did not tweet** in 2022.*
false-premise	*In which round did Novak Djokovic lose at the 2022 Australian Open?*	*He **was not allowed to play** at the tournament due to his vaccination status.*

FRESHLLMS: REFRESHING LARGE LANGUAGE MODELS WITH SEARCH ENGINE AUGMENTATION, Arxiv 2023, Vu et al.

Why RAG

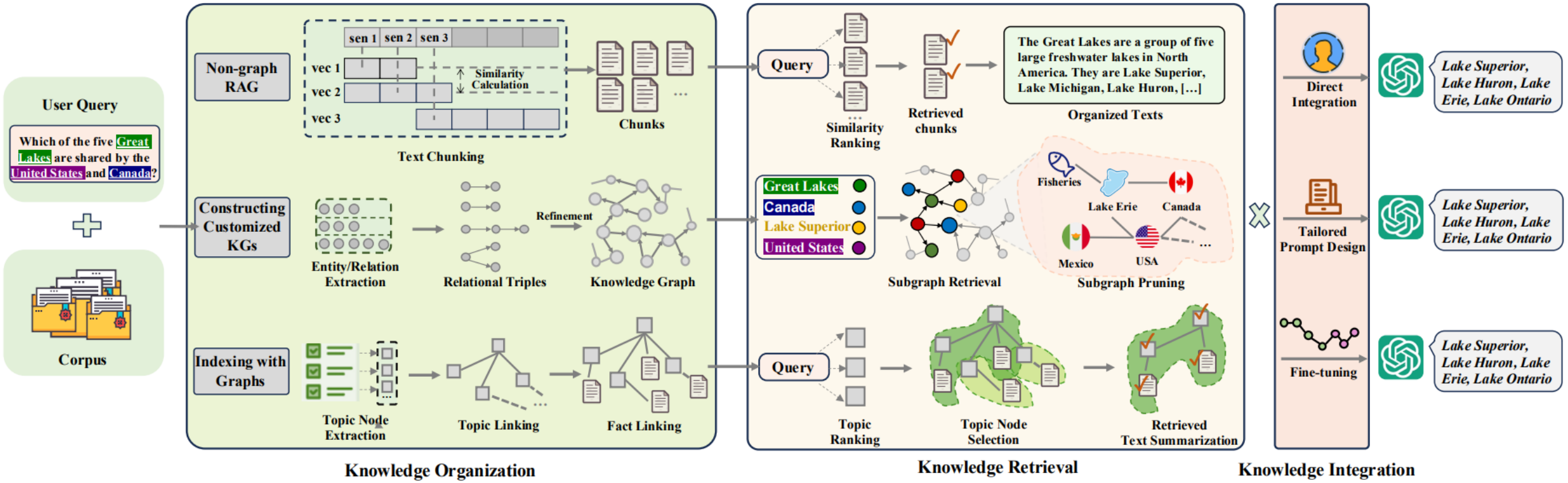
- 答案缺少来源，用户难以验证



Enabling Large Language Models to Generate Text with Citations , EMNLP Main2023, Gao et al.

Previous RAGs

- 早期 RAG 采用“检索—拼接—生成”的线性流程，从分块语料中检索 Top-(k) 文本作为外部知识输入模型。随后研究逐步发展到优化分块与检索，并引入层次化索引和知识图谱，以支持跨文档聚合和多跳推理。



A Survey of Graph Retrieval-Augmented Generation for Customized Large Language Models, Arxiv 2025, Zhang et al.

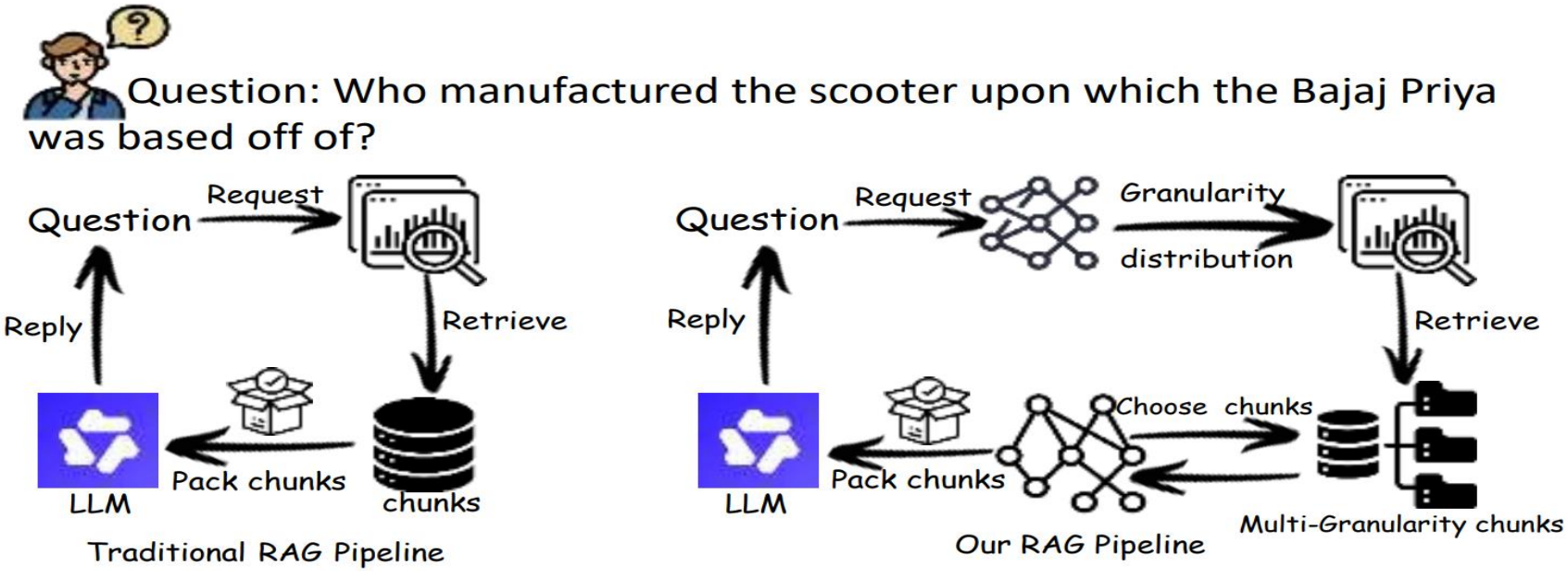
DMG-RAG: Dynamic Multi-Grained Retrieval Augmented Generation for Multi-Hop QA

Yu Pei[1,2,3,4], Mieradilijiang Maimaiti[1,2,3,4,*], Yi Chen[1,2,3,4], Wu Le[5], Zhuofei Xie[5], Jiawei Chen[5]
[1]School of Computer Science and Technology, Xinjiang University
[2]Xinjiang Laboratory of Multi-Language Information Technology, Xinjiang University
[3]Xinjiang Multilingual Information Technology Research Center
[4]Joint International Research Laboratory of Silk Road Multilingual Cognitive Computing
No. 777, Huarui Street, Shuimogou District, Urumqi 830017, Xinjiang, China
[5]Integrated Laboratory for Space, Air, and Ground Systems
Email: peiyu@stu.xju.edu.cn, miradeljan51@xju.edu.cn, 107552501327@stu.xju.edu.cn,
alaia@richnetworks.hk, zhx34@pitt.edu, syscjw@zohomail.cn

Pei et al., IJCNN 2026

Dynamic RAG

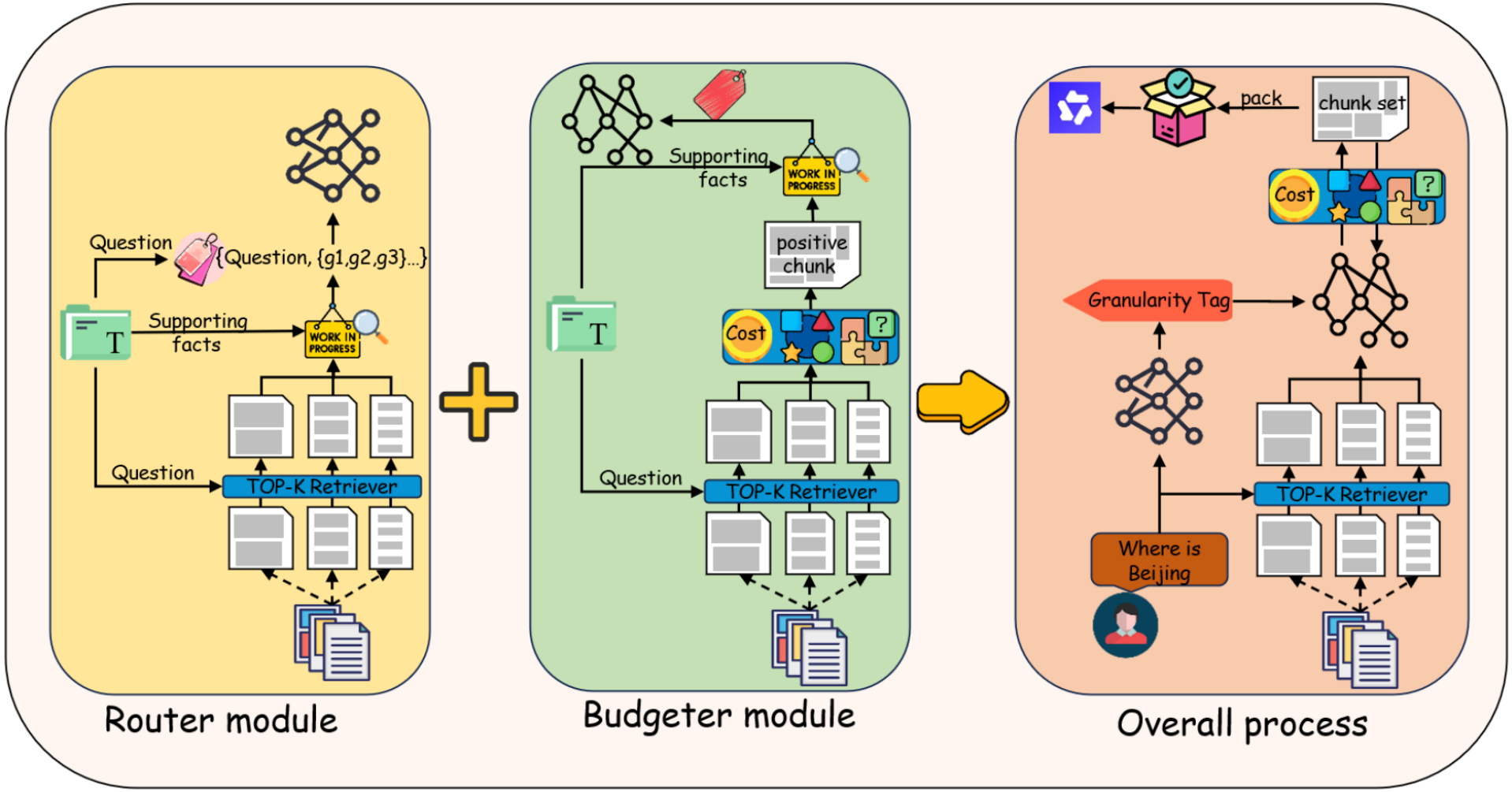
- Motivation
- DMG-RAG针对传统 RAG 使用固定分块粒度、**固定 Top-k 检索**和按排名直接拼接上下文的问题，同时建立句子、段落和文档三种粒度的语料索引。
- 系统根据不同问题动态分配各粒度的检索数量，再在固定 Token 预算内选择相关、互补且低冗余的证据，使最终上下文更适合多跳推理。



Pei et al., IJCNN 2026

Dynamic RAG

- DMG-RAG 整体分为三个部分：语料库的多粒度索引；负责不同粒度tokens配额的 router ；负责选择与问题最相关，花费最小，所选择语句内部最多样的budgeter



Pei et al., IJCNN 2026

Dynamic RAG

- Experiment

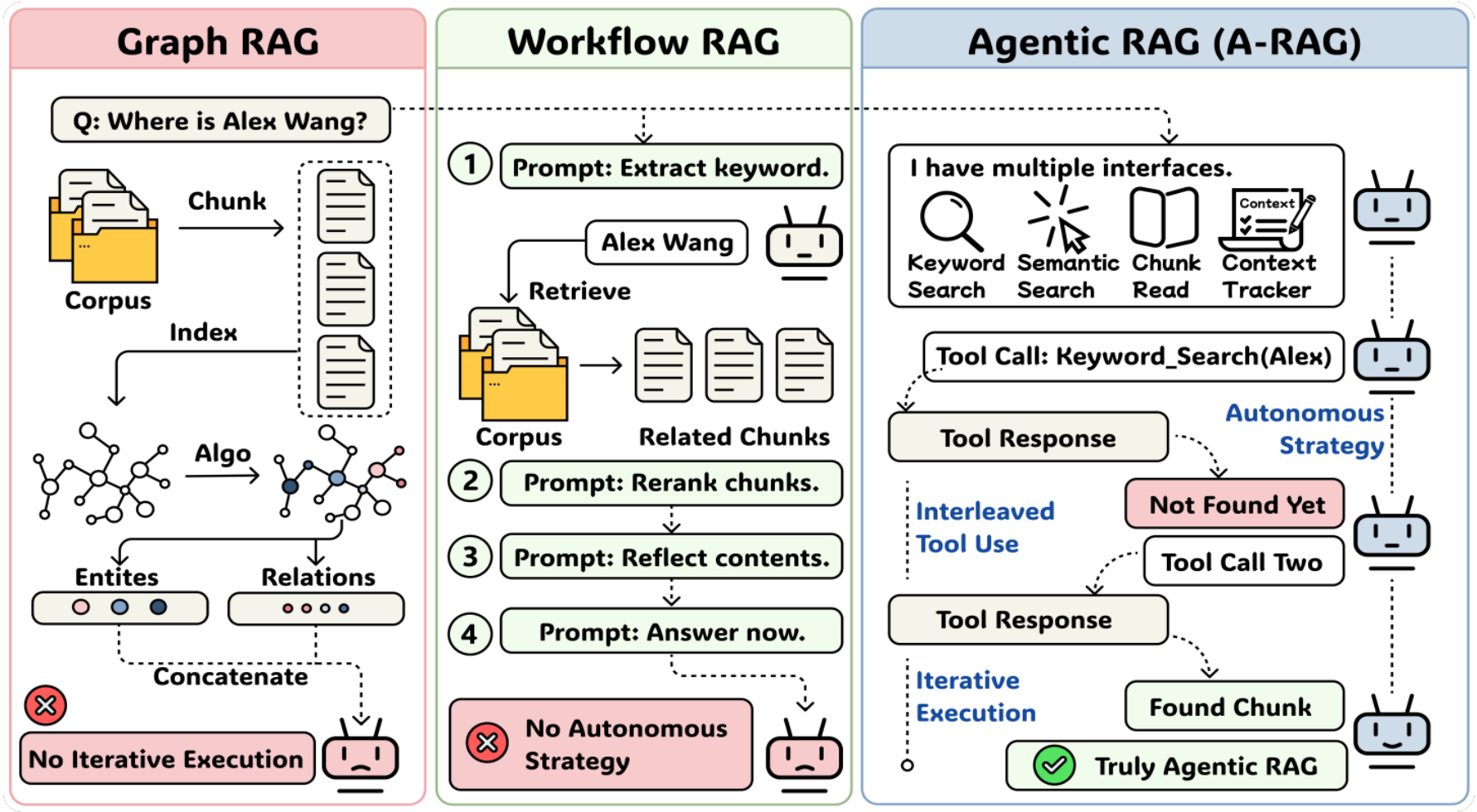
Methods	HotpotQA		2WikiMultiHopQA		Average	
	EM	F1	EM	F1	EM	F1
Direct(qwen2.5-14B) [7]	20.6	27.4	23.2	25.49	21.9	26.5
Naive RAG [1]	53.1	67.1	39	46.5	46.05	56.8
MoGRAG [6]	40.4	52.37	19	23.61	29.7	37.99
GenGroundRAG [27]	36.59	45.91	21.2	26.42	28.9	36.17
Ours	**56.8**	**71.3**	**49.3**	**58.1**	**53.05**	**64.7**

TABLE II: Answer EM and F1 (%) on HotpotQA [3] and 2WikiMultiHopQA [4]. We compare DIRECT (LLM-only, no retrieval) with RAG baselines and our query-adaptive multi-granularity retrieval with budget-aware evidence packing. **Average** reports the mean score over HotpotQA and 2WikiMultiHopQA (higher is better). Underlined values indicate the best baseline performance.

Hou et al., IJCNN 2026

Agentic RAG

- 现阶段 RAG 正由固定的“检索—生成”流程发展为 **Agentic RAG**：LLM 可自主判断何时检索、选择工具、改写查询并评估证据，再通过多轮推理与反馈完成回答。代表性工作包括 **Self-RAG**、**CRAG** 和 **A-RAG**。



A-RAG: Scaling Agentic Retrieval-Augmented Generation via Hierarchical Retrieval Interfaces, Arxiv 2026, Du et al.

Future Works

- Low-resource LLMs: Text-only methods
- Multilingual, Multimodal, LLM -based MT
- Retrieval-Augmented Intelligence: RAG for MT
- Reasoning-based MT
- Tool-augmented LLMs
- ……

Hou et al., IJCNN 2026

Outline

- 大语言模型时代的低资源语言处理
- 面向低资源语言的大语言模型机器翻译探索
- 多模态大语言模型驱动的低资源翻译探索
- 面向低资源语言大模型的未来探索
- 总结

Conclusion

Text-only v.s. Multi-modal LLM-based MT

1.Unified paradigm: LLMs enable multilingual MT with reduced deployment cost.

2.Text-only strategies: Techniques such as CoT, RAG, and word sense disambiguation help compensate for limited parallel data.

3.Multimodal MT: Visual memory and cross-modal fusion provide additional support for low-resource translation.

4.Data bottleneck: High-quality instruction data remains critical; automated pipelines offer a scalable solution.

5.Future directions: Expand instruction corpora for more language pairs, integrate multimodal signals more effectively, and develop better evaluation metrics.

6.Underlying cause: The suboptimal performance of both **unimodal** and **multimodal** LLMs in **low-resource settings** ultimately stems from their **pre-trained foundation models**.

Thank you all Contributors ~

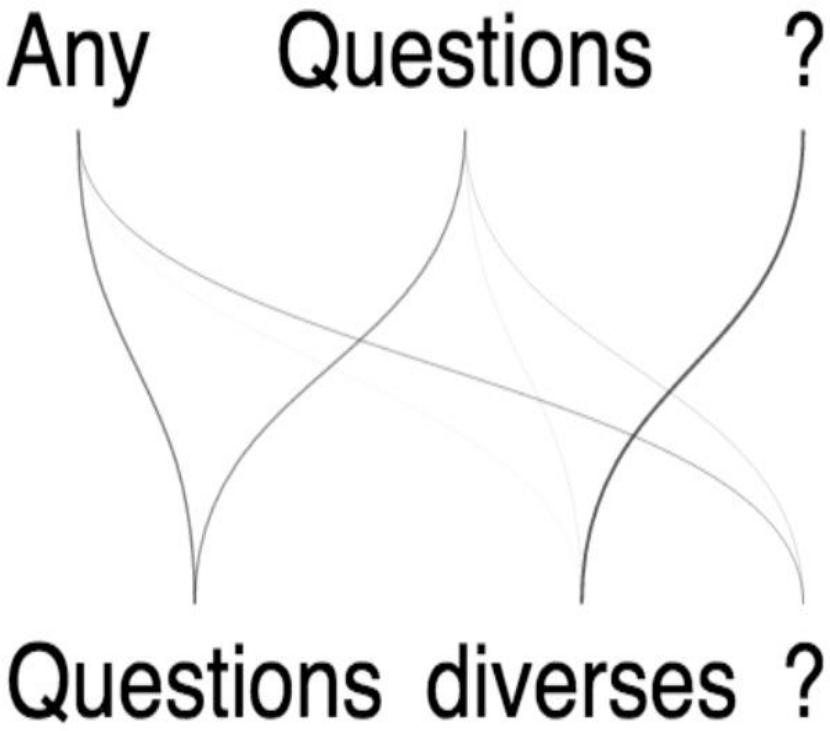






Thank You!

Q & A



This inspiration comes from Dzmitry Bahdanau @ ICLR2014

Join Us!

Email: miradel_51@hotmail.com； miradeljan51@xju.edu.cn

- I am recruiting self-motivated Master's and PhD students to join my research team as Research Interns (RIs).
- I welcome research collaborations with researchers and students worldwide, both online and on-site.

- Computational resources：
  - H100 80G 10
  - 5090 32G 10
  - 4090 24G 10
  - 3090 16G 10