School of Computer Science and Technology
(School of Cyberspace Security)

Large Language Models for Under-Resourced Languages: From Text Understanding to Multimodal Retrieval

Mieradilijiang Maimaiti
https://www.miradeljan.com

MSBC2026,
Tashkent, Uzbekistan
20260923

Outline

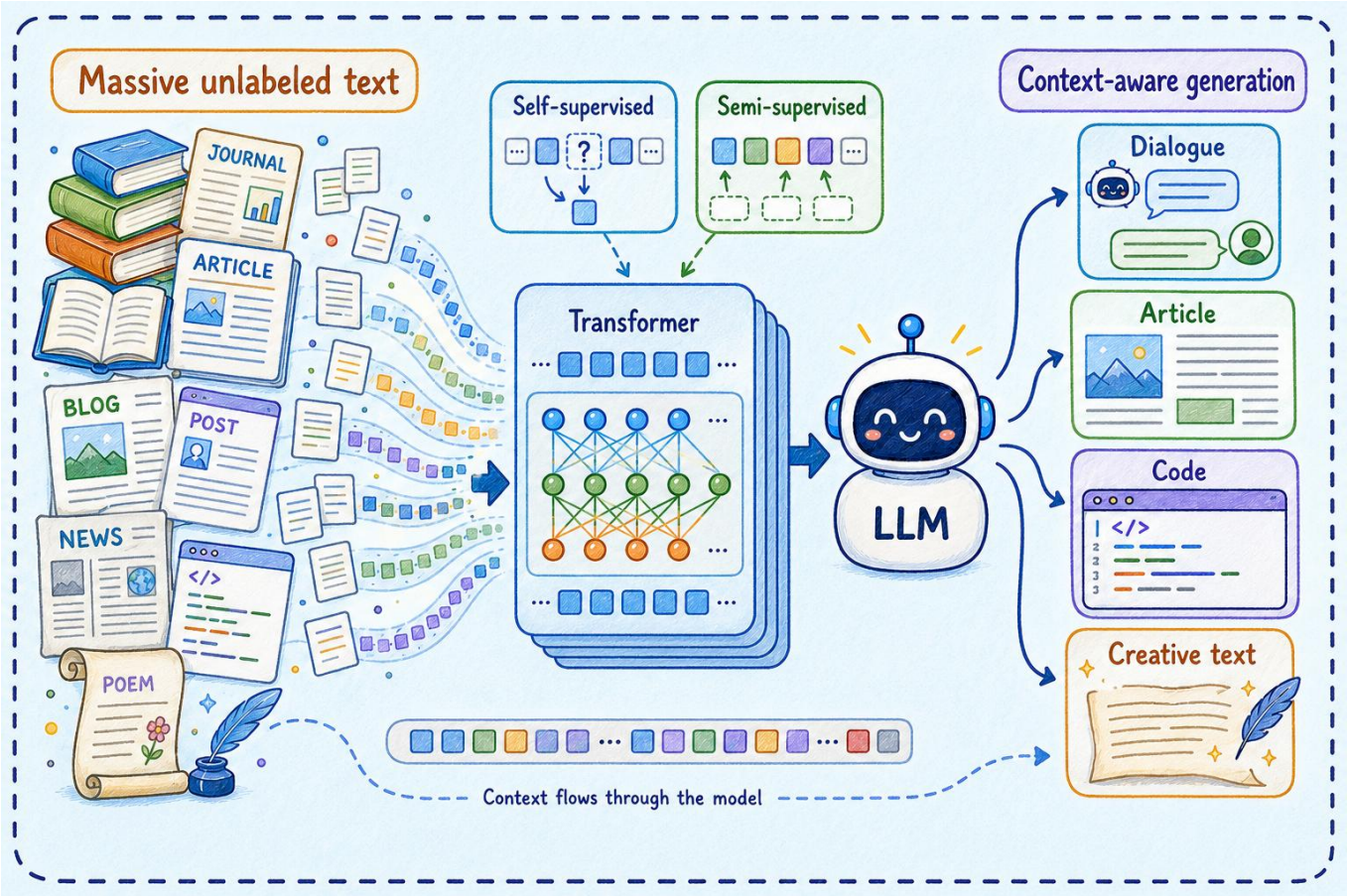
- Low-Resource Language Processing in the Era of LLMs
- Strategies for Text-only LLM-based MT
- Findings on Multi-modal LLM-based MT
- Exploring RAG for Low-Resource Languages
- Conclusion

Low-Resource Language Processing in the Era of LLMs

Actually, the LLM is ….

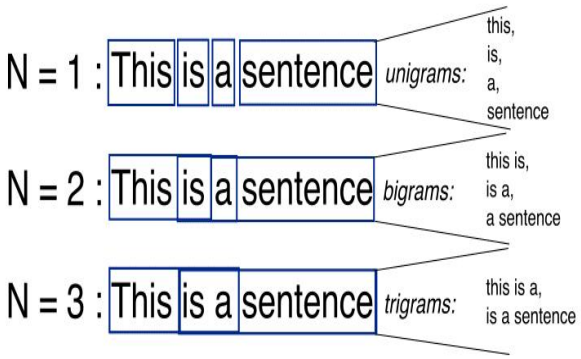
Large language models (LLMs) are based on the **Transformer** architecture and are extensively trained on vast amounts of unlabeled text through **self-supervised** or **semi-supervised learning**. They aim to generate natural-sounding, contextually relevant text in a variety of styles and formats.

Large-scale Pre-trained Language Model

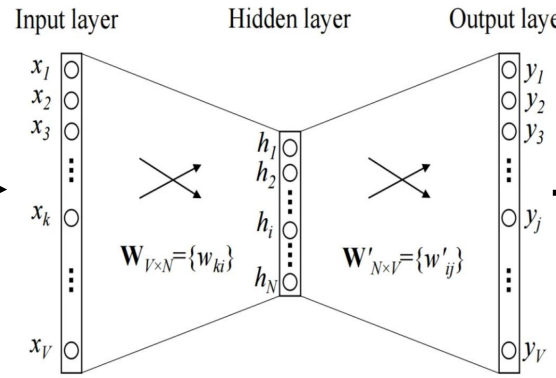


N-gram LM ➡ LLM

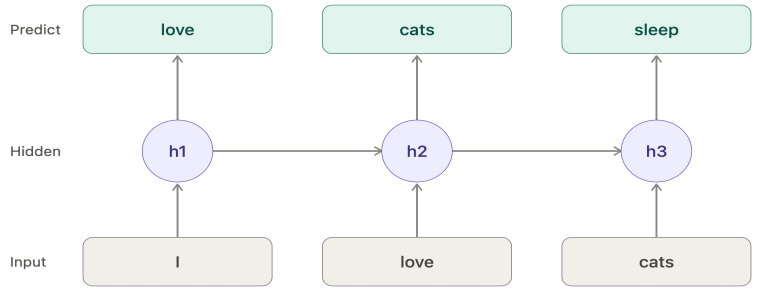
Language Models



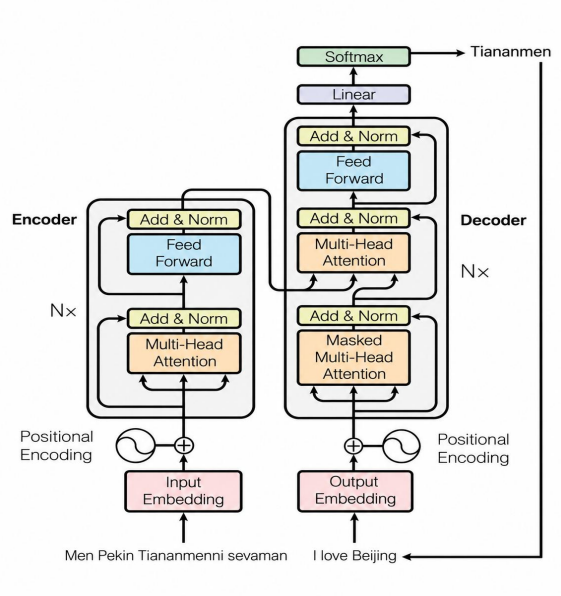
N-gram LM



Word2Vec



RNNLM



Transformer

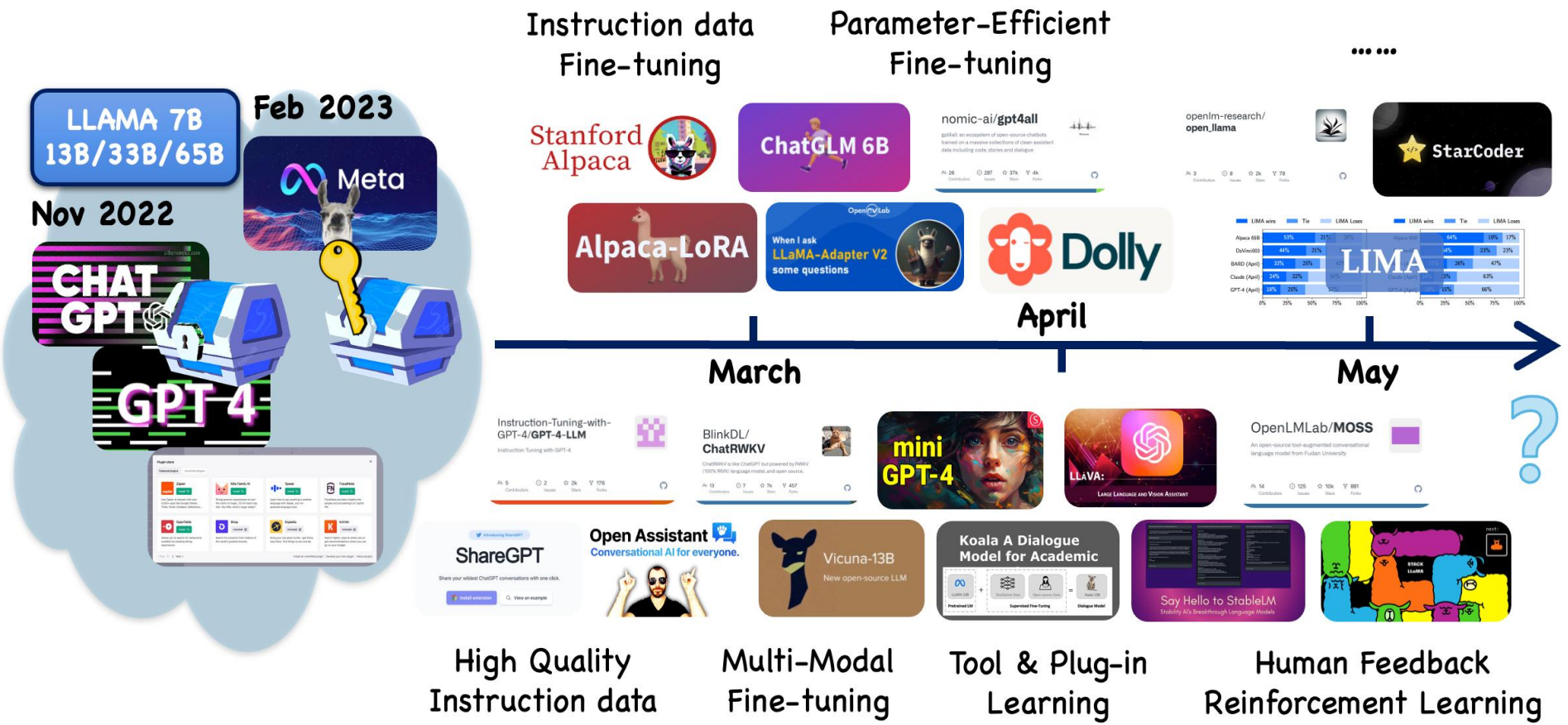
LLM

PLM

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 Language Understanding

Language Understanding

Understand the semantics, intent, and contextual meaning of low-resource languages

Urdu

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

براہ کرم آئندہ ہفتے کی تیاری کیے کریں.

▶ **Interpretation:** Thank the participants and remind them to prepare for next week.

Uzbek

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

▶ **Interpretation:** Thank the participants and remind them to prepare for next week.

2 Instruction Following

Understand and strictly follow user instructions to complete the assigned task

User Instruction (Urdu)

مندرجہ ذیل جملے کو انگریزی میں ترجمہ کریں

میں آج بازار گیا کیونکہ مجھے کچھ سامان خریدنا پڑا۔

Model Output (English)

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

User Instruction (Uzbek)

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

Model Output (Chinese)

I went to the market today because I needed to buy some food.

The 4 Core Capabilities of LLMs

3 In-context Learning

Learn new tasks from a few examples, with strong generalization ability

Example (Urdu → English Translation)

مثال 1: میں گھر جایا ہوں. → I am going home.
مثال 2: وہ کتاب پڑھ رہی ہے. → She is reading a book.
مثال 3: ہمیں کل سفر کرنا ہے. → We have to travel tomorrow.
.....

New Sentence (Urdu) → وہ اپنے دوست سے ملنے پارک گیا.

Model Output (English) → He went to the park to meet his friend.

Example (Uzbek → Chinese Translation)

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。
.....

New Sentence (Uzbek) → U do'kon bilan suhbatlashdi.

Model Output (Chinese) → 他和店主聊了天。

4 Reasoning Ability

Reasoning

Perform logical reasoning, causal analysis, and complex problem solving

Question (Urdu)

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

Model Reasoning Process (Urdu)

1. اس شخص کے پاس لہتذا میں 3 سیب ہیں
2. 2. اس نے اپنے دوست کو 2 سیب دیتے
3. 1 = 3 - 2 = باقی سیب

جواب اس کے پاس 1 سیب باقی رہنے گے.

Question (Uzbek)

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.

Model Reasoning Process (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.

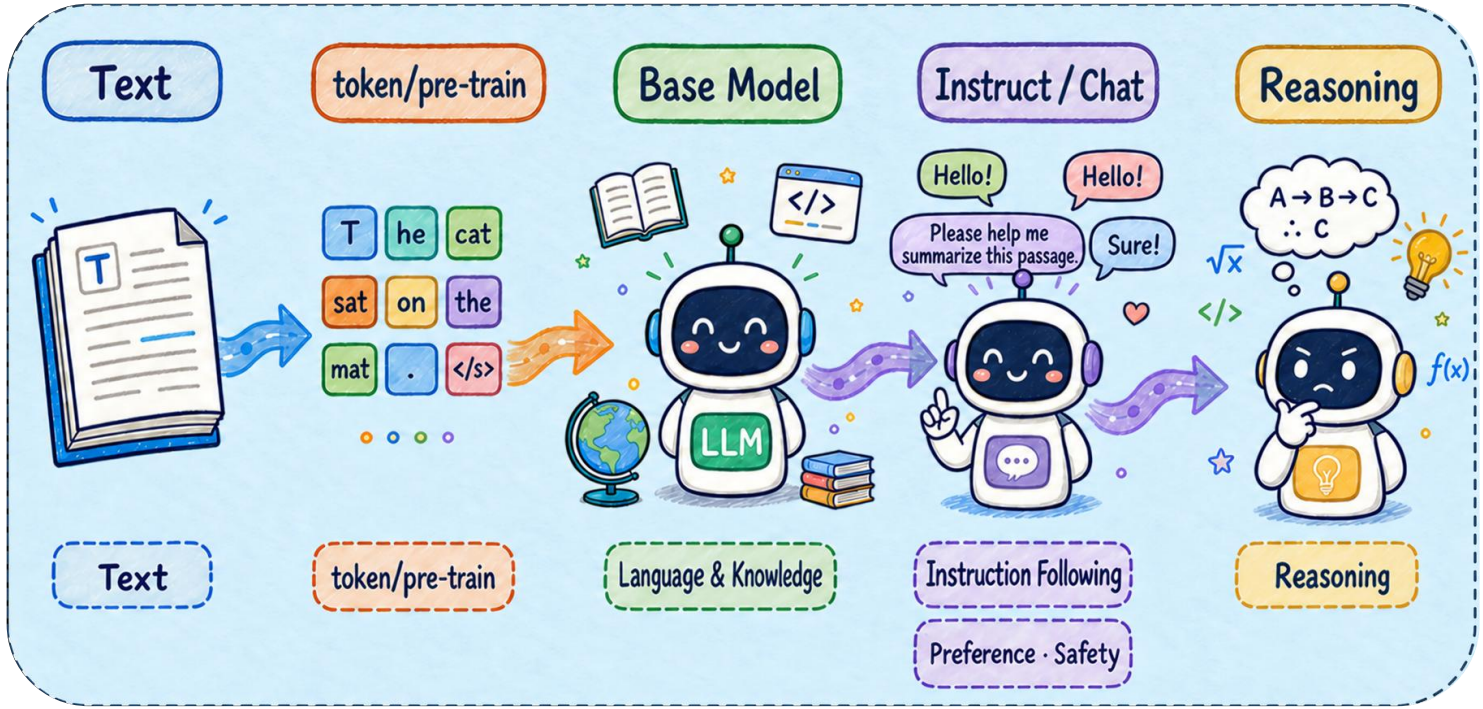
Javob: Yo'q, yetarli bo'lmaydi.

LLM Capabilities Across Different Training Stages

Each stage addresses a distinct problem

Text is first encoded into tokens. Pre-training teaches language and knowledge, while post-training shapes instruction following, preferences, reasoning, and safety.

- Pre-training answers: Can the model do it?
- Post-training answers: Can the model do it as requested?
- Evaluation answers: Is the model ready for use?



Core thread: the tokenizer determines what the model sees, pre-training determines what it knows, and post-training determines how it responds.

Tokenizer: Converting Natural Language into Discrete Tokens

A vocabulary is not a manually compiled list of all words. It is a set of characters, subwords, and special tokens learned from representative corpora.

Prepare the Corpus → **Learn the Vocabulary** → **Freeze and Use**

01 Normalize and Pre-tokenize

Standardize encoding and whitespace, then apply language-specific rules to produce initial segments for further merging.

02 Train BPE or Unigram

Iteratively merge frequent segments or select the optimal subword set to create a fixed vocabulary and ID mapping.

03 Verify Encoding and Decoding

Check compression, multilingual coverage, and code coverage. Pre-training and inference must use the same tokenizer.

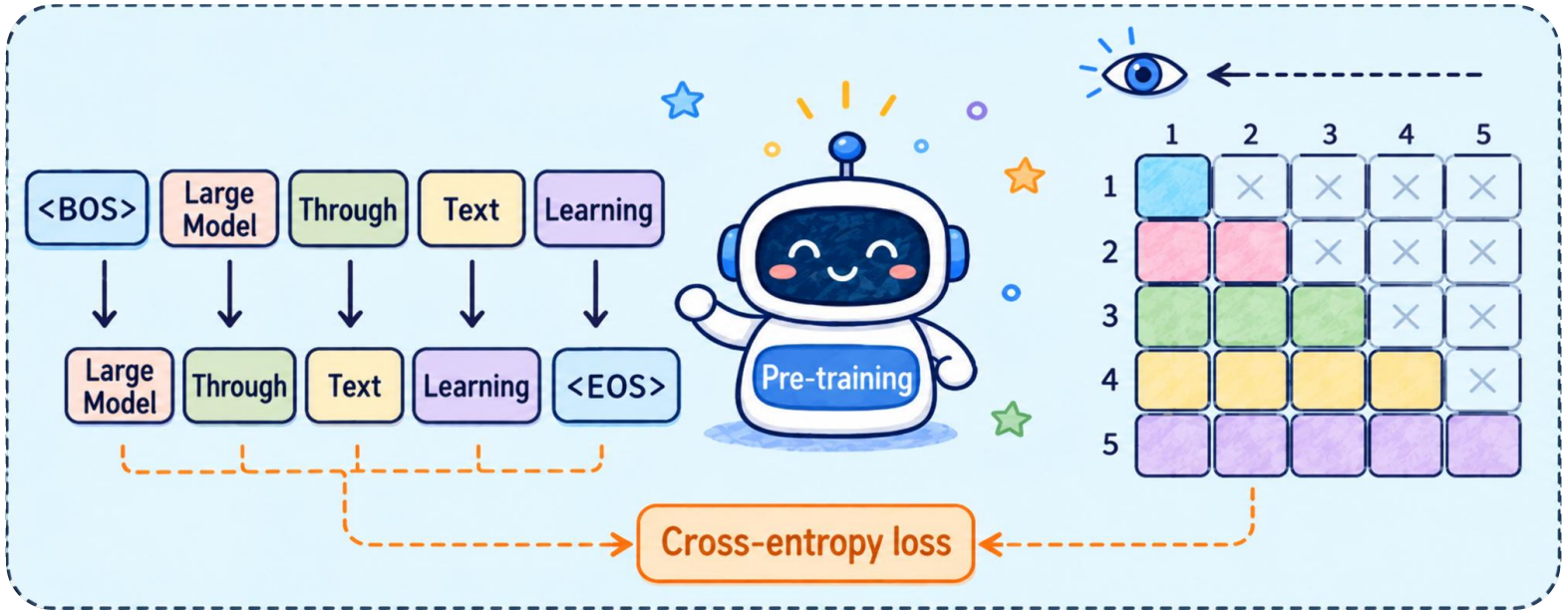
The tokenizer changes sequence length and word boundaries: overly fine segmentation increases computation, while insufficient coverage weakens multilingual and coding capabilities.

Pre-training: Autoregressive Next-Token Prediction

Training Objective

Shift the text one position to the right to obtain supervision signals

Given $x_1 \ldots x_t$, the model learns to increase the probability of the true next token x_{t+1}. Prediction errors across all positions form the cross-entropy loss.



Input

Token sequence
Left context only

Objective

Minimize cross-entropy
Raise true-token probability

Output

Base Model
Text completion + knowledge

Key to self-supervision: the training material serves as both input and label, so text at scale can be directly converted into learning signals.

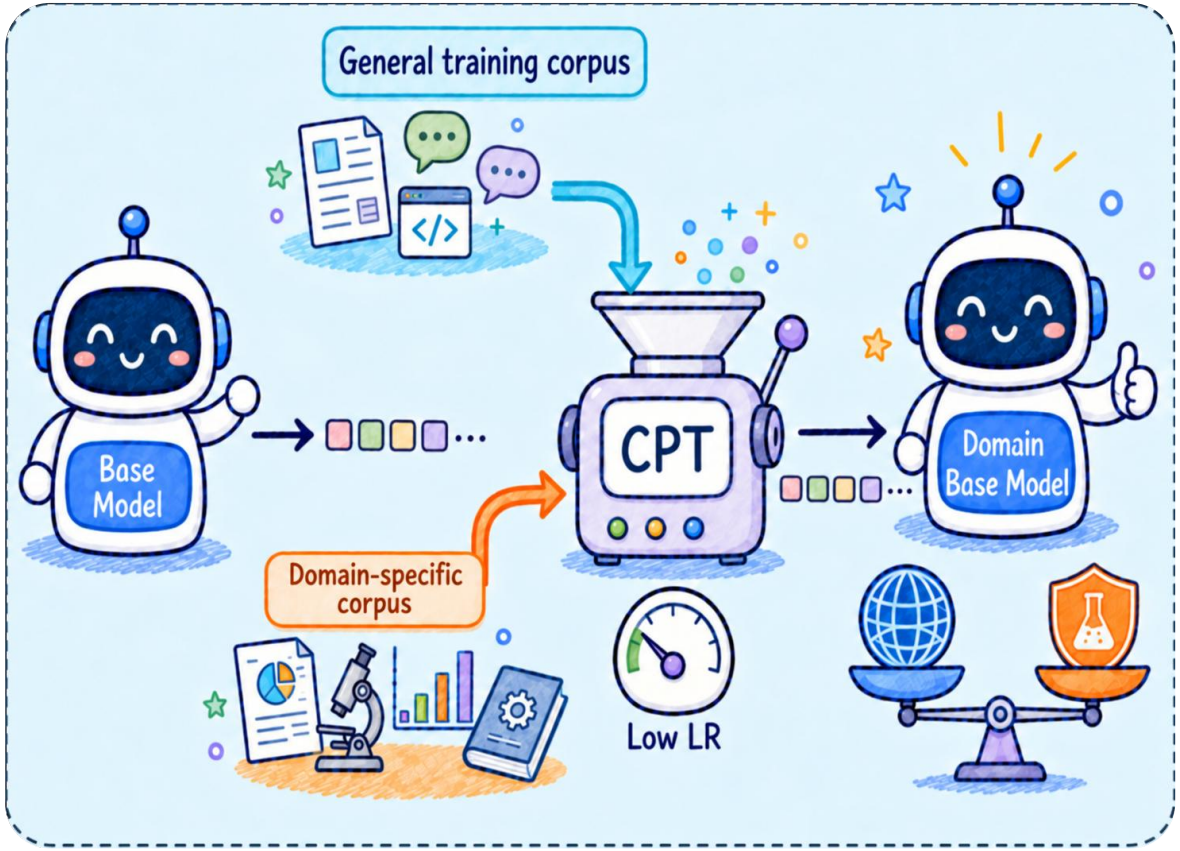
Continued Pre-training: Enhancing Domain && Language-Specific Capabilities

Continued Pre-training (CPT)

Continue using the language-modeling objective on a general base model to add domain, language, up-to-date knowledge, or longer-context capabilities.

Keep the objective, change only the training-data distribution

Continue making predictions on the next token; by incorporating mixed domains, multiple languages, code, or long documents, the base model continues to evolve toward the target capability.



Rebalanc

More domain data
Keep general replay

Low Rate

Lower learning rate
Control token budget

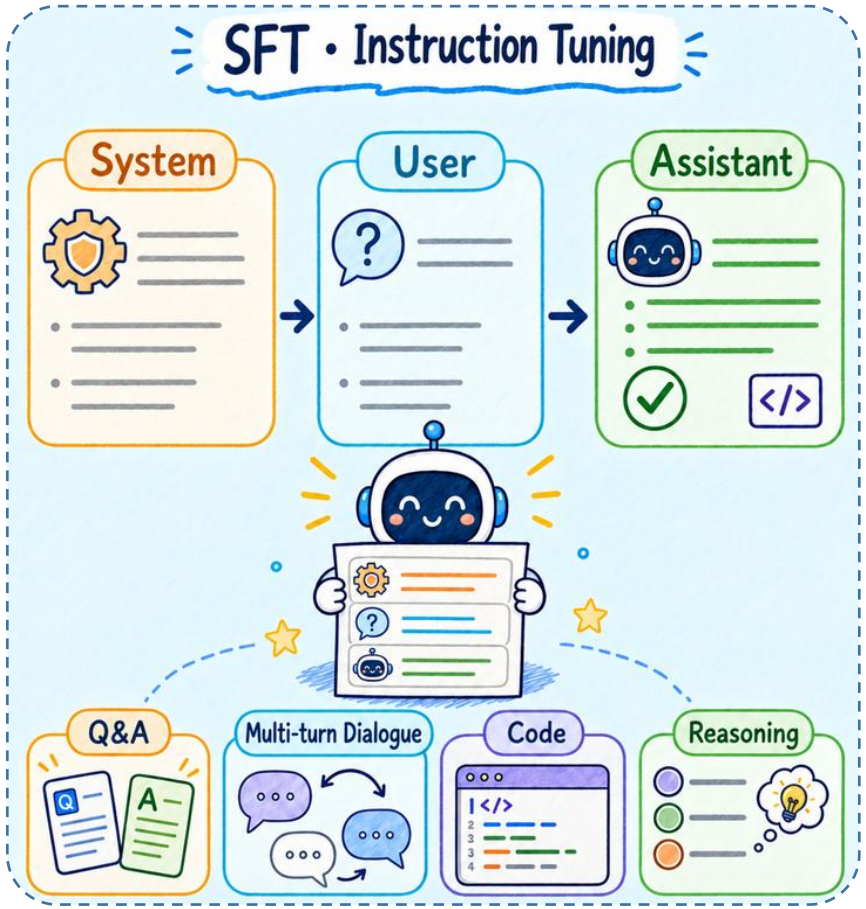
Regressi

Measure domain gains
Monitor forgetting

A base model excels at completing text but may not consistently follow instructions. It must undergo post-training to become a useful assistant.

SFT: Learning Instructions, Dialogues, and Desired Responses

The training format remains next-token prediction, but the data become system prompts, user instructions, and assistant responses. The loss usually covers only the target response.



Examples → **Template** → **Instruct**

01 Collect and Filter

Human, expert, or reviewed synthetic data.
Cover Q&A, code, reasoning, tools, and dialogue.

02 Format and Compute Loss

Use one chat template. Mask system and user text; supervise the assistant response.

03 Train and Regressions

Use full fine-tuning (FFT) or PEFT, such as prefix tuning or LoRA.

SFT shows desired responses but does not rank viable alternatives.

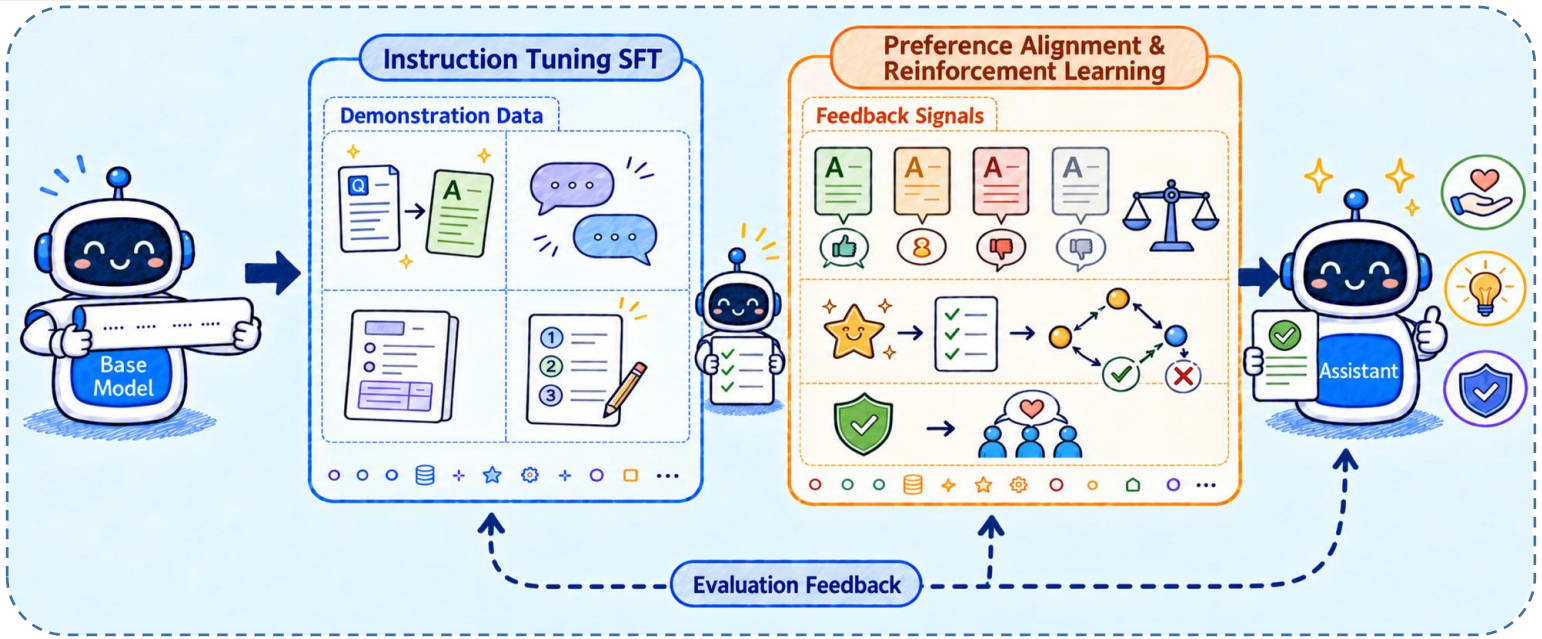
Post-training: From Next-Token Prediction to Helpful AI Assistance

Post-training builds on a pre-trained model and uses techniques such as instruction fine-tuning and reinforcement learning to enable instruction following, human-preferred content generation, and reasoning.

SFT
Instruction pairs
Dialogue format

Preferences
Rank answers
DPO / RLHF

Reasoning
Sample and verify
RLVR / GRPO



Shared goal of post-training: shift the response distribution so the model more readily generates high-quality, helpful, and safe responses.

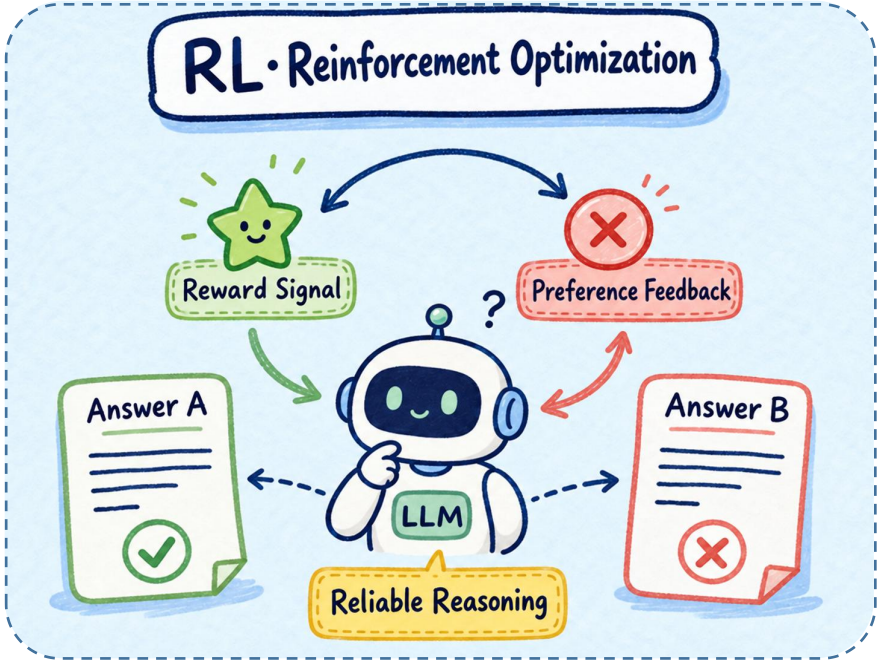
RL: Enhancing Reasoning and Preference Alignment after SFT

The model receives feedback through verifiable rewards or preference rewards, continually increasing the probability of generating high-quality responses.

Generate Responses
Sample answers and reasoning traces.

Compute Rewards
Rules, reward models, or preferences score answer quality.

Update the Policy
Favor high-reward responses; suppress low-quality behavior.



Stronger Reasoning

Math, code, logic, and complex problems
Route: RLVR
Algorithms: GRPO, PPO

Preference Alignment

Helpfulness, instruction following, and safety
Routes: RLHF, RLAIF
PPO or DPO

Reinforcement learning improves reasoning through verifiable rewards and optimizes helpfulness and safety through preference rewards.

TianChi LLM

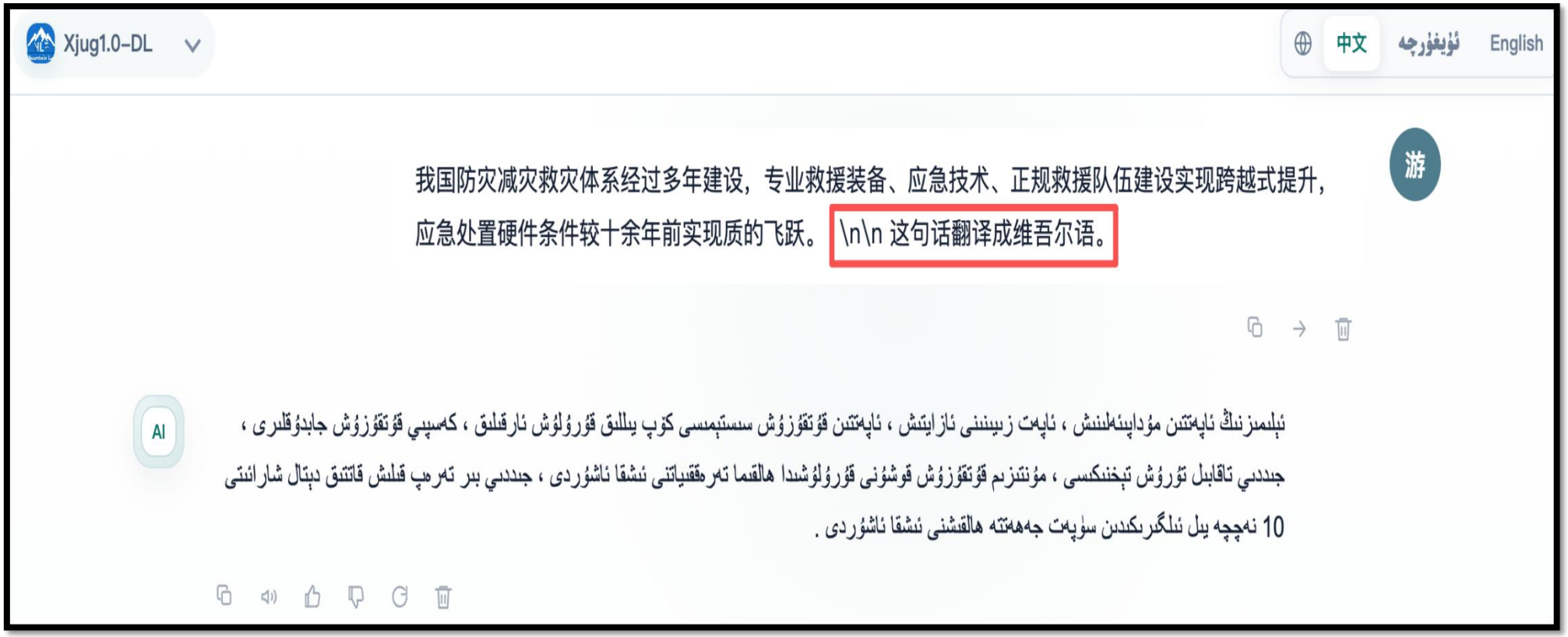

新对话
对话记录
新对话
07/26 01:30
游客用户
未登录状态，可注册或登录后保存个人信息。
Xjug1.0-DL
中文
ئۇيغۇرچە
English
NLP
Mountain Lab
你好，有什么可以帮你?
我是你的智能助手，可以回答问题、协助写作与翻译。
给维语智能助手发送消息
深度思考

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 يارەمچىڭىز، سوئاللارغا جاۋاب بېرىش، يېزىش ۋە تەرجىمىگە يارەملىشىمەن.
ئۇيغۇرچە ئەقلىي يارەمچىگە ئۇچۇر يېزىڭ
چوڭقۇر تەپەككۇر
مېھمان
تىزىمغا كىرمىگەن.
تىزىملىتىپ ياكى كىرىپ
شەخسىي
ئۇچۇرلىرىڭىزنى
ساقلىيالايسىز.

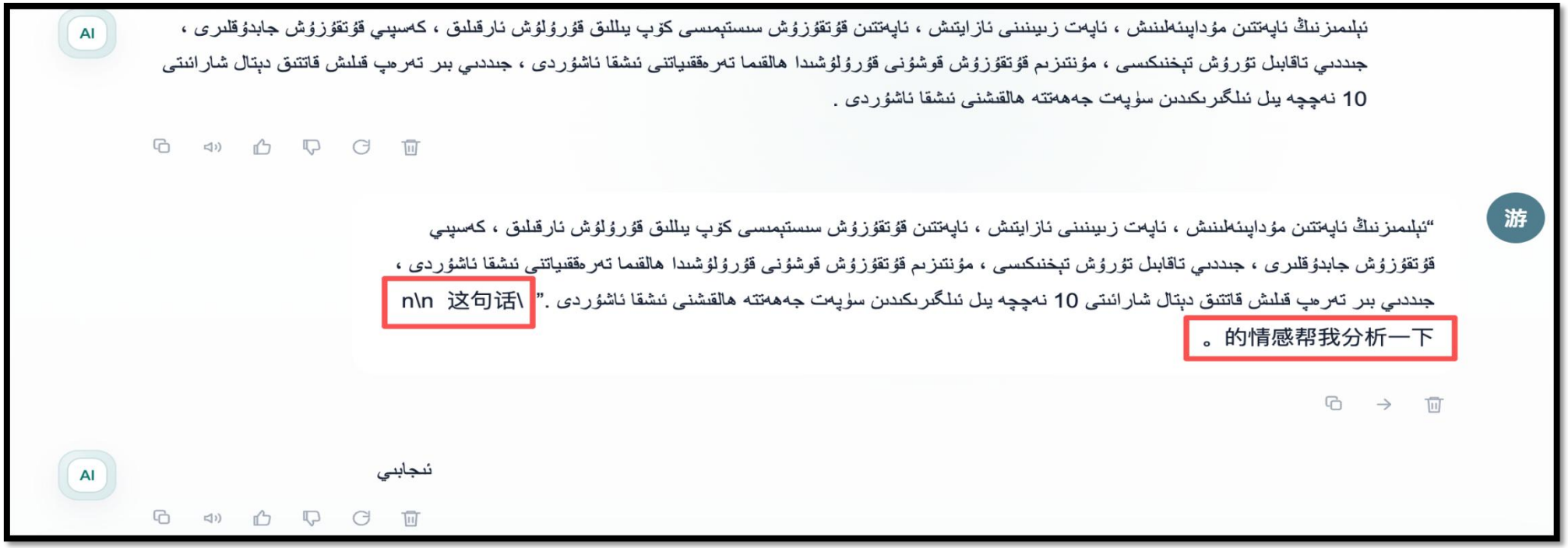
TianChi LLM

MT Task

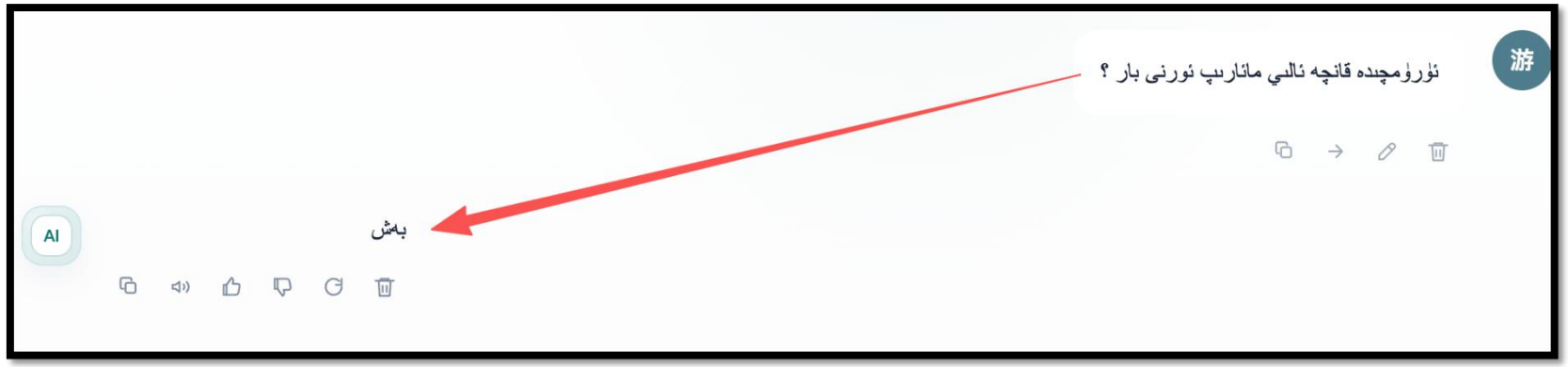


TianChi LLM

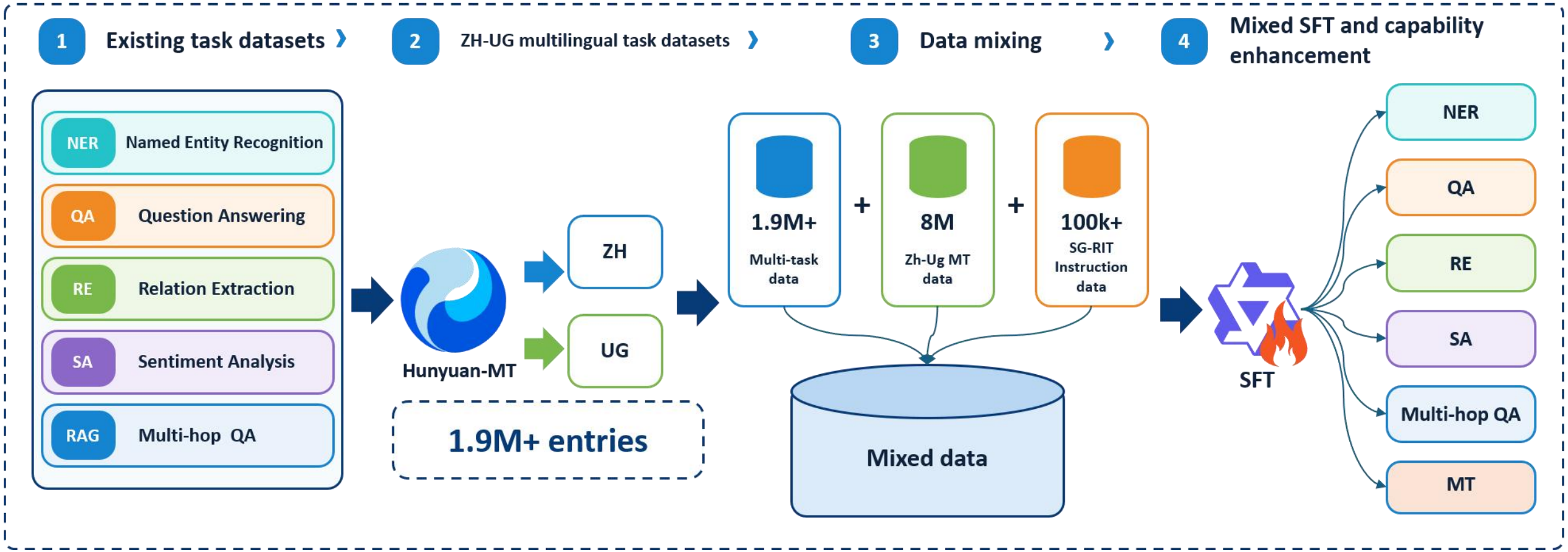
SA Task



QA Task



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



Objective: improve key-task capabilities for Chinese, Uyghur, and other languages with high-quality mixed corpora.

Evaluation Results

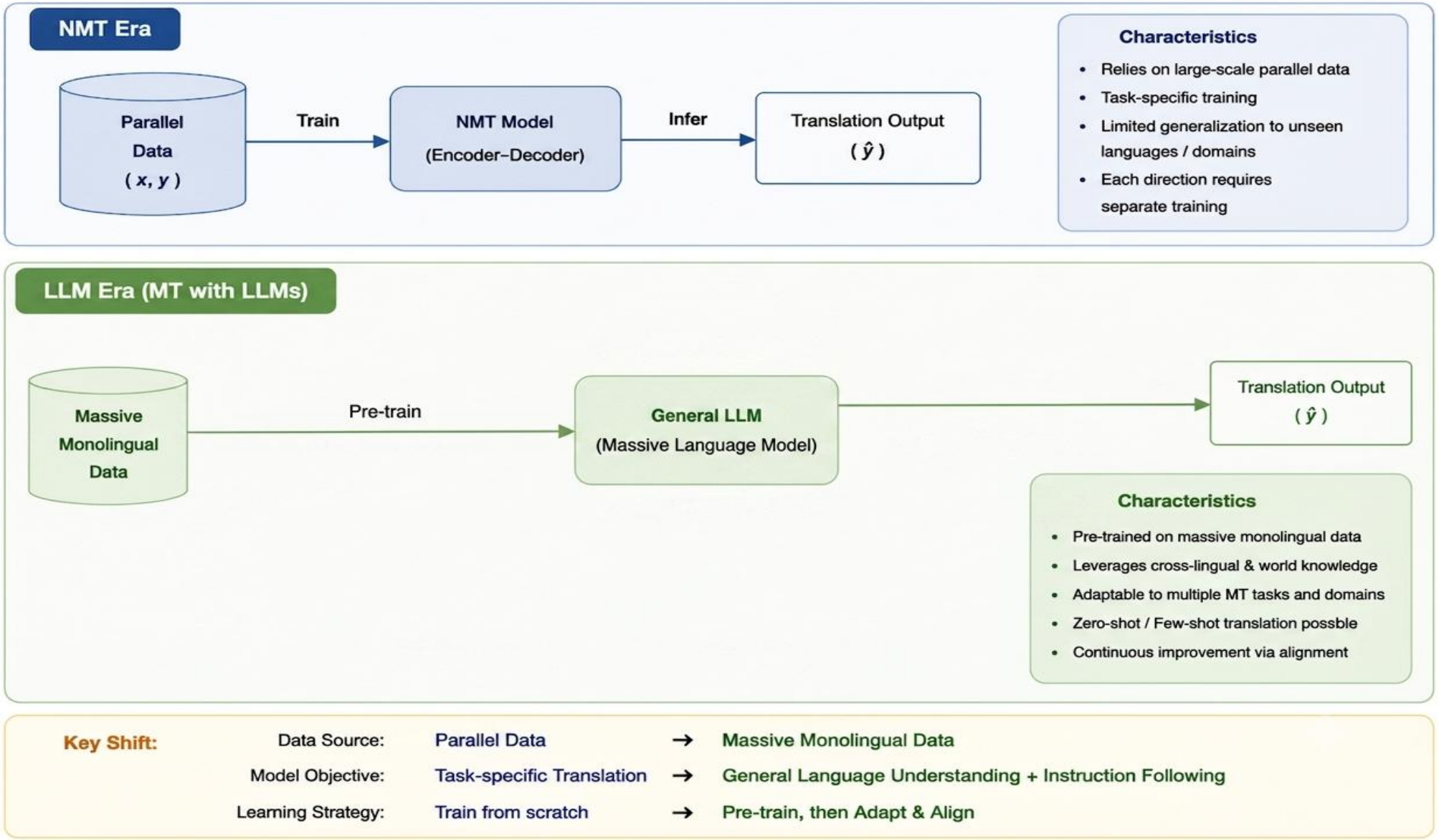
Machine Translation

Model	Chinese → Uyghur					
	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**

Question Answering, Multi-hop QA, Named Entity Recognition, Relation Extraction

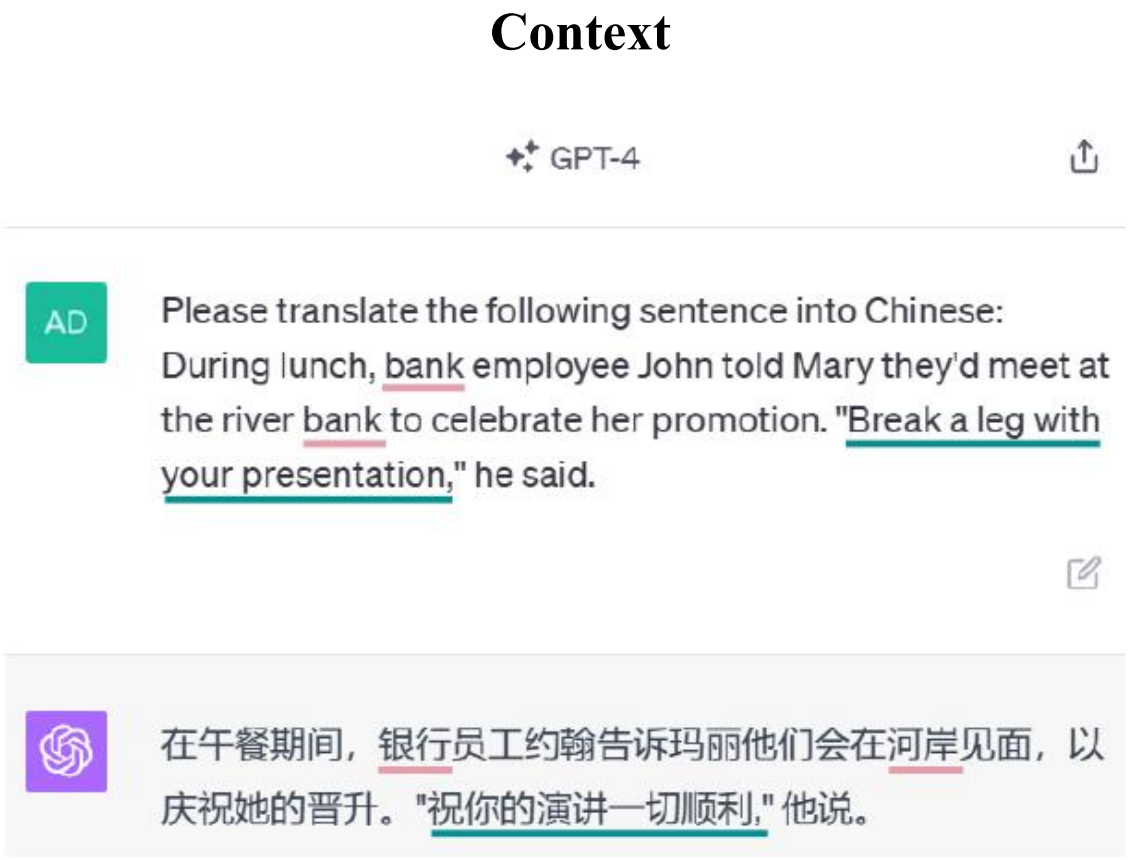
Language	Model	Task			
		QA	Multi-hop QA	NER	RE
ZH	Qwen3-32B	23.65	3.98	4.00	5.78
	Ours	**62.43**	**73.20**	**87.31**	**42.64**
UG	Qwen3-32B	10.91	2.15	6.80	10.47
	Ours	**51.12**	**68.08**	**64.11**	**21.13**

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



Pros of LLM-based MT

- **Context Understanding**
  - Improves translation accuracy and fluency
- **Interactive Translation**
  - Users can provide real-time feedback or guidance
- **Versatile Text Processing**
  - Can translate formal text, colloquial language, and slang
- **Document-level Translation**
  - Maintains consistency and context in long and complex documents
- **Multilingual Capability**
  - Can translate across multiple languages



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
   - Language distribution imbalance during pre-training step
3. Evaluation
   - Traditional metrics cannot fully measure translation quality
4. Efficiency
   - High cost for fine-tuning and deployment

MT Performance Comparison

Zh →Ug translation sample

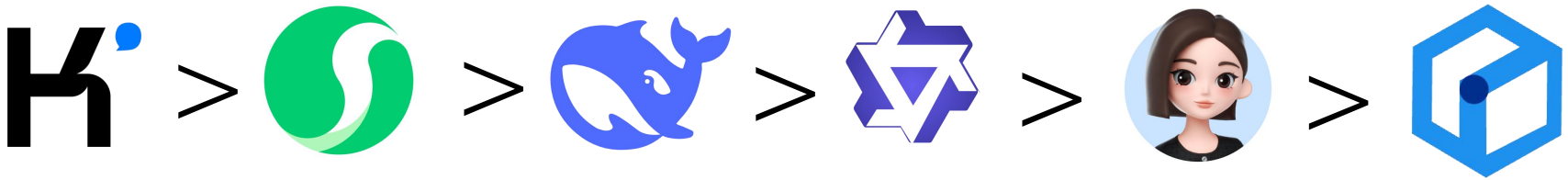
 >  >  > 

Ug→Zh translation sample

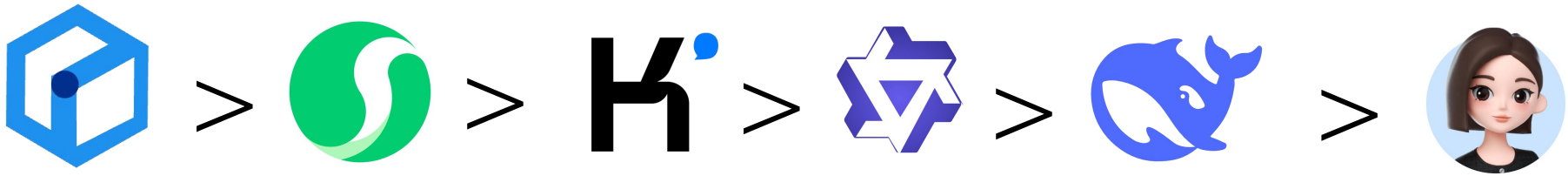
 >  >  > 

MT Performance Comparison

Ug→Zh translation sample



Zh →Ug translation sample



Main Challenges of LLM-based MT on LRLs

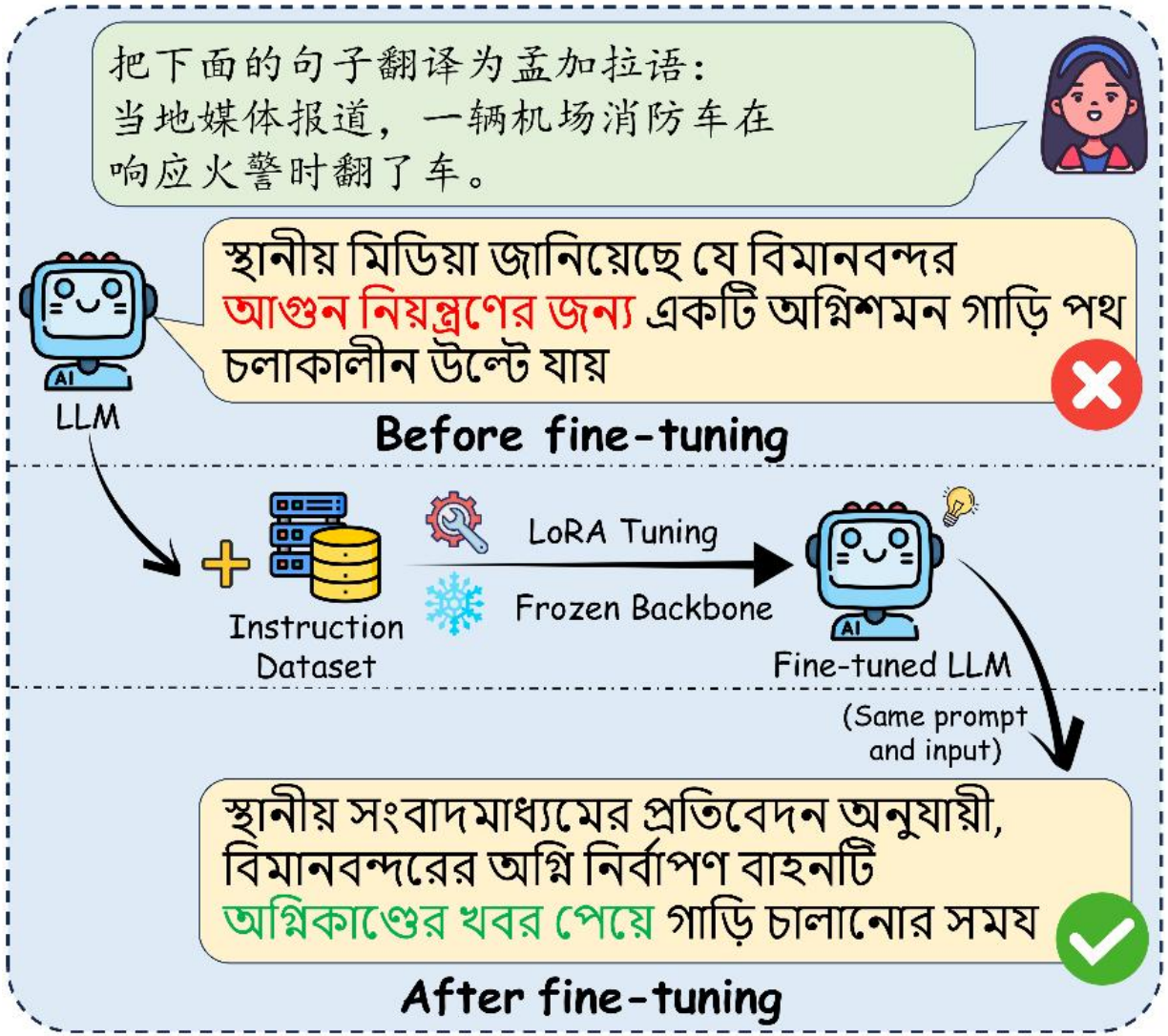
- **Omission** – Missing content or phrases
- **Over-translation** – Unnecessary or redundant content
- **Semantic Errors** – Incorrect meaning or misinterpretation
- **Syntactic Errors** – Grammatical or structural issues
- **Word-level Errors** – Inaccurate lexical choices
- ……

Outline

- Low-Resource Language Processing in the Era of LLMs
- Strategies for Text-only LLM-based MT
- Findings on Multi-modal LLM-based MT
- Exploring RAG for Low-Resource Languages
- Conclusion

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

NAACL2027 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

NAACL2027 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 $p < 0.01$.

NAACL2027 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).

NAACL2027 under review

Multi-Source Knowledge

- Motivation

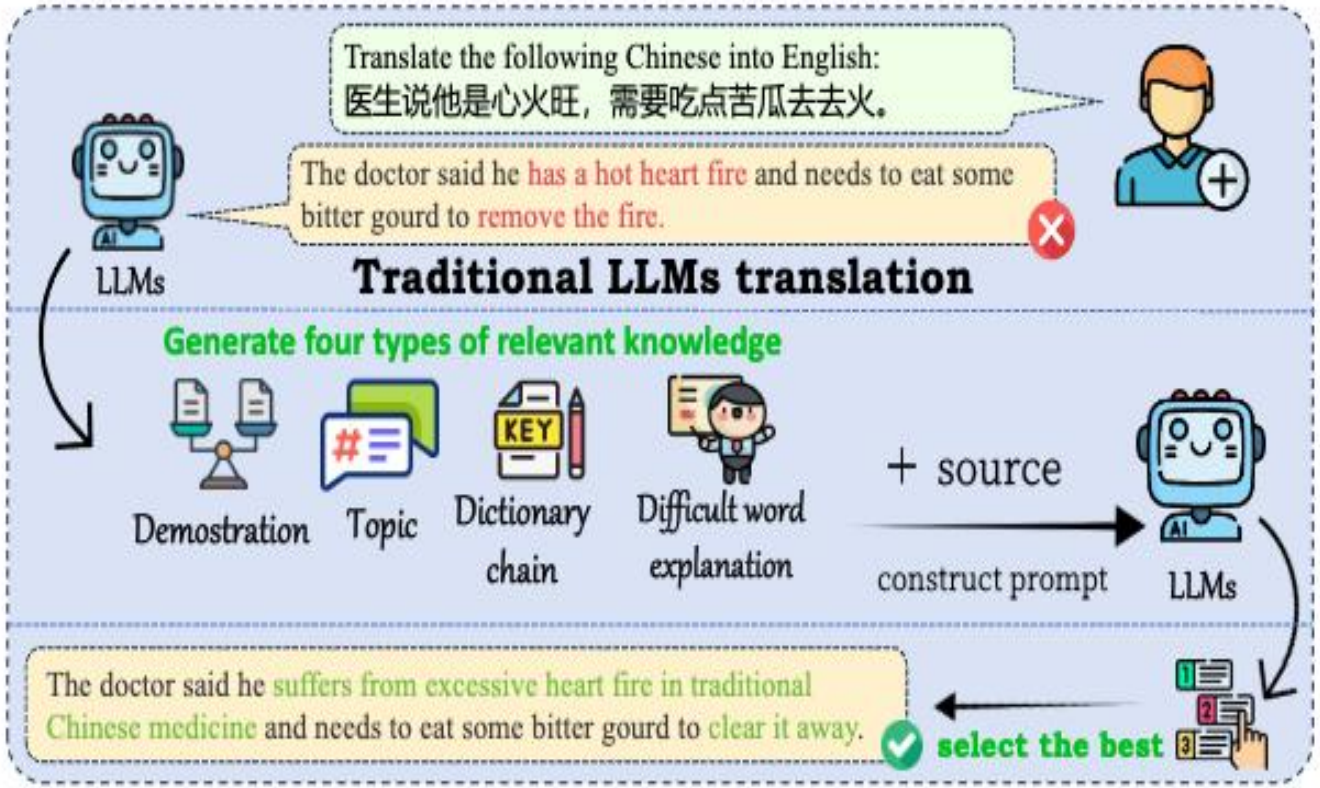


Figure 1: Multi-source reasoning for low-resource languages translation

- 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

Multi-Source Knowledge

- Experiment
  - Main experiment

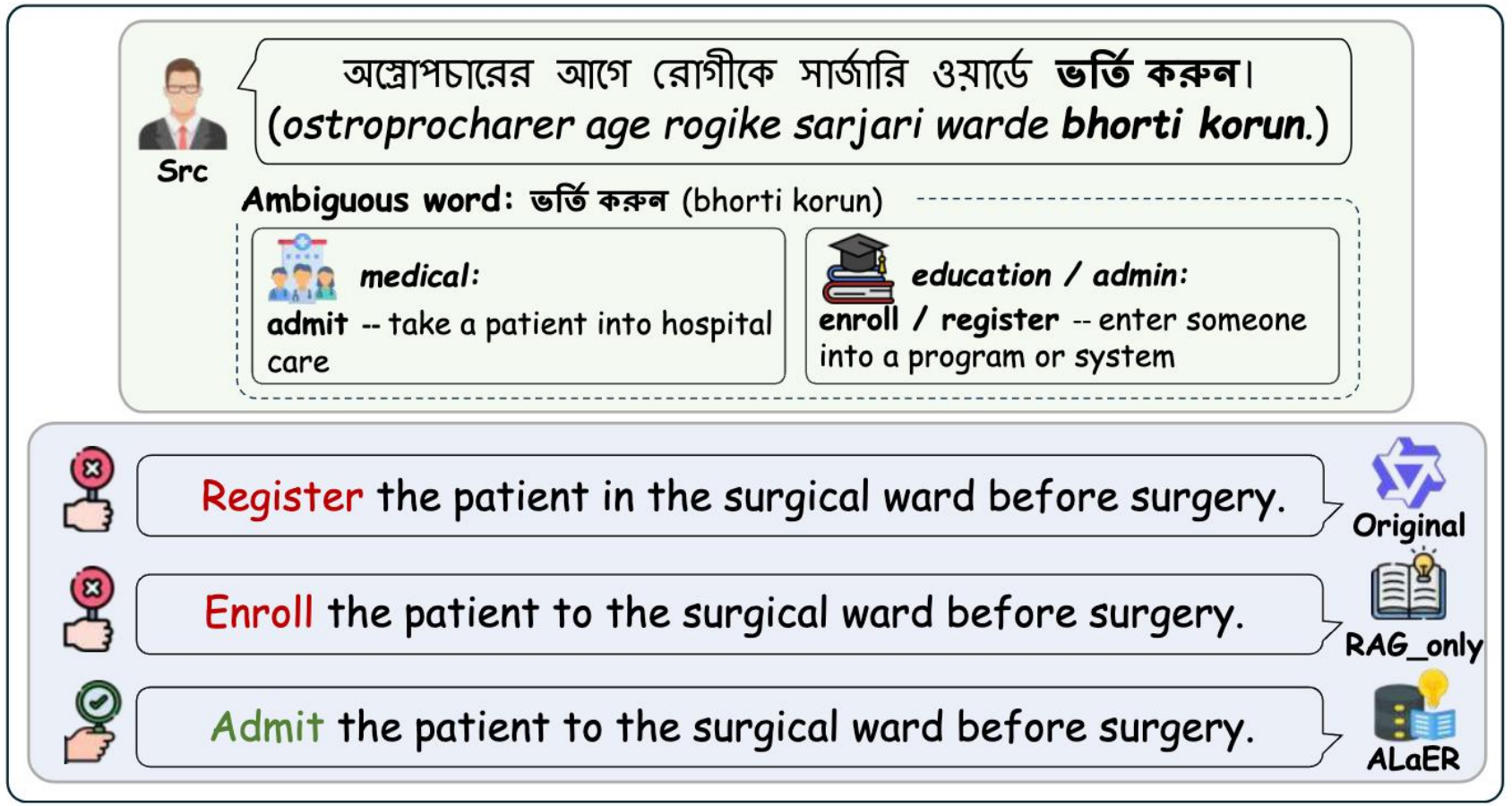
Table 2: Results on the IWSLT dataset (**Chinese** → **XX**) of Qwen3-32B evaluated by XCOMET-XL.

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

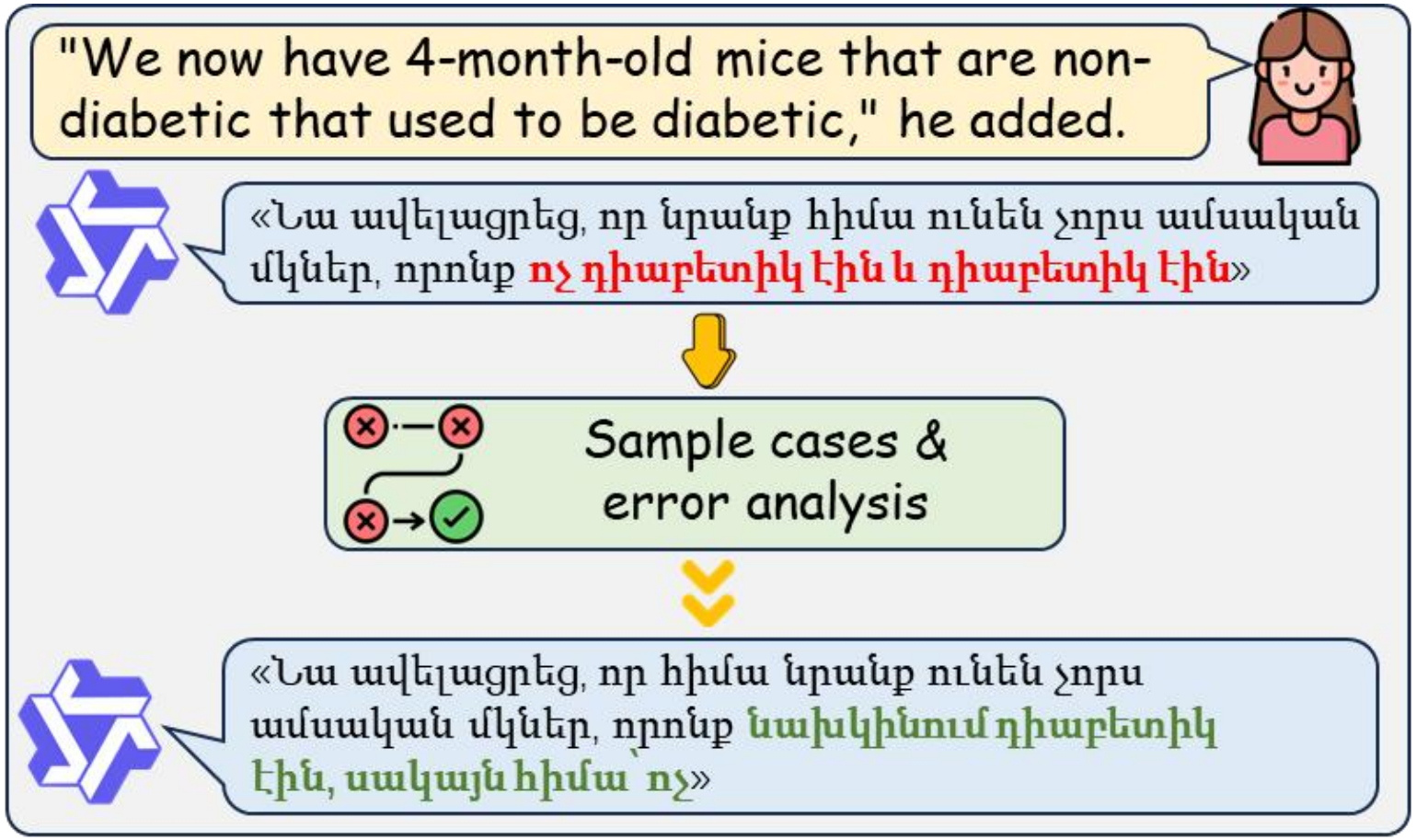
Shibo Zhang[1,2,3,4], Mieradilijiang Maimaiti[1,2,3,4,*], Zhengyi Guo[1,2,3,4], Dezhi Wang[1,2,3,4], Wu Le[5], Zhuofei Xie[5], Jiawei Chen[5], Wushouer Silamu[1,2,3,4],

[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

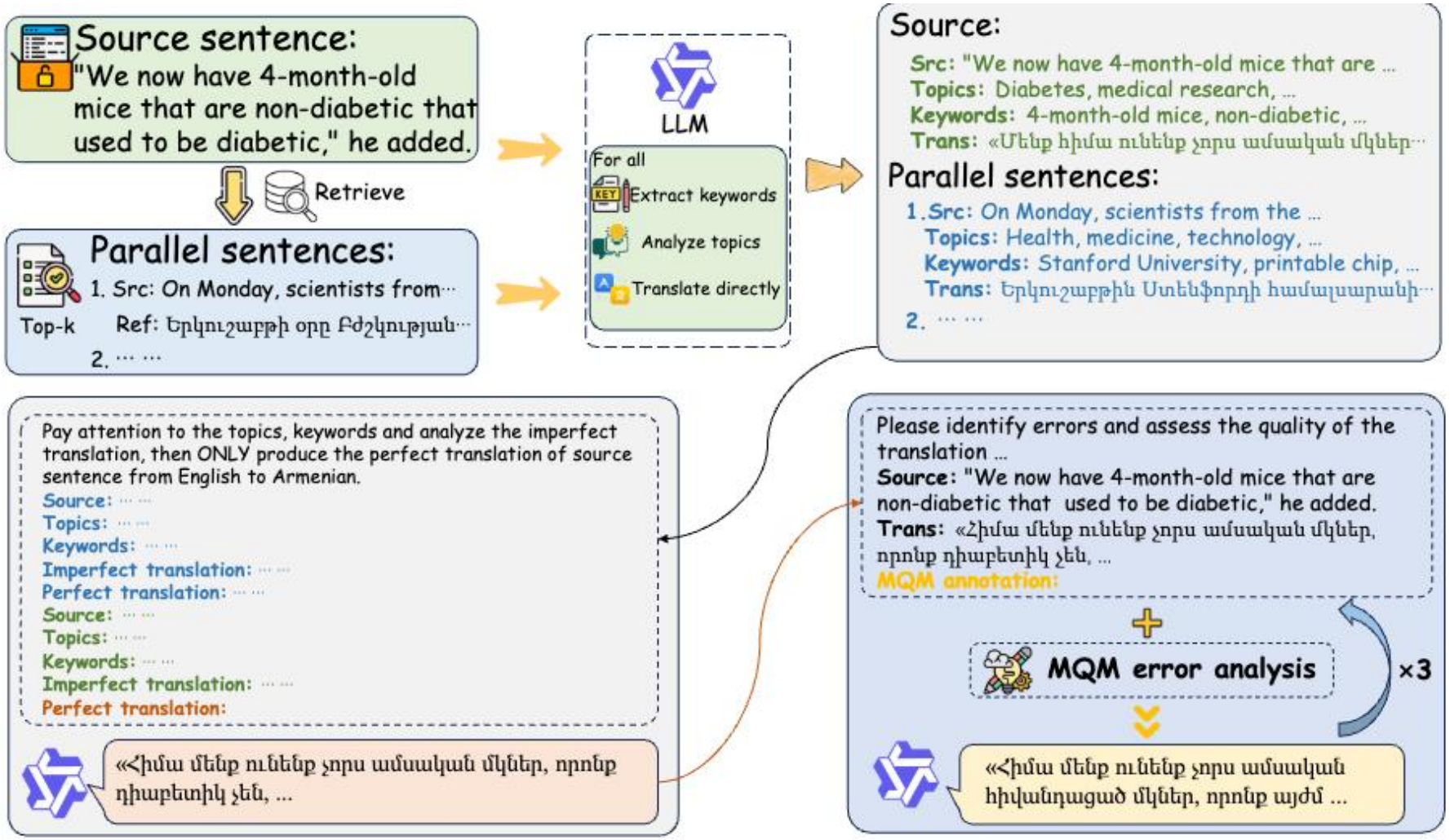


Fig. 2. Overview of the proposed framework. It integrates Implicit Refinement using retrieved knowledge and semantic constraints with Explicit Refinement via iterative ($T = 3$) MQM-based self-correction.

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

- Low-Resource Language Processing in the Era of LLMs
- Strategies for Text-only LLM-based MT
- Findings on Multi-modal LLM-based MT
- Exploring RAG for Low-Resource Languages
- Conclusion

Findings on Multi-modal LLM-based MT

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

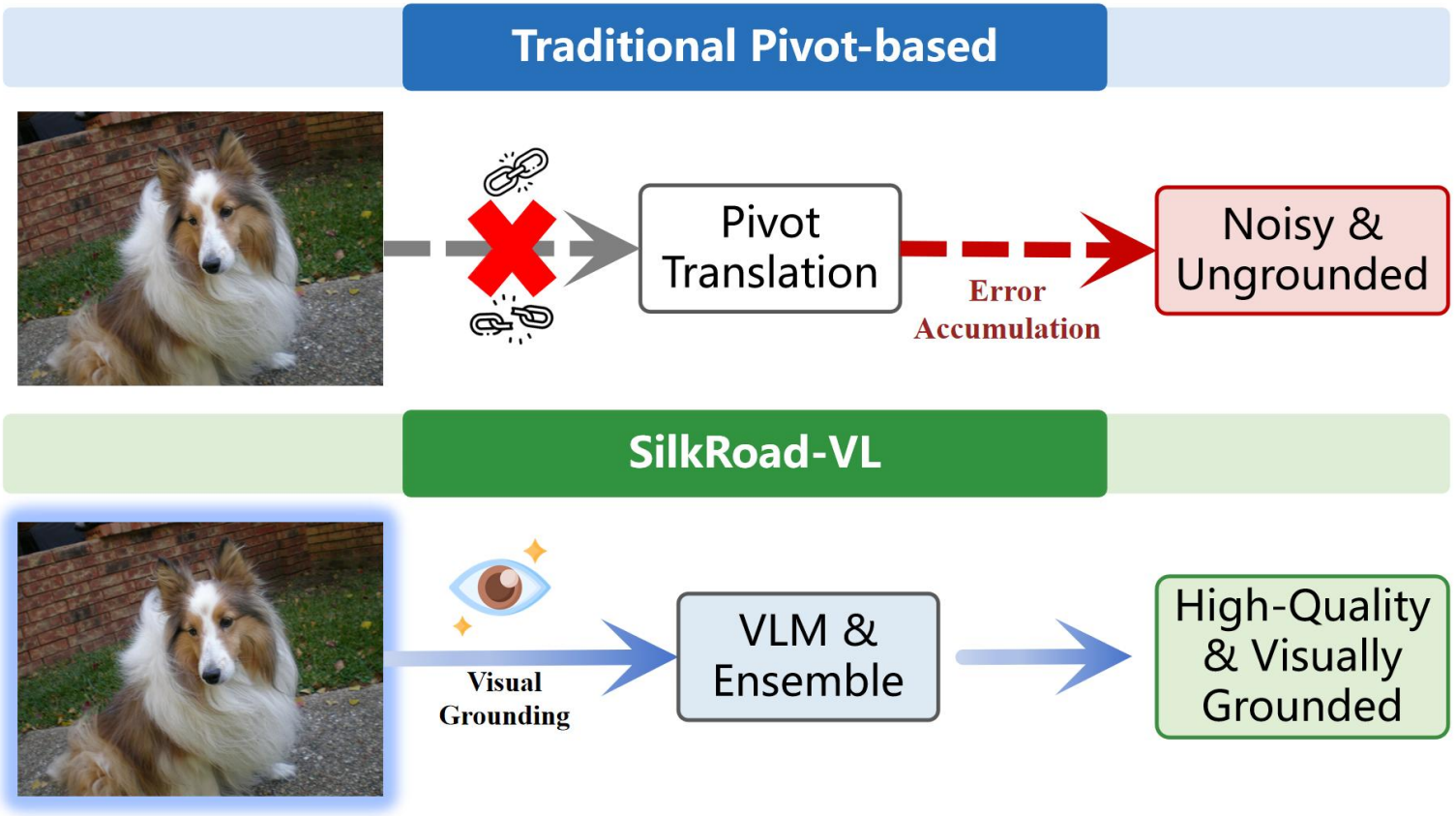
Shuoshuo Hou[1,2,3,4], Mieradilijiang Maimaiti[1,2,3,4,*], Zhexin Li[1,2,3,4], Jiaxin Wang[1,2,3,4], Ahmad Hassan[1,2,3,4], Kaishaer Jiapaer[1], Nilufar Abdurakhmonova[5], Roza Urinbayeva[5], Madina Mansurova[6], Shormakova Assem[6], Gulnar Murat[7], Le Wu[8], and Wushouer Silamu[1,2,3,4]

[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.

AAAI2027 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.

AAAI2027 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

AAAI2027 under review

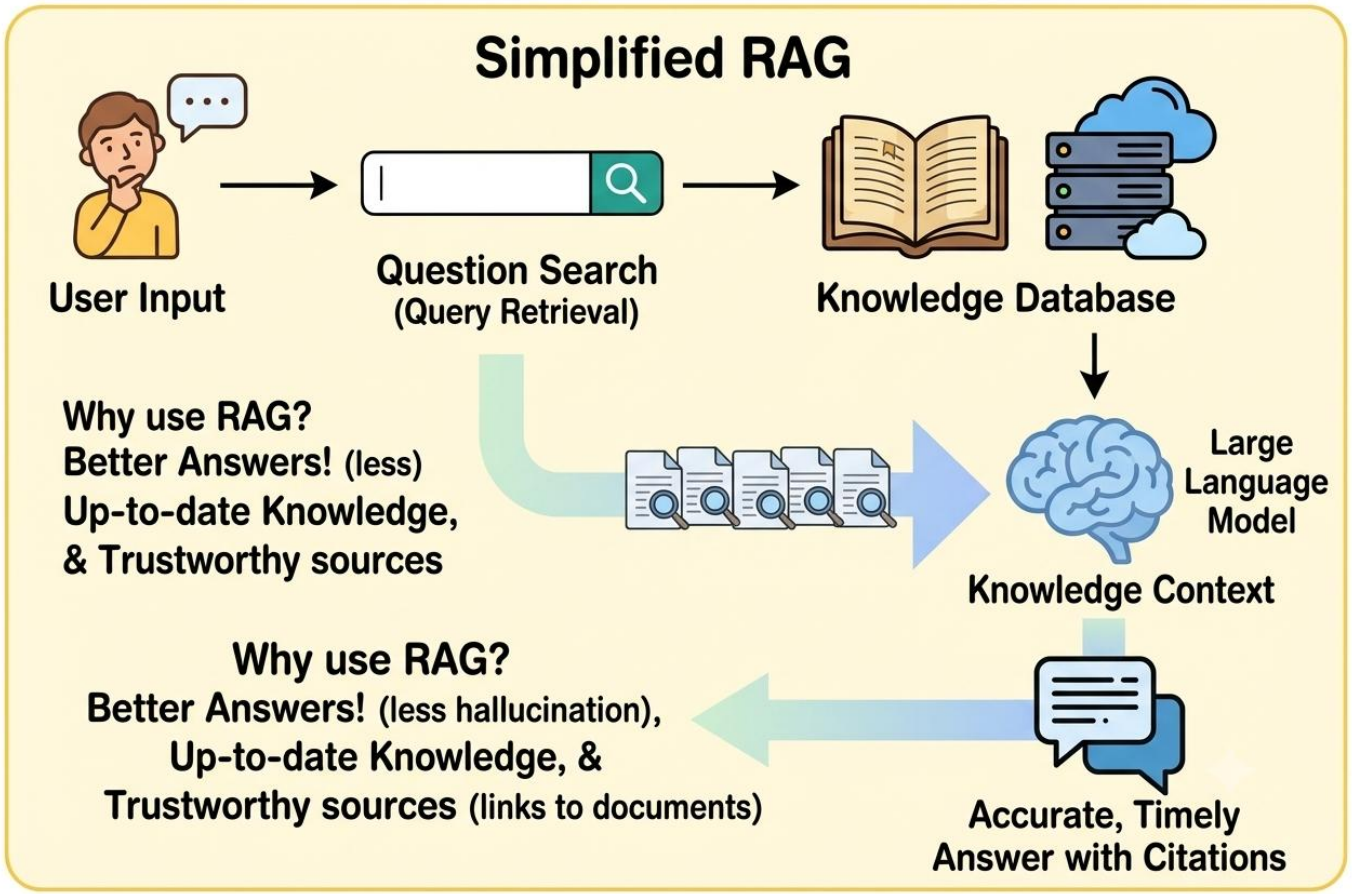
Outline

- Low-Resource Language Processing in the Era of LLMs
- Strategies for Text-only LLM-based MT
- Findings on Multi-modal LLM-based MT
- Exploring RAG for Low-Resource Languages
- Conclusion

Exploring RAG for Low-Resource Languages

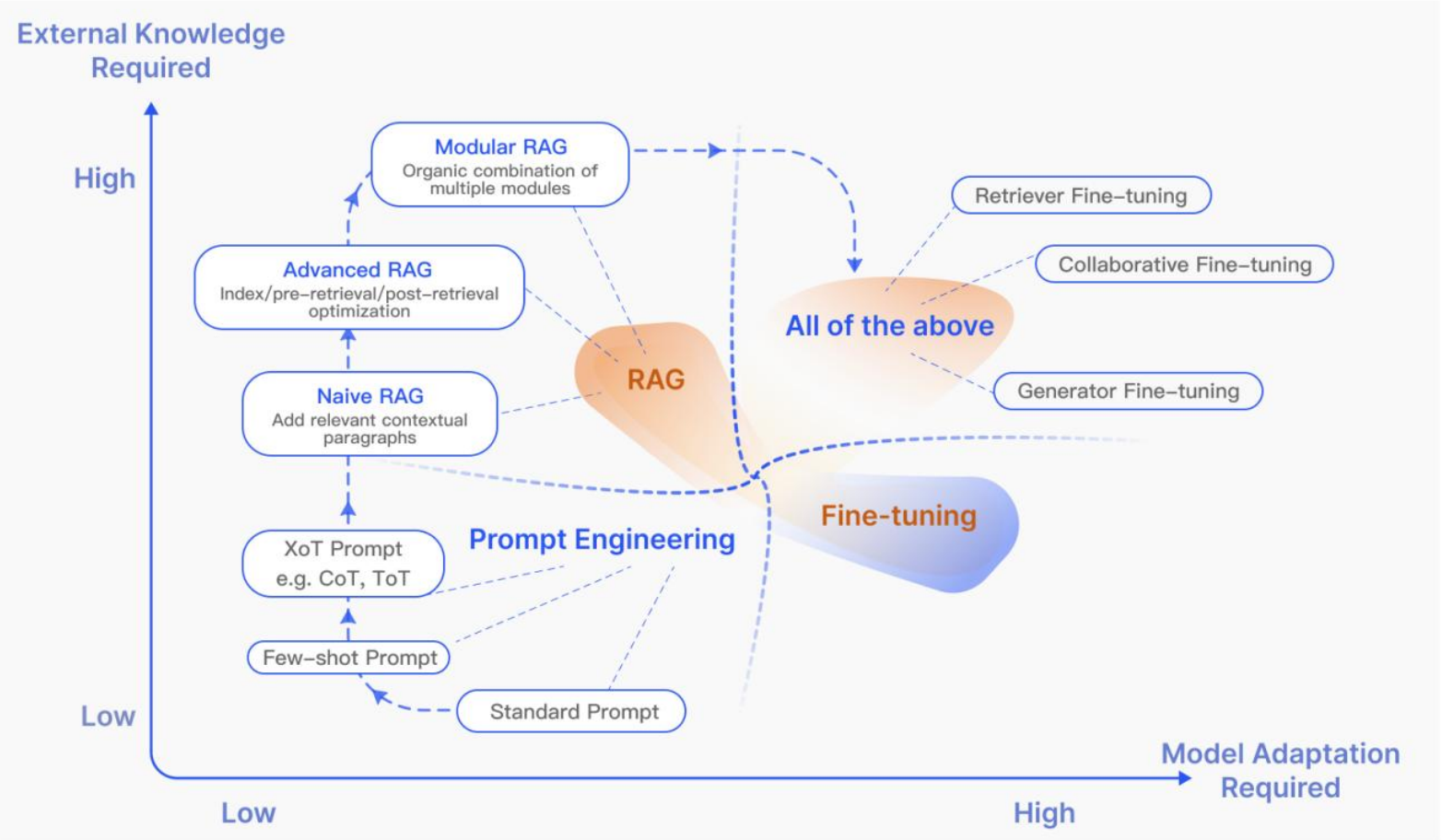
What Is RAG?

- Retrieval-augmented generation (RAG) combines external knowledge retrieval with text generation by large language models (LLMs). It first retrieves relevant documents and then generates an answer based on the retrieved information.



RAG

- Compared with other methods for improving LLMs, RAG places greater demands on the quality of external knowledge bases, as retrieved content directly informs the model's reasoning and responses.



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

Why RAG?

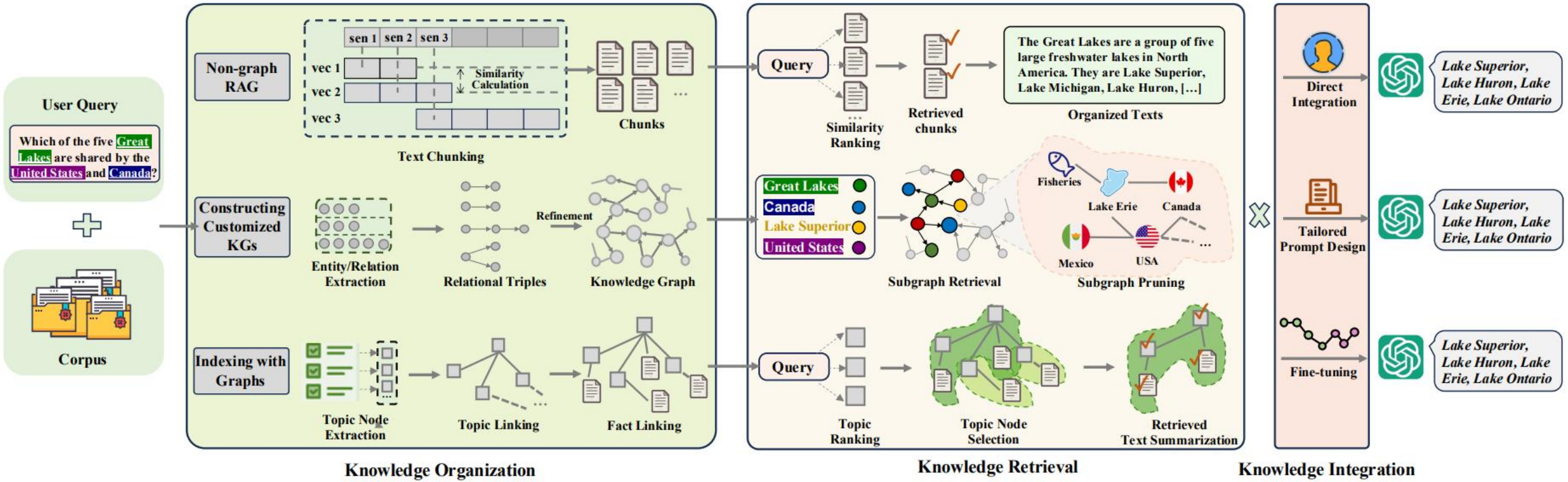
- Outdated LLM knowledge: LLMs cannot keep up with rapidly changing information.

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.

Early RAG Approaches

- Early RAG used a retrieve–concatenate–generate pipeline with **top-k text** chunks as external knowledge. Later approaches improved chunking and retrieval, adding hierarchical indexing and knowledge graphs for cross-document aggregation and multi-hop reasoning.



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

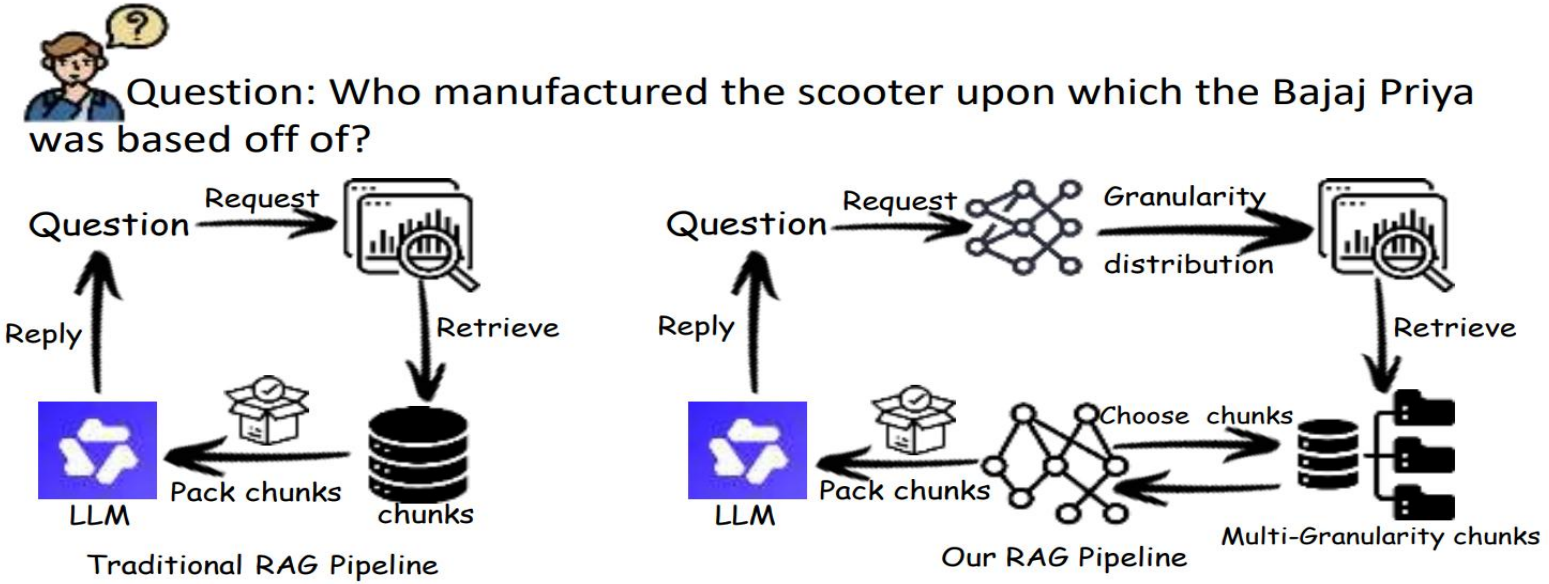
DMG-RAG: Dynamic Multi-Grained Retrieval Augmented Generation for Multi-Hop QA

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 addresses the limitations of conventional RAG, including fixed chunk granularity, fixed **Top-*k*** retrieval, and direct concatenation of retrieved text in ranked order. It builds corpus indexes at three granularities: sentence, paragraph, and document.
- For each question, the system dynamically allocates the number of retrieved chunks at each granularity. It then selects relevant, complementary evidence with low redundancy within a fixed token budget, producing a context better suited to multi-hop reasoning.

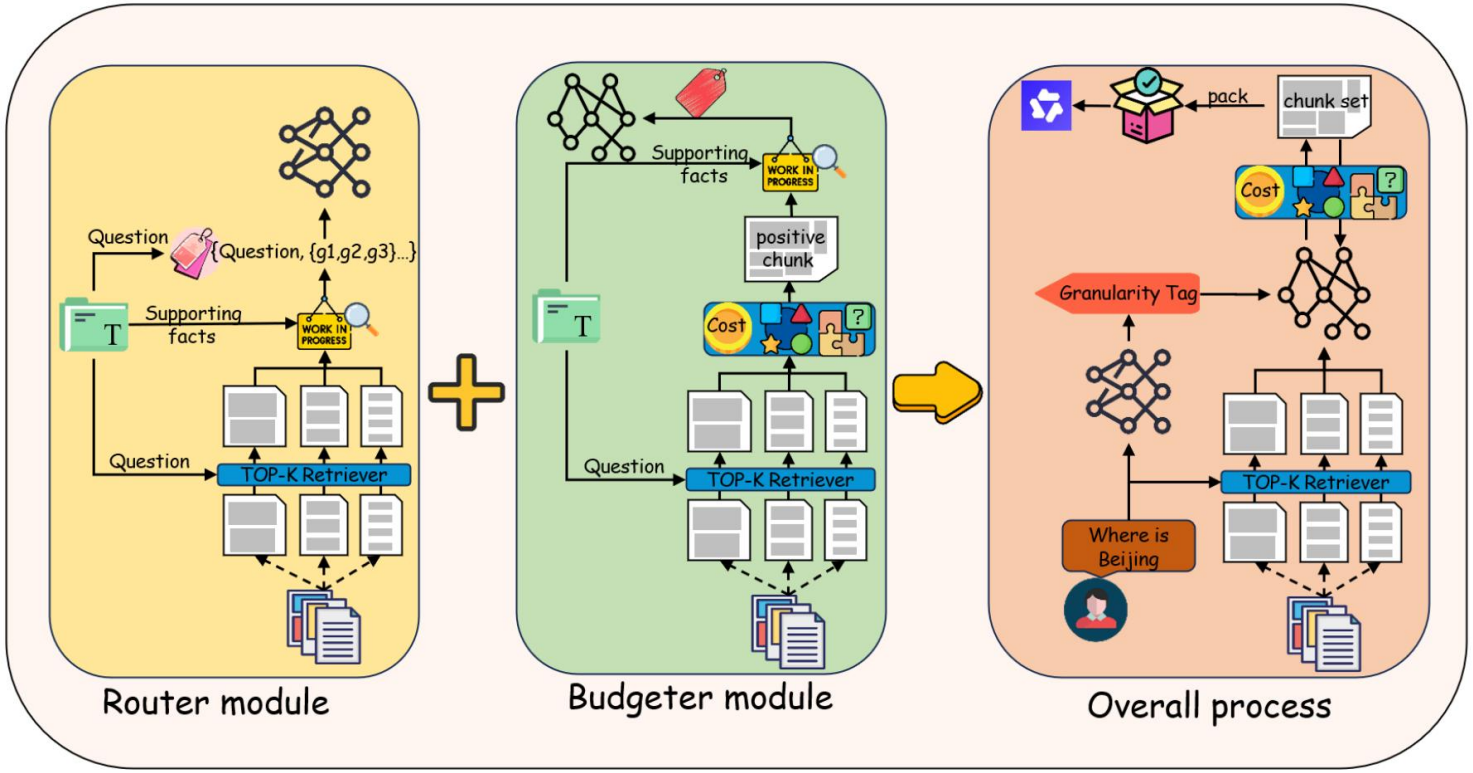


Pei et al., IJCNN 2026

Dynamic RAG

DMG-RAG consists of three components:

- Multi-granularity indexing of the corpus.
- A router that allocates retrieval counts and token budgets across granularities.
- A budgeter that selects text chunks to maximize relevance to the question, minimize token cost, and ensure diversity across granularities.



Pei et al., IJCNN 2026

Dynamic RAG

- Experiments

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]	<u>53.1</u>	<u>67.1</u>	<u>39</u>	<u>46.5</u>	<u>46.05</u>	<u>56.8</u>
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). <u>Underlined</u> values indicate the best baseline performance.

Hou et al., IJCNN 2026

Outline

- Low-Resource Language Processing in the Era of LLMs
- Strategies for Text-only LLM-based MT
- Findings on Multi-modal LLM-based MT
- Exploring RAG for Low-Resource Languages
- Conclusion

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**.

Thanks for all Contributors ~

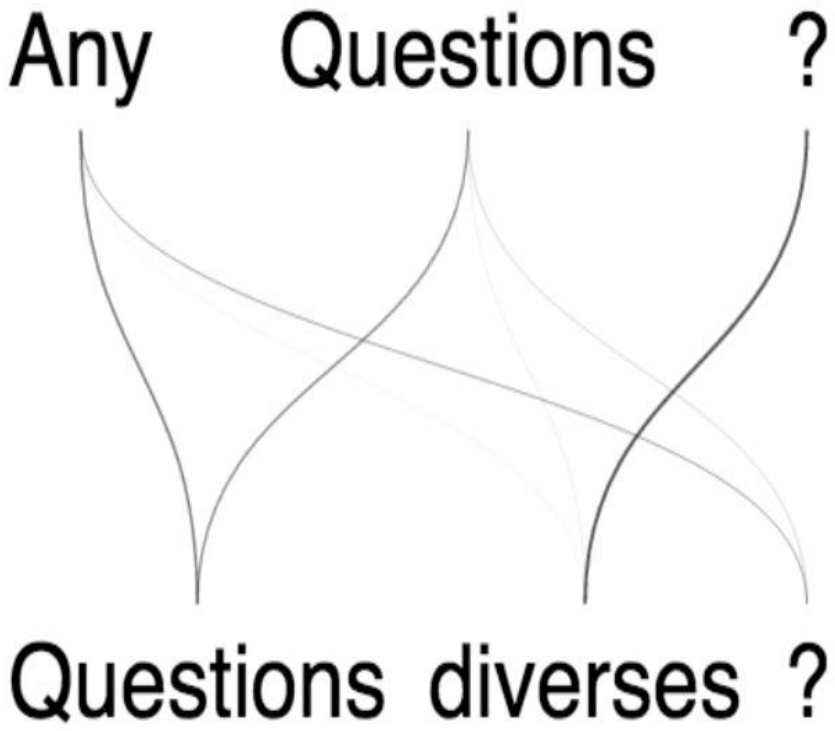






Thank You!

Q & A



This inspiration comes from Dzmitry Bahdanau @ ICLR2014

Contact me~

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

- I’m hiring **self-motivated** master's and PhD candidates to join my team as a research intern (RI).
- I am open to collaborations both **online** and **on-site**.
- Computational resources：
  - H100 80G 50
  - 5090 32G 10
  - 4090 24G 10
  - 3090 16G 8