评估大型语言模型在中文成语翻译上的表现

Cai Yang[1], **Yao Dou**[2], **David Heineman**[2], **Xiaofeng Wu**[2], **Wei Xu**[2]

[1]Independent Contributor [2]Georgia Institute of Technology

caiyang.cy@outlook.com, { douy,david.heineman,xwu414 } @gatech.edu, wei.xu@cc.gatech.edu

Abstract

习语通常具有与其字面解释不同的比喻意义，它们在日常语言中很常见，尤其是在中文中，习语常常包含历史引用并遵循特定的结构模式。尽管近年来在使用大型语言模型进行机器翻译方面取得了一定进展，但对中文习语翻译的研究却很少。在这项工作中，我们引入了 IDIOMEVAL，一个具有全面错误分类法的中文习语翻译框架。我们对从九个现代系统（包括 GPT-4o 和 Google 翻译）中收集的 900 个翻译对进行了注释，这些翻译来自四个领域：网络、新闻、维基百科和社交媒体。我们发现这些系统在习语翻译上表现不佳，产生了错误的、字面的、部分的甚至遗漏的翻译。表现最好的系统 GPT-4 在 28 % 的情况下出现错误。我们还发现现有的评估指标对习语质量的测量很差，与人工评分的 Pearson 相关性低于 0.48。因此，我们开发了改进的模型，在检测习语翻译错误方面达到了 F_1 评分 0.68。

1 介绍

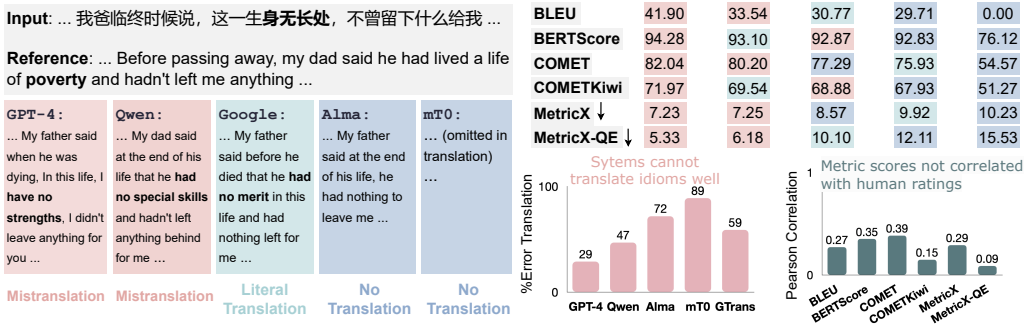

Figure 1: 左上: 中文输入和来自五个系统的翻译示例。右上: 每个系统中良好翻译的百分比。左下: 系统翻译的评价指标分数。右下: 人类评分和指标分数之间的 Pearson's r 。

成语是指其含义通常与其各组成部分的字面意义不同的短语。它们在日常语言中被普遍使用，因此理解成语对于任务如神经机器翻译 (Baziotis et al., 2023) 和写作辅助 (Tan & Jiang, 2021; Dankers et al., 2022a) 至关重要。特别是，成语为翻译带来了挑战，因为它们的含义通常是不可组合的，无法从字面部分推断。例如，my two cents 意指某人的观点，而不是实际的钱。近年来，越来越多的关注被放在成语翻译上。Dankers et al. (2022b) 发现，transformer 编码器倾向于将比喻性的成语作为单个单位更加强烈地处理，而 Baziotis et al. (2023); Liu et al. (2023) 显示不同的训练方法对成语翻译可能是有帮助的。

然而，大量这类研究集中在英语、日语和欧洲语言的成语上，而对汉语成语关注较少。从语言学的角度来看，前者属于印欧语系，而汉语则属于汉藏语系 (Katzner & Miller, 2002) 。此外，各种语言间的成语在其定义和分类上存在差异。例如，虽然英语成语通常强调句法结构，而中文成语则更注重语义和语用方面 (Wang, 2021) 。同时，中文成语通常包含谚语、寓言和格言，这些大多来源于古代文学 (Wang, 2021; 2022) 。一个例子是成语 UTF8gbsn 明日黄花 ，它最早出现在宋代（960 年至 1279 年）并被字面地翻译为 “明日黄花”，但隐喻意义是 “已成过时的事物”。由于文化和语言的差异 (Wang, 2022) ，汉语成语给翻译系统带来了独特的挑战。

UTF8gbsn

Error Type & Description	Example (zh)	Example (en)
No Error: Translation is of high quality.	... 科学家们欢呼雀跃，却发现...	...Scientists cheered and rejoiced, but they discovered that...
	... 油霸们再捅刀美元，项庄舞剑意在沛公...	...oil tycoons stab the US dollar again, Xiang Zhuang's sword dance is aimed at Pei Gong...
Mistranslation: Translation is incorrect and disrupts idiom understanding.	... 问得太多只会自寻烦恼...	Asking too many questions will only bring them problems. (bring yourself problems)
Unnatural: Translation is understandable yet suboptimal, and it can be improved.	... 善战者无赫赫之功...	A good warrior does not achieve a glaring victory... (a remarkable victory)
Literal: Translation is literal and coherent with the surrounding context.	... 这一生身无长处...	...I have no merits in this life... (live in poverty)
Addition: Translation contains irrelevant content beyond the accurate meaning.	... 不仅于剧学素有深造，无所不通...	...not only well-versed in drama studies, knowledgeable and passionate in all aspects... (knowledgeable in all aspects)
Partial: Given idiom is translated partially.	...此处不留人，自有留人处，对于这类无良民宿...	...there is no place to stay, and against these bad inns... (if you don't stay here, there is a place to stay elsewhere)
Repetition: The translation contains repeated correct content or synonyms.	... 最佳的父子关系应该如兄如弟...	...the best relationship between a father and son should be like that of a brother and a brother... (like that of brothers)
No Translation: Translation lacks the idiom's meaning in the output.	... 正从布隆泉飞向普利托利亚，除了不速之客的蛇外，机上还载了 4 名乘客。	...was flying from Bloemfontein to Pretoria and had four other passengers on board, including the snake. (the unwelcome snake)

Table 1: IDIOMEVAL 错误分类法，其中包含来自我们在 Section 3.1 中收集的数据集的每种错误类型的示例翻译。修正的部分以斜体显示。如果存在多个错误，参与者可以选择 More than One 。

在这项工作中，我们旨在深入研究现代翻译系统处理中文成语的有效性，以及当前的自动评估指标是否准确反映成语翻译的质量。我们首先介绍了 IDIOMEVAL ，一种错误分类法，捕获了中文成语翻译的七种不同失败模式，例如误译和部分翻译（见表 1 ）。为了探讨领域上下文的影响，我们从四个不同领域收集包含成语的中文文本：网页、新闻、维基百科和社交媒体。使用 IDIOMEVAL 分类法，我们收集了来自包括 GPT-4o (Achiam et al., 2023) 在内的 9 个现代翻译系统的 900 个翻译对的注释。通过这些注释，我们对系统翻译进行错误分析，并调查当前评估指标在评估中文成语翻译质量方面的表现。我们的主要发现如下：

为了提高成语翻译好坏的检测能力，我们对 Qwen2.5 模型进行 instruction-tuning，并实现了一个 Macro F_1 得分为 0.68 的结果，优于所有现有指标并提示 GPT-4o。此外，我们发现随着模型规模的扩大，性能有所提升，除了在 72B 时例外。我们希望我们的工作能够突出成语在翻译中的影响，并为开发更好的模型和评估指标铺平道路。我们将发布所有注释、代码和模型。总之，我们的贡献是：

1. 我们介绍了 IDIOMEVAL ，这是一种新的中文成语翻译错误分类法。
2. 我们收集了一个包括四个领域和九个现代机器翻译系统的 900 个人工标注翻译的高质量数据集。
3. 我们提供了首个关于现代系统在习语翻译上的表现以及现有指标与人工评判相关性的综合分析。
4. 我们表明，经过指令调整的大型语言模型在检测习语翻译错误方面显著优于现有的度量标准和提示方法。

2 IDIOMEVAL：评估框架

在本节中，我们讨论了我们的习语翻译评估框架的详细信息。IDIOMEVAL 包含两个步骤：标注者在上下文中选择习语的翻译范围，并将其分类为 9 个高层次的语言学基础类别，这些类别进一步细分为 13 个子类别；然后，标注者评估翻译错误的严重程度，并对标注的置信度进行评估。我们的标注界面是使用 Thresh (Heineman et al., 2023) 构建的，我们将发布我们的配置供公众使用。

选择翻译范围和类别。框架开始时需要选择习语的相应翻译范围。然后，标注者被要求选择一个最能描述习语翻译的类别。对于某些类别，标注者还需要选择一个子类别。IDIOMEVAL 类别在 Table 1 中定义。我们在 Appendix A 提供了每个类别、子类别及示例的详细描述。

评分严重性和置信度。由于每个翻译类别对整体句子质量的影响程度不同，我们要求标注者对错误的严重性进行评分。根据 Karpinska et al. (2022); Dou et al. (2022) ，我们定义了三

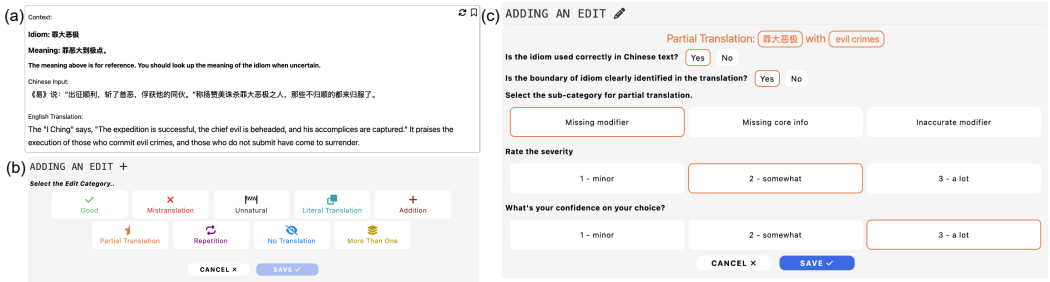

Figure 2: 成语注释框架。

个严重性级别：轻微 (1)、中等 (2) 和严重 (3)。最后，我们要求标注者在相同的尺度上评估他们选择的正确性置信度。标注界面的概览显示在 Figure 2 中。

3 数据收集和标注

我们描述了自动构建一个包含 623K 个包含成语的中文句子的四个领域数据集的方法。然后，使用该数据的一个子集，我们收集并注释了来自 9 个现代系统的 900 个翻译，分析结果见 §4 和 §5 。

Domain	Source	Period	Instances	Idioms	Avg. occurrence by freq. range				
					VH	H	M	L	N
News	Common Crawl News	Mar.-Apr.,Oct	50,845	5,333	16.7	3.9	2.5	1.5	1.5
Web	Common Crawl	Jan.-Feb.,Dec.	463,642	15,319	61.1	2.6	1.4	1.4	9.3
Wikipedia	Wikipedia Meta History	Jul.,Dec.	39,699	5,947	12.1	2.1	1.4	1.2	1.1
Social Media	Weibo Search Results	Jan.-Jul.,Jan.-Dec.	55,315	351	519.0	274.25	84.2	22.4	5.0

Table 2: 按领域统计收集的中文数据。来源列出了数据收集来源。时间段显示了 2023 年和 2024 年收集数据的月份。实例数表示每个领域的数据样本数量。成语列显示了收集数据中成语的总数。最右侧的列显示每个频率范围内成语的平均出现次数。每个频率范围内成语的数量可以在 Appendix B.2 中找到。

3.1 收集含有习语的文本

我们使用由 Tan & Jiang (2021) 提供的习语词汇，因为它具有较高的覆盖率，其中包含 30,999 个具有定义的习语。为了评估在不同语境下的习语翻译，我们首先收集包含习语的中文文本，这些文本来自四个领域：新闻、网页、维基百科和社交媒体。我们后续的分析涉及在 2023 年和 2024 年发布的大型语言模型。为了尽量减少这些文本出现在用于翻译的 LLM 预训练语料库中的机会（§3.2 ），我们收集在这些 LLM 各自发布日期之后才出现的文本（即 `new corpus` 所指的文本）。

对于每个领域，我们根据整体频率的 1^{st} 、2^{nd} 和 3^{rd} 分位数将习语分为五组：VH（非常高）、H（高）、M（中等）、L（低）和 N（在数据中从未出现）。这使我们能够衡量习语频率在预训练语料库中的影响。我们使用 `old corpus` （同一领域中 2023 年之前的数据，详情见 Appendix B.1 ）来计算习语频率，因为这一时期的习语更有可能被包括在 LLM 的训练数据中。鉴于习语在语言中往往具有长期存在性，其在同一领域内的频率分布不太可能随时间显著变化。表 2 展示了收集数据的统计结果。每个领域的数据收集详情在附录 C.1 中描述。

3.2 收集系统的翻译

我们精心挑选了一组多样的翻译系统，以便我们可以评估不同类型系统在习语翻译上的性能：GPT-4（一个领先的大型语言模型） (Achiam et al., 2023) ，Alma-13B（一个经过微调的翻译模型） (Xu et al., 2024) ，Qwen-14B-Chat (Bai et al., 2023) （一个用于中文和英文的双语大型语言模型），mT0-13B（一个多语言的大型语言模型） (Muennighoff et al., 2022) ，Google Translate（一个广泛使用的商业服务），GPT-4o（一个领先的大型语言模型） (Hurst et al., 2024) ，以及 QWen2.5 经过指令微调的模型（多语言大型语言模型，拥有 7B, 14B 和 72B 参数） (Yang et al., 2024a;b) 。

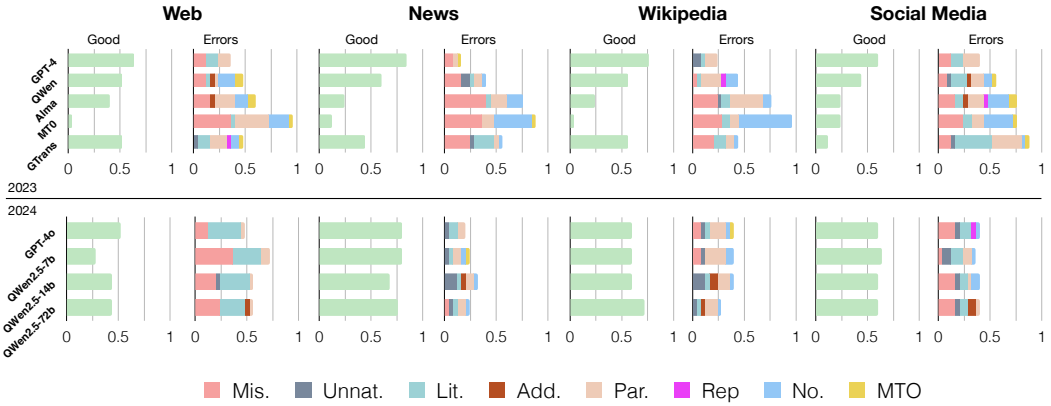

Figure 3: 系统中心视角下的每个领域的翻译结果。每个数据集都有两个子图。左侧子图：优质翻译的比率。右侧子图：每个系统的错误类别组成。

我们从每个频率范围中抽取 5 个实例，每个领域总共翻译 25 篇中文文本。总共，这从九个系统中产生了 900 对翻译对。这些作者都是中文和英文的母语者，他们手动将每篇中文文本翻译成英文。这些翻译在我们的分析中用作参考。

3.3 人工标注

我们通过 Prolific 招募精通英语的中文母语者。标注过程分为两个阶段：试运行阶段和主要阶段。在试运行阶段，向 20 名参与者提供标注指南和小测验。得分最高的 5 名将被选入主要阶段。在主要阶段，每对翻译都会分配给三名标注者，以确保工作量均等分配。最终的标注结果通过多数投票决定，对于任何平票情况，由第一作者手动解决。关于标注者之间一致性和质量控制的详细信息，请参见 Appendix C.2 。

4 翻译系统评估

我们对现代系统的成语翻译质量进行了详细分析。我们讨论了每个系统在各个领域的表现及其相对比较。接着，我们研究了领域和成语频率如何影响翻译质量，并讨论了翻译错误的严重性。我们在 Figure 3 中展示了我们的结果，该结果显示了每个系统在四个领域中不同翻译类别的比例。进一步的分析，如常见错误、所选子类别、按频率范围的错误分布以及每个类别的平均严重性得分，可以在 Appendix D 中找到。以下是主要发现：

1. GPT-4 和 Qwen 系列整体翻译质量较高，而 Alma 和 mT0 一直表现不佳。在 2023 年的数据中，GPT-4 在所有领域中拥有最多的优质翻译，而 Qwen 获得了相当于第二名的表现。在 `News` 和 `Wikipedia` 上，GPT-4 分别超出第二名系统（Qwen）0.24 和 0.2。这个差异在 2024 年的数据中被最小化，两种模型的质量接近。尽管在所有模型中表现最佳，GPT-4 和 Qwen 系列仍未能达到令人满意的结果，超过 20 % 的翻译存在问题。对 Alma 和 mT0 来说，不到 25 % 的翻译被评为良好（除了 Alma 在 `Web` 上的表现）。这种差异可能是由于在预训练和微调阶段缺乏习语相关的中文数据所致。例如，Alma 的骨干 LLaMA-2 主要在英文中进行预训练，限制了其性能。相比之下，在中文和英文双语数据上进行预训练的 Qwen 达到了更好的翻译质量。

2. 系统的翻译质量因领域而异，它们在 `News` 上的表现较好，但在 `Web` 和 `Social Media` 上的表现较差。尽管 GPT-4 和 Qwen 总体上提供了最高的翻译质量，但它们的优秀翻译数量在不同领域间差异显著。对于 GPT-4 和 GPT-4o，超过 80 % 的翻译在 `News` 上被评为优秀，但在 `Web` 上降至 50 %，在 `Social Media` 上降至 60 %。Qwen 模型表现出了类似的波动。总体而言，翻译质量在 `Web` 和 `Social Media` 上较差。这可能是因为这两个领域经常出现新兴背景，其中的习语使用方式相比其他领域有新的变化。这些变化突显了各领域中习语使用的内在差异，并强调了多样化训练数据的重要性，以确保持续高质量的翻译。

3. 各系统的翻译错误类型有所不同，但误译和部分翻译是常见的错误。GPT-4 和 GPT-4o 的错误主要是误译和部分翻译。在 `Web` 上，GPT-4o 还产生了一些字面翻译。Qwen（包括 Qwen2.5）和 Alma 产生了更多样化的错误。它们是唯一产生添加错误的两个模型（`Web` 和 `Social Media` ）。mT0 经常不翻译，特别是在 `Wikipedia` 上，至少有 40 % 的习语在不同频

率范围内未被翻译。谷歌翻译经常产生字面翻译（36 % 在 Social Media 上）以及在三个领域内（4 %）的不自然翻译。

5 自动指标评估

我们评估当前自动翻译指标在成语翻译评价中的表现，重点考察它们对含有成语的翻译的评价能力、对成语翻译错误的敏感性，以及是否能够检测特定的成语翻译错误范围。

Scope	Category	Size	Reference-Based Metrics					Reference-Free Metrics		
			BLEU	BERTScore	COMET	MetricX -23-XXL	MetricX -24-XXL	COMETKIWI	MetricX-QE -23-XXL	MetricX-QE -24-XXL
Full	Mistranslation	128	0.071	0.119	0.170	0.118	0.170	0.215	0.080	0.140
	Literal	90	0.250	0.325	0.376	0.305	0.270	0.100	0.144	0.083
	Partial	105	0.254	0.208	0.284	0.163	0.244	0.136	0.322	0.341
	No Translation	74	0.186	0.271	0.348	0.198	0.226	0.081	0.138	0.268
Idiom	Mistranslation	128	-0.074	0.056	0.239	0.251	0.166	0.174	-0.021	0.055
	Literal	90	0.113	0.140	0.217	0.219	0.119	-0.081	-0.095	-0.122
	Partial	105	0.131	0.109	0.107	-0.112	-0.261	-0.034	-0.332	-0.326
	No Translation	74	*	*	-0.052	-0.025	-0.512	-0.088	-0.502	-0.536
Full	All errors	452	0.277	0.322	0.376	0.264	0.241	0.208	0.186	0.175
	All categories	900	0.274	0.352	0.386	0.341	0.292	0.151	0.131	0.090
Idiom	All errors	452	0.227	0.229	0.343	0.250	-0.122	-0.004	-0.277	-0.274
	All categories	900	0.343	0.399	0.473	0.483	0.036	0.026	-0.161	-0.205

Table 3: Pearson 的 r 在指标和人工注释之间的表现。评估分为完整翻译和习语翻译两种情况。由于所有的实例都被评定为最高严重性，因此在此略去 More Than One 。由于样本量小，略去了 Unnatural 、Addition 和 Repetition 。对于习语翻译中的 No Translation ，BLEU 和 BERTScore 的输出为 0，因此被省略。肯德尔的 τ 结果（见附录 E.1 中的表 9 ）显示出类似的模式。

5.1 指标性能

我们考虑了六种最常用且最强的自动化指标 (Freitag et al., 2023)：BLEU（Papineni et al., 2002），BERTScore Zhang et al. (2019)，COMET（Rei et al., 2022a），COMETKIWI（Rei et al., 2022b），MetricX-XXL（Juraska et al., 2023; 2024）和 MetricX-QE-XXL（Juraska et al., 2023; 2024）。在这些指标中，COMETKIWI 和 MetricX-QE 是无需参考的。为了便于与人工注释进行比较，我们将每个注释折叠为一个单一的数值评分，该评分是严重程度分数（从 1 到 3 ）直接作为总体评分。良好的翻译被标记为 0 。

为了了解这些指标在上下文中和独立评估习语的效果，我们基于完整上下文和仅习语部分计算它们的分数。Table 3 报告了不同错误类型的指标与人工评估之间的皮尔逊相关系数（r ）。我们发现，无论在完整上下文还是习语层面，指标都难以提供有效的细粒度测量，完整上下文的最高相关系数为 0.386，而习语为 0.483。在大多数情况下，习语翻译的相关性比全文更强，因为习语的周围上下文也会影响指标得分。然而，现有指标与人工评分的相关性较弱，因此无法从类似人类的视角衡量习语翻译的质量。类似的趋势也在检查与 Kendall 的 τ 的排名相关性时观察到（详情见 Appendix E.1 ）。

鉴于这些度量标准无法有效衡量错误的严重程度，我们接下来探讨一个更简单的问题：它们能否区分好的习语翻译和不好的翻译？我们报告了每个度量标准的 ROC-AUC 得分以及在 Table 4 中（左），好翻译比坏翻译得分更高的句子对的百分比（针对同一个中文句子）。BERTScore 和 MetricX 在这两个指标上表现最佳。然而，在超过 20 % 的情况下，好翻译的评分仍然低于那些带有错误的翻译，这表明这些度量标准在区分好坏习语翻译时并不总是可靠的。

5.2 度量灵敏度

我们基于类似于 Karpinska et al. (2022) 的方法来衡量指标的敏感性，通过比较每个指标对黄金翻译与扰动翻译的评分。对扰动敏感的指标能够准确识别扰动。我们关注无参考指标，因为 COMETKIWI 和 MetricX-QE 等基于参考的指标使用黄金翻译作为参考。为了获得扰动翻译，我们使用 GPT-4o (OpenAI, 2024) 来将现有的参考翻译编辑为不同的错误。任务被分解为两个步骤：1）提取：从英文句子中提取成语翻译，2）扰动：通过错误扰动成语翻译

Metric	ROC-AUC		Good > Bad	
	Full	Idiom	Full	Idiom
BLEU	0.63	0.69	68 %	54 %
BERTScore	0.67	0.71	72 %	77 %
COMET	0.67	0.75	64 %	56 %
MetricX-23-XXL	0.67	0.78	67 %	78 %
MetricX-24-XXL	0.66	0.57	71 %	76 %
COMETKIWI	0.58	0.51	58 %	51 %
MetricX-QE-23-XXL	0.58	0.47	65 %	60 %
MetricX-QE-24-XXL	0.59	0.45	64 %	62 %

Perturbation	COMETKIWI		MetricX-QE -23-XXL		MetricX-QE -24-XXL	
	Full	Idiom	Full	Idiom	Full	Idiom
Mistranslation	0.73	0.66	0.80	0.62	0.79	0.67
Unnatural	0.59	0.57	0.56	0.52	0.66	0.52
Literal	0.65	0.65	0.73	0.76	0.72	0.73
Addition	0.66	0.57	0.54	0.52	0.72	0.48
Partial	0.59	0.58	0.61	0.57	0.72	0.65
Repetition	0.67	0.54	0.74	0.71	0.72	0.69
No Translation	0.68	0.28	0.75	0.75	0.72	0.77

Table 4: 现有指标在成语翻译中的表现。左表：基于 ROC-AUC 分数和全文与成语翻译的成对比较准确性，区分好与差的翻译。右表：基于无参考指标对全文和成语翻译的敏感性，测量未干扰翻译优于干扰翻译的实例比例。

（参见 Appendix E.2 中的提示）。对于每种错误类型，我们随机抽取 10 个实例并检查编辑后的翻译是否符合预期的错误类别。每个类别中超过 70 % 的实例被正确编辑。由于模型被特别指示引入错误，黄金翻译应始终优于扰动翻译。

我们计算扰动精度为金译文得分高于扰动译文的实例比例，如 Table 4 （右）所示。对于全上下文翻译，表现最好的指标 MetricX-QE-24 在大多数扰动下的精度约为 0.7。相比之下，COMETKIWI 和 MetricX-QE-23 在某些扰动下的精度低于 0.6。然而，当评估习语部分的翻译时，这些指标在几乎所有扰动下表现较差，显示出在仅给出习语时，它们检测错误的能力较弱。

5.3 现有指标能识别成语错误范围吗？

我们使用 xCOMET Guerreiro et al. (2023)，这是最先进的错误范围检测模型，以检查其是否能准确检测成语翻译错误的范围。我们使用字符级别的 F_1 得分评估性能，并仅考虑目标范围。与 Blain et al. (2023) 类似，我们还通过将严重性得分转换为二进制权重来计算加权 F_1 得分。Table 5 显示了相应的性能。我们发现 xCOMET 在检测成语翻译错误的范围时表现不佳，最佳模型仅实现了 0.3 F_1。

	P	R	F_1	WP	WR	W F_1
xCOMET-XL	12.5	44.7	16.6	0.4	2.2	0.7
xCOMET-XXL	10.2	30.3	12.4	0.5	1.6	0.6
xCOMET-XL (src)	13.3	49.7	18.1	0.5	2.4	0.7
xCOMET-XXL (src)	9.3	11.4	30.0	0.5	1.6	0.5

Table 5: 使用 xCOMET 进行错误跨度检测的精度和 F_1 分数。src 表示无参考版本。W 表示加权指标输出。为了便于阅读，所有值均乘以 100。

6 改进对成语翻译好坏的检测

我们的语料库旨在开发一种模型，以判断中文成语是否被正确翻译。此任务被表述为一个二分类问题：令 i 为成语，s 为包含 i 的源中文文本，t 为 s 的英文翻译。我们希望预测 i 在 t 中是否被正确翻译。在我们的实验中，我们仅对基于参考的指标（如 BLEU）使用参考。所有其他方法仍然是无参考的，因为在现实世界场景中参考可能不可用。

在实际操作中，指标用于评估新发布的模型，因此我们按模型划分数据：训练集和验证集包含来自较旧模型（GPT-4、Qwen、Alma、mT0 和 Google Translate）的翻译，而测试集包含来自较新模型（GPT-4o、Qwen2.5-7B、Qwen2.5-14B 和 Qwen2.5-72B）的翻译，从而导致不同采样的习语，其重叠最小。训练集、验证集和测试集分别包含 450、50 和 400 个实例。

我们考虑了两种基线：基于回归的度量和基于大型语言模型的提示。回归度量生成一个连续分数，因此我们使用训练集从 ROC 曲线中找到一个最佳阈值，然后将其应用于测试集。对于提示，我们评估了 GPT-4o 和 Qwen2.5 的零样本直接回答和 CoT 提示 (Wei et al., 2022)。

我们探讨了两种微调方法。首先，我们微调了 MetricX-24，这是 WMT24 中表现最好的机器翻译度量指标。其次，我们对 Qwen2.5 系列模型进行指令微调，这是中文和英文中最好的开源大型语言模型。

Model	# Params.	F_1-Good	F_1-Bad	F_1-Macro
Dummy	∼	0.76	0.00	0.38
基于 ROC 曲线的阈值度量				
Reference-Based				
BLEU	∼	0.68	0.43	0.56
BERTScore	355M	0.72	0.43	0.57
COMET	580M	0.65	0.50	0.58 *
MetricX-23	13B	0.68	0.54	0.61 *
MetricX-24	13B	0.66	0.54	0.60 *
Reference-Free				
COMETKIWI	3.5B	0.49	0.52	0.50
MetricX-QE-23	13B	0.60	0.37	0.49

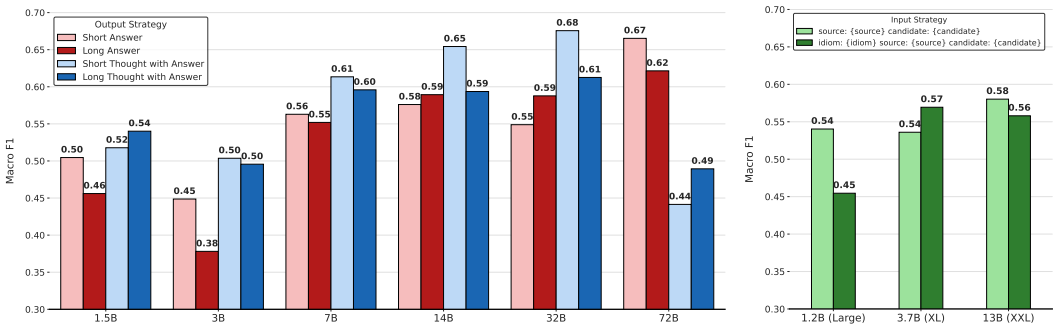

Figure 4: 对比 Qwen2.5（左）和 MetricX（右）训练中的模型尺寸和策略的性能表现。对于 Qwen2.5（1.5B–32B），短思考加最终答案获得了最佳结果，而在 72B 时，较短的答案格式表现相近。对于 MetricX，省略成语在 1.2B 和 13B 时能产生更好的结果。总体而言，除了在 Qwen2.5-72B 之外，模型尺寸的扩展提高了性能。

微调 MetricX。MetricX 基于 mT5 (Xue et al., 2020)，并通过 LM 头的特殊标记对数预测一个从 0 到 25 的回归分数。为了将其适应二分类任务，我们将 LM 头替换为一个分类头，该头输入解码器的起始标记的最终层表示。我们比较用于训练的两种输入格式：一种包含成语（"idiom: { idiom } source: { source } candidate: { candidate } "），另一种不包含成语（"source: { source } candidate: { candidate } "），后者遵循原始 MetricX-24 的训练格式。这使我们能够研究在有限的训练数据设置中，添加额外信息如何与现有格式交互。

指令微调 Qwen2.5。我们将每个实例格式化为一个指令，其中输入包括成语、中文原文和英文翻译。对于输出，即模型计算损失的部分，我们测试了四种策略：(1) 简单的是/否回答，(2) 略长的回答，(3) 简短的思考过程和最终答案，以及 (4) 较长的思考过程和最终答案。后面三种策略使用了附录 F 中描述的模板，我们还在其中提供了所有方法的附加细节。

6.1 结果

表 6 显示了好坏翻译类别的 F_1 分数及其宏平均值。我们进行单边配对自助显著性检验（10,000 重采样）以比较每个指标的宏 F_1 分数与表现最佳的模型，即微调的 Qwen2.5-72B 的分数。星号 (*) 表示统计显著性 ($p < 0.05$)。我们仅展示每种方法类别中表现最好的指标。图 4 提供了不同模型规模和微调 MetricX 以及指令调优 Qwen2.5 的训练策略的更详细比较。

指令微调 Qwen2.5 的表现优于所有基线，其次是 Prompting GPT-4o，然后是阈值参考基础上的度量。微调带有简短思考过程和最终答案的 Qwen2.5-32B 实现了最高的宏观 F_1，达到 0.68。微调 MetricX-QE-24 将其宏观 F_1 从 0.46 提升到 0.58。就提示而言，GPT-4o 达到了宏观 F_1 为 0.63，而 Qwen2.5-72B 则为 0.56。有趣的是，链式思维提示降低了这两种模型的性能，表明这些模型在使用 CoT 时更倾向于判断翻译为正确。对于阈值度量，基于参考的度量明显优于无参考的度量，简单的 BLEU 超出 COMETKIWI 约 0.06 F_1。模型规模的扩大通常有助于 Qwen2.5 和 MetricX，72B 是个例外。在 Qwen2.5 的 1.5B 到 32B 版本中，使用简短的思考过程加上最终答案可产生最佳结果。然而，在 72B 上，训练简短答案表现最佳，与最佳 32B 变体相当。对于 MetricX，使用其原始训练格式（省略成语）在 1.2B 和 13B 规模上效果更好，这表明当训练数据有限时，数据格式的一致性可能是至关重要的。

我们进一步调查为什么在 Qwen2.5 72B 时性能没有提升。在训练损失（交叉熵）方面，我们发现当在思维上进行微调时，72B 模型大约停滞在 0.2 到 0.3，而较小的模型（32B、14B、7B）更平稳地收敛到更低的损失（~ 0.02）。相反，仅在答案上训练 72B 模型，尤其是短答案，导致更低的损失和更好的性能。

在模型行为方面，我们发现即便在更长的格式（如长答案或基于思考的回复）上进行了微调，72B 模型始终输出简短答案格式（是/否）。这与所有较小尺寸的模型（1.5B 到 32B）不同，它们遵循训练输出格式。虽然我们并不完全明白这种行为的根本原因，但这可能解释了为什么在简短答案上训练 72B 模型会产生最佳性能。

7 相关工作

带有习语的机器翻译。我们将多词表达（MWE）视为与习语密切相关的工作领域，并且已经有很多关于 MWE 的研究。(Sag et al., 2002; Calzolari et al., 2002) 。Zaninello & Birch (2020) 探索了 MWE 翻译的注释和数据增强技术。Fadaee et al. (2018) 通过预设特殊标记来指示习语的存在，增加了额外的输入特征。Gamallo & Garcia (2019) 提出了使用跨语言嵌入的组合翻译。更近的研究集中在习语识别 (Tedeschi et al., 2022) ，特征化 (Socolof et al., 2021) 和表示 (Zeng & Bhat, 2022) 。Dankers et al. (2022b) 和 Baziotis et al. (2023) 在神经机器翻译背景下提供了对习语处理的变形金刚模型分析。Liu et al. (2023) 对习语翻译及其问题提供了表征，并引入了增强权重和检索模块以改进翻译。Rezaeimanesh et al. (2025) 评估了波斯语-英语习语翻译，发现 LLMs 在比较模型性能方面十分有效。Li & Goyal (2025) 发现 LLMs 在处理习语时依赖于整合上下文线索和推理的混合方法。与这些工作相反，我们的研究引入了一个细粒度的习语翻译评估框架，并突出了现代 LLMs（如 GPT-4）及自动评估指标在生成和评估习语翻译方面的局限性。

关于汉语成语的研究。目前涉及汉语成语的研究相对有限。Zheng et al. (2019) 提出了一个大规模的汉语成语完形填空数据集 ChID，这项工作首次以机器阅读理解的形式研究汉语成语理解。Tan & Jiang (2021) 构建了一个数据集以促进汉语成语嵌入的评估。Tan et al. (2021) 通过在线抓取电子书收集了一个大规模的汉字预训练语料库。Liao & Cheng (2023) 研究了使用汉语成语的文本润饰，其中句子被改写得更为优雅。其他研究则集中于创建包含汉语成语的平行语料库。Shao et al. (2018) 构建了 CIBB 数据集，包含了 50 个不同汉语成语的英文翻译，以及字面翻译的黑名单。Wang & Yu (2010) 创建了 CIKB 数据集，包含大约 38,000 个成语及其不同的英文翻译。然而，只有 28 % 的样本有完整翻译，并且该数据集未公开。Tang (2022) 构建了 PETCI 数据集，收集了来自字典的成语及谷歌和 DeepL 的翻译结果。最近，Tang et al. (2024) 使用 GPT-4 生成了语境感知的成语翻译，但依然缺少外部数据集和专家评估。Li et al. (2024) 介绍了 IdiomKB，以改善语言模型的成语翻译。作者展示了 GPT-4 可以作为评估员，但与人类评级的相关性仍低于 0.7。Liu et al. (2024) 对多种任务中的中文大型语言模型进行了基准测试，发现其在成语理解任务上的表现较差。Fu et al. (2025) 对中文成语理解和使用进行了大型语言模型的基准测试，发现模型在理解成语意义方面存在根本误解。

8 结论

我们引入了 IDIOMEVAL ，这是一种经过精心设计的分类系统框架，便于评估中文成语翻译。我们认为该框架可以扩展到不同的语言。我们对多个领域的各种系统的评估显示，它们在生成高质量的成语翻译方面存在困难。我们还发现现有的指标无法与人工评分相关联，也无法可靠地区分出好的翻译和糟糕的翻译。此外，我们证明了经过微调的 Qwen2.5 模型能够以 0.68 的宏观 F_1 识别出好的和差的成语翻译，超越了现有的指标和基于提示的大型语言模型。我们的工作为未来的研究奠定了基础，以改善成语翻译性能并开发针对成语的指标。

9

致谢 这项研究部分得到了 NSF CAREER 奖 IIS-2144493 的支持。本文中包含的观点和结论是作者的观点和结论，不应被解释为代表 NSF 或美国政府的官方政策，无论是明示的还是暗示的。尽管其中有任何版权标注，美国政府被授权为政府目的复制和分发重印件。

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Christos Baziotis, Prashant Mathur, and Eva Hasler. Automatic evaluation and analysis of idioms in neural machine translation. In Andreas Vlachos and Isabelle Augenstein (eds.), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 3682–3700, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.267. URL https://aclanthology.org/2023.eacl-main.267.

Frederic Blain, Chrysoula Zerva, Ricardo Rei, Nuno M Guerreiro, Diptesh Kanojia, José GC de Souza, Beatriz Silva, Tânia Vaz, Yan Jingxuan, Fatemeh Azadi, et al. Findings of the wmt 2023 shared task on quality estimation. In *Proceedings of the Eighth Conference on Machine Translation*, pp. 629–653, 2023.

Nicoletta Calzolari, Charles J. Fillmore, Ralph Grishman, Nancy Ide, Alessandro Lenci, Catherine Macleod, and Antonio Zampolli. Towards best practice for multiword expressions in computational lexicons. In *International Conference on Language Resources and Evaluation*, 2002.

Verna Dankers, Christopher Lucas, and Ivan Titov. Can transformer be too compositional? analysing idiom processing in neural machine translation. *ArXiv*, abs/2205.15301, 2022a.

Verna Dankers, Christopher Lucas, and Ivan Titov. Can transformer be too compositional? analysing idiom processing in neural machine translation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3608–3626, Dublin, Ireland, May 2022b. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.252. URL https://aclanthology.org/2022.acl-long.252.

Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A. Smith, and Yejin Choi. Is GPT-3 text indistinguishable from human text? scarecrow: A framework for scrutinizing machine text. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7250–7274, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.501. URL https://aclanthology.org/2022.acl-long.501.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. Examining the tip of the iceberg: A data set for idiom translation. *ArXiv*, abs/1802.04681, 2018.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frédéric Blain, Daniel Deutsch, Craig Stewart, et al. Results of wmt23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pp. 578–628, 2023.

Yicheng Fu, Zhemin Huang, Liuxin Yang, Yumeng Lu, and Zhongdongming Dai. CHENGYU-BENCH: Benchmarking large language models for chinese idiom understanding and use. In *ACL 2025 Student Research Workshop*, 2025. URL https://openreview.net/forum?id=Ku0p1ZXkWA.

Pablo Gamallo and Marcos Garcia. Unsupervised compositional translation of multiword expressions. In *MWE-WN@ACL*, 2019.

Nuno M Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André FT Martins. xcomet: Transparent machine translation evaluation through fine-grained error detection. *arXiv preprint arXiv:2310.10482*, 2023.

David Heineman, Yao Dou, and Wei Xu. Thresh: A unified, customizable and deployable platform for fine-grained text evaluation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 336–345, 2023.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz (eds.), *Proceedings of the Eighth Conference on Machine Translation*, pp. 756–767, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.63. URL https://aclanthology.org/2023.wmt-1.63.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz (eds.), *Proceedings of the Ninth Conference on Machine Translation*, pp. 492–504, Miami, Florida, USA, November 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.wmt-1.35.

Marzena Karpinska, Nishant Raj, Katherine Thai, Yixiao Song, Ankita Gupta, and Mohit Iyyer. DEMETR: Diagnosing evaluation metrics for translation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 9540–9561, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.649. URL https://aclanthology.org/2022.emnlp-main.649.

Kenneth Katzner and Kirk Miller. *The languages of the world*. Routledge, 2002.

Aochong Oliver Li and Tanya Goyal. Memorization vs. reasoning: Updating LLMs with new knowledge. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 25853–25874, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1326. URL https://aclanthology.org/2025.findings-acl.1326/.

Shuang Li, Jiangjie Chen, Siyu Yuan, Xinyi Wu, Hao Yang, Shimin Tao, and Yanghua Xiao. Translate meanings, not just words: Idiomkb' s role in optimizing idiomatic translation with language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 18554–18563, 2024.

Junwei Liao and Shuai Cheng. Text polishing with chinese idiom: Task, datasets and pre-trained baselines. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 2023.

Chuang Liu, Linhao Yu, Jiaxuan Li, Renren Jin, Yufei Huang, Ling Shi, Junhui Zhang, Xinmeng Ji, Tingting Cui, Liutao Liutao, Jinwang Song, Hongying Zan, Sun Li, and Deyi Xiong. OpenEval: Benchmarking Chinese LLMs across capability, alignment and safety. In Yixin Cao, Yang Feng, and Deyi Xiong (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 190–210, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-demos.19. URL https://aclanthology.org/2024.acl-demos.19/.

Emmy Liu, Aditi Chaudhary, and Graham Neubig. Crossing the threshold: Idiomatic machine translation through retrieval augmentation and loss weighting. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical*

Methods in Natural Language Processing, pp. 15095–15111, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.933. URL https://aclanthology.org/2023.emnlp-main.933.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*, 2022.

OpenAI. Hello gpt-4o, May 2024. URL https://openai.com/index/hello-gpt-4o/.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri (eds.), *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pp. 578–585, Abu Dhabi, United Arab Emirates (Hybrid), December 2022a. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.52.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri (eds.), *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pp. 634–645, Abu Dhabi, United Arab Emirates (Hybrid), December 2022b. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.60.

Sara Rezaeimanesh, Faezeh Hosseini, and Yadollah Yaghoobzadeh. Large language models for Persian-English idiom translation. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7974–7985, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.405. URL https://aclanthology.org/2025.naacl-long.405/.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann A. Copestake, and Dan Flickinger. Multiword expressions: A pain in the neck for nlp. In *Conference on Intelligent Text Processing and Computational Linguistics*, 2002.

Yutong Shao, Rico Sennrich, Bonnie Webber, and Federico Fancellu. Evaluating machine translation performance on Chinese idioms with a blacklist method. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1005.

Michaela Socolof, Jackie Chi Kit Cheung, Michael Wagner, and Timothy J. O' Donnell. Characterizing idioms: Conventionality and contingency. *ArXiv*, abs/2104.08664, 2021.

Minghuan Tan and Jing Jiang. Learning and evaluating chinese idiom embeddings. In *Recent Advances in Natural Language Processing*, 2021.

Minghuan Tan, Jing Jiang, and Bingtian Dai. A bert-based two-stage model for chinese chengyu recommendation. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 20:92:1–92:18, 2021.

Kenan Tang. Petci: A parallel english translation dataset of chinese idioms. *ArXiv*, abs/2202.09509, 2022.

Kenan Tang, Peiyang Song, Yao Qin, and Xifeng Yan. Creative and context-aware translation of east asian idioms with gpt-4. *arXiv preprint arXiv:2410.00988*, 2024.

Simone Tedeschi, Federico Martelli, and Roberto Navigli. Id10m: Idiom identification in 10 languages. In *NAACL-HLT*, 2022.

Lei Wang and Shiwen Yu. Construction of Chinese idiom knowledge-base and its applications. In *Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications*, pp. 11–18, Beijing, China, August 2010. Coling 2010 Organizing Committee. URL https://aclanthology.org/W10-3703.

Ran Wang. The cultural differences between english and chinese idioms and translation. In *2022 International Conference on Creative Industry and Knowledge Economy (CIKE 2022)*, pp. 610–614. Atlantis Press, 2022.

You Wang. A contrastive research on the definitions and categories of chinese and english idioms. In *7th International Conference on Humanities and Social Science Research (ICHSSR 2021)*, pp. 423–426. Atlantis Press, 2021.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. A paradigm shift in machine translation: Boosting translation performance of large language models. In *ICLR*, 2024.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*, 2020.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024a.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024b.

Andrea Zaninello and Alexandra Birch. Multiword expression aware neural machine translation. In *International Conference on Language Resources and Evaluation*, 2020.

Ziheng Zeng and S. Bhat. Getting bart to ride the idiomatic train: Learning to represent idiomatic expressions. *Transactions of the Association for Computational Linguistics*, 10:1120–1137, 2022.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

Chujie Zheng, Minlie Huang, and Aixin Sun. Chid: A large-scale chinese idiom dataset for cloze test. *ArXiv*, abs/1906.01265, 2019.

A 习语翻译示例

以下是不同翻译类别和子类别的输入和输出，以及相应的解释。我们略去简单的情况。

UTF8gbsn

No Error: Correct Underlying Meaning

Chinese : 主断层绕过核电站，停了下来。美国地质调查局的科学家们欢呼雀跃，却发现断层并未被完全阻止。

English : The main fault bypassed the nuclear power plant and came to a stop. Scientists from the United States Geological Survey cheered and rejoiced, but they discovered that the fault had not been completely halted.

Explanation : 欢呼雀跃 is translated correctly into "cheered and rejoiced".

No Error: Literal and Coherent

Chinese : 此举指向性明确——挑衅美元，直接让联储局骑虎难下。（更多内容见：《沙特为首的油霸们再捅刀美元 项庄舞剑意在沛公》）3 月，沙特一连串的‘出牌’，是沙特战略转型的重要一步。

English : This move has a clear direction - provoking the US dollar and putting the Federal Reserve in a difficult position. (For more content, see: "Saudi-led oil tycoons stab the US dollar again, Xiang Zhuang's sword dance is aimed at Pei Gong") In March, Saudi Arabia's series of 'moves' is an important step in its strategic transformation.

Explanation : The original meaning of 项庄舞剑 means the actual intention is not what it appears to be. However, in this case, the idiom is used to refer to the historical story and a literal translation is more appropriate.

Mistranslation

Chinese : 另一半有时需要一个人安静地思考，此时你不必过多打扰，问得太多只会自寻烦恼。

English : The other half sometimes needs someone to think quietly by themselves, at which point you don't need to disturb them too much. Asking too many questions will only bring them problems.

Explanation : 自寻烦恼 means asking for trouble or trouble someone him/herself, not "bringing other people trouble".

Unnatural

Chinese : 善战者无赫赫之功。

English : A good warrior does not achieve a glaring victory.

Explanation : 赫赫之功 means "a great achievement" or "a great victory". "A glaring victory" conveys a rough meaning but the usage of "glaring" is not perfect enough.

Literal Translation

Chinese : 我爸临终时候说，这一生身无长处，不曾留下什么给我

English : My father said when he was dying, 'I have no merits in this life, I have not left anything for you.'

Explanation : 身无长处 is translated literally here. 身: in this life. 无: no. 长处: merits.

Addition

Chinese : 程长庚不仅“于剧学素有深造，无所不通”，而且怀有崇高的戏剧理想和抱负。

English : Cheng was not only "well-versed in drama studies, knowledgeable and passionate in all aspects", but also harbored lofty theatrical ideals and ambitions.

Explanation : 无所不通 means knowledgeable in all aspects, and it is not related to one's passion.

Partial Translation: Missing Modifier

Chinese : 《易》说：“出征顺利，斩了首恶，俘获他的同伙。”称扬赞美诛杀罪大恶极之人，那些不归顺的都来归服了。

English : The "I Ching" says, "The expedition is successful, the chief evil is beheaded, and his accomplices are captured." It praises the execution of those who commit evil crimes, and those who do not submit have come to surrender.

Explanation : 罪大恶极 refers to extremely evil crimes. Here, the translation has conveyed the core idea of the idea, but the degree of “evil” is missing.

Partial Translation: Missing Core Info Chinese : 再者，平台繁多，此处不留人，自有留人处，对于这类无良民宿，仅靠平台按现行规则管制，并不足以纠偏。

English : Also, there are many agencies, there is no place to stay, and against these bad inns, only the agencies with the current rules cannot adequately curb the abuse.

Explanation : Only the first half of 此处不留人，自有留人处 is translated. The second half of the idiom contains important information for us to understand the idiom, too.

Partial Translation: Inaccurate Modifier

Chinese : 被俘俄军惊恐万状，其实他比地上的同伙幸运多了。

English : The captured Russian soldiers are a bit terrified, but in reality, they are much luckier than their comrades on the ground.

Explanation : 惊恐万状 means “very terrified”. The translation inaccurately used “a bit” as a modifier.

Repetition

Chinese : 我认为最佳的父子关系应该如兄如 弟。

English : I think the best relationship between a father and son should be like that of a brother and a brother.

Explanation : 如兄如弟 means the relationship is like brothers. In this case, the translation is repeated.

No Translation: Missing from Output

Chinese : 伊拉斯谟当时正从布隆泉飞向普利托利亚，除了不速之客的蛇外，机上还载了 4 名乘客。

English : Irascible was flying from Bloemfontein to Pretoria and had four other passengers on board, including the snake.

Explanation : 不速之客 is not translated and is missing from the English sentences.

No Translation: Chinese Idioms Copied or Paraphrased

Chinese : 他指平日儿子和媳妇都早出晚归，只有他们两老和两名孙女在家，他担心对方会变本加厉。

English : He 指平日儿子和媳妇都早出晚归，只有他们两老和两名孙女在家，他担心对方会变本加厉。

Explanation : 早出晚归 is not translated and remains in Chinese text.

B 语料统计

B.1 旧语料库的统计数据

Table 7 显示了旧语料库的统计数据。

Domain	Source	Period	Instances	Idioms	Avg. occurrence by freq. range			
					VH	H	M	L
News	News Crawl	2019	17,931,873	10,830	425.6	38.6	8.0	1.5
Web	mc4	2008-22	54,542,308	30,652	8135.5	220.0	53.0	26.0
Wikipedia	wiki2019zh	2018-19	1,043,224	11,754	95.1	13.0	3.4	1.0
Social Media	PchatbotW	2008-22	139,448,339	15,089	2313.1	189.9	16.2	1.8

Table 7: 旧语料库的统计。Source 列展示了数据收集的来源。Period 列列出了数据收集的年份。Instances 列展示了每个领域中的数据样本数量。Idiom 列显示了成语的总数量。最右边的列显示了在每个频率范围内成语的平均出现次数。

B.2 各频率范围内的成语数量

Table 8 显示了我们新收集的数据中每个频率范围的成语数量。

Domain	VH	H	M	L	N
News	2,486	1,768	692	225	162
Web	7,310	4,344	2,117	1,442	106
Wikipedia	2,814	1,702	772	269	390
Social Media	60	60	60	104	67

Table 8: 我们收集的数据中每个频率范围内的成语数量。

C 数据收集和标注细节

C.1 每个领域的数据收集

新闻。我们收集了来自 Common Crawl News（2023 年 4 月和 2024 年 10 月快照）的数据，这些数据包含 41,354 篇中文新闻文章，为了一致性我们将其转换为简体中文。对于每篇文章，我们提取了包含成语的句子，以及前后句子以提供上下文，总共得到 50,845 个实例。

网页。我们使用 Common Crawl（2023-06 和 2024-12 快照）来收集网页数据。从每个 URL 获取 HTML 内容后，我们提取了包含习语及其周围上下文的句子，类似于新闻数据。总共得到了 463,642 个实例。

维基百科。我们选择中文维基百科元历史（2023-07 和 2024-12 的快照）进行数据收集。我们保存首次修订在截止日期之后的文章的修订 ID。我们通过 Wikimedia REST API [1] 获取每次修订的 HTML 页面。然后，我们从修订后的文本中提取带有成语及其上下文的句子。总共有 39,699 个实例。

[1] https://zh.wikipedia.org/api/rest_v1/

社交媒体。我们使用微博，一个类似于 Twitter 的流行中国社交媒体平台，来收集包含成语的帖子。我们依赖微博的搜索功能来收集包含成语的帖子。我们检查每个帖子的发布日期以确保它是在 2023 年之后发布的。与之前的领域不同，我们是从 `old corpus` 中收集包含所有成语的新数据，查询每个成语将导致较高的 API 使用开销。我们改为从 `old corpus` 中基于频率范围抽取 400 个成语（`VH`，`H`，`M` 各 60 个，`L` 和 `N` 各 110 个）。我们加倍了 `L` 和 `N` 的数量以缓解数据的稀缺性。我们设法收集到了 400 个成语中 351 个的帖子，总共有 55,315 个实例。

C.2 质量控制

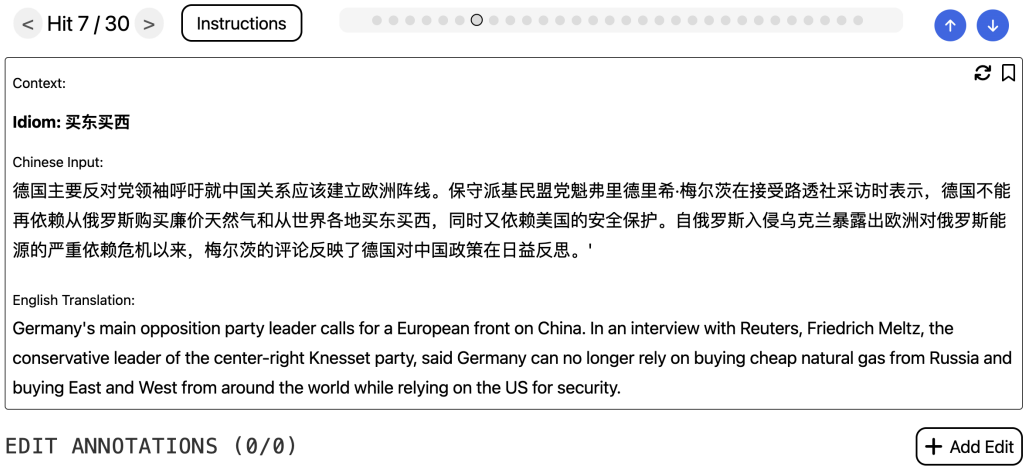

Figure 5: 示例测验问题。

我们使用 Prolific 招募以中文为母语且流利掌握英文的标注人员。我们仅收集我们提供的数据上的注释，不收集关于标注人员本身的信息。我们已获得标注人员在本次任务上的完全同意。标注过程分为两个阶段：试验阶段和主要阶段。在试验阶段，选择 20 名参与者进行预筛选。他们会得到详细的指南，必须完成一个测验以展示他们的理解。测验包含 30 个问题，每个类别和子类别有 2-3 个问题。一个示例测验问题如 Figure 5 所示。从试验阶段中表现最好的 5 名人员被选入主要阶段。在主要阶段，我们随机将每对翻译分配给三名标注员，同时我们确保每个标注员看到的对数相等。我们抽取 100 个注释并与作者自己的注释进行对比评估，实现了 66 % 的类别准确性和 88 % 的子类别准确性。对于最终的注释，我们采用多数投票并由第一作者手动解决任何打平情况。这种解决方案将标注人员在类别（子类别）一致性上的平均 Cohen's κ 从 0.47 （0.38 ）提高到 0.73 （0.69 ），表明尽管某些翻译可能有多个答案，标注人员仍然有很强的一致性。例如，将 UTF8gbsn 虚情假意（虚假情感和意图，表明意图不诚实）翻译为“虚假”可以被视为误译或部分翻译。标注人员的报酬为每小时 $ 22。

D 对翻译系统的进一步分析

D.1 类别分解

Figure 7 显示了良好翻译、误译和部分翻译的子类别组成。可以看出，子类别在不同领域和系统中有所不同。例如，对于部分翻译，在 `News` 和 `Wikipedia` 上，Alma 可能会省略修饰语（100 % 和 75 %），而在 `Web` 和 `Social Media` 上，它在大多数情况下错过了核心含义（100 % 和 67 %）。mT0 在每个领域中都有大量未翻译，导致其翻译大多在输出中缺失。

D.2 频率分解

Figure 9 展示了每个系统在各个领域中每个频率范围的性能。可以看出，习语的频率范围与翻译质量无关。对于出现在 `old corpus` 的习语，其翻译质量不会随着频率范围的变化而下降。例如，来自 M 范围（100 % 良好翻译）的 `Web` 的习语在 GPT-4 中得到了最佳翻译，而那些来自 H 和 L 范围（60 % 良好翻译）的习语则实现了相当的翻译质量。在其他模型中也可以发现类似的模式。有趣的是，从未出现在 `old corpus` （N 范围）的习语并不总是导致

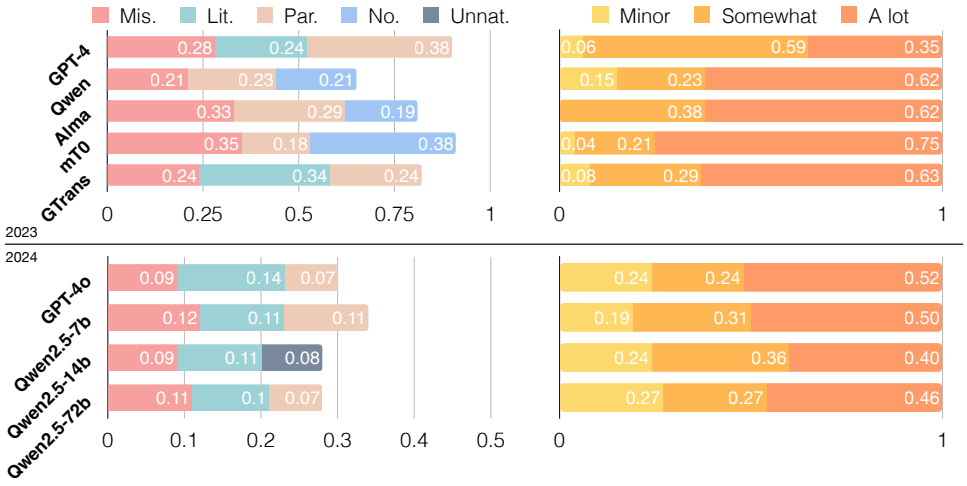

Figure 6: 左侧：每个系统最常见的前三个错误。`Mis.`：Mistranslation；`Lit.`：Literal Translation；`Par.`：Partial Translation；`No.`：No Translation。右侧：每个系统的各严重性评分比例，计算了所有的错误和领域。

较差的翻译。例如，在 `Wikipedia` 上，Alma 为 N 范围的习语提供了比其他频率更高质量的翻译。这一观察在 Social Media 的数据中尤为显著，其中一些从未出现的习语在 2023 年的数据中几乎在所有模型中都得到了最佳的翻译质量。这些习语在 2024 年的数据中也得到了相对较好的翻译质量。由于 N 范围是基于该领域中习语在 `old corpus` 中的出现来定义的，未在一个领域中出现的习语仍然可以出现在另一个领域中并被包括在预训练语料库中。这表明语言模型能够转移来自不同领域的知识并正确地翻译习语。

D.3 常见错误和严重性评分

Figure 6 显示了每个系统中最常见的前三个错误以及每个系统的每个严重性评分的比例。Figure 8 展示了每个系统在各个领域中的每种错误类别的严重性评分细分。我们注意到，会破坏意义的错误比保留整体意义的错误更为严重。当模型出现错误时，其中一半以上会对文本的理解产生重大影响。对于 2023 年的数据，尽管 GPT-4 表现最好，但只有 6 % 的错误翻译对文本的影响较小，使其在 Qwen 和 Google 翻译之后排名第三。Alma 犯的错误最为严重，没有任何错误被归类为轻微错误。在 2024 年的数据中，所有模型中超过 40 % 的错误被标记为最高严重性。

误译和未翻译是最严重的错误类型，因为它们会显著扰乱意思。除了 2024 年数据中的 `Wikipedia` 之外，误译的最低平均严重性评分为 2.43。同时，未翻译在许多领域的严重性评分为 3。相比之下，不自然翻译（最大严重性为 2）、添加（最大严重性为 2.5）和重复（最大严重性为 2）则较不严重，因为它们通常保留了一些正确或接近正确的惯用语部分。

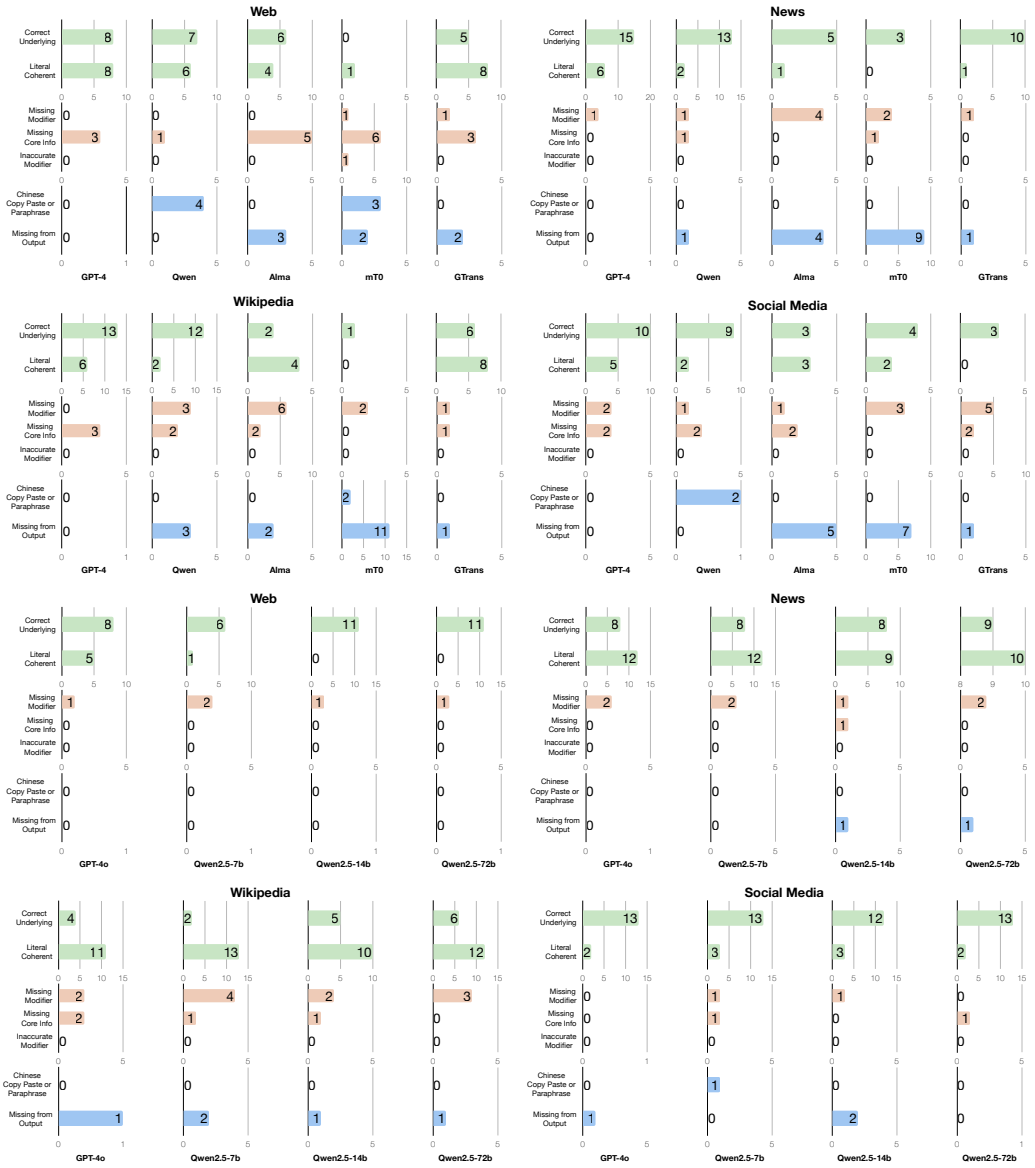

Figure 7: 子类别组成对于 Good Translation 、Partial Translation 和 No Translation 。

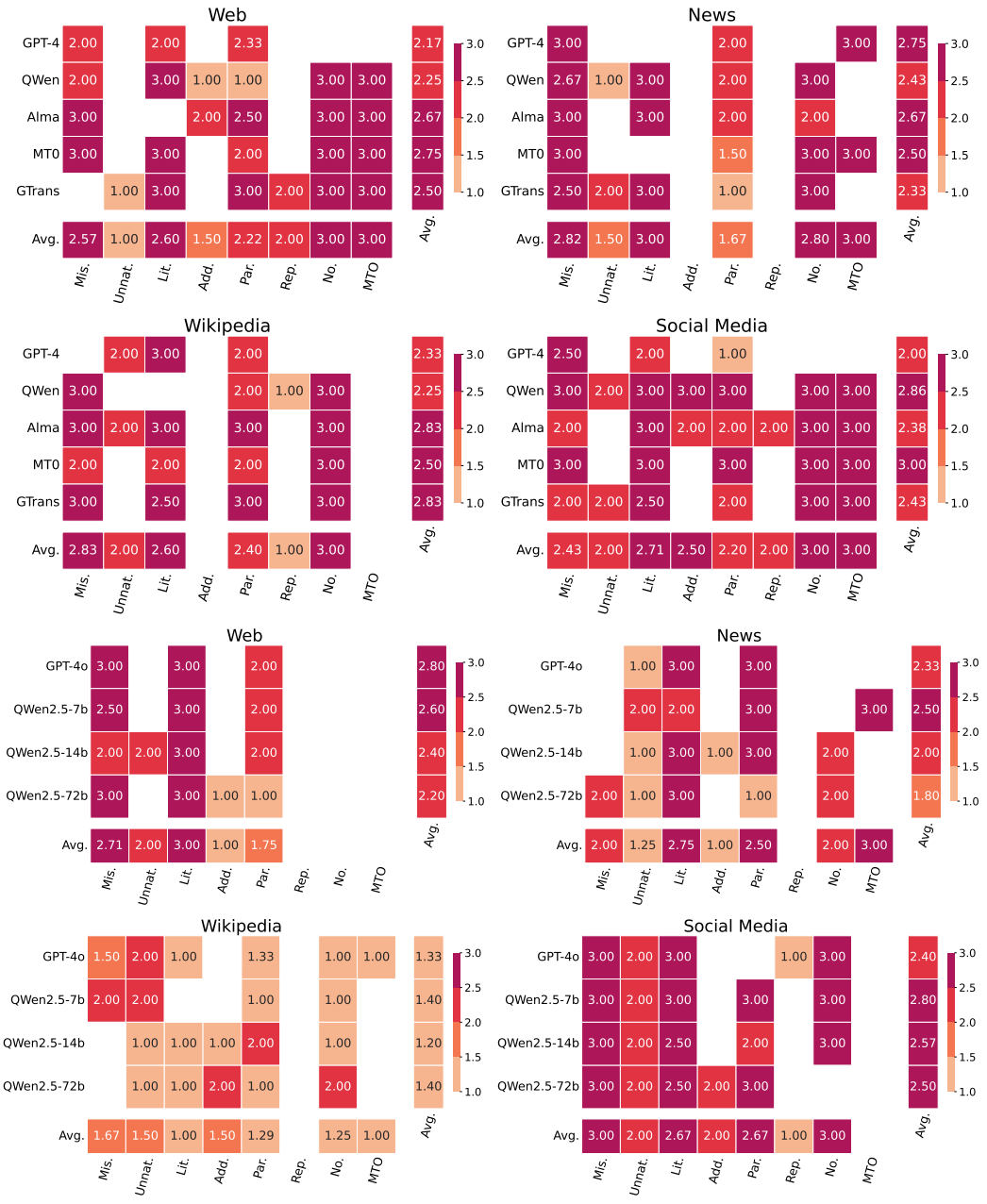

Figure 8: 每个领域中的每个翻译类别的平均严重程度分数。最右边的列表示每个系统所有错误的平均严重程度分数。底部的行表示每个错误所有系统的平均严重程度分数。

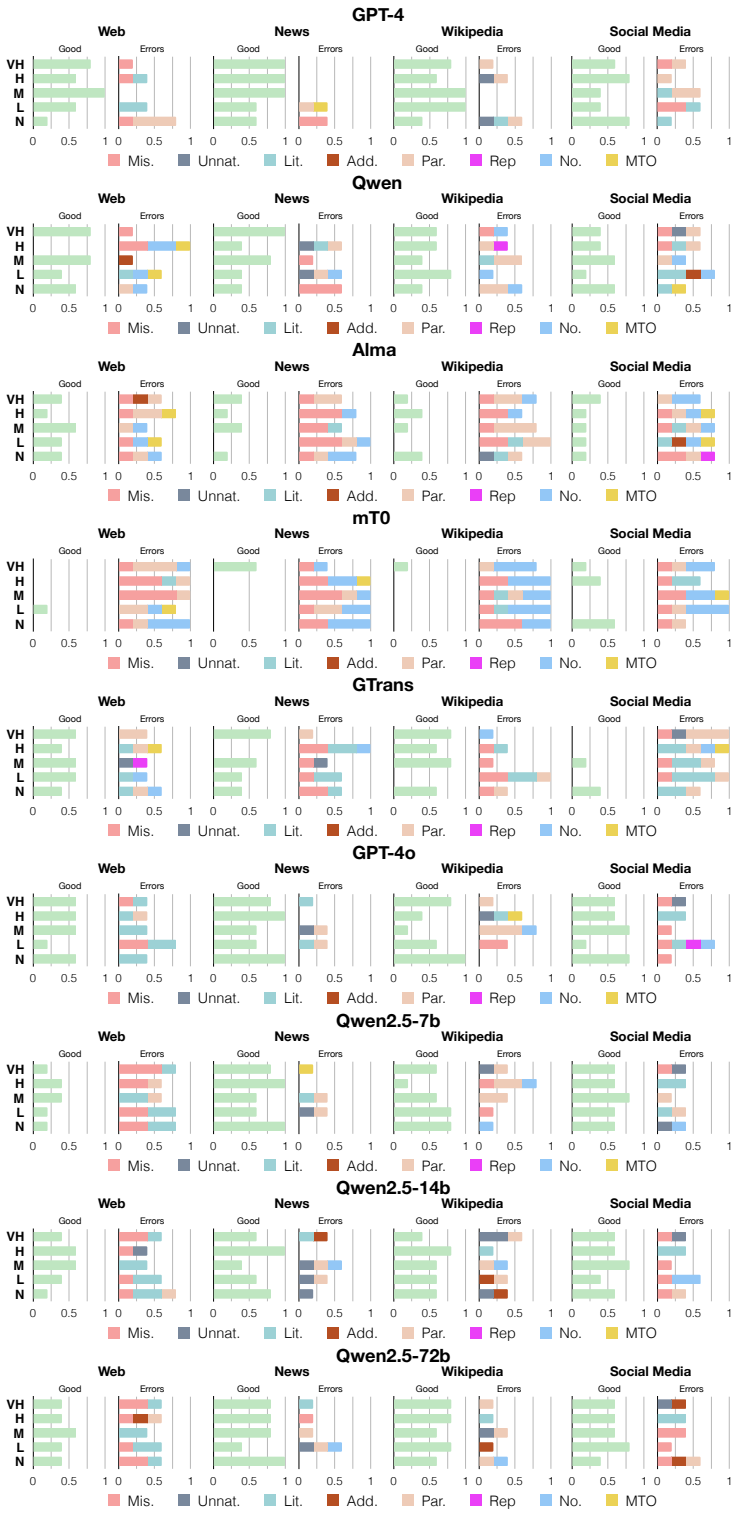

Figure 9: 以频率为中心的各领域翻译结果视图。在每个数据集中有两个子图。左子图：所有系统中好翻译的平均比例。右子图：所有系统中错误类别的平均比例。

E 进一步分析评价指标

E.1 Kendall 的 τ 在评价指标与人工标注之间的关系

Table 9 显示了评价指标与人工标注之间的 Kendall 的 τ。与 Table 3 类似，Kendall 的 τ 在各个类别中仍然较弱，这表明基于指标得分的排名与人工评分之间的一致性较低。

Scope	Category	Size	Reference-Based Metrics					Reference-Free Metrics		
			BLEU	BERTScore	COMET	MetricX -23-XXL	MetricX -24-XXL	COMETKIWI	MetricX-QE -23-XXL	MetricX-QE -24-XXL
Full	Mistranslation	128	0.064	0.099	0.141	0.081	0.129	0.192	0.049	0.097
	Literal	90	0.153	0.212	0.314	0.248	0.210	0.091	0.158	0.040
	Partial	105	0.176	0.175	0.238	0.098	0.158	0.097	0.240	0.247
	No Translation	74	0.182	0.271	0.345	0.204	0.232	0.072	0.156	0.273
Idiom	Mistranslation	128	-0.011	0.030	0.196	0.210	0.181	0.184	0.120	0.123
	Literal	90	-0.005	0.078	0.143	0.216	0.082	0.016	-0.052	-0.102
	Partial	105	0.122	0.104	0.072	0.034	-0.107	-0.043	-0.171	-0.195
	No Translation	74	*	*	-0.016	-0.019	-0.269	0.031	-0.273	-0.252
Full	All errors	452	0.199	0.269	0.323	0.199	0.211	0.173	0.155	0.162
	All categories	900	0.208	0.274	0.285	0.264	0.246	0.140	0.137	0.147
Idiom	All errors	452	0.178	0.218	0.259	0.269	0.006	0.012	-0.078	-0.140
	All categories	900	0.309	0.311	0.373	0.414	0.093	0.016	-0.054	-0.090

Table 9: Kendall's τ 在评价指标和人工注释之间。评估分为完整翻译和成语翻译。因为没有严重性评分，所以省略了 Good。这里省略了 More Than One，因为其所有实例都被评定为最高严重性。由于样本量小，省略了 Unnatural、Addition 和 Repetition。*: 对于成语中的 No Translation，BLEU 和 BERTScore 输出为 0，因此被省略。现有指标未能产生与人工注释强烈相关的度量。

E.2 提示

图 10 显示了用于提取成语翻译的提示。图 11 显示了用于编辑成语翻译的提示。图 12 显示了用于识别成语翻译的提示。

F 用于检测好坏习语翻译的实现细节

我们使用最新的 GPT-4o-20241120（简称 GPT-4o）和 Qwen2.5 指令模型进行提示。生成是通过贪婪解码完成的。图 13 显示了零样本提示，图 14 显示了带有思维链的零样本提示。

F.1 微调方法

我们使用 LoRA 对 Qwen2.5 Base 模型进行微调 (Hu et al., 2022)，并在推理期间使用贪婪解码。对于 MetricX-24，我们在解码器中添加一个分类头，并在训练期间同时更新 LoRA 参数和分类头。我们对学习率进行调优，在 { 5e-5, 1e-4, 3e-4, 5e-4 } 范围内进行 3 个周期的训练，并选择验证集上表现最好的检查点进行最终测试。

Qwen2.5 微调的输入格式：

```
Evaluate whether the Chinese idiom is correctly translated in the following text:

- Chinese Idiom: {idiom}
- Chinese Text: {zh_sentence}
- English Translation: {en_sentence}
```

Qwen2.5 微调的不同策略的输出格式：

1. 简短回答
   模板:
   ```
   yes / no
   ```
   描述: 该模型输出一个表示正确性（“yes”）或错误性（“no”）的单一标记。

2. 长答案
   模板：

   ```
   the idiom { idiom } got translated { correctly / incorrectly } .
   ```

   描述：模型输出一句简短的话，描述成语翻译是正确还是错误的，但不包括任何推理。
3. 简短的思考过程 + 最终答案
   模板：

   ```
   The translation of the idiom is “ { idiom_translation } ”. Given the
   context, I think this is a { category } , so my final answer is: the
   idiom got translated { decision } .
   ```

   描述：该模型提供了一个简短的推理（例如，提到习语翻译），然后说明它是否正确或不正确。
4. 长时间思考过程 + 最终答案
   模板：

   ```
   Analyzing the translation of “ { idiom } ” as “ { idiom_translation }
   ”. When examining this translation: { definition based on category }
   Therefore, I conclude this is (not) a good translation. The idiom “ {
   idiom } ” has been translated { decision } .
   ```

   描述：模型生成一个更详尽的思考，参考了相关错误类别（或“良好翻译”）的预定义定义。最后，决定翻译是否正确或不正确。

Prompt for Extracting Idiom Translations

```
## Task
Analyze the provided Chinese idiom, Chinese sentence containing the idiom
 ↪  and its English translation to extract corresponding idiom
 ↪  translation.

## Input to Analyze
* Chinese idiom: [PLACEHOLDER]
* Chinese sentence: [PLACEHOLDER]
* English translation: [PLACEHOLDER]

Please note:
1. Ensure the extracted idiom translation is short and concise, and do
 ↪  not include irrelevant translation.
2. If the given idiom appears multiple times in the Chinese sentence,
 ↪  only analyze the first occurrence.
3. If no corresponding translation is found, output an empty string.

Read the input carefully. Write down a brief thought process first, with
 ↪  the 3 notes in mind. Then extract the translation of the idiom.
Respond in a JSON format {"Idiom translation": }. The key must be "Idiom
 ↪  translation", and the value must be the corresponding idiom
 ↪  translation in the English sentence.
```

Figure 10: 提取成语翻译的提示

Prompt for Perturbing Idiom Translations

```
## Task
Analyze the provided Chinese idiom, idiom meaning, Chinese sentence
↪  containing the idiom and its English translation.
The goal is to edit the given idiom translation to comply with the given
↪  category.


## Categories
There are 7 categories to consider:
* Mistranslation: idiom translation is incorrect due to wrong choices of
↪  words or phrases, and it affects our understanding of the translated
↪  sentence.
* Partial Translation: idiom is translated partially. Part of the idiom
↪  meaning is missing or the extremity is inaccurate.
* Repetition: translation of the idiom contains repeated words or
↪  phrases, or their synonyms.
* Unnatural: the translation is ok but not perfect enough due to improper
↪  choices of words or grammar errors. In other words, the translation
↪  can be improved by more appropriate choices of words.
* Literal Translation: idiom is translated literally and translation is
↪  not coherent with context.
* Addition: In addition to good translation of idiom, the translation
↪  also contains non-present content in the source.
* No Translation: there is no translation of the idiom in the output.

## Input to Edit
* Chinese idiom: [PLACEHOLDER]
* Idiom meaning: [PLACEHOLDER]
* Chinese sentence: [PLACEHOLDER]
* English translation: [PLACEHOLDER]
* Idiom translation: [PLACEHOLDER]

Please note:
1. Given category only applies to the idiom translation itself.
2. An idiom can have multiple meanings. Find the most appropriate meaning
↪  under the given context.


Read the input carefully. YOU NEED TO MAKE THE IDIOM TRANSLATION A
↪  [PLACEHOLDER].
First, modify the idiom translation to comply with the given category.
Then, rewrite the given translation by replacing the original idiom
↪  translation with the new translation.

Do not include your thought process.

Output the corresponding edited translation only. Do not include anything
↪  else in your output.
```

Figure 11: 促使习语翻译的扰动

Prompt for Editting Idiom Translations

```
## Task
Analyze the provided Chinese idiom, idiom meaning, Chinese sentence
↪  containing the idiom and its English translation to identify idiom
↪  translation category.

## Categories
There are 8 categories to consider:
* Good Translation: the idiom is translated perfectly.
* Mistranslation: idiom translation is incorrect due to wrong choices of
↪  words or phrases, and it affects our understanding of the translated
↪  sentence.
* Partial Translation: idiom is translated partially. Part of the idiom
↪  meaning is missing or the extremity is inaccurate.
* Repetition: translation of the idiom contains repeated words or
↪  phrases, or their synonyms.
* Unnatural: the translation is ok but not perfect enough due to improper
↪  choices of words or grammar errors. In other words, the translation
↪  can be improved by more appropriate choices of words.
* Literal Translation: idiom is translated literally and translation is
↪  not coherent with context.
* Addition: In addition to good translation of idiom, the translation
↪  also contains non-present content in the source.
* No Translation: there is no translation of the idiom in the output.

## Input to Analyze
* Chinese idiom: [PLACEHOLDER]
* Idiom meaning: [PLACEHOLDER]
* Chinese sentence: [PLACEHOLDER]
* English translation: [PLACEHOLDER]
* Idiom translation: [PLACEHOLDER]

Please note:
1. Focus on analyzing the translation of the idiom under the context of
↪  the sentence.
2. Compare the idiom meaning and provided idiom translation to make the
↪  judgement. Note an idiom can have multiple meanings.
3. If the given idiom appears multiple times in the Chinese sentence,
↪  only analyze the first occurrence.
4. If the idiom translation only captures the essence of the meaning, it
↪  is a "Partial Translation".

Read the input carefully. Write down a brief thought process first, with
↪  the 4 notes in mind. Then identify the translation category of the
↪  idiom.
Respond in a JSON format: the keys must be "Category", and the values
↪  must be the corresponding category of the idiom translation.
```

Figure 12: 编辑习语翻译的提示

Zero-shot Prompt for Idiom Translation Correctness Detection

```
Evaluate whether the Chinese idiom is correctly translated in the
↪  following text:

- Chinese Idiom: {idiom}
- Chinese Text: {zh_sentence}
- English Translation: {en_sentence}

Note: only output Yes or No in your response. Do not include anything
↪  else.
```

Figure 13: 零样本提示用于成语翻译正确性检测

Zero-shot with CoT Prompt for Idiom Translation Correctness Detection

```
Evaluate whether the Chinese idiom is correctly translated in the
↪  following text:

- Chinese Idiom: {idiom}
- Chinese Text: {zh_sentence}
- English Translation: {en_sentence}

Note: Analyze this step by step with the following output format:
── Thought Process: {{your analysis of the idiom's meaning and
↪  translation accuracy}}
── Final Answer: {{correct translation / wrong translation}}
```

Figure 14: 零样本与链式思维提示用于成语翻译正确性检测