DELTA: An Online Document-level Translation Agent Based On Multi-level Memory

Yutong Wang[1], Jiali Zeng[2], Xuebo Liu[1], Derek F. Wong[3], Fandong Meng[2], Jie Zhou[2], Min Zhang[1]

[1]Institute of Computing and Intelligence, Harbin Institute of Technology, Shenzhen, China
[2]Pattern Recognition Center, WeChat AI, Tencent Inc, China
[3]NLP^{2}CT Lab, Department of Computer and Information Science, University of Macau, China

ICLR 2025

Challenges in Document-Level Translation

- **Translation inconsistency**

It's a story about this woman, Natalia Rybczynski.

nà tǎ lì yà léi bù qín sī jī
这是关于这个女人娜塔莉亚·雷布琴斯基的故事。

Natalia Rybczynski: Yeah, I had someone call me "Dr. Dead Things."

nà tǎ lì yà lì qín sī jī
娜塔莉亚·丽琴斯基：是的，有人叫我“死物博士”。

- **Low translation quality**

And we make decisions about where to live, who to marry and even who our friends are going to be, based on what we already believe.

我们根据自己已有的信念来做决定，包括选择居住的地方，结婚对象，甚至决定谁会成为我们的朋友。

Window=1

我们决定居住地、婚姻对象，甚至我们的朋友根据我们已经相信的事情。

window=10

- **Sentence omission**

But here's the truth. //Here's the epiphany that I had that changed my thinking. //From 1970 until today, the percentage of the...

但事实是，//*Missing Translation*//自1970年……

Challenges in Document-Level Translation

- **Proper Noun Translation Consistency Metrics**

$$\text{LTCR-1} = \frac{\sum_{p\in P}\sum_{i=2}^{k_p}\mathbb{1}(\mathcal{T}_i(p) = \mathcal{T}_1(p))}{\sum_{p\in P}(k_p - 1)}$$

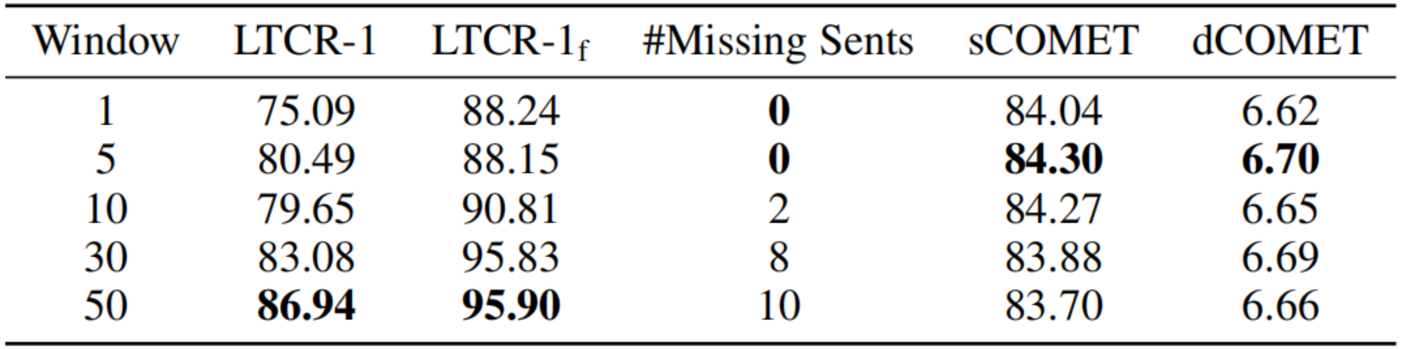
$$\text{LTCR-1}_\text{f} = \frac{\sum_{p\in P}\sum_{i=2}^{k_p}\mathbb{1}(\mathcal{T}_i(p) \subseteq \mathcal{T}_1(p) \vee \mathcal{T}_1(p) \subseteq \mathcal{T}_i(p))}{\sum_{p\in P}(k_p - 1)}$$

$p \in P$: proper noun

k_p: occurence times of p

$\mathcal{T}_i(p)$: the translation of p for the i-th occurence

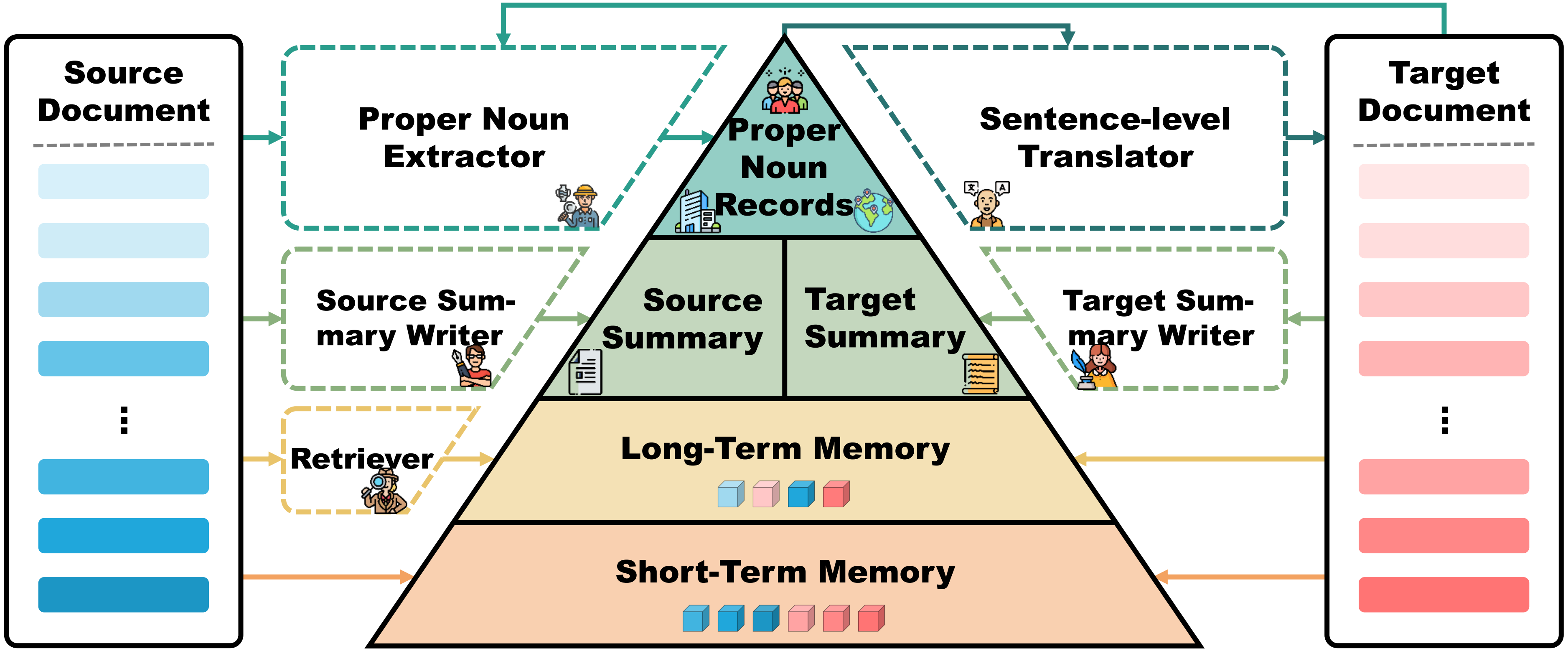
- **Translate more sentence at once (window↑), consistency↑, but omission↑, quality↓**



Window	LTCR-1	LTCR-1$_\text{f}$	#Missing Sents	sCOMET	dCOMET
1	75.09	88.24	**0**	84.04	6.62
5	80.49	88.15	**0**	**84.30**	**6.70**
10	79.65	90.81	2	84.27	6.65
30	83.08	95.83	8	83.88	6.69
50	**86.94**	**95.90**	10	83.70	6.66

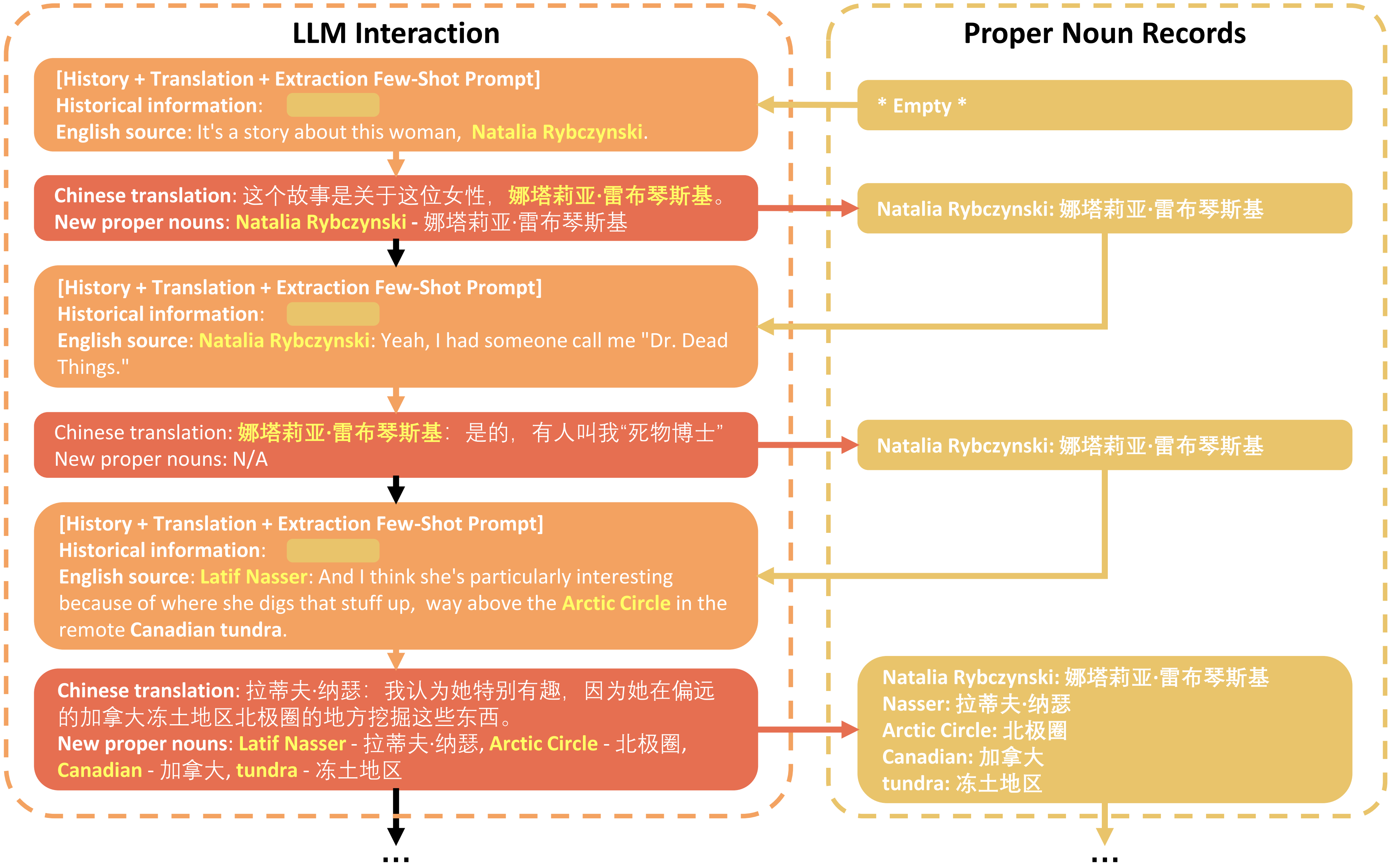
Methodology

- **DELTA: An Online Document-level Translation Agent Based On Multi-level Memory**
  - Memory Modules: Proper None Records, Bilingual Summary, Long & Short-Term Memory
  - LLM Modules: Proper Noun Extractor, Summary Writer, Long-Term Memory Retriever



Methodology

- **Proper Noun Records**: Store historical translations of proper nouns to maintain consistency



LLM Interaction

[History + Translation + Extraction Few-Shot Prompt]
Historical information:
English source: It's a story about this woman, Natalia Rybczynski.

Chinese translation: 这个故事是关于这位女性，娜塔莉亚·雷布琴斯基。
New proper nouns: Natalia Rybczynski - 娜塔莉亚·雷布琴斯基

[History + Translation + Extraction Few-Shot Prompt]
Historical information:
English source: Natalia Rybczynski: Yeah, I had someone call me "Dr. Dead Things."

Chinese translation: 娜塔莉亚·雷布琴斯基：是的，有人叫我“死物博士”
New proper nouns: N/A

[History + Translation + Extraction Few-Shot Prompt]
Historical information:
English source: Latif Nasser: And I think she's particularly interesting because of where she digs that stuff up, way above the Arctic Circle in the remote Canadian tundra.

Chinese translation: 拉蒂夫·纳瑟：我认为她特别有趣，因为她在偏远的加拿大冻土地区北极圈的地方挖掘这些东西。
New proper nouns: Latif Nasser - 拉蒂夫·纳瑟, Arctic Circle - 北极圈, Canadian - 加拿大, tundra - 冻土地区

...

Proper Noun Records

* Empty *

Natalia Rybczynski: 娜塔莉亚·雷布琴斯基

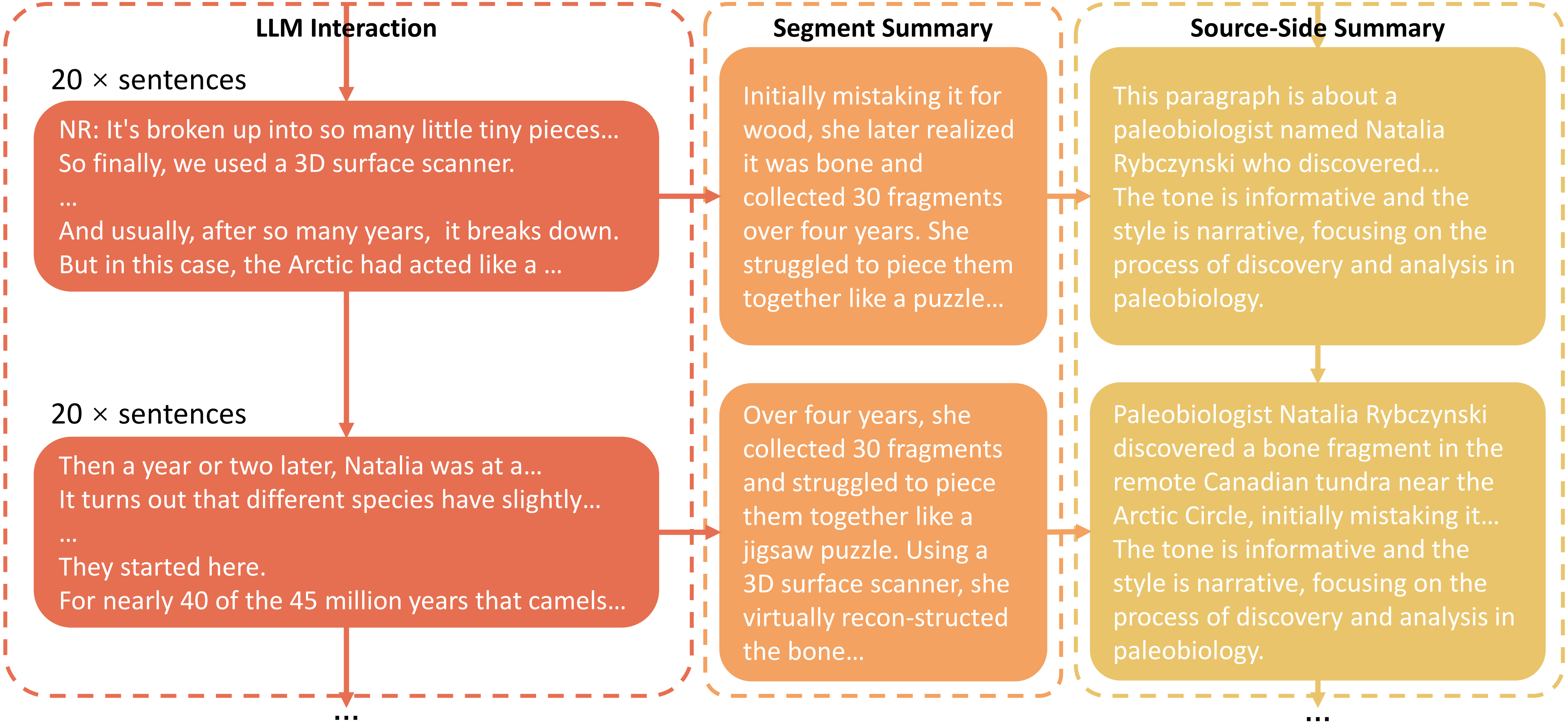
Natalia Rybczynski: 娜塔莉亚·雷布琴斯基

Natalia Rybczynski: 娜塔莉亚·雷布琴斯基
Nasser: 拉蒂夫·纳瑟
Arctic Circle: 北极圈
Canadian: 加拿大
tundra: 冻土地区

...

Methodology

- **Bilingual Summary**: Generate a segment summary for current source/target window first, then merge it into the source/target side summary (updated iteratively)



LLM Interaction

20 × sentences

NR: It's broken up into so many little tiny pieces…
So finally, we used a 3D surface scanner.
…
And usually, after so many years, it breaks down.
But in this case, the Arctic had acted like a …

20 × sentences

Then a year or two later, Natalia was at a…
It turns out that different species have slightly…
…
They started here.
For nearly 40 of the 45 million years that camels…

…

Segment Summary

Initially mistaking it for wood, she later realized it was bone and collected 30 fragments over four years. She struggled to piece them together like a puzzle…

Over four years, she collected 30 fragments and struggled to piece them together like a jigsaw puzzle. Using a 3D surface scanner, she virtually recon-structed the bone…

Source-Side Summary

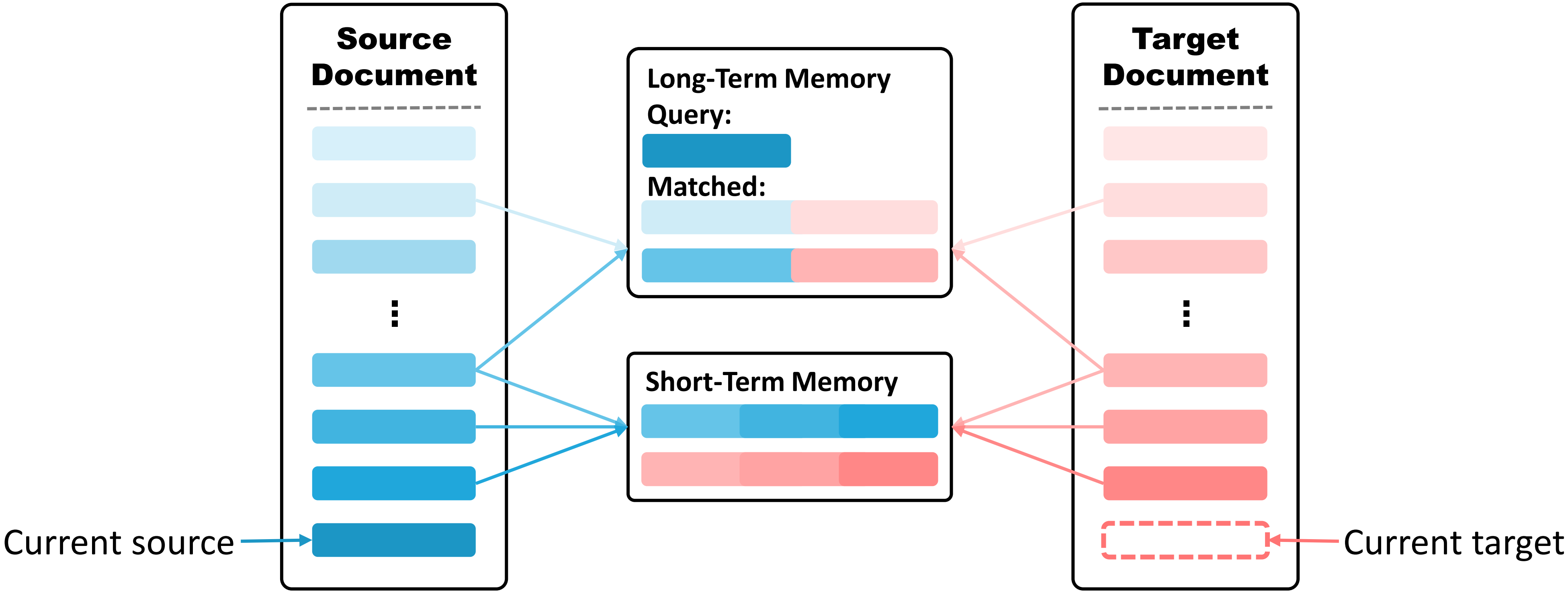
This paragraph is about a paleobiologist named Natalia Rybczynski who discovered…
The tone is informative and the style is narrative, focusing on the process of discovery and analysis in paleobiology.

Paleobiologist Natalia Rybczynski discovered a bone fragment in the remote Canadian tundra near the Arctic Circle, initially mistaking it…
The tone is informative and the style is narrative, focusing on the process of discovery and analysis in paleobiology.

…

Methodology

- **Long & Short-Term Memory**
  - Short-Term Memory: Store the recent source-target pairs (smaller windows)
  - Long-Term Memory: LLMs retrieve the most relevant sentence pairs (larger windows)

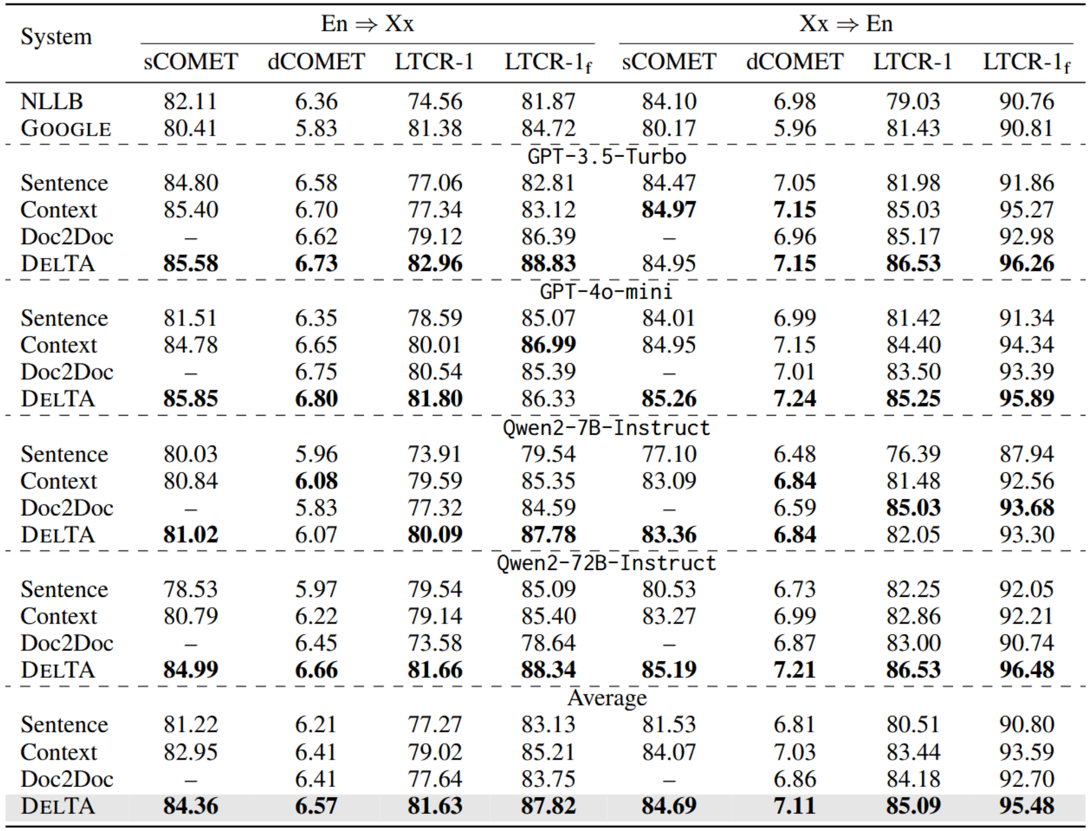


Main Results

- Experiment Settings
  - Datasets: ①IWSLT2017 (Speech, En⇔Zh, De, Fr, Ja) ②Guofeng (Web novel, Zh⇒En)
  - Models: ①GPT-3.5-Turbo ②GPT-4o-mini ③Qwen2-7B-Instruct ④Qwen2-72B-Instruct
  - Metrics: ①sCOMET, dCOMET (Quality) ②LTCR-1, LTCR-1_f (Consistency)
  - Long-Term Memory window size: 20, retrieved sentence number: 2
  - Short-Term Memory window size: 3
  - Update Bilingual Summary every 20 sentences
- Baselines
  - Sentence: Translate sentence by sentence
  - Context: Translate with recent 3 source-target sentence pairs as in-context information
  - Doc2Doc: Translate 10 sentences at once, all previous context stored in chat history

Main Results

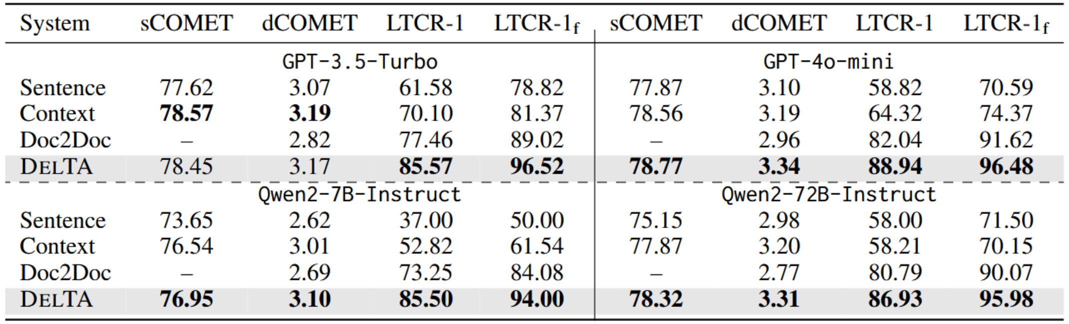
- IWSLT2017



System	En ⇒ Xx				Xx ⇒ En			
	sCOMET	dCOMET	LTCR-1	LTCR-1_f	sCOMET	dCOMET	LTCR-1	LTCR-1_f
NLLB	82.11	6.36	74.56	81.87	84.10	6.98	79.03	90.76
GOOGLE	80.41	5.83	81.38	84.72	80.17	5.96	81.43	90.81
GPT-3.5-Turbo								
Sentence	84.80	6.58	77.06	82.81	84.47	7.05	81.98	91.86
Context	85.40	6.70	77.34	83.12	**84.97**	**7.15**	85.03	95.27
Doc2Doc	–	6.62	79.12	86.39	–	6.96	85.17	92.98
DELTA	**85.58**	**6.73**	**82.96**	**88.83**	84.95	**7.15**	**86.53**	**96.26**
GPT-4o-mini								
Sentence	81.51	6.35	78.59	85.07	84.01	6.99	81.42	91.34
Context	84.78	6.65	80.01	**86.99**	84.95	7.15	84.40	94.34
Doc2Doc	–	6.75	80.54	85.39	–	7.01	83.50	93.39
DELTA	**85.85**	**6.80**	**81.80**	86.33	**85.26**	**7.24**	**85.25**	**95.89**
Qwen2-7B-Instruct								
Sentence	80.03	5.96	73.91	79.54	77.10	6.48	76.39	87.94
Context	80.84	**6.08**	79.59	85.35	83.09	**6.84**	81.48	92.56
Doc2Doc	–	5.83	77.32	84.59	–	6.59	**85.03**	**93.68**
DELTA	**81.02**	6.07	**80.09**	**87.78**	**83.36**	**6.84**	82.05	93.30
Qwen2-72B-Instruct								
Sentence	78.53	5.97	79.54	85.09	80.53	6.73	82.25	92.05
Context	80.79	6.22	79.14	85.40	83.27	6.99	82.86	92.21
Doc2Doc	–	6.45	73.58	78.64	–	6.87	83.00	90.74
DELTA	**84.99**	**6.66**	**81.66**	**88.34**	**85.19**	**7.21**	**86.53**	**96.48**
Average								
Sentence	81.22	6.21	77.27	83.13	81.53	6.81	80.51	90.80
Context	82.95	6.41	79.02	85.21	84.07	7.03	83.44	93.59
Doc2Doc	–	6.41	77.64	83.75	–	6.86	84.18	92.70
DELTA	**84.36**	**6.57**	**81.63**	**87.82**	**84.69**	**7.11**	**85.09**	**95.48**

Table 2: Test results on the IWSLT2017 dataset. Since the translations produced by the Doc2Doc method are not aligned at the sentence level with the source text, we do not report the sCOMET scores for this method. The highest score in each block is highlighted in **bold font** The results in the "Average" block represent the mean scores across the four backbone models.

- Guofeng



System	sCOMET	dCOMET	LTCR-1	LTCR-1_f	sCOMET	dCOMET	LTCR-1	LTCR-1_f
	GPT-3.5-Turbo				GPT-4o-mini			
Sentence	77.62	3.07	61.58	78.82	77.87	3.10	58.82	70.59
Context	**78.57**	**3.19**	70.10	81.37	78.56	3.19	64.32	74.37
Doc2Doc	–	2.82	77.46	89.02	–	2.96	82.04	91.62
DELTA	78.45	3.17	**85.57**	**96.52**	**78.77**	**3.34**	**88.94**	**96.48**
	Qwen2-7B-Instruct				Qwen2-72B-Instruct			
Sentence	73.65	2.62	37.00	50.00	75.15	2.98	58.00	71.50
Context	76.54	3.01	52.82	61.54	77.87	3.20	58.21	70.15
Doc2Doc	–	2.69	73.25	84.08	–	2.77	80.79	90.07
DELTA	**76.95**	**3.10**	**85.50**	**94.00**	**78.32**	**3.31**	**86.93**	**95.98**

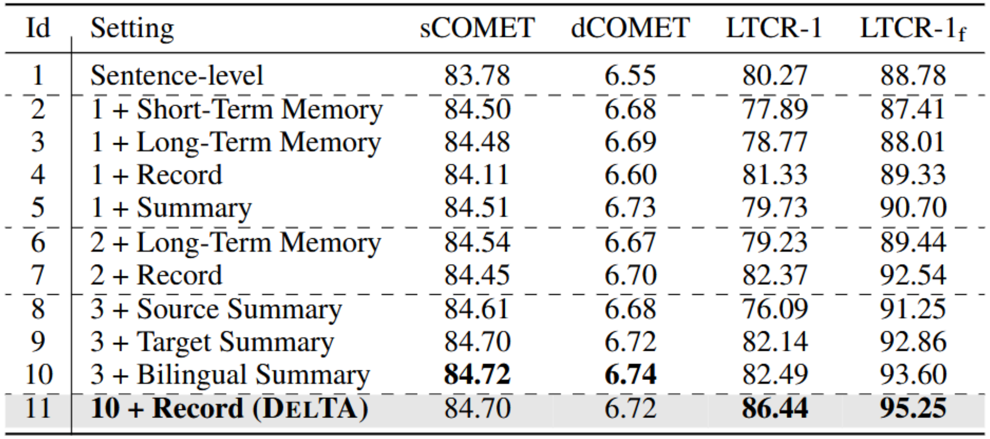
Table 3: Test results on the Guofeng dataset.

✓ **DELTA** Improves both translation consistency and quality.

✓ In Guofeng, translation consistency is more difficult to maintain, but **DELTA** still works well.

Analysis

- Ablation Study

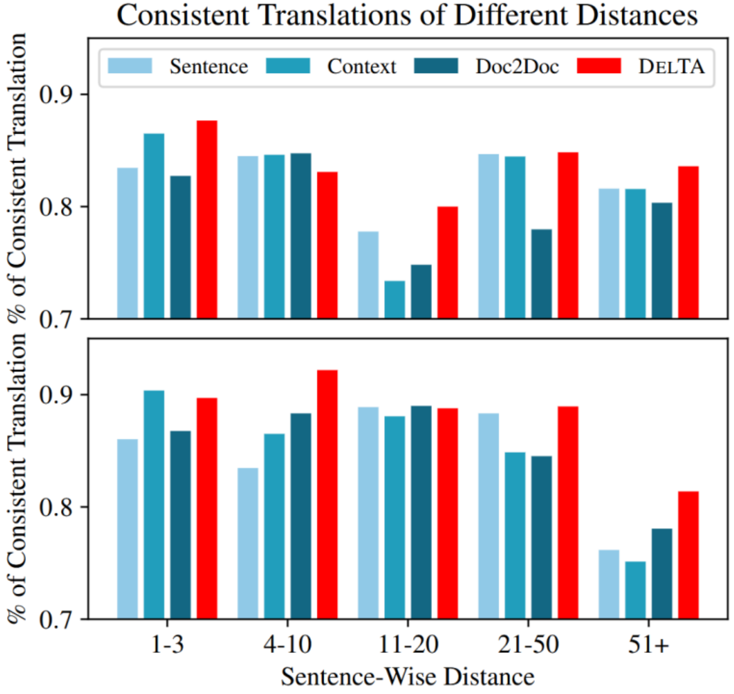


Id	Setting	sCOMET	dCOMET	LTCR-1	LTCR-1_f
1	Sentence-level	83.78	6.55	80.27	88.78
2	1 + Short-Term Memory	84.50	6.68	77.89	87.41
3	1 + Long-Term Memory	84.48	6.69	78.77	88.01
4	1 + Record	84.11	6.60	81.33	89.33
5	1 + Summary	84.51	6.73	79.73	90.70
6	2 + Long-Term Memory	84.54	6.67	79.23	89.44
7	2 + Record	84.45	6.70	82.37	92.54
8	3 + Source Summary	84.61	6.68	76.09	91.25
9	3 + Target Summary	84.70	6.72	82.14	92.86
10	3 + Bilingual Summary	**84.72**	**6.74**	82.49	93.60
11	**10 + Record (DELTA)**	84.70	6.72	**86.44**	**95.25**

Table 10: More detailed results of the ablation study.

- **Long & short-term** memory improves **quality**.
- **Proper noun records** improves **consistency**.
- **Bilingual Summaries** is better than monolingual summaries.

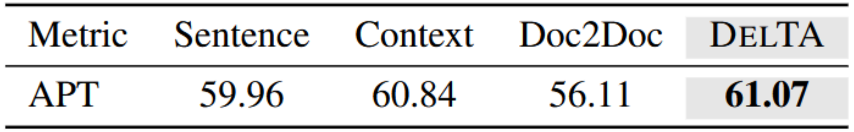
➢ Consistency over Long Distances



- **DELTA** is able to maintain **consistency** of proper nouns **over a longer span**

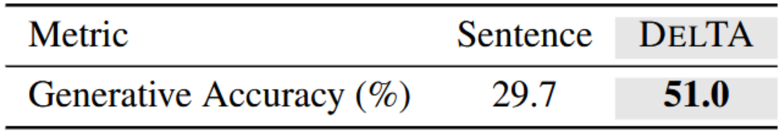
Analysis

- Accuracy of pronoun translation (APT)



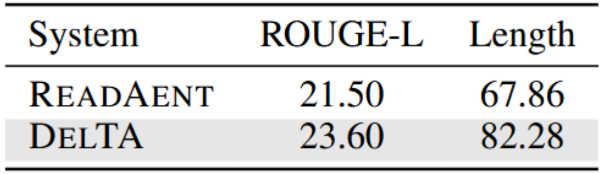
Metric	Sentence	Context	Doc2Doc	DELTA
APT	59.96	60.84	56.11	**61.07**

- Context-dependent translation



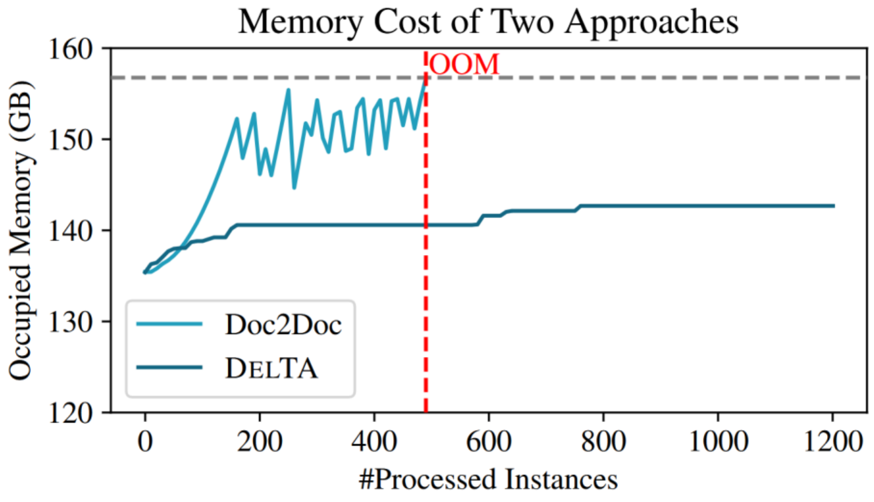
Metric	Sentence	DELTA
Generative Accuracy (%)	29.7	**51.0**

- Use DELTA's summary module to solve the query-based summarization task



System	ROUGE-L	Length
READAENT	21.50	67.86
DELTA	23.60	82.28

- GPU memory costs



- The Doc2Doc baseline method reserves all previous context stored in chat history, resulting in high GPU memory consumption.
- **DELTA** achieves higher quality and consistency **with lower memory usage**.
- Lower expansion and deployment costs, higher feasibility.

Thanks for your listening!