人工智能推理模型应用于法律文本汉英翻译的效能研究

魏榕[1]

（1.西南政法大学，重庆 401120）

摘　要：随着人工智能技术的持续演进，人工智能推理模型逐渐被引入法律翻译实践中。然而，该领域的系统性实证研究仍较为稀缺，尤其是对模型输出质量的量化分析与对比研究尚待深入。本研究《中华人民共和国刑法》总则为语例，选取四种主流推理模型——ChatGPT o1、DeepSeek R1、Grok 3 with Think 和 Gemini 2.0 Flash Thinking Experimental，对模型生成的译文进行 BLEU 评分，并结合 Python 与 SPSS 进行统计检验与多维度对比分析。数据表明，四种推理模型在法律文本翻译中的表现存在显著差异，其中 Grok 3 与 Gemini 2.0 在 BLEU 均值、高质量翻译频次和低分率等维度上相对优越，而 ChatGPT o1 与 DeepSeek R1 则在术语处理、句法结构及稳定性方面存在较大提升空间。进一步的分析表明，推理模型在应对法律术语、复杂句式和法律逻辑方面仍存在准确性挑战。人工智能推理模型虽具备赋能法律翻译的潜力，但在专业语域的深层理解与表达上仍难以取代人工译者。为提升模型译文质量，建议加强本土法律语料建设、优化术语一致性机制、引入译后人工校对流程，并推动构建具有中国特色的法律翻译模型。

关键词：人工智能推理模型；法律翻译；BLEU 指标；翻译质量评估

Doi：doi.org/10.70693/rwsk.v1i6.682

一、引言

随着人工智能技术的快速发展，特别是大语言模型和推理模型的不断演进，人工智能翻译工具已广泛应用于各类翻译实践中，在提升翻译效率、降低人工成本的同时，也推动了翻译研究的范式转型。在此背景下，如何科学评估人工智能在专业领域中的翻译表现，成为应用翻译研究亟待回应的重要议题。法律文本作为一种高度规范化、专业化和制度化的文本类型，其语言形式严谨、术语密集、逻辑结构复杂，对翻译的准确性、规范性与文化适应性提出了极高要求。

近年来，人工智能推理模型在自然语言处理领域表现出强大的语言理解与生成能力，被广泛应用于对话生成、文本摘要和语言推理等任务。然而，在法律翻译这一特定语境中，推理模型的应用仍处于探索阶段，其译文是否具备专业性与可接受性，尚缺乏系统的实证研究和客观量化评估。党和政府一直高度重视外宣工作[1]。在此背景下，推动中国法律制度及法治理念的有效对外传播，亟需借助技术手段提升法律翻译质量，以实现精准表达与国际传播之间的平衡。

鉴于此，本研究选取《中华人民共和国刑法》总则部分作为语例，测试并比较四款当前具有代表性的人工智能推理模型——ChatGPT o1、DeepSeek R1、Grok 3 with Think (以下简称 Grok 3) 和 Gemini 2.0 Flash Thinking Experimental (以下简称 Gemini 2.0) 的翻译表现。研究采用 BLEU (Bilingual Evaluation Understudy) 作为量化评价工具，借助 Python 编程语言和 SPSS 统计软件对模型输出进行量化分析，并结合典型译例从术语准确性、句法结构、逻辑推理等方面进行微观对比，力图多维度呈现各模型在法律文本翻译中的优势与不足。本研究有望为人工智能推理模型应用于法律文本翻译提供借鉴和启示，为翻译在人工智能时代提供新的视角和方向。

二、相关研究

不少学者已经关注到人工智能大语言模型给翻译工作带来的机遇和挑战，并且主要集中在以下两个方

作者简介：魏榕（1993—），女，广东深圳人，西南政法大学外语学院硕士研究生，研究兴趣为法律翻译理论与实践。

面：一是使用 BLEU 等自动化指标和专题数据库测试各个大语言模型的翻译质量。有学者使用 BLEU 和 TER 量化 ChatGPT 和其他机器翻译翻译中国共产党二十大报告，发现 ChatGPT 相对于其他翻译工具表现出一定的优势，但是 ChatGPT 在处理涉及意识形态、复杂结构、文化负载词、隐转喻等内容以及在翻译准确性上的局限性仍然明显[2]。亦有学者以专题数据库和 ChatGPT 为研究对象，指出基于 ChatGPT 的智能翻译结果相比，平台检索的翻译结果更加精准，能够弥合知识域与语言域之间的差距，为翻译教学及研究提供可靠的翻译数据支持[3]。

二是有学者关注大语言模型对语言和翻译带来的优势和困境。首先是研究人员指出，机器翻译将成为人工翻译的好朋友和得力助手，机器翻译和人工翻译应当和谐共生，相得益彰[4]。再有学者基于 ChatGPT 的翻译实例研究证明，人工智能辅助译后编译可提供汉译英语言对中译文校对、润色、评估、反馈及建议，同时可从译文批改等方面协助翻译教学[5]。最后是部分学者关注 ChatGPT 赋能翻译过程中的伦理挑战，如数据偏见与歧视、侵犯隐私权和知识产权，以及质量问题与责任归属，并且提出有效建议，以确保翻译服务的质量、公正性和可持续性，最终实现 ChatGPT 在翻译领域的潜力和价值[6]。

以上研究分析了人工智能大语言模型对翻译领域带来的巨大影响，但是几乎没有学者关注人工智能推理模型应用于法律文本翻译的效能表现，这为我们进行相关研究提供了必要性和可行性。

三、研究设计

（一）研究问题

本研究旨在讨论以下三个问题：第一，人工智能推理模型生成的法律翻译质量如何？第二，人工智能推理模型能否替代人工译者在法律文本翻译中的功能？第三，在人工智能时代，法律翻译应该何去何从？

（二）文本和模型选择

中文文本选自《中华人民共和国刑法》总则部分，总计 101 条，共获取中文文本 9488 个汉字，参考英文翻译选取的是全国人民代表大会常务委员会法制工作委员会官方网站提供的译文，共 7222 个单词。本文测试的是市面上四款主流的人工智能推理模型，它们分别是美国人工智能研究实验室 OpenAI 开发的 ChatGPT o1，杭州深度求索人工智能基础技术研究有限公司开发的 DeepSeek R1，美国 xAI 公司开发的 Grok 3 以及谷歌公司开发的 Gemini 2.0。

这四款推理模型代表了当前人工智能领域中的领先技术，涵盖了不同公司的创新成果。ChatGPT o1 具有强大的自然语言处理能力；DeepSeek R1 则专注于深度学习与语义理解，并且具备完整的思维链；Grok 3 结合了最新的推理技术，表现出较高的智能推理能力；Gemini 2.0 则利用谷歌在大规模数据处理和模型优化上的优势，致力于提供高精度的翻译结果。选择这四款模型，有助于全面评估不同技术框架下法律翻译的质量与效果。

（三）研究方法

BLEU（Bilingual Evaluation Understudy）指标是一种用于评估机器翻译质量的自动化评价方法[7]。它通过比较机器生成的翻译结果与参考翻译之间的 n-gram 重合度来量化翻译质量。BLEU 可以用来衡量术语短语和句子结构的一致性，因此非常适合高度规范化和格式化的法律文本。根据李洛克（2020），BLEU 的计算公式如下[8]：

$$P_n = \frac{\sum_i^E \sum_k^K min(h_k(c_i), min_{j \in M} h_k(s_{i,j}))}{\sum_i^E \sum_k^K min(h_k(c_i))} \#(1)$$

（1）式中，S_j表示参考译文，其中$j \in M$，M 表示共有 M 个参考答案。C_i表示机器译文，其中

$i \in E$E，E 表示共有 E 个翻译。N-gram 表示 n 个单词长度的词组集合，令 k 表示第 k 个词组；$h_k(c_i)$表示第 k 个词组在机器译文c_i中出现的次数;$h_k(s_{i,j})$表示第 k 个词组在参考翻译$s_{i,j}$中出现的次数。

$$BP = \begin{cases} 1 & ,if\ l_c > l_s \\ e^{1-\frac{l_s}{l_c}} & ,if\ l_c \le l_s \end{cases} \#(2)$$

l_c表示机器译文的长度，l_s表示参考翻译的有效长度。

$$BLEU = \mathrm{BP} \times \exp\left(\sum_{n=1}^{N} W_n log(P_n)\right) \#(3)$$

BLEU 的原型系统采用的是均匀加权，即$W_{n=}\frac{1}{N}$。N 的上限取值为 4，即最多只统计 4-gram 的精度。

BLEU 计算结果得出以后，使用 SPSS 27 进行正态性检验，而后使用 Python 进行（非）参数检验，其中使用到的 Python 库包括读取 Excel 文件的 pandas，进行曼-惠特尼 U 检验的 scipy.stats 以及进行邦费罗尼校正的 statsmodels.stats.multitest 等。此外，代码还使用了 Python 内置库。

（四）研究过程

人工智能推理模型的使用和指令息息相关，不同的指令会产生不同的效果。因此，为了保证实验的严谨性以及确保测试结果的信度与效度，我们对各个模型都使用同一指令。

在指令相同的情况下，我们对比了 ChatGPT o1、DeepSeek R1、Grok 3 和 Gemini 2.0 所产出的译文质量。我们首先使用 Python 分别计算各推理模型翻译刑法总则每一条的 BLEU 值，结合 SPSS 27 和 Python 对 BLEU 值进行统计分析，以判断这四者在同一测评指标下的翻译质量是否存在统计学意义上的显著差异，最后从多维度对比分析这四种人工智能推理模型刑法总则的 BLEU 值。这样多角度、多层次的评估方法更能揭示人工智能推理模型的优缺点，进一步推动机器翻译技术的优化和发展。

四、研究结果与分析

我们在获取译文后将四者生成的译文进行对比，并使用 Python 计算出它们翻译刑法总则每一条的 BLEU 分数。在进行正态性检验的同时，获取各推理模型描述信息，如表 1 和表 2 所示。

表 1 各推理模型描述信息

推理模型	个案数	BLEU 均值	良好翻译频次	最值		平均值的 95%置信区间	
				最小值	最大值	下限	上限
ChatGPT o1	101	17.77	11	2.48	100	15.21	20.33
Grok 3	101	32.75	50	3.76	100	29.61	35.89
DeepSeek R1	101	13.18	14	0	80.53	10.11	16.26
Gemini 2.0	101	30.55	40	4.28	66.04	27.95	33.14

表 2 正态性检验

推理模型	柯尔莫戈洛夫-斯米诺夫 [a]			夏皮洛-威尔克		
	统计	自由度	显著性	统计	自由度	显著性
ChatGPT o1	.159	101	<.001	.778	101	<.001
Grok 3	.089	101	.047	.948	101	<.001
DeepSeek R1	.199	101	<.001	.797	101	<.001
Gemini 2.0	.097	101	.020	.975	101	.051

a.里利氏显著性修正

根据表 2 所示，在柯尔莫戈洛夫-斯米诺夫和夏皮洛-威尔克检验下，ChatGPT o1、Grok 3 和 DeepSeek R1 的 p 值均小于 0.05，因此可以认为三者均不服从正态分布，虽然 Gemini 2.0 在夏皮洛-威尔克检验下的 p 值大于 0.05，但是靠近边界，并且在柯尔莫戈洛夫-斯米诺夫检验下 p 值小于 0.05，因此没有十足的把握确定 Gemini 2.0 是否服从正态分布。因此，我们继续使用独立样本的非参数检验，这种方法适用于比较多组独立样本之间的差异，且对数据的分布没有严格假设要求，常用于数据不服从正态分布或含有异常值的情况。检验结果如表 3 所示。

表 3 检验统计 [a,b]

	BLEU
克鲁斯卡尔-沃利斯 H	126.166
自由度	3
渐进显著性	<.001

a. 克鲁斯卡尔-沃利斯检验

b.分组变量：推理模型

根据表 3 所示，渐进显著性<.001，这表明四种推理模型存在统计学意义上的显著差异。为了进一步明晰究竟是哪两种推理模型之间存在显著差异，我们使用曼-惠特尼 U 检验对四个推理模型之间的性能进行了成对比较。检验基于每个模型翻译的 BLEU 分数进行，以评估不同模型在翻译效果上的差异性。为了避免多重比较问题的影响，我们采用了邦费罗尼校正对 p 值进行了调整。它是一种用于多重比较中的 p 值调整方法，用于减少因进行多次假设检验而增加的第一类错误（假阳性）的风险。通过邦费罗尼校正，原始的显著性水平被除以检验的次数，从而使得每个独立检验的显著性水平变得更严格。检验结果如表 4 所示。

表 4 各推理模型成对分析

比较模型	原始 p 值	校正后 p 值	是否有显著差异
ChatGPT o1 vs Grok 3	<.001	<.001	是
ChatGPT o1 vs DeepSeek R1	<.001	<.001	是
ChatGPT o1 vs Gemini 2.0	<.001	<.001	是
Grok 3 vs DeepSeek R1	<.001	<.001	是
Grok 3 vs Gemini 2.0	0.38	1	否
DeepSeek R1 vs Gemini 2.0	<.001	<.001	是

根据曼-惠特尼 U 检验及其邦费罗尼校正结果，四个推理模型之间存在显著的性能差异，尤其是 ChatGPT o1 与其他模型之间的差异更为显著，而 Grok 3 与 Gemini 2.0 之间未显示出显著的差异。

对 BLEU 值进行上述统计分析仅能大致判断各个推理模型是否存在显著的性能差异，根据表 1，我们将从以下几个方面从宏观分析对四种推理模型在 BLEU 测评中的表现，试图挖掘更多的信息：

（1）整体充分性和流畅性。倘若一个翻译工具生成译文的 BLEU 值达到 31.4%，就表明其译文质量良好并且达到了机器翻译的基本要求[9]。我们分析 Python 计算得出的数据发现，Grok 3 的平均 BLEU 值最高，为 32.75%；Gemini 2.0 次之，平均 BLEU 值为 30.55%；ChatGPT o1 的平均 BLEU 值为 17.77%；DeepSeek R1 的平均 BLEU 值最低，仅为 13.18%。这些数据表明，在译文的充分性和整体性方面，只有 Grok 3 达到了机器翻译的效果，其余三者均未达到基本要求，在充分表达原意和语言流畅性方面还有较大的上升空间。

（2）高质量翻译频次。我们发现在这四种推理模型中，Grok 3 译文在具有较高 BLEU 值的频次达到 50 次，Gemini 2.0 为 40 次，DeepSeek R1 为 14 次，ChatGPT o1 为 11 次。从高质量翻译出现频次这一评判指标来看，仍然是 Grok 3 表现最好。

（3）低分率。值得注意的是，四种推理模型译文的 BLEU 值均出现了分数极低的情况，这表明推理模型生成的译文和人工生成的参考译文差异较大。因此，我们还应当关注推理模型的低分率。BLEU 值低分频次越多，表明该推理模型的翻译质量越不稳定。经过计算，Grok 3 和 Gemini 2.0 各译文的 BLEU 值低于 10% 的频次最少，仅有 5 次，ChatGPT o1 为 26 次，DeepSeek R1 为 59 次，其中有 18 次趋近于 0。这进一步证明了 Grok 3 相对优越的性能。

在对各个推理模型进行上述宏观分析之后，我们发现，Grok 3 的表现最佳，Gemini 2.0 紧随其后，ChatGPT o1 和 DeepSeek R1 各有优劣。我们拟从以下几方面对它们的翻译表现进行微观分析：

(1)最值表现。首先看最大值。Grok 3 和 ChatGPT o1 的 BLEU 最大值都为 100%，该最大值分别出现在

刑法第二十八条“对于被胁迫参加犯罪的，应当按照他的犯罪情节减轻处罚或者免除处罚”和第五十二条“判处罚金，应当根据犯罪情节决定罚金数额”。 DeepSeek R1 的最大值 80.53%出现在第十一条“享有外交特权和豁免权的外国人的刑事责任，通过外交途径解决”，Gemini 2.0 的最大值 66.04%则出现在第五十四条“剥夺政治权利是剥夺下列权利：（一）选举权和被选举权；（二）言论、出版、集会、结社、游行、示威自由的权利；（三）担任国家机关职务的权利；（四）担任国有公司、企业、事业单位和人民团体领导职务的权利”。由此可见，这四个推理模型处理句式较为简单以及术语密度较低的法条的表现出色。

接着再看最小值。这四个推理模型的最小值都小于 10%，DeepSeek R1 的最小值 0%出现在第八十四条“被宣告假释的犯罪分子，应当遵守下列规定：（一）遵守法律、行政法规，服从监督；（二）按照监督机关的规定报告自己的活动情况；（三）遵守监督机关关于会客的规定；（四）离开所居住的市、县或者迁居，应当报经监督机关批准”，ChatGPT o1 的最小值 2.48%则出现在第二十二条“为了犯罪，准备工具、制造条件的，是犯罪预备。对于预备犯，可以比照既遂犯从轻、减轻处罚或者免除处罚”，Grok 3 和 Gemini 2.0 的最小值同时出现在第七十四条“对于累犯和犯罪集团的首要分子，不适用缓刑”。即便在句式较为简单的情况下，倘若术语出现的频率较高，比如“犯罪预备”、“ 累犯”、“ 首要分子”、“缓刑”和“假释”等，这四个推理模型表现不尽如人意，由此可见，它们对法律术语的翻译能力还有较大的提升空间。

(2)句子结构。刑法第五条“刑罚的轻重，应当与犯罪分子所犯罪行和承担的刑事责任相适应”，这是罪责刑相适应原则，原文看似将“刑法的轻重”、“罪行”和”刑事责任”三个概念相提并论，其实“刑法的轻重”对应的是“罪行”的大小以及“刑事责任”的大小。注重意合的汉语原文当然可采此句式，注重形合的英语译文却绝不可依样画葫芦，但是这四个推理模型都不约而同处理将句式结构为 The severity of punishment shall……。因此，推荐译文应当是：An offender shall be sentenced to a punishment in accordance with his criminal conduct and criminal responsibility[10]。由此看来，尽管这四个人工智能推理模型在翻译时加入了推理的方法，可以正确处理法律汉语的简单句型结构，但并不是对所有句型都了如指掌，因此人工译者在译后编辑中的作用仍不可小觑。

再如刑法第六条“凡在中华人民共和国领域内犯罪的， 除法律有特别规定的以外，都适用本法”，这里说的是刑法的属地管辖权。ChatGPT o1 根据汉语语序逐字逐句翻译，其余三者则先译“都适用本法”，其余成分处理为插入语，最后译“除法律有特别规定的以外”。除 ChatGPT o1 以外，其余三个模型的译文采用 “主语+动词+条件”的结构，这样能够突出该法条的重点。

(3)术语准确性。刑法总则中出现次数较多的术语主要有“有期徒刑”、“ 无期徒刑”和“死刑”，这三者现在最常见的译法分别是 fixed-term imprisonment、life imprisonment 和 death penalty（capital punishment），四个推理模型都能翻译出来。需要指出的是，考虑到约定俗成的术语翻译原则，除了“死刑”的英译 death penalty（capital punishment）没有什么争议之外，这些都是最常见但并非最准确的译法。原因在于，“有期徒刑”是剥夺罪犯一定期限的人身自由，而不是固定期限（fixed-term），“无期徒刑”则是指无限期地剥夺罪犯的人身自由，但是根据我国刑法的规定，无期徒刑适用假释和减刑，从实践层面来看，大部分的罪犯并没有关押到死，“无期徒刑”并不是真正意义上的“终身监禁（life imprisonment）”，因此在保留上述翻译的同时，应把“有期徒刑”翻译为 imprisonment with work for a definite term，把“无期徒刑”翻译为 imprisonment with work for a indefinite term（屈文原文如此），这样就可避免上述问题[11]。

对“管制”的处理，ChatGPT o1 译为 control，其余三者译为 public surveillance。首先，control 和“管制”的意思相去甚远，“管制”不是单纯的“控制”，再说 surveillance，Black's Law Dictionary (9th ed.)对 surveillance 的解释是：close observation or listening of a person or place in the hope of gathering evidence[12]，人民法院对罪犯判处管制，并非处于“收集证据”的目的，因此也不能将“管制”译为 public surveillance。根据刑法第三十八条的规定，“对判处管制的犯罪分子，依法实行社区矫正”，“同时禁止犯罪分子在执行期间从事特定活动，进入特定区域、场所，接触特定的人”，因此比较合适的英译为 non-custodial correction。

五、挑战与风险

（一）意识形态风险

人工智能推理模型普遍依赖大规模语料库进行训练，这些语料库多来源于英语语境中的主流媒体、法律文献、社交网络和互联网文本，其语料的选择、组织和标注过程不可避免地蕴含意识形态偏向。当此类模型被应用于中国法律文本的英译实践时，存在“默认”采用西方语言表达方式、价值取向甚至法律逻辑的风险。这种潜在的意识形态渗透可能在无形中削弱中国法律话语的主体性表达，甚至对我国法治理念的国际传播产生误导。具体而言，在刑法总则等高度政治化和制度化的文本翻译中，众多具有中国特色的法律

政治术语，极易在推理模型的处理过程中被“中和”或“转化”为符合西方自由主义话语体系的词汇体系，从而导致法律语义的偏移乃至误读。此外，由于训练语料中可能包含意识形态偏见、文化刻板印象以及西方中心主义叙事框架，人工智能推理模型在输出翻译时，也可能将中国法律制度中的独特构造错误类比为西方法律结构，形成文化误译或制度误译。这一风险不仅关涉法律语言的准确传达问题，更涉及国家话语权与文化主权在跨语言传播中的维护问题。因此，在人工智能推理模型参与法律文本翻译的过程中，应强化对意识形态风险的识别与应对机制。例如，可通过构建带有中国特色法学术语的专属法律语料库、开发融入中国法治价值的定向微调模型，或引入人工译者进行意识形态敏感词汇的审校，从而有效防范模型输出中潜藏的意识形态偏差，确保法律文本翻译既符合目标语言表达习惯，又忠实于源语言的制度语义与文化内涵。

（二）准确性挑战

法律语言以其高度的规范性、严密性和精确性著称，任何翻译偏差都可能引发语义歧义、法律适用错误乃至制度性误解。人工智能推理模型虽然在自然语言生成方面取得了显著进展，但其在法律文本翻译中的准确性仍面临诸多挑战，尤其是在术语处理、句法结构、逻辑推理和法律概念对等方面暴露出严重不足。首先，法律术语的特殊性决定了其不可随意替代。然而，在实际翻译过程中，部分模型将“有期徒刑”译为 fixed-term imprisonment 或将“管制”译为 control 与 public surveillance，忽略了术语背后的制度含义与文化内涵。这些翻译虽然在字面上看似合理，但在法律体系的语境中却可能造成概念错置或法律效力的误判。其次，复杂句法结构的处理能力不足亦是当前推理模型准确性不足的重要表现。如罪责刑相适应原则等含有多重并列与逻辑关系的条文，要求译文既要保持逻辑清晰，又要符合法律英语的表达习惯。然而，多数推理模型倾向于逐字直译，忽视了英汉法律语言在表达方式上的形合与意合差异，导致句意模糊和逻辑割裂，甚至结构紊乱。此外，推理模型在处理法律中的隐含推论与制度性前提时亦存在短板。法律文本往往依赖高度凝练的表达方式传递复杂的法律逻辑与权利义务关系，要求译者具备相应的法学知识结构以进行合理还原。而当前主流模型在缺乏深层次法律知识支撑的情况下，容易对概念间的逻辑关系做出错误推断，进而影响整体语义的准确传达。

六、结语

随着人工智能技术的迅猛发展，推理模型作为机器翻译新兴工具，正逐步渗透至法律语言这一高度专业化的语域。本文以《中华人民共和国刑法》总则为语例，选取四种主流人工智能推理模型进行对比研究，通过 BLEU 量化指标和多维度翻译表现分析，系统评估了推理模型在法律文本翻译中的表现差异与局限，旨在为人工智能翻译技术在法律领域的应用提供实践参考。

研究结果表明，尽管人工智能推理模型在语法结构处理、法律术语识别等方面已表现出初步能力，其中 Grok 3 与 Gemini 2.0 在 BLEU 指标中相对优越，但整体仍未达到替代人工译者的质量要求，尤其在处理复杂法律句式、专业术语、制度内涵及意识形态负载表达方面存在明显不足。此外，模型间在翻译稳定性和术语准确性等方面差异显著，提示未来的模型设计应更加重视语言风格、专业知识与语境敏感性的协同嵌入。

从宏观层面来看，人工智能推理模型在赋能法律翻译的同时，也带来了意识形态输出、语义误读和责任归属模糊等风险。因此，在拥抱技术进步的同时，亟需构建人机协同的翻译机制，发挥人工译者在语义阐释与文化中介中的不可替代作用。此外，应推动建立覆盖法律语料预处理、术语标准化、模型微调及译后编辑等全流程的专业化翻译生态，以此提高人工智能翻译系统在法律语境下的适配度和可信度。

未来研究可在以下几个方向展开：一是拓展研究对象至其他部门法律以验证推理模型的通用性与适应性；二是引入人工评价量表与用户视角的接受度研究，补充 BLEU 等量化指标的局限；三是关注模型语料构建过程中的文化立场与价值选择，推进具有中国法治话语表达能力的本土化法律翻译模型研发。

总之，人工智能推理模型在法律文本翻译中的应用尚处于探索阶段，其发展潜力与现实挑战并存。唯有在技术迭代与法律语言研究的双重推动下，才能真正实现智能翻译与准确传播的深度融合，更好地服务国家法律对外传播战略与国际法治话语体系建设。

参考文献：

[1]胡开宝. 国家外宣翻译能力：构成、现状与未来[J]. 上海翻译, 2023(4): 1-7+95.
[2]文旭,田亚灵. ChatGPT 应用于中国特色话语翻译的有效性研究[J]. 上海翻译, 2024(2): 27-34+94-95.
[3]宋丽珏.法律翻译的数字人文转型研究——以专题数据库和 ChatGPT 为中心[J]. 外语学刊, 2024(2): 51-57.

[4]冯志伟,张灯柯. 机器翻译与人工翻译相辅相成[J]. 外国语(上海外国语大学学报), 2022, 45(6): 77-87.
[5]耿芳,胡健. 人工智能辅助译后编辑新方向——基于 ChatGPT 的翻译实例研究[J]. 中国外语, 2023, 20(3): 41-47.
[6]于浩,郭赟赟.风险与超越: ChatGPT 赋能翻译的伦理分析[J]. 中国翻译, 2024, 45(4): 115-122.
[7]Papineni, K., Roukos, S., Ward, T., *et al*. BLEU：A method for automatic evaluation of Machine[A]In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*[C]. Philadelphia, 2002: 311-318.
[8]李洛克. BLEU 算法（例子和公式解释）[EB/OL].(2020-2-8)[2025-2-1]. https://blog.csdn.net/qq_30232405/article/details/104219396
[9]周成彬,刘忠宝. 基于语义信息共享 Transformer 的古文机器翻译方法[J]. 情报工程, 2022, 8(6): 114-127.
[10]]滕超, 孔飞燕. 英汉法律互译：理论与实践[M]. 杭州：浙江大学出版社, 2008.
[11]屈文生. 中国法律术语对外翻译面临的问题与成因反思——兼谈近年来我国法律术语译名规范化问题[J]. 中国翻译, 2012, 33(6): 68-75.
[12]Bryan A. Garner. *Black's Law Dictionary* (9th ed.) [Z]. St. Paul: West Publishing Co., 2009.

A Study on the Effectiveness of Applying Artificial Intelligence Reasoning Models to Legal Text Translation

Wei Rong[1]

1.School of Foreign Languages, Southwest University of Political Science and Law, Chongqing

Abstract: With the continuous advancement of artificial intelligence technologies, reasoning models have gradually been introduced into the practice of legal translation. However, systematic empirical research in this field remains limited, particularly in terms of quantitative analysis and comparative evaluation of model outputs. This study takes the *General Principles of the Criminal Law of the People's Republic of China* as a case corpus and selects four mainstream AI reasoning models—ChatGPT o1, DeepSeek R1, Grok 3 with Think, and Gemini 2.0 Flash Thinking Experimental—to generate translations for evaluation. BLEU scores were calculated and analyzed using Python and SPSS, incorporating statistical testing and multidimensional comparison. The results reveal significant differences in the performance of the four models in legal translation. Grok 3 and Gemini 2.0 outperform the others in terms of average BLEU scores, frequency of high-quality translations, and lower incidence of poor scores, while ChatGPT o1 and DeepSeek R1 show greater room for improvement in handling legal terminology, syntactic structures, and output stability. Further analysis indicates that current reasoning models still face challenges in accurately translating legal terms, complex sentence structures, and legal logic. Although AI reasoning models demonstrate potential in empowering legal translation, they remain unable to replace human translators in the deep comprehension and nuanced expression required in specialized legal discourse. To enhance translation quality, this study recommends strengthening domestic legal corpora, improving terminology consistency mechanisms, incorporating post-editing by human translators, and developing localized legal translation models with Chinese characteristics.

Keywords: AI reasoning models; legal translation; BLEU metric; translation quality evaluation