

从多个视角看机器翻译研究

宗成庆

中国科学院自动化研究所

cqzong@nlpr.ia.ac.cn

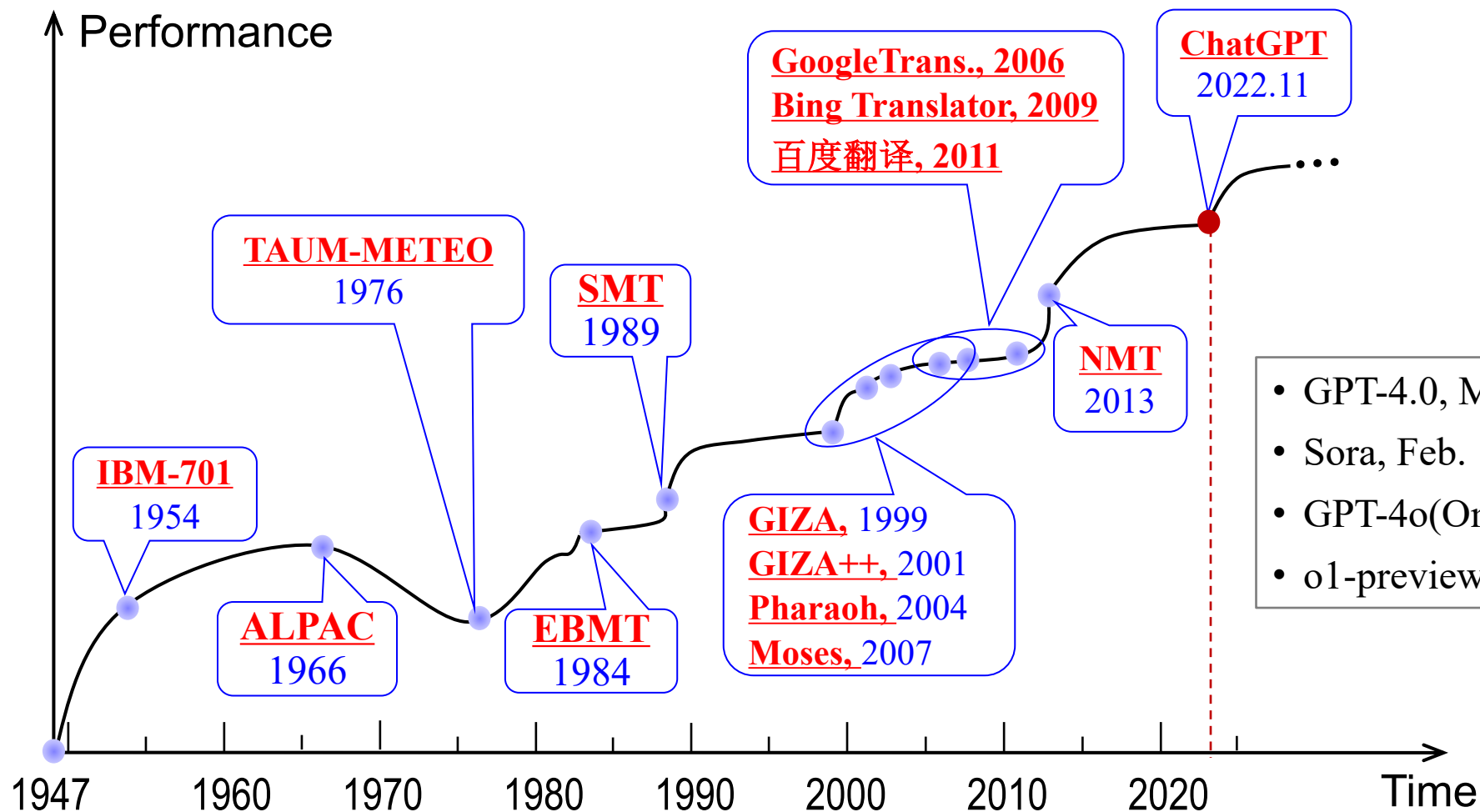
提 纲

- ➔ 1. 回顾
- 2. 现状分析
- 3. 相关研究
- 4. 未来展望

1. 回顾

Expectation/

Performance



1. 回顾

- ~ 2000: Rule-Based Machine Translation (RBMT)
- 2000: RBMT → Statistical MT (SMT)

“基于规则的机器翻译是傻子，基于统计的机器翻译是疯子。”

原文: Beijing made a third solemn **representation** to Manila and warned that it is hard to be optimistic about a territorial impasse over an island. Authorities say they have prepared for any escalation of the situation by Manila.

(<http://www.chinadaily.com.cn/>, 8 May 2012)

Results given by Google Translator:

1. 回顾

- ✧北京提出了第三次严正交涉，马尼拉，并警告说，这是大约一个多岛屿的领土僵局难以乐观。当局说，他们已经准备为任何升级的情况下由马尼拉。[\(2012.5.8\)](#)
- ✧北京做了第三马尼拉严正交涉，并警告说，这是很难被看好在一个小岛的领土僵局。当局说，他们已经准备任何马尼拉局势升级。[\(2013.5.7\)](#)
- ✧北京做第三严正交涉到马尼拉，并警告说，这是很难约为领土僵局在一个小岛乐观。当局说，他们已经准备的情况下由马尼拉的任何升级。[\(2014.5.5\)](#)
- ✧北京做第三严正交涉到马尼拉，并警告说这是很难约领土僵局的一个岛屿乐观。当局说，他们已经准备了马尼拉的情况有任何升级。[\(2015.4.28\)](#)
- ✧北京由第三严正交涉到马尼拉，并警告说这是很难约了一个岛领土僵局持乐观态度。当局说，他们已经为马尼拉局势的升级准备。[\(2016.5.1\)](#)

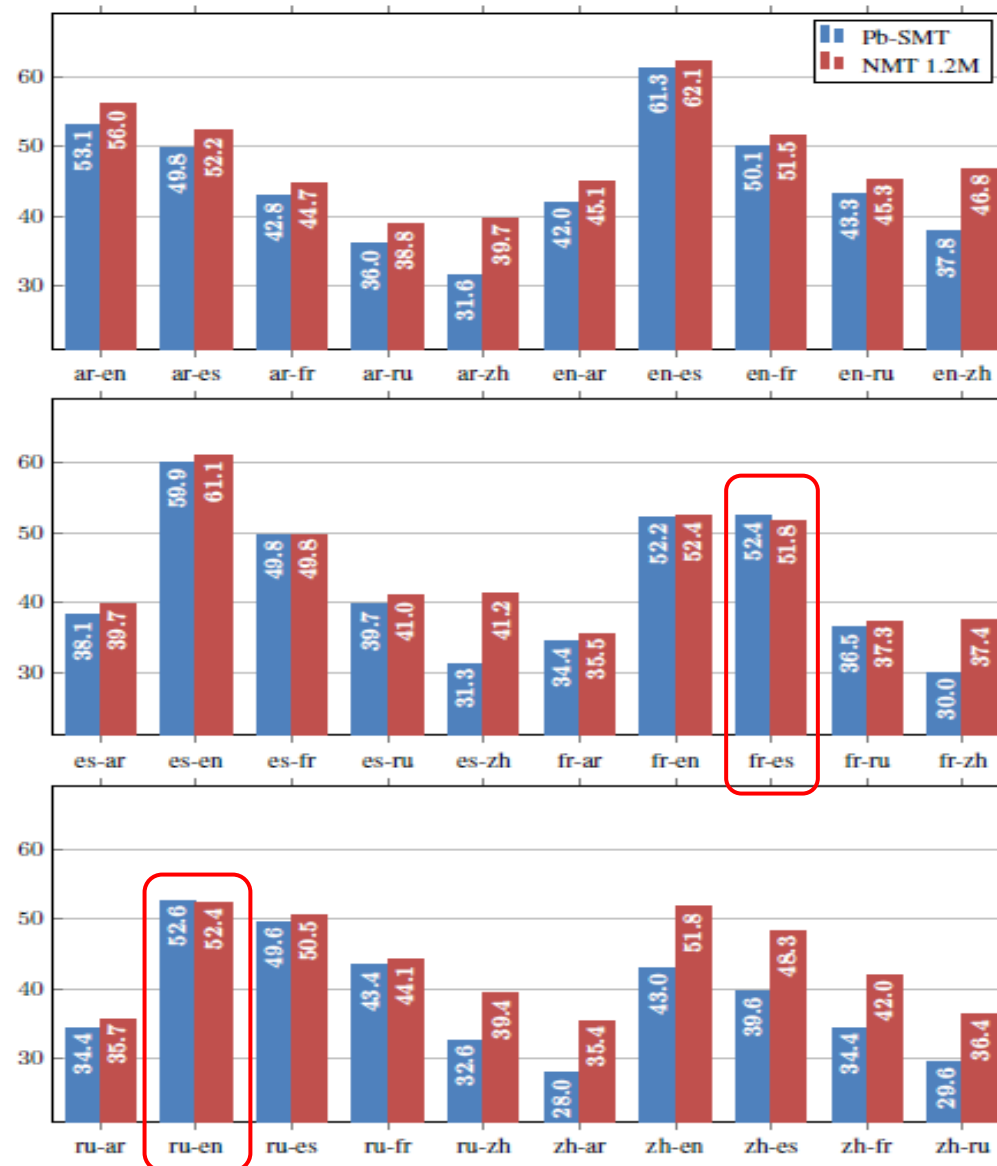
1. 回顾

- ✧北京对马尼拉进行了第三次庄严的代表，并警告说，对岛上的领土僵局很难看好。当局表示，他们为马尼拉的情况升级做好了准备。[\(2017.5.15\)](#)
- ✧北京向马尼拉提出了第三次严正交涉，并警告说很难对一个岛屿的领土僵局持乐观态度。当局表示，他们已经为马尼拉局势升级做好了准备。[\(2018.5.20\)](#)
- ✧北京向马尼拉提出了第三次庄严的代表，并警告说，对一个岛屿的领土僵局很难保持乐观。有关当局表示，他们已经为马尼拉的局势升级做好了准备。[\(2019.5.6\)](#)
- ✧北京对马尼拉进行了第三次庄严的代表，并警告说，对一个岛屿的领土僵局不容乐观。当局说，他们已经为马尼拉局势的任何升级做好了准备。[\(2020.10.21\)](#)
- ✧北京第三次向马尼拉提出严正交涉，并警告说，对一个岛屿的领土僵局很难乐观。当局表示，他们已为马尼拉局势升级做好准备。[\(2021.8.24\)](#)
- ✧北京向马尼拉提出了第三次严正交涉，并警告说，很难对一个岛屿的领土僵局感到乐观。当局表示，他们已为马尼拉局势升级做好准备。[\(2022.5.24\)](#)
- ✧北京第三次向马尼拉提出严正交涉，并警告说，岛屿领土争端的僵局不容乐观。当局表示，他们已做好应对马尼拉局势升级的准备。[\(2023.12.12\)](#)

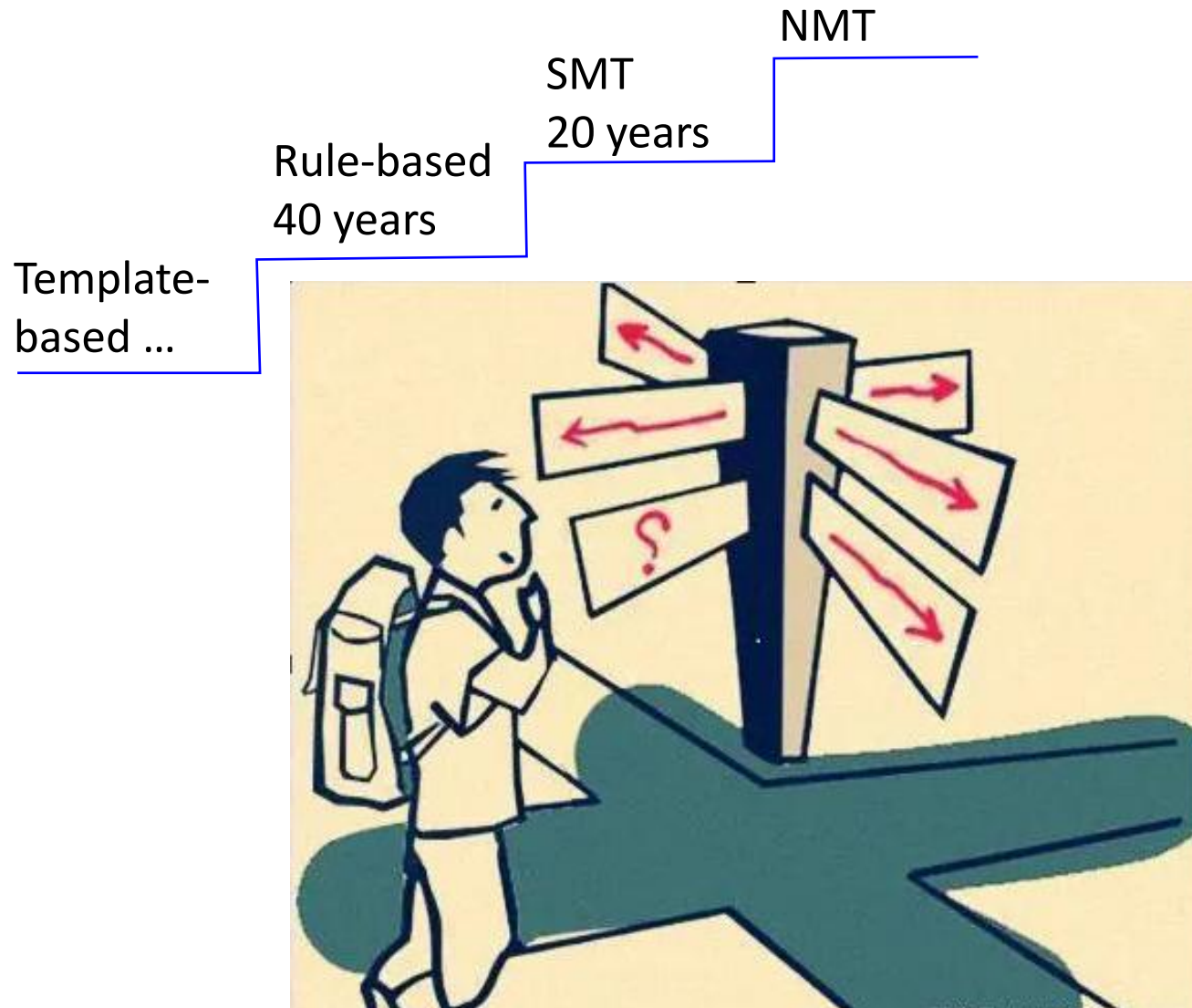
4. 神经机器翻译

● 2016: SMT → Neural MT (NMT)

Marcin Junczys-Dowmunt,
 Tomasz Dwojak and Hieu Hoang, [2016](https://arxiv.org/pdf/1610.01108.pdf). Is
 Neural Machine Translation Ready for
 Deployment? A Case Study on 30
 Translation Directions.
<https://arxiv.org/pdf/1610.01108.pdf>



1. 回顾



To be or not to be,
this is a question.

活着，还是不活，这是个问题。

提 纲

1. 回顾

→ 2. 现状分析

3. 相关研究

4. 未来展望



2. 现状分析

◆ 译文质量客观评价

- **2023.5:** (1)高资源语言：中文，德语
(2)中等资源语言：爱沙尼亚语，立陶宛语
(3)低资源语言：僧伽罗语，尼泊尔语

**Flores-200
测试集**

语言对	COMET			BLEU		
	ChatGPT	Google	比值	ChatGPT	Google	比值
汉语→英语	88.7	89.4	99.3	29.6	38.6	76.7
德语→英语	90.7	91.0	99.7	43.0	47.3	90.9
爱沙尼亚语→英语	90.1	91.7	98.3	36.9	45.5	81.1
立陶宛语→英语	86.1	89.2	96.6	32.6	41.3	78.9
尼泊尔语→英语	87.4	93.2	93.8	24.5	51.3	47.7
僧伽罗语→英语	61.3	90.5	67.7	3.5	45.5	7.7

2. 现状分析

➤ 引入世界知识的能力对比(adding world knowledge)

模型	翻译实例
Source	On Dec. 17, according to disclosure documents, Paul Pelosi bought 10 Alphabet call options with a strike price of \$2,000 with an expiration date set for mid-September 2022.
ChatGPT	根据披露文件，保罗·佩洛西于12月17日购买了10份 谷歌母公司 的认购期权，行使价格为2000美元，到期日定于2022年9月中旬。
Google Translate	根据披露文件，12月17日，保罗·佩洛西 (Paul Pelosi) 购买了 10 份 Alphabet 看涨期权，行使价为 2,000 美元，到期日定为 2022 年 9 月中旬。
Source	Visa was the sole Dow component to decline, with the other 29 stocks in the benchmark index finishing the day in the green .
ChatGPT	Visa是唯一一只道琼斯指数下跌的成分股，其他29只股票在基准指数中 收于涨势 。
Google Translate	Visa 是唯一一下跌的道琼斯指数成份股，该基准指数中的其他 29 只股票当天 以绿色收盘 。

2. 现状分析

➤ 关于幻觉表现的对比(hallucinations)

模型	翻译实例
Source (尼泊尔语)	चराहरूमा प्वाँखको विकासको अन्तरदृष्टि पनि यस खोजले प्रदान गर्छ।
ChatGPT	This search also provides insight into the development of wings in insects .
Google Translate	The discovery also provides insight into the evolution of feathers in birds .
Reference	The find also grants insight into the evolution of feathers in birds .
Source (僧伽罗语)	ඔබ සැතපුම් දහස් ගණනක් දුරින් ඉන්නා කෙනෙකුට කථා කරන විට, ඔබ චන්ද්‍රිකාවක් භාවිතා කරයි.
ChatGPT	When you speak with someone who understands your language , you use Chinese .
Google Translate	When you call someone who is thousands of miles away , you are using a satellite .
Reference	When you speak to someone thousands of miles away , you use a satellite .

武俊宏等, ChatGPT 能力分析与未来展望, 中国科学基金, 2023年05期 v.37 第735-742页

2. 现状分析

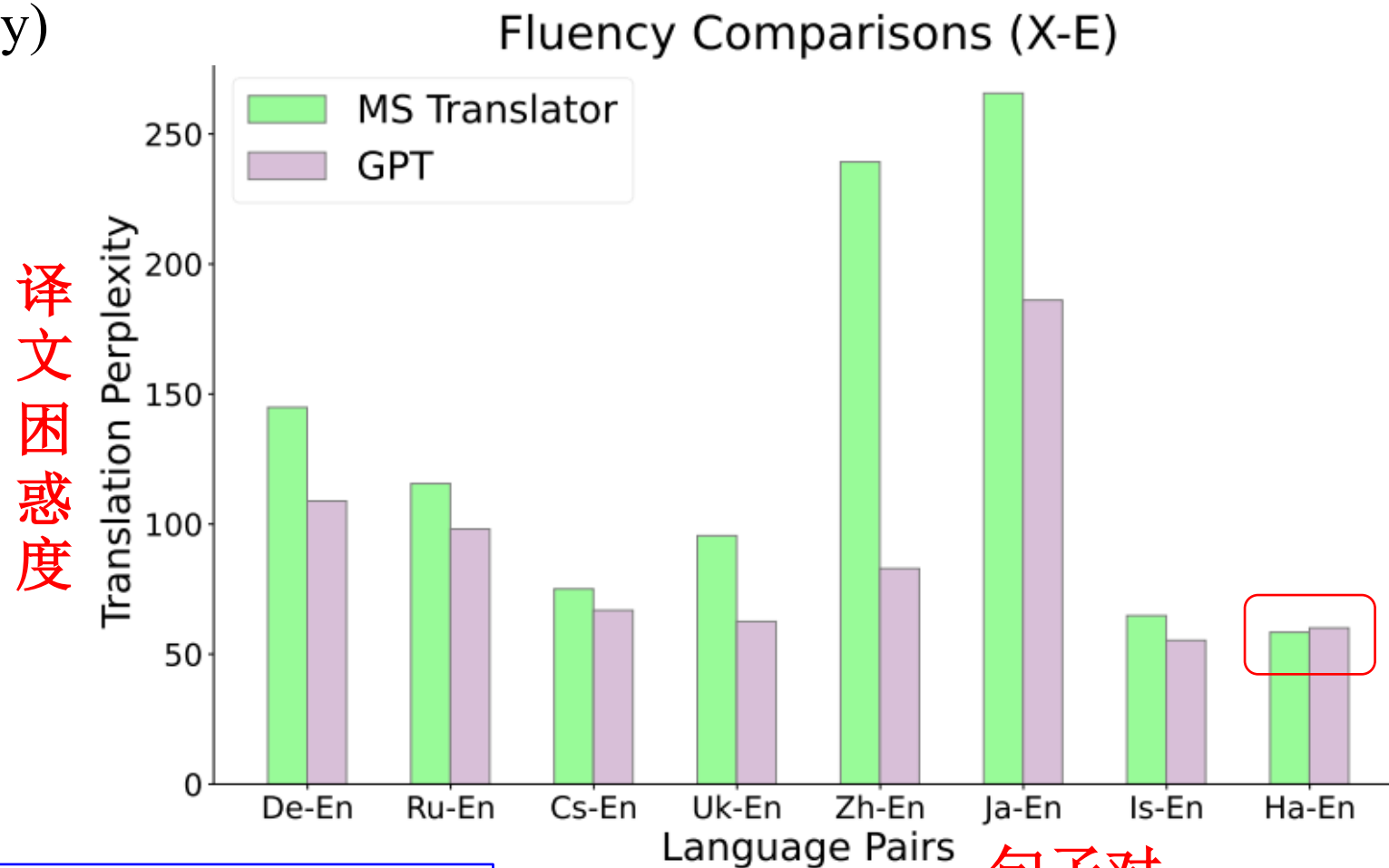
● GPT 译文特点分析

Sequence Type	Translation Instance	Phenomenon
Source	Bis auf die E 95 02 wurden <u>alle Lokomotiven zerlegt</u> .	Non-Monotonicity (NM)
MS Translator	With the exception of E 95 02, all locomotives were dismantled .	
GPT	<u>All locomotives were dismantled</u> except for the E 95 02.	
Source	Oder ist sie ganz aus dem Sortiment genommen?	Fluency (F)
MS Translator	Or is it completely removed from the range?	
GPT	<u>Or has it been</u> completely removed from the range?	
Source	Sehen Sie bitte im Screenshot was der Kollege geschrieben hat	Punctuation Insertion (PI)
MS Translator	Please see in the screenshot what the colleague wrote	
GPT	Please see the screenshot for what the colleague wrote.	
Source	Die Email zur Stornierung wurde am 26.12. <u>#NUMBER#</u> versendet.	Dropped Content (USW)
MS Translator	The cancellation email was sent on 26.12. #NUMBER# .	
GPT	The cancellation email was sent on December 26th.	
Source	"We won't accept the <u>CAA</u> and that is for sure.	Inserted Content (UTW)
MS Translator	“我们不会接受 CAA ，这是肯定的。	
GPT	“我们不会接受 <u>《公民法》</u> ，这是肯定的。	

Amr Hendy et al., Good Are GPT Models at Machine Translation? A Comprehensive Evaluation,
<https://arxiv.org/abs/2302.09210>

2. 现状分析

➤ 译文流畅性对比(fluency)



De- 德语 Ru- 俄语 CS- 捷克语 Uk- 乌克兰语
Zh- 中文 Is- 冰岛语 Ja- 日语 Ha- 豪萨语 En- 英语

2. 现状分析

● 2024.10

➤ 数据集

- WMT23: 德语(de), 中文(zh), 俄语(ru), 日语(ja), 乌克兰语(uk), 希伯来语(he)
- WMT22: 冰岛语(is)

➤ 翻译方向: $EN \rightarrow X$, $X \rightarrow EN$

	高资源	中等资源	低资源
与英语同语言族	de, ru	uk	is
与英语不同语言族	zh	ja	he

2. 现状分析

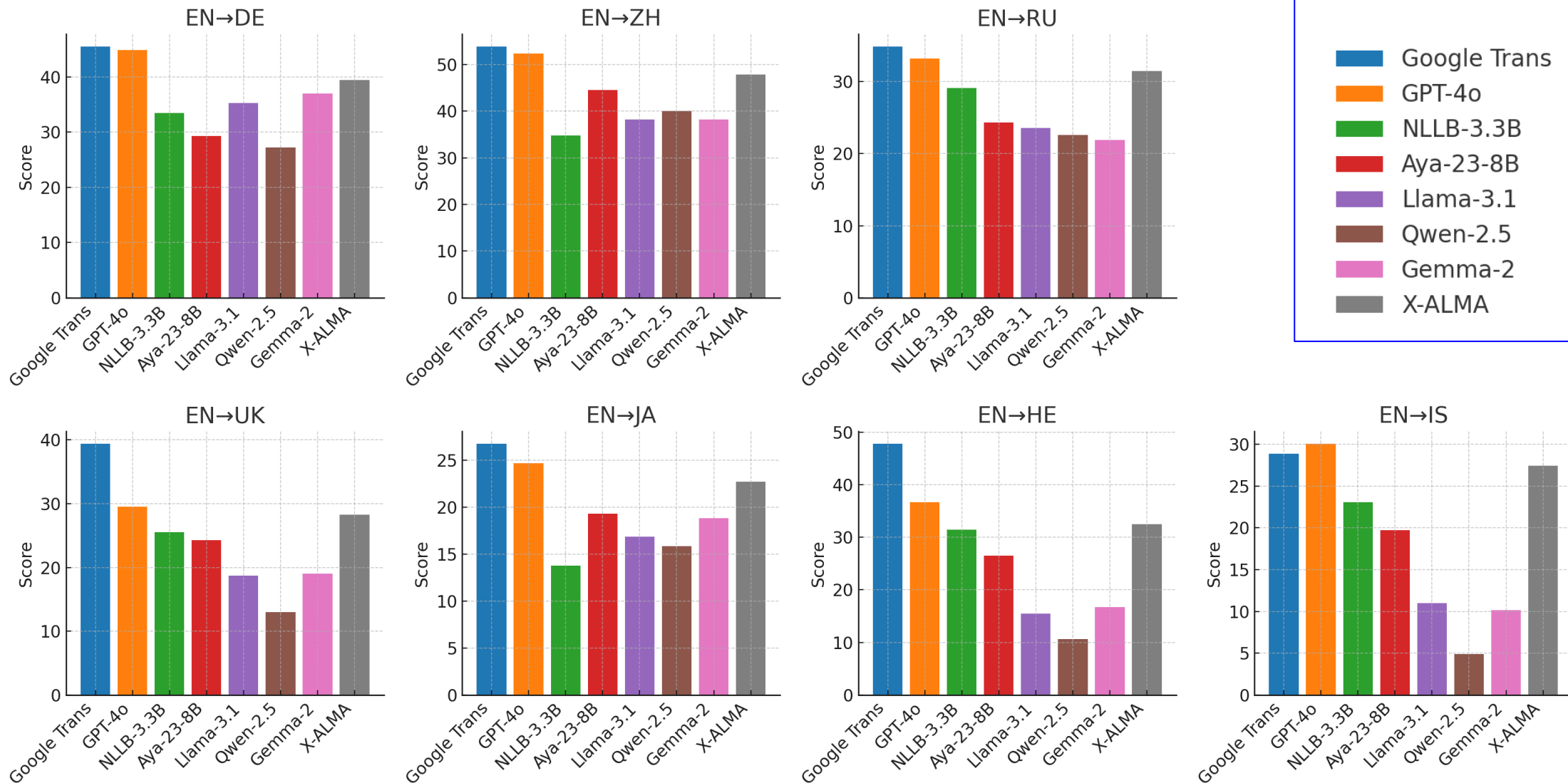
➤ 实验设置

➤ 模型/ 系统

- API: Google Translation, GPT-4o
- 开源指令模型: NLLB-3.3B, Aya-23-8B, Llama-3.1-8B,
Qwen-2.5-7B, Gemma-2-9B
- 基于LLM的翻译模型: X-ALMA

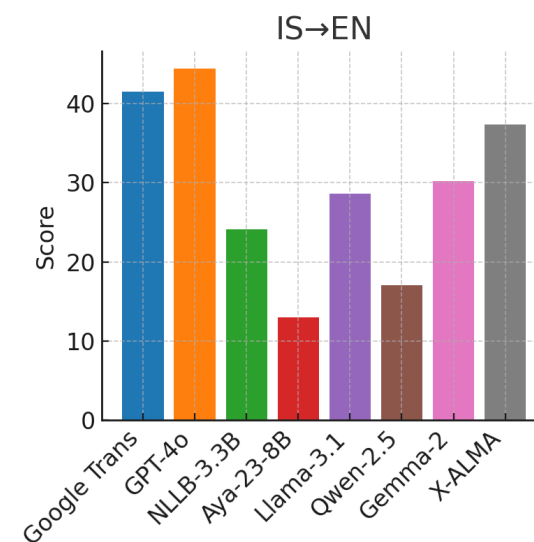
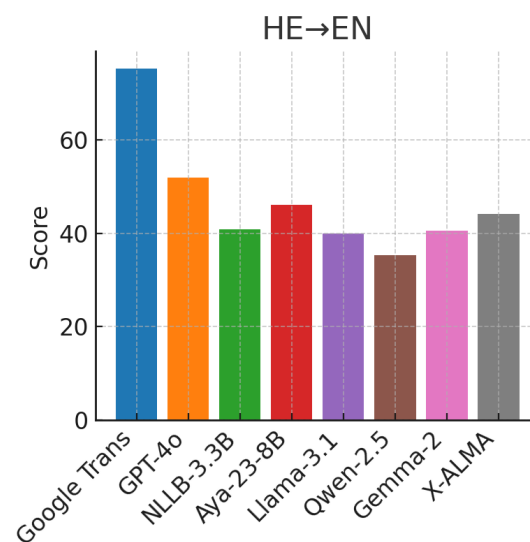
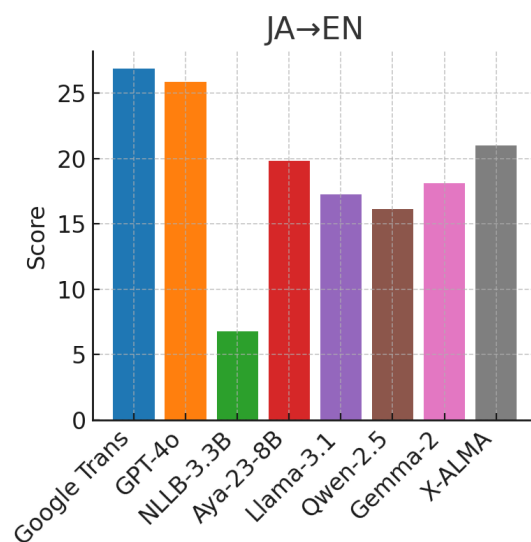
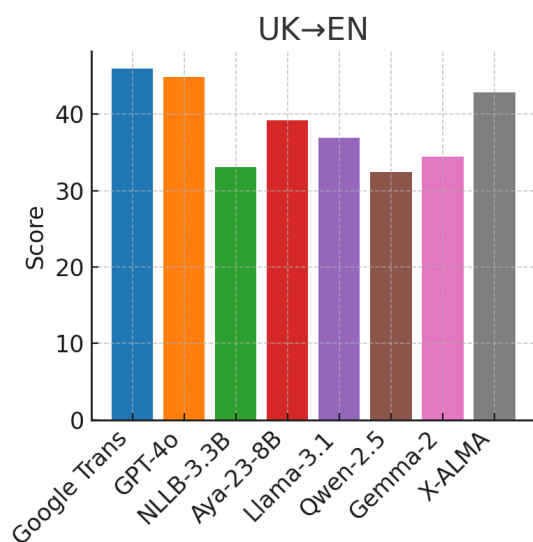
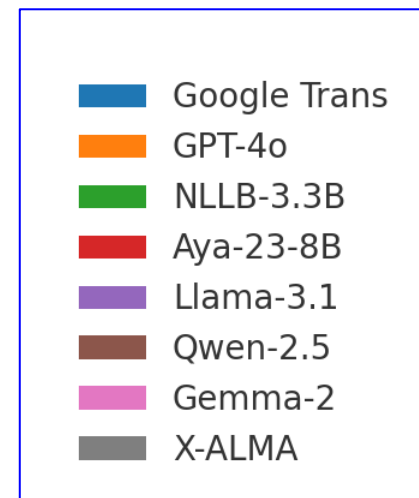
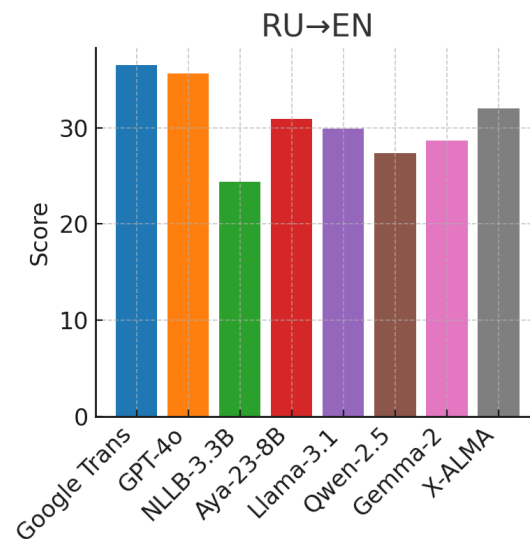
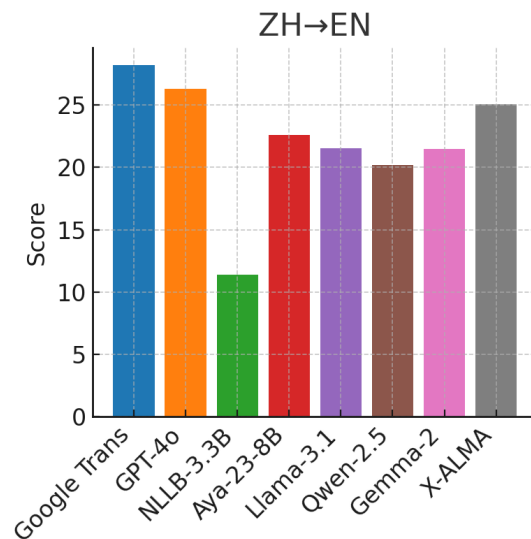
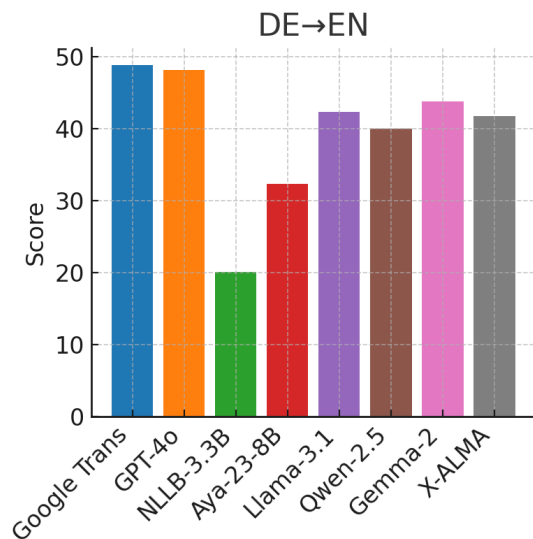
2. 现状分析

B
L
E
U



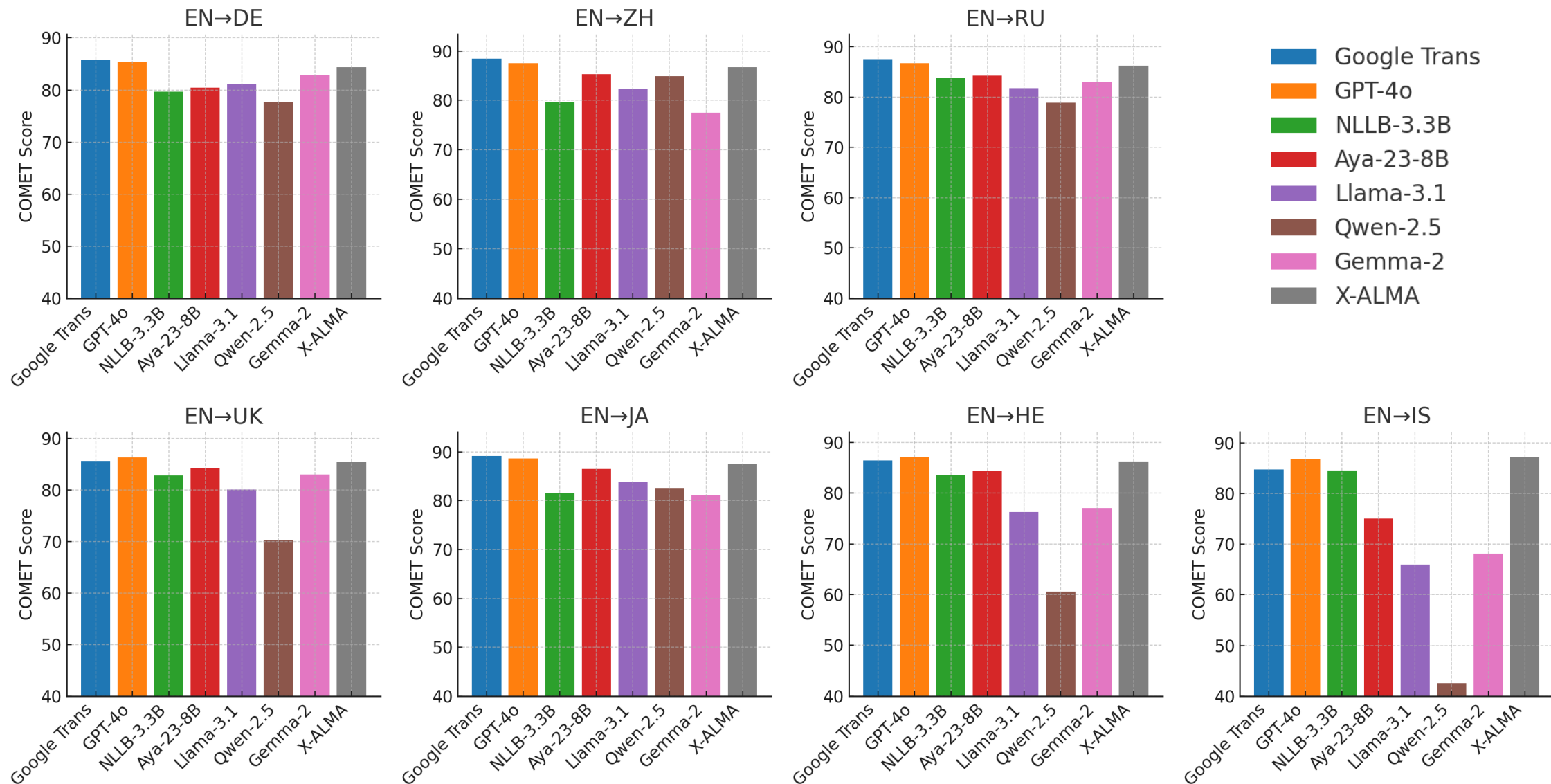
2. 现状分析

B
L
E
U



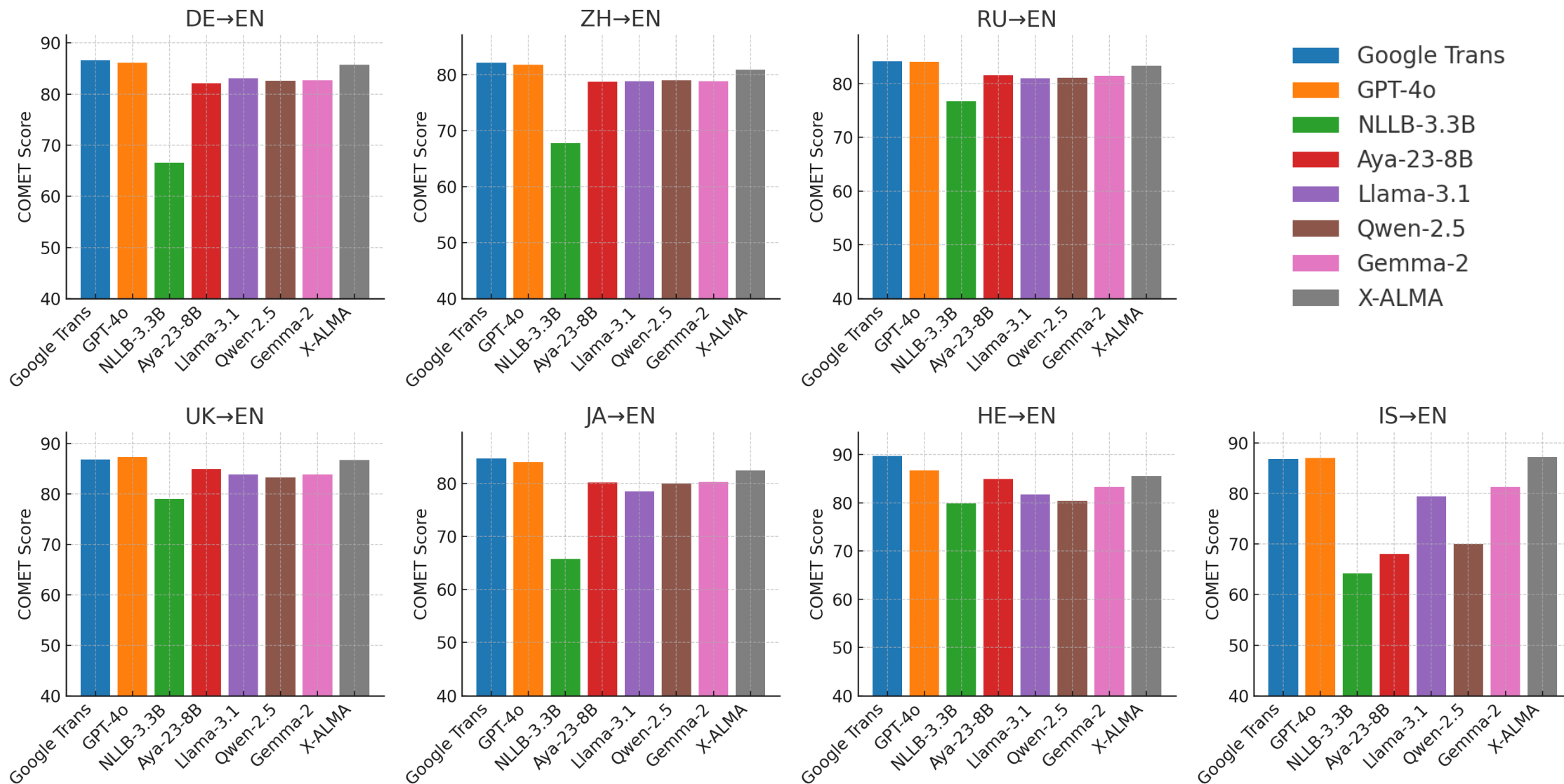
2. 现状分析

COMET-22



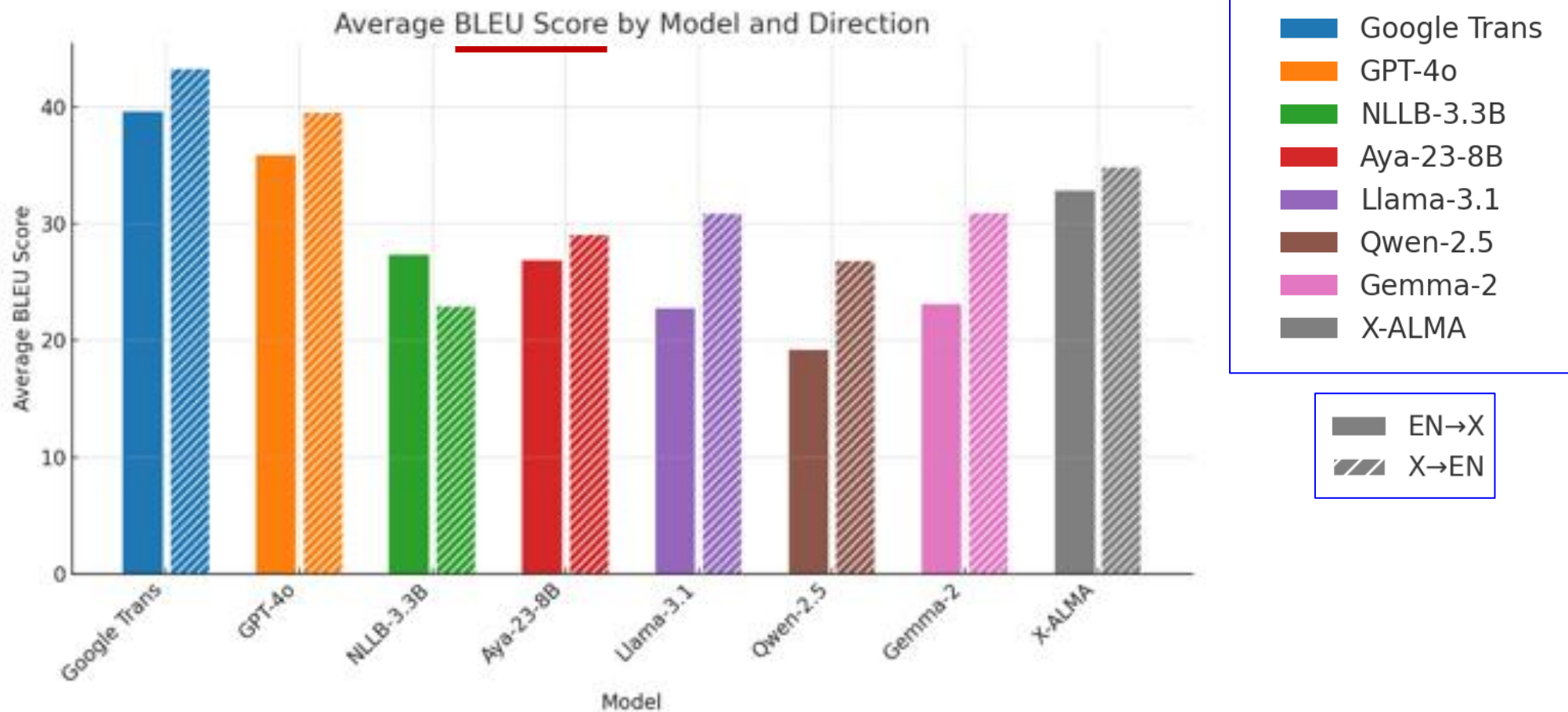
2. 现状分析

COMET-22



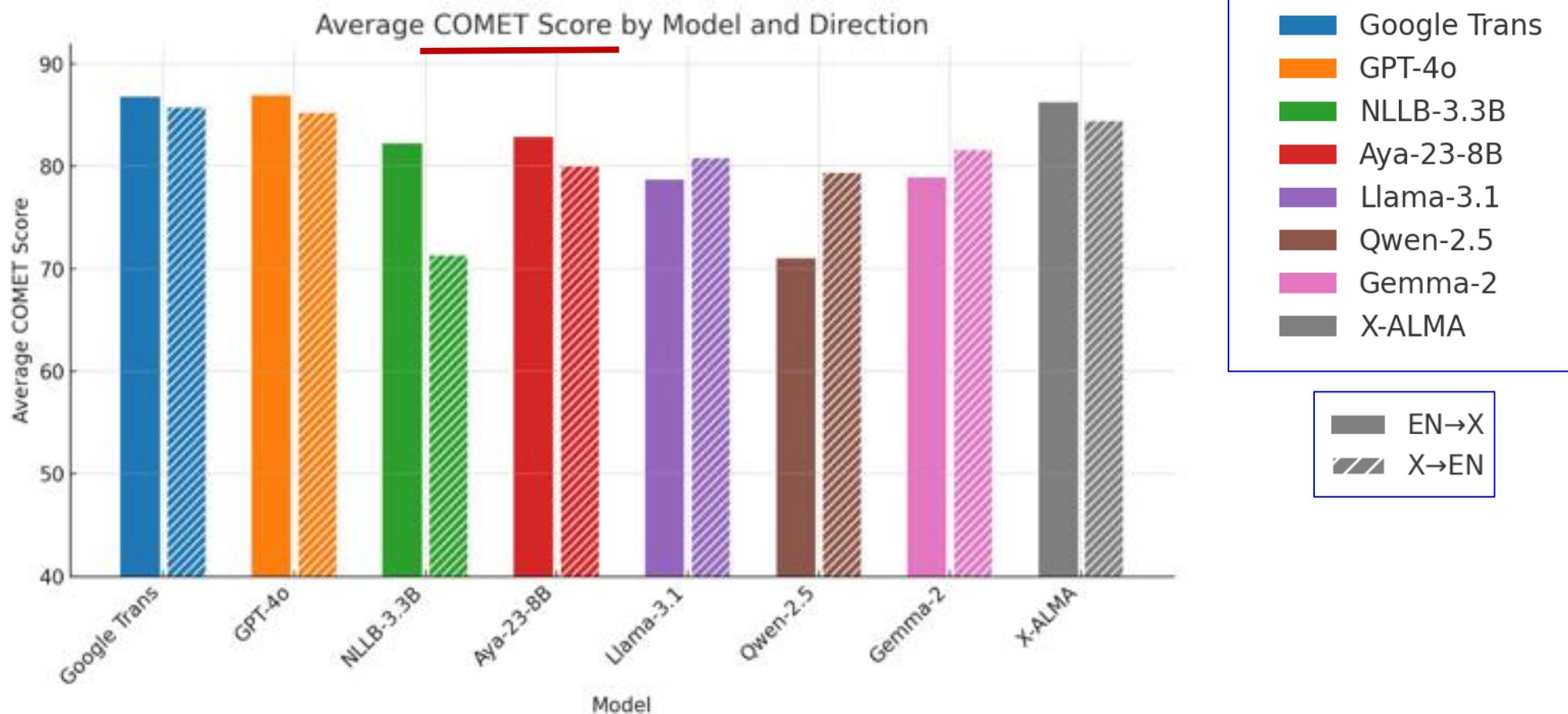
2. 现状分析

● 平均得分



2. 现状分析

● 平均得分



2. 现状分析

● 分析

- GPT-4o的翻译能力已与Google Translator相当；
- 开源大模型，如Llama-3.1、Qwen-2.5等在翻译性能上已经接近Encoder-Decoder结构的SOTA多语言翻译模型(NLLB)，但仍与闭源模型有一定差距；
- 大模型在英语、中文、德语等资源丰富的语言对上表现优异；在处理低资源语言时，通过迁移学习和知识共享，大模型可以在有限数据下实现相对合理的翻译效果，但仍逊于专门训练出来的基于Transformer架构的机器翻译模型；
- 大模型在相同语系的语言之间，如印欧语系的英语与德语、俄语之间表现较好，但在不同语系间，如英语与汉藏语系的中文之间表现略逊；
- 大模型在处理非英语为目标语言的翻译任务时 ($EN \rightarrow X$)，通常表现稍弱，这可能与训练时英语作为中心语言的偏向有关。

2. 现状分析

◆ 译文质量主观评价

✧ 梁茂成教授，北京航空航天大学

- ① 现在在本科阶段设有翻译专业，而翻译专业离开扎实的语言基本功显然是不可行的。从该专业的现状看，**翻译专业本科生的翻译水平比不过机器翻译。**
- ② 研究生阶段设有翻译专业硕士(MTI, Master of Translation and Interpretation)，从MTI的现状看，全国布点过多，导致MTI毕业生水平参差不齐，大多数翻译专业硕士学位点的毕业生质量堪忧，**很多毕业生的翻译水平比不上机器翻译。**



✧ 胡开宝教授，上海外国语大学

机器翻译(大语言模型)的译文质量超过我们培养的翻译专业的研究生，甚至博士生大部分学生翻译的水平。

2. 现状分析

黛玉自在枕上感念宝钗……又听见窗外竹梢焦叶之上，雨声淅沥，消寒透幕，不觉又滴下泪来。

(《红楼梦》第45回)

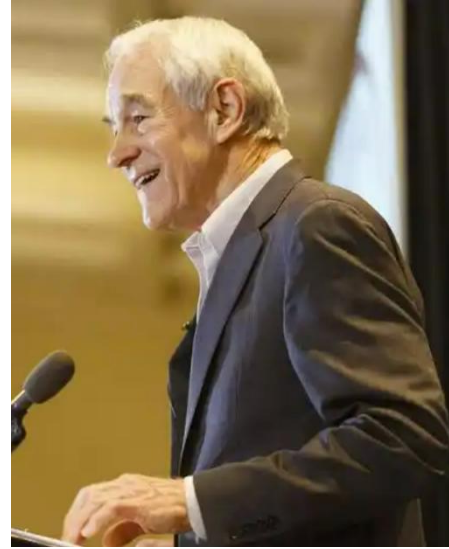
As she lay there alone, Dai-Yu's thoughts turned to Bao-chai ...Then she listened to the insistent rustle of the rain on the bamboos and plantains outside her window. The coldness penetrated the curtains of her bed. Almost without noticing it she had begun to cry.

— 文学翻译家 David Hawkes

摘自冯志伟著《机器翻译研究》，2004

Daiyu felt sorry for Baochai on the pillow..... **I heard** the sound of rain pouring on the scorched leaves of the bamboo shoots outside the window, and the cold was through the curtain, and **I didn't** realize that **I was** dripping tears again. (MS Translator, 2024.11.2)

2. 现状分析



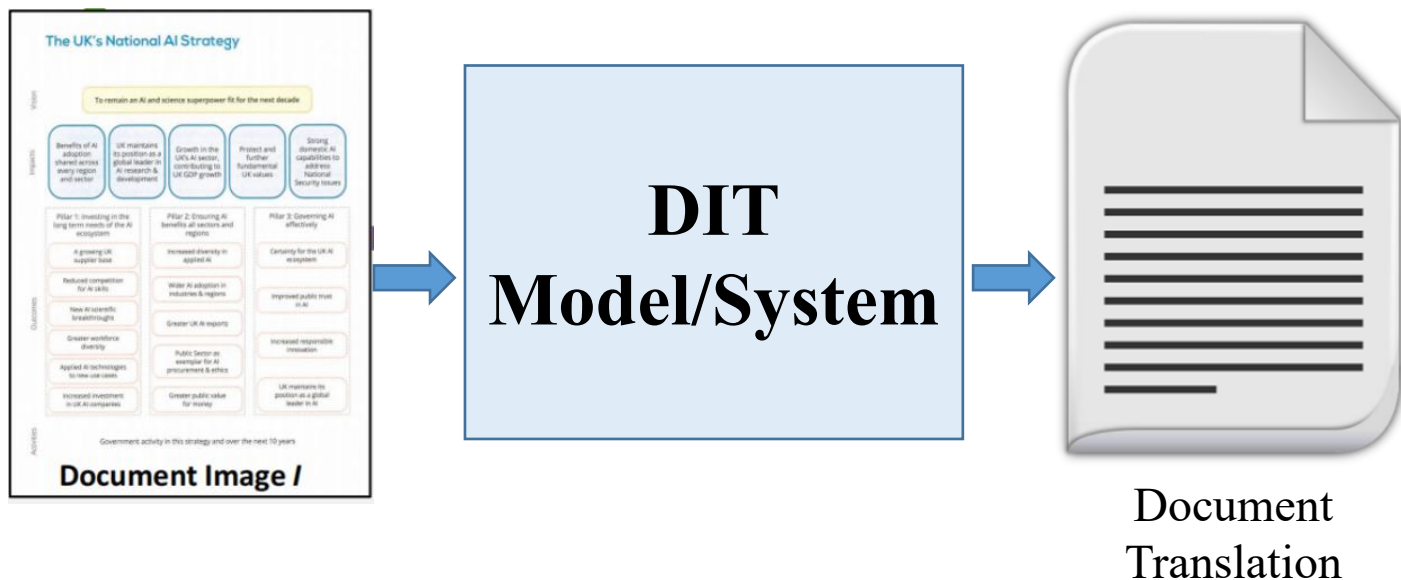
提 纲

1. 回顾
2. 现状分析
- ➔ 3. 相关研究
4. 未来展望

3. 相关研究

(1) 文档图像翻译

文档图像翻译(Document Image Translation, DIT)任务是感知和识别文档图像中包含的文字,并将其从源语言翻译成指定的目标语言。



$$\hat{Y} = \arg \max_Y P_{\theta} (Y | I, T, B)$$

$$= \arg \max_Y \prod_{j=1}^{|Y|} P_{\theta} (y_j | y_{<j}, I, T, B)$$

I - 文档图像

T - OCR识别出的单词

B - OCR识别出的单词版面位置坐标

Y - 目标语言的文档译文

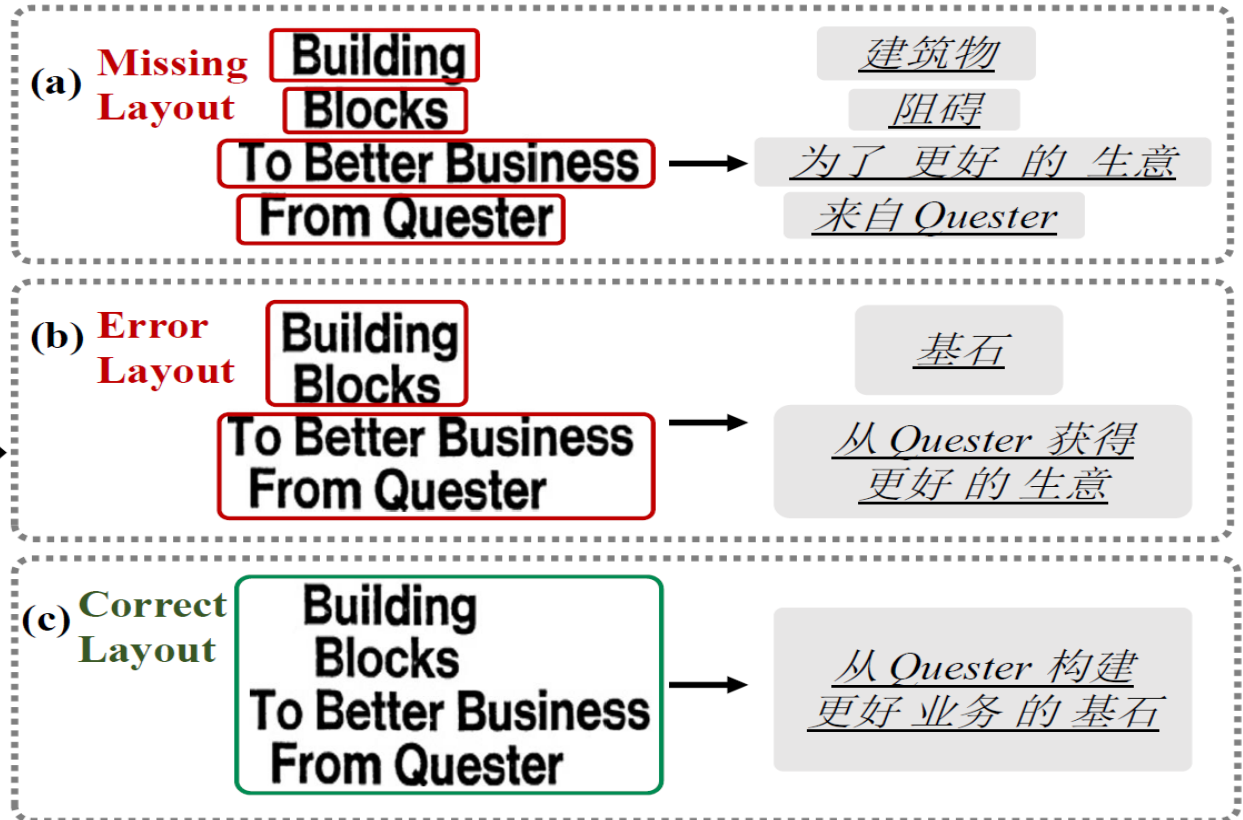
3. 相关研究

● 版面信息对文档图像翻译不可或缺

➤ 对版面信息的忽视（如图a）或错误理解（如图b）会导致翻译错误。



Building
Blocks
To Better Business
From Quester



Document Image

Raw OCR Results

OCR with Visual Layout

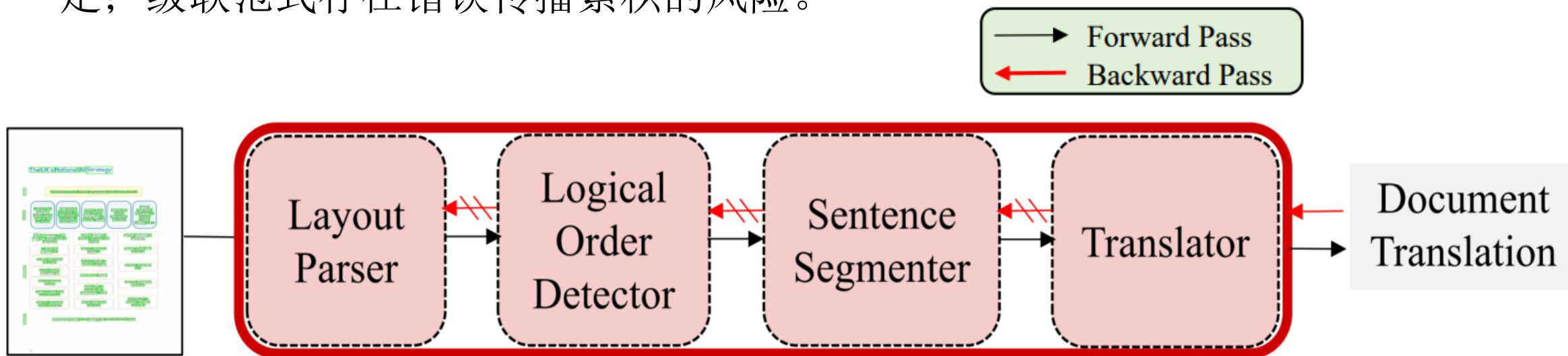
Translation Results

3. 相关研究

● 已有方法

传统的级联式文档图像翻译方法为版面理解和翻译分别部署独立的模型

- 优点：简化了版面和翻译各个子任务的建模难度
- 缺点：版面理解和翻译各个子任务独立无交互，可能导致翻译对版面的利用不足；级联范式存在错误传播累积的风险。



3. 相关研究

● 我们的方法

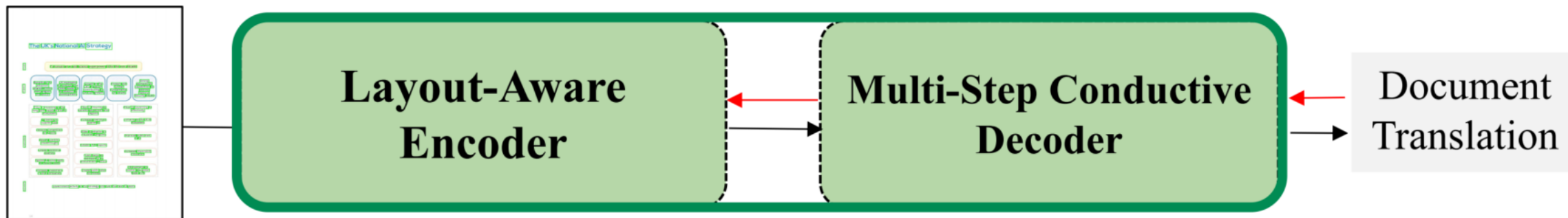
包含版面信息感知编码器和多步传导解码器两个主要模块：

➤ 版面感知编码器

- 提取和编码版面特征，并将版面和文本特征融合；
- 对融合后的版面-文本联合特征进行深度上下文融合。

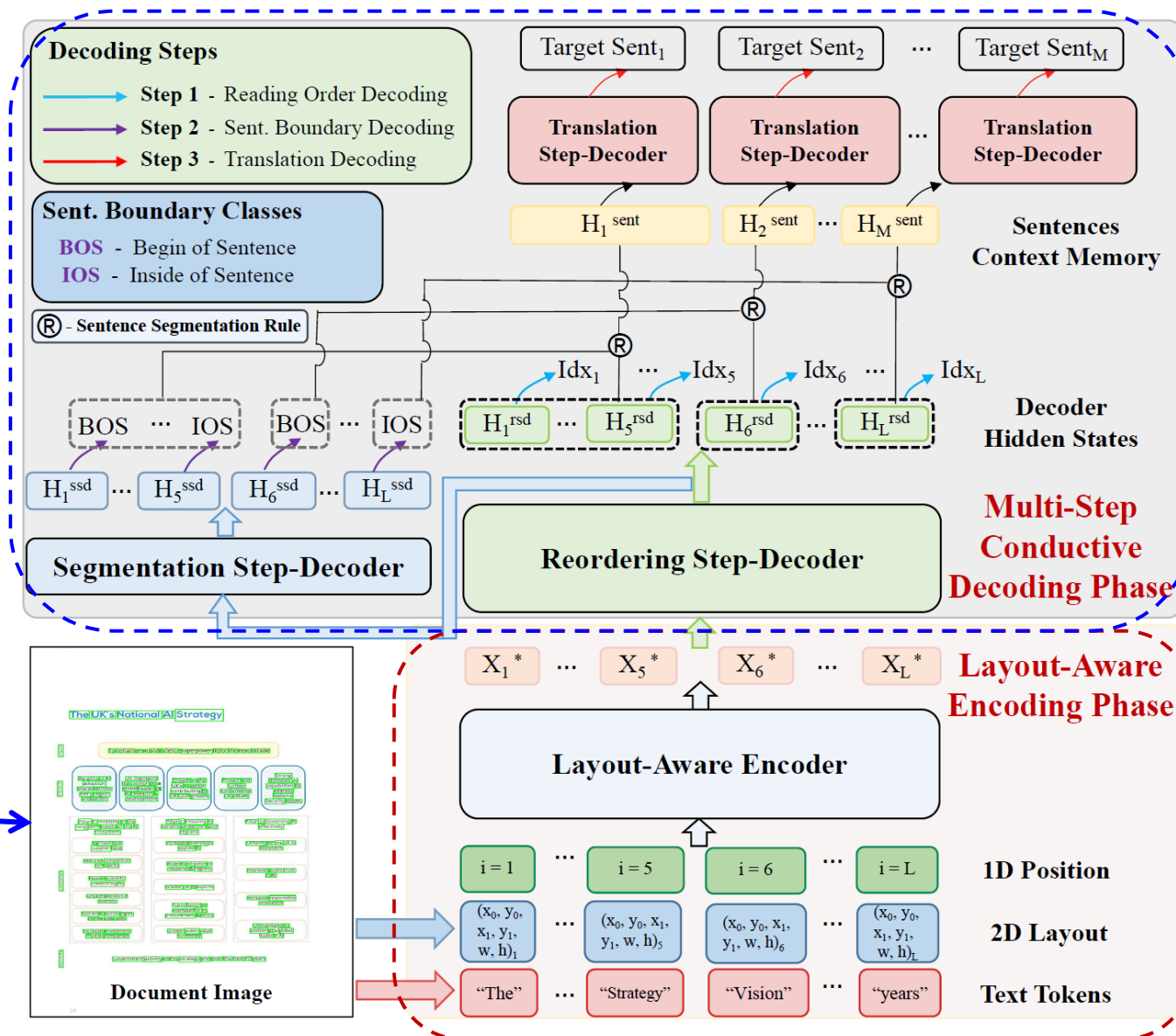
➤ 多步传导解码器

- 分而治之：多步骤解码逐步推理出必要的中间结果直至最终译文；
- 版面感知编码阶段与译文解码阶段端到端联合，最大化模块之间的协作，降低错误累积。



3. 相关研究

● 模型框架



- ① 解码文本顺序和句子边界，获得有逻辑的句子特征；
- ② 基于句子特征，逐句解码出目标语言译文；
- ③ 特征传导连接编码器和各个子解码器，端到端联合优化。

← 多步传导解码器

← 版面感知编码器

- ① 编码版面特征；
- ② 在输入端融合版面和文本特征；
- ③ 对版面-文本特征进行上下文深度融合。

提取出每个单词的文本和版面位置坐标。

光学字符识别OCR

3. 相关研究

- 实验

Table 1: Dataset Statistics

Dataset	ReadingBank	DITrans		
Domain	General	Political Report	Newspaper	Advertisement
# Page	410K	902	396	377
# Word Avg.	196	245	219	123
# Sent. Avg.	21	24	16	16
Train/Test (# Page)	400K/10K	790/112	347/49	330/47
Train/Test (# Sent.)	8,467K/211K	19,210/2,698	5,501/929	5,303/815

3. 相关研究

● 实验

Table 2: Results on the General-Domain Dataset

#	Method	End-to-End	# Params (M)	BLEU
1	DocHandler-1 (Sable et al., 2023)	-	142	30.47
2	DocHandler-2 (Sable et al., 2023)	-	172	37.75
3	LayoutLM-Dec	✓	232	45.54
4	LiLT-Dec	✓	250	45.79
5	LayoutDIT-Cascade	-	293	48.29
6	LayoutDIT	✓	206	48.20

- 所提方法LayoutDIT在通用领域场景的文档翻译质量显著超越传统的级联式方法DocHandler^[2].
- 同时显著优于其他的具备版面感知能力的端到端方法LayoutLM^[3]-Dec和LiLT-Dec^[4].

3. 相关研究

● 实验

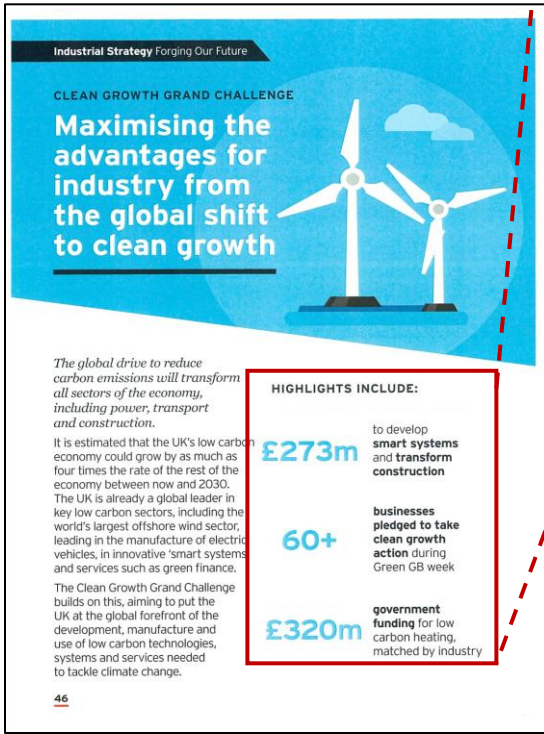
Table 3: Results in the Three Specific Domains of Our DITrans Dataset

#	Method	End-to-End	# Params (M)	Multi-Domain Learning			Zero-Shot Cross-Domain Transfer		
				Political Report	Newspaper	Advertisement	Political Report (SD)	Newspaper (TD.1)	Advertisement (TD.2)
1	DocHandler-1 (Sable et al., 2023)	-	142	21.47	26.28	23.14	21.80	16.87	13.55
2	DocHandler-2 (Sable et al., 2023)	-	172	24.80	27.34	24.99	25.19	16.88	13.60
3	MGTrans-DETR (Hinami et al., 2021)	-	212	28.30	27.61	25.76	27.34	16.26	13.09
4	MGTrans-Conv (Hinami et al., 2021)	-	238	28.99	28.84	25.63	29.51	17.78	14.07
5	LayoutLM-Dec	✓	232	45.86	40.66	36.90	45.71	27.56	21.40
6	LiLT-Dec	✓	250	46.11	40.37	36.22	46.61	23.71	18.14
7	LayoutDIT-Cascade	-	293	36.78	34.56	35.36	39.85	21.89	16.62
8	LayoutDIT	✓	206	46.97	38.82	45.72	46.70	29.01	29.46

- 在报告、新闻报纸和广告这三个特定的文档领域上，所提方法LayoutDIT显著优于最好的级联式方法MGTrans^[1]和端到端方法LiLT-Dec^[4]。
- 所提方法具备较好的零样本跨领域迁移能力，适用于样本稀缺领域的文档翻译。

3. 相关研究

● 实例



HIGHLIGHTS INCLUDE:

£273m

to develop
smart systems
and **transform**
construction

60+

businesses
pledged to take
clean growth
action during
Green GB week

£320m

government
funding for low
carbon heating,
matched by industry

Findings of
EMNLP 23

Model	Source Text (S) & Translation (T)
MGTrans	(1) S: HIGHLIGHTS INCLUDE: to develop smart systems T: 亮点 包括 开发 智能 系统 ✗
	(2) S: £273m and transform construction businesses pledged to take T: 在 绿色 GB 周 期间, 273 万英镑 的 转型 和 建筑 企业 承诺 采取 60 多名 清洁 增长 行动 ✗
	(3) S: 60+ clean growth action during Green GB week government funding for low T: 政府 对 低收入者 的 资助 ✗
	(4) S: £320m carbon heating, matched by industry T: 按 行业 匹配, 3.2 亿英镑 的 碳 供暖 ✗
LayoutDIT (Ours)	(1) S: HIGHLIGHTS INCLUDE: £273m T: 亮点 包括 2730 万英镑 ✗
	(2) S: to develop smart systems and transform construction T: 开发 智能 系统 和 改变 建设 ✓
	(3) S: 60+ T: 60+ ✓
	(4) S: businesses pledged to take clean growth action during Green GB week T: 企业 承诺 在 绿色 GB 周 期间 采取 清洁 的 增长 行动 ✓
	(5) S: £320m T: 3.2 亿英镑 ✓
	(6) S: government funding for low carbon heating, matched by industry T: 政府 对 低碳 加热 的 资金, 与 行业 匹配 ✓

3. 相关研究

◆ 数据集构建

序号	数据集名称	发表/在投的论文	规模	简介
1	E2ETIMT	ICPR 2022	3,016,282	端到端文本行级别图像文本翻译数据集，标注信息由<图片，源语言文本，目标译文>三元组组成，包含合成领域、字幕领域、以及街景领域，覆盖英中、英德、中英三个翻译方向。
2	DITtrans	EMNLP 2023	1,675	多粒度端到端文档图像翻译数据集，人工标注了高质量的英中翻译的多粒度信息，如单词内容、单词位置、句子分隔标识、句子阅读顺序、版面区域类型、版面位置以及句子级别的多粒度标注信息，可用于句子阅读顺序、篇章级别OCR、篇章级文档图像翻译等任务，覆盖政府报告、新闻、广告等不同领域。
3	DITtrans v2	TPAMI (under review)	411,796	多粒度端到端文档图像翻译数据集，在DITtrans的基础，进行数据规模、翻译方向以及领域的扩展，包含通用领域和报告、新闻、广告、书籍四个特定领域，覆盖英中、英德、英法三个翻译方向，每条数据包含<图片，源文本，源文本单词版面坐标和逻辑顺序，源文本句子边界，目标译文>细粒度标注信息。



3. 相关研究

序号	数据集名称	发表/在投的论文	规模	简介
4	DIT700K	COLING 2025 submitted	718,000	网络文档领域端到端文档图像文本翻译数据集，在DITtransv2基础上进一步增加标注数据的多样性，提出了一种全新的自动数据构建方式，覆盖英中、英德、中英三个翻译方向，每条数据包含<图片，源文本，源文本单词版面坐标，源文本句子顺序，目标译文>细粒度标注信息。
5	DoTA	NACCAL 2024	126,345	大规模学术论文结构化文档图像翻译数据集，让机器可以结构化输出学术论文的markdown译文结构，具体随机选取了2020年至2023年间在arXiv上发布的18,496篇论文，覆盖物理、数学、生物、金融、统计学、电子工程、经济学以及计算机等多个学科领域、每条数据包含<图片，译文文本markdown>标注信息。

3. 相关研究

(2) 同声传译

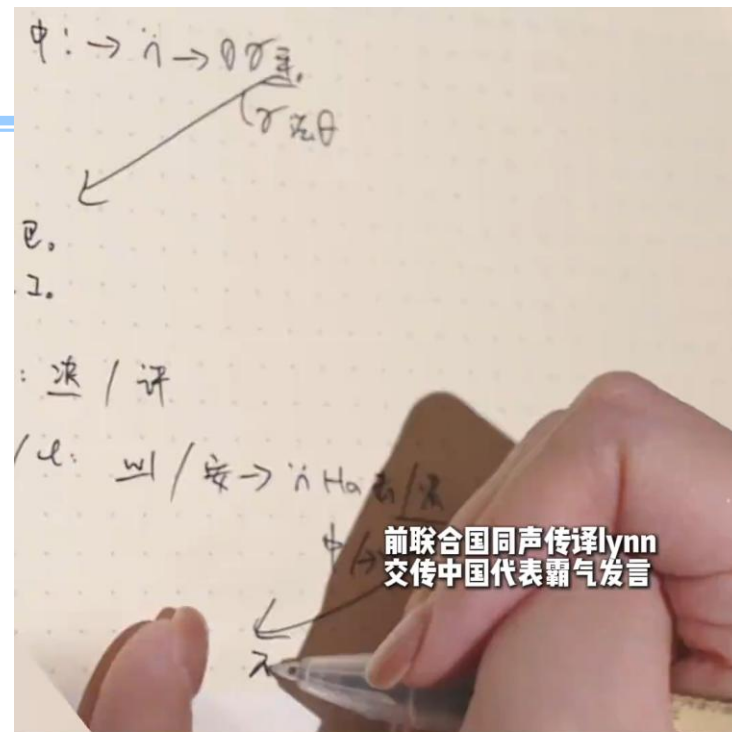
≠ 交互翻译(交译)

同声传译场景的特殊性:

- 除了译文的准确性, 听众希望译文便于理解;
- 尽量减少翻译的延迟, 尽快听到/看到译文。

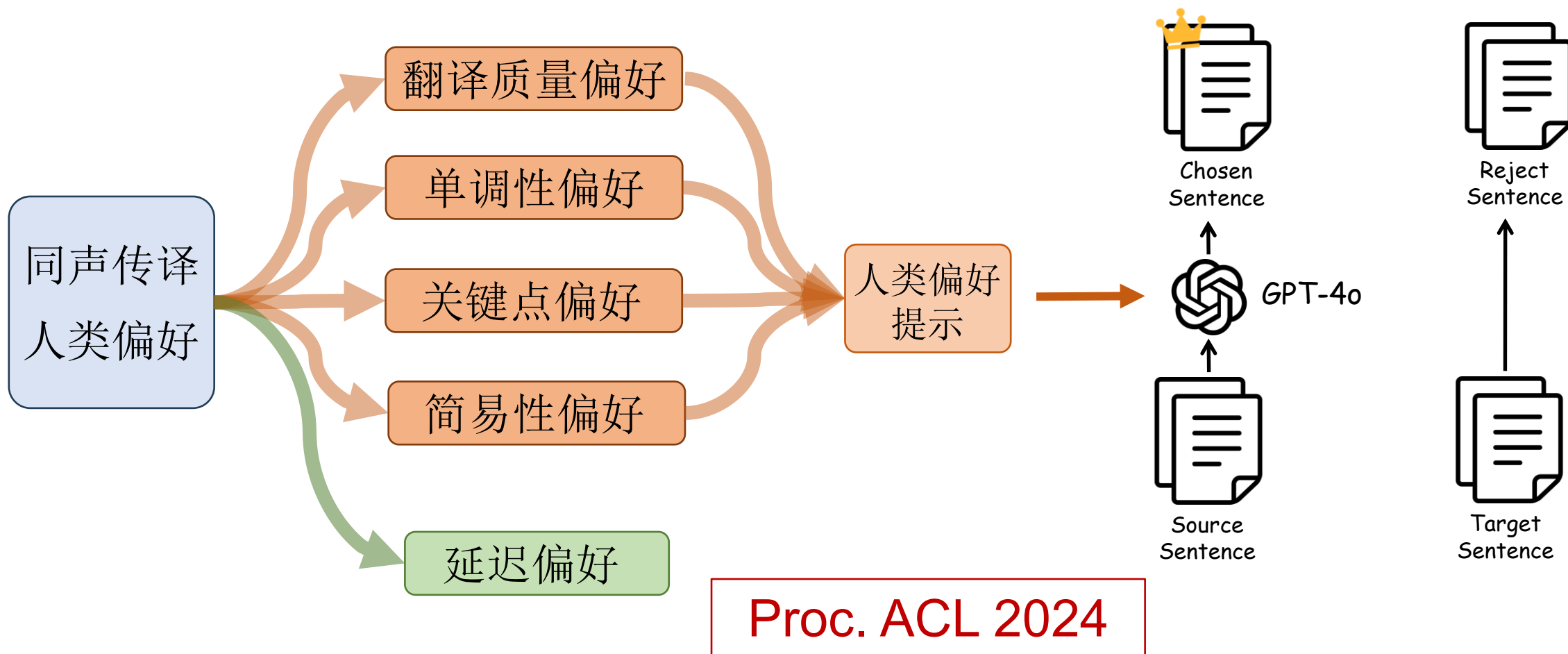
偏好优化的独特性:

- 需要同时考虑对译文偏好和对延迟时间的偏好;
- 同时优化翻译能力和读/写决策能力。



3. 相关研究

◆ 实时偏好学习框架



3. 相关研究

◆ TriFine数据集构建

视频引导的多模态机器翻译(Video-guided Machine Translation, VMT)任务是指在视频片段的帮助下将一个源语言句子（通常是字幕）翻译成目标语言的句子。

已有的数据集存在如下问题：

- **使用粗粒度信息。** 当前的VMT方法几乎全是使用视觉模型从视频的帧画面中提取整体的粗粒度信息，然后将这些信息和文本融合进而完成翻译任务。但这些粗粒度视觉特征会导致**信息冗余**和较高的**计算开销**；
- **不考虑视频中的语音。** 当前的VMT方法普遍忽略了视频中的语音信息，缺少对此的分析。

3. 相关研究

- 源字幕: A lot of bugs.
- 目标字幕: 很多虫子。

bug: 故障/ 错误;
窃听器;
虫子。

多模态细粒度标签

动作: run away

语音情感: negative

实体: people, beach, net

表情: none

地点: beach

语音轻重音: 0.97, 1.02, 0.91, 1.08

视频描述: Two people on a beach by the sea, one of them runs away quickly after touching a fishing net, while the other one has been standing on the right side.



3. 相关研究

◆ TriFine 数据集和已有VMT对比

Dataset	Language	Domain	# Clip	Duration	# FG	Audio	Amb	Info-spec	A-S Align
How2 (2018)	En-Pt	instruction	189K	5.8s	0	✓	✗	✗	✓
VATEX (2019)	EN-Zh	caption	41K	10s	0	✓	✗	✗	✗
VISA (2022b)	En-Ja	subtitle	40K	10s	0	✗	✓	✗	✗
MSCTD (2022)	En-Zh/De	subtitle	172K	-	1	✗	✗	✗	✗
EVA (2023b)	En-Zh/Ja	subtitle	1.4M	10s	0	✗	✓	✗	✗
BigVideo (2023)	En-Zh	subtitle	3.3M*	8s	0	✓	✓	✗	✗
MAD-VMT (2024)	En-Zh	caption	193K	-	0	✗	✗	✗	✗
TriFine (Ours)	En-Zh	subtitle	2.4M	10s	7	✓	✓	✓	✓

“# FG”、“Amb”、“Info-spec”、“A-S Align”分别表示细粒度标签的种类、歧义测试集、专门特征测试集、语音和字幕对齐。

TriFine数据集中的每条数据均包含音频并且音频可以和字幕对齐。每条数据均包含7种来自视觉和语音模态的细粒度标签。

3. 相关研究

从数据集中随机抽样400个样例人工评估的准确率

modality	Category	Accuracy	# Samples
Visual	Location	89.50%	400
	Entity	88.00%	
	Expression	86.50%	
	Action	93.25%	
	Caption	93.75%	
Audio	Audio Sentiment	79.50%	400

Boyu Guan et al., TriFine: A Large-Scale Dataset of Vision-Audio-Subtitle for Tri-Modal Machine Translation and Benchmark with Fine-Grained Annotated Tags.

Submitted to COLING 2025

提 纲

1. 回顾
2. 现状分析
3. 相关研究
- ➔ 4. 未来展望

4. 未来展望

- 从提高译文质量，追求翻译最终目标“信、达、**雅**”的角度，仍然有太多的问题需要研究，如术语、俚语、标题、文化差异等；

原文：美方提出的所谓“合作、竞争、对抗”三分法就是“障眼法”。合作是权宜之计，竞争是话语陷阱，对抗遏制是本质。有求于中方时就要求合作，在有优势的领域就脱钩制裁，为了遏制不惜冲突对抗。**美国的企图注定不会得逞，在台湾问题上，中国是有底气的。**

译文： The so-called "cooperation, competition, and confrontation" tripartite approach proposed by the United States is a "deceptive tactic". Cooperation is a temporary solution, competition is a discourse trap, and confrontation and containment are essential. When seeking cooperation from the Chinese side, **they** demand cooperation, decouple sanctions in areas where they have advantages, **and are willing to engage in conflict and confrontation to contain them.** (2024.11.9)

4. 未来展望

- 我们要传承**红色基因**，赓续红色血脉，肩负起中华民族

We must inherit the **red genes**, carry on the **red bloodline**, and shoulder the responsibility of the Chinese nation...

I will pull my socks up to do that well. 我会把我的袜子做好。

It is just a Spanish castle. 这只是一座西班牙城堡。

- **Transformers**: 变压器; 变形金刚; 改革者; 变换模块(器)
- **Captain**: 上尉(陆、空、海陆); 上校(海军、海岸警卫队)

4. 未来展望

- 从实际应用系统研发的角度，同样有很多技术难题需要解决，机器翻译系统接收的输入往往不是干净、规范、准确的普通文本(plain text);

印尼语、马来语、菲律宾语三种语言的主流媒体中新闻文本分别有20%、14%和10%以上的词汇存在拼写错误。

如印尼语中表示“失败”的词“gagal”，常被“ga2l”替代；为了强调语气，有时会重复输入字符，“cemungudh”常被写为“cemunguuuudh”。

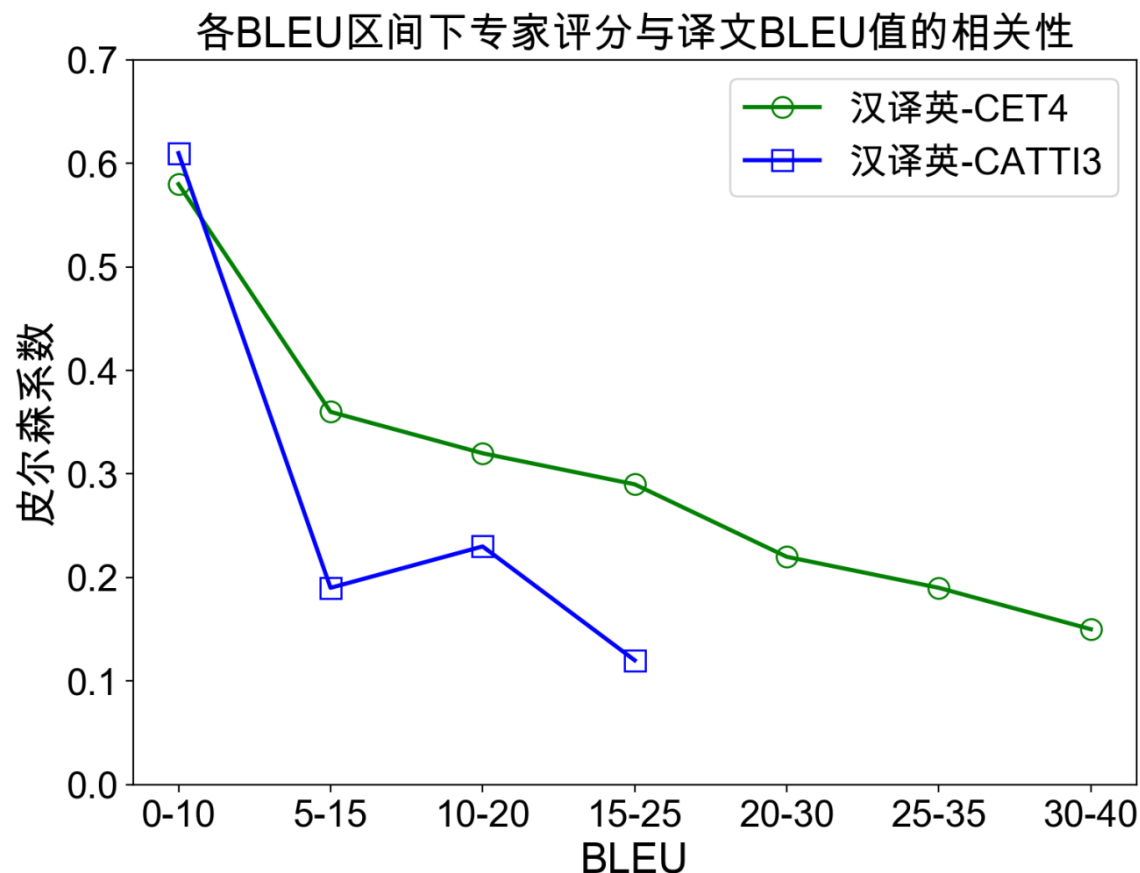
Arrive at tks at 0800hrs → 在08时00分到达塔克斯。

whr cldy → The weather is cloudy

4. 未来展望

- 从译文评价质量的角度，BLEU，COMET等常用的评价指标只是针对文本翻译在一定程度上反映译文质量的大概水平；

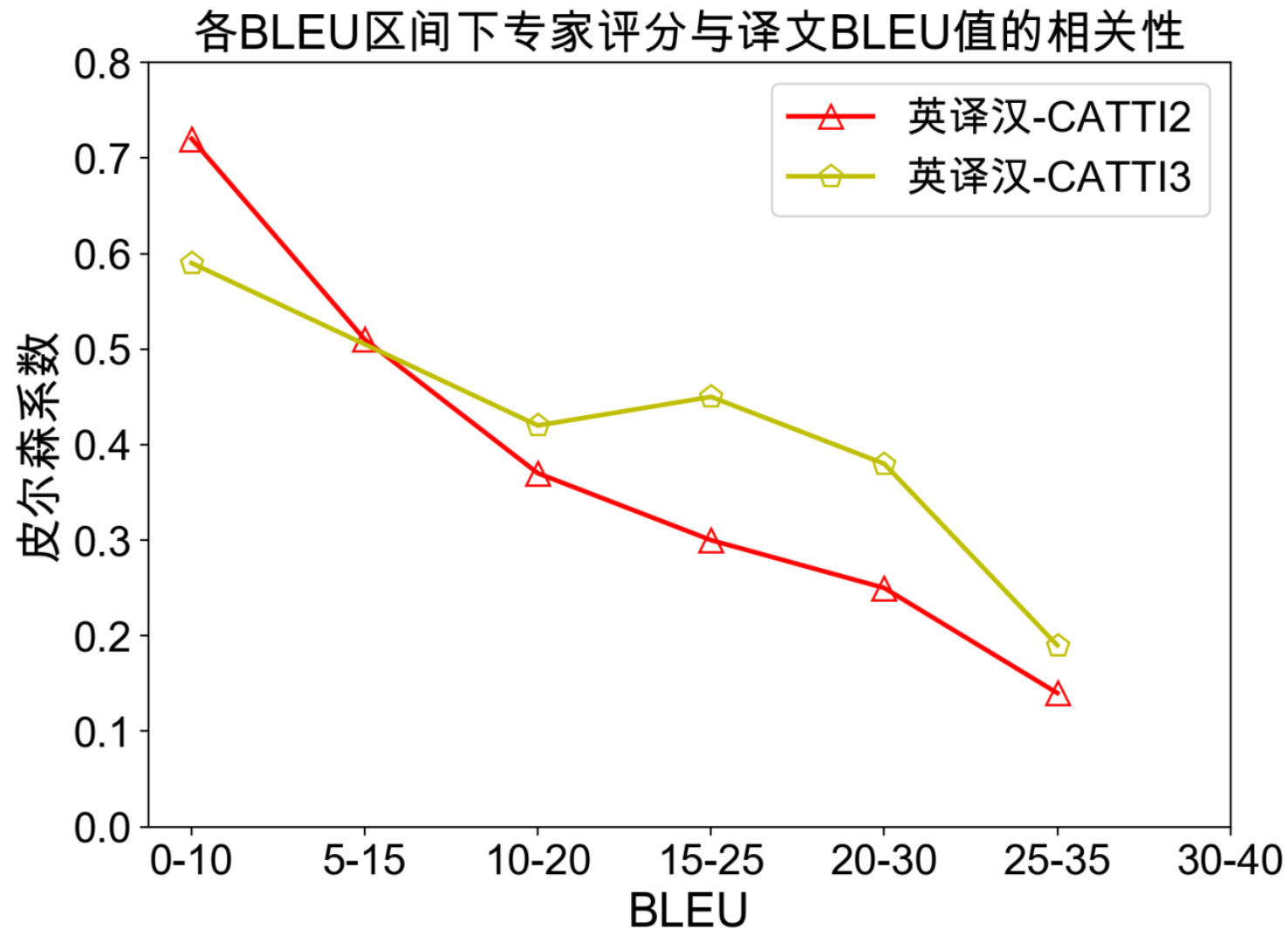
➤ 汉译英



- 随着BLEU值的增加，BLEU值和专家评分的相关性越低。
- CET4的评分相关性大于CATTI3的评分相关性。

4. 未来展望

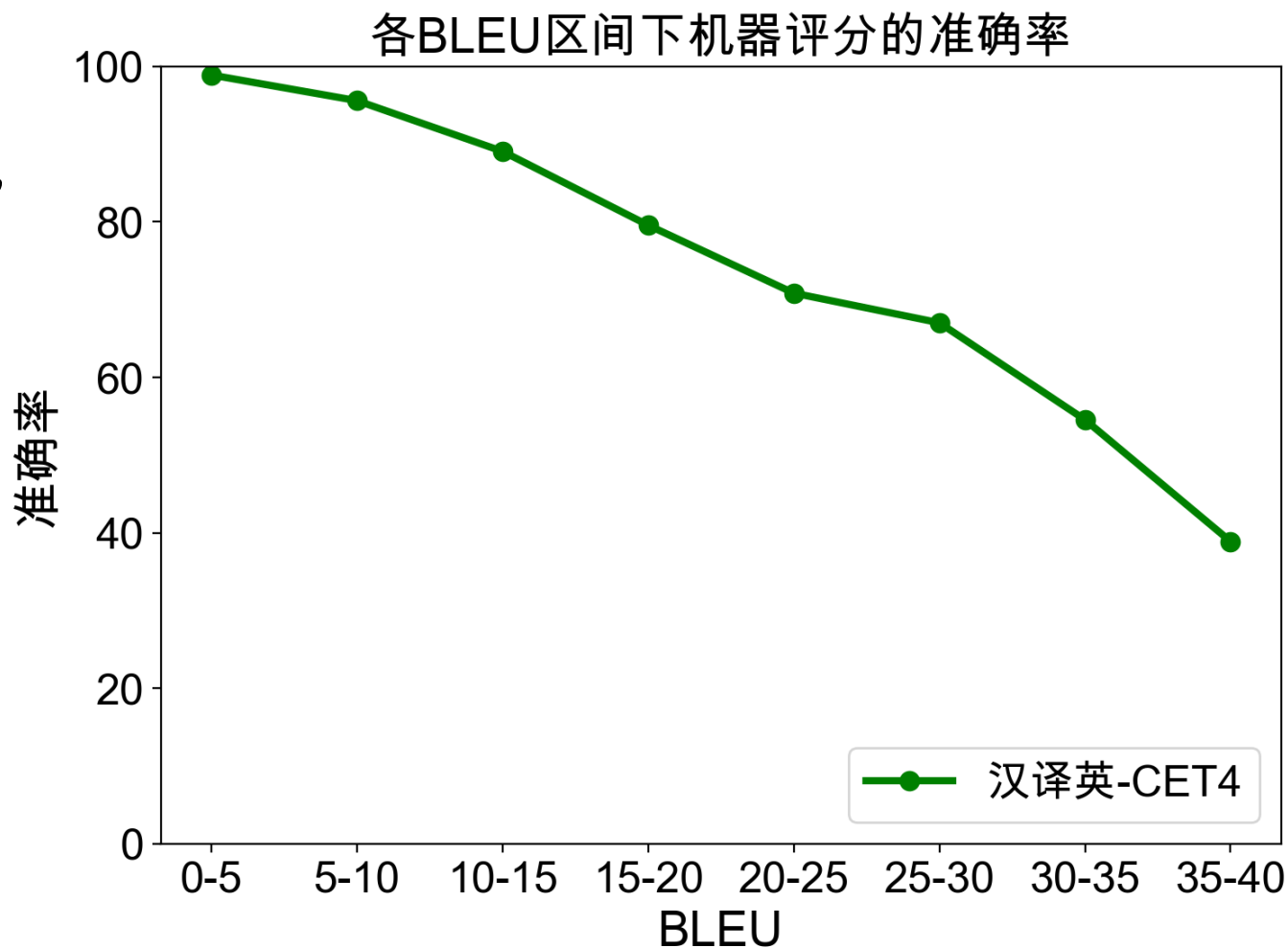
➤ 英译汉



4. 未来展望

➤ 根据BLEU值预测专家评分的准确率

- 根据BLEU值预测专家评分，随着**BLEU值的增加**，预测准确率**越低**。



4. 未来展望

例如1: BLEU值高而专家评分低 (英译汉)

原文 (CATTI2)

Angkor is one of the most important archaeological sites of Southeast Asia. For several centuries, Angkor was the center of the Khmer Kingdom. **With impressive monuments,** several different ancient urban plans and large water reservoirs, the site is a unique concentration of features testifying to an exceptional civilization. **The architecture and layout of the successive capitals bear witness to a high level of social order and ranking within the Khmer Empire.** The Angkor complex encompasses all major architectural buildings and hydrological engineering systems from the Khmer period. **All the individual aspects illustrate the intactness of the site very much reflecting the splendor of the cities that once were.**

4. 未来展望

参考译文：

吴哥是东南亚最重要的考古遗址之一。它作为高棉王国的中心，延续了数个世纪。吴哥古迹，**包括令人敬仰的遗迹**、不同时代的古代城市规划和大贮水池，独特且集中地体现了其非凡文明的特点。**历届政府首都的建筑风格和布局见证了高棉王国社会秩序和等级的有序和谐**。吴哥建筑群涵盖了高棉时期的所有重要建筑和水文工程系统。**这个建筑群的方方面面构成了完整的遗址，充分体现了城市的昔日辉煌。**...

BLEU值为0.17分（较高），而专家评分仅为34分（百分制）的译文：

吴哥窟是东南亚地区最重要的考古地点之一,几个世纪以来,吴哥窟一直是高棉王国的中心。**此地有令人叹为观止的纪念碑**、各种各样的古代城市规划和大型水库。吴高窟是各种特色的独特汇聚地,证明其文明的独特性。**后来的首都的结构和设计见证了高棉社会秩序水平交高及高棉国的排名**。吴新窟建筑群包含了高棉时期所有重大建筑和水工程系统。**每个单独的建筑物都能诠释该地曾经一度的完整性和城市的繁华。**...

4. 未来展望

例2: BLEU值低而专家评分高 (英译汉)

原文 (CET-4)

Both WHO's constitution and the Universal Declaration of Human Rights assert that health is a human right, **not a privilege for those who can afford it**. Over time, that right has made its way into both national and international law. Importantly, the right to health is not simply a noble idea on a piece of paper. It has been a platform for major improvements in global health. Since 1948, life expectancy has increased by 25 years. Smallpox has been eradicated and polio is on the brink. **We have turned the tide on the HIV/AIDS epidemic**. Deaths from malaria have dropped dramatically. And **there are many other causes for celebration**. ...

4. 未来展望

参考译文：

世界卫生组织的《宪章》和《世界人权宣言》都声称，健康是一项人权，**不是有钱人才享有的特权**。随着时间的推移，这项权利已经进入国家和国际法律。重要的是，健康权不仅仅是停留在纸上的高尚想法，它始终是全球大力改善健康的重要平台。自1948年以来，预期寿命增加了25年。天花已经根除，小儿麻痹症即将被消灭，**艾滋病毒/艾滋病蔓延的趋势也得以扭转**。疟疾死亡人数急剧下降，新的疫苗使那些过去谈虎色变的疾病容易预防。**值得庆祝的进步还有很多**。…

BLEU分数仅为0.07，而专家评分却为60（百分制）的译文：

世卫组织章程和世界人权宣言明确规定健康是人的权利，**任何人都没有特权**。早前，此项权利已经被列入国家和国际法。更重要的是，健康权利不简单的是写在纸上的伟大想法，它是提高全球健康的主要平台。自1948年起，生命增加了25年。天花已被根除，小儿麻痹症快要被消灭，**我们开始对抗艾滋病病毒**。因炎症而死的已很少很少，新型疫苗很容易预防疫病，**并且有很多其他喜事**。…

4. 未来展望

对于多模态翻译、同声传译等各类复杂翻译任务，如何准确地评价？

- 从撰写和发表论文的角度，针对纯文本的机器翻译方法往往面临更大的难度，而融合多模态、低资源、复杂场景的翻译方法拥有更大的优势；
- 从机器翻译方法角度，大模型是总体趋势，具有通用性好、处理生词能力强、译文流畅性好等优点，但针对不同用户的具体需求是否一定采用大模型，需要综合权衡，研究方法 with 最终目标密切相关（理论探索/应用）。



- [1] R. Hinami, S. Ishiwatari, K. Yasuda, and Y. Matsui, Towards fully automated manga translation, *Proc. of AAAI*, 2021, pp. 12 998–13 008.
- [2] N. P. Sable, P. M. Shelke, N. Deogaonkar, N. Joshi, R. Kabadi, and T. Joshi, Doc-handler: Document scanner, manipulator, and translator based on image and natural language processing, *Proc. of ESCI*, 2023, pp. 1–6.
- [3] Xu, M. Li, L. Cui, S. Huang, F. Wei, and M. Zhou, Layoutlm: Pretraining of text and layout for document image understanding, *Proc. of KDD*, 2020, pp. 1192–1200.
- [4] J. Wang, L. Jin, and K. Ding, Lilt: A simple yet effective language_independent layout transformer for structured document understanding, *Proc. of ACL*, 2022, pp. 7747–7757.

感谢赵阳、张亚萍、张志扬、武俊宏、管博宇、梁雨普、于东磊 提供的帮助！





谢谢!
Thanks!



清华大学出版社，2024.1