估计机器翻译难度

Lorenzo Proietti[1,*] **Stefano Perrella**[1,*] **Vilém Zouhar**[2,*]
Roberto Navigli[1] **Tom Kocmi**[3]
[1]Sapienza NLP Group, Sapienza University of Rome [2]ETH Zurich [3] Cohere
{ lproietti,perrella,navigli } @diag.uniroma1.it vzouhar@ethz.ch kocmi@cohere.com

Abstract

机器翻译的质量在某些设置中已经达到了近乎完美的翻译。这些高质量的输出使得区分最先进的模型变得困难，并难以识别未来改进的领域。自动识别机器翻译系统困难的文本在发展更具判别力的评估和指导未来研究方面具有潜力。我们形式化了翻译难度估计的任务，基于翻译的预期质量定义文本的难度。我们引入了一种新的指标来评估难度估计器，并使用它来评估基本线和新方法。最后，我们通过使用难度估计器构建更具挑战性的机器翻译基准来展示其实用价值。我们的结果表明，专门的模型（称为 Sentinel-src ）优于基于启发式的方法（例如，词汇稀有度或句法复杂度）和以 LLM 为评判的方式。我们发布了两个改进的难度估计模型，Sentinel-src-24 和 Sentinel-src-25 ，可以用于扫描大型文本集合，并选择最有可能挑战当代机器翻译系统的文本。

1 介绍

并不是所有的数据样本都是平等的：机器学习模型可能在处理某些样本时比其他样本更困难。能够自动评估样本的难度在模型开发的各个阶段都是不可或缺的。例如，在训练期间，从最容易到最难对样本进行组织，称为课程学习，可以提高性能和训练效率 (Bengio et al., 2009; Wang et al., 2022; Soviany et al., 2022) 。即使在推理阶段，通过对容易的样本提前退出，计算成本也可以减少 (Teerapittayanon et al., 2016; Schwartz et al., 2020) 。

评估模型时也需要对样本难度进行估计，因为过于简单或过于困难的基准无法有效区分模型 (Lalor et al., 2018; Rodriguez et al., 2021) 。这个问题在机器翻译（MT）中尤为重要，因为最近的最先进模型已经获得了近乎完美的分数，表现接近人类水平 (Kocmi et al., 2024a) 。在简单的测试集中，实践者可能难以区分顶级表现的模型，并评估是否还有进一步改进模型的空间。此外，尽管 WMT 的 MT 测试套件子任务针对特定的复杂翻译现象，但尚未系统地调查广义翻译难度的概念。

为了解决这一空白，我们在机器翻译中探讨样本难度的概念。首先，我们提出了翻译难度的定义，并正式介绍将翻译难度估计作为一项新任务，其中自动预测源文本的难度。然后，我们提出了难度估计相关性（DEC），这是一种旨在评估难度估计方法性能的度量。最后，我们测试了基线和新提出的难度估计方法，并验证了它们在下游应用中创建一个具有挑战性的基准的实际效用，这涉及从大型语料库中自动选择具有挑战性的样本子集。

我们发现，诸如词汇稀有度、句法复杂性，甚至使用大型语言模型作为判定者的方法在捕捉翻译难度方面表现不如专门的解决方案。具体来说，我们发现 Sentinel-src ——一种仅基于源文本本身来预测给定文本的预期翻译质量的模型——在评估翻译难度时优于其他方法。因此，我们训练了两个改进版本，称为 Sentinel-src-24 和 Sentinel-src-25 ，并公开发布了它们。[*]

2 相关工作

以往的工作可以根据其侧重点是人类翻译还是机器翻译难度分为两类。

人类翻译难度。 最早的研究 (Fang, 1959; Hale and Campbell, 2002) 尝试将一般文本复杂性与人类翻译的困难程度联系起来。Mishra et al. (2013) 的一个更现代的调查将人类翻译困难框定为翻译一个句子所需的时间，并使用译者眼动数据来估计。他们将文本长度、词汇多义程度和句法复杂性作为翻译困难的预测因素。Vanroy et al. (2019) 研究了错误数量、单词翻译熵和句法等效性与翻译时长、注视以及其他人类翻译困难的代理之间的相关性。更近些时候，Lim et al. (2023, 2024) 使用单词对齐分布和解

[*]Equal contribution.

[*]模 型：hf.co/collections/Prosho/translation-difficulty-estimators-6816665c008e1d22426eb6c4
代码：github.com/zouharvi/translation-difficulty-estimation

码器困惑度来预测人类翻译困难。

机器翻译的困难。 为了实现一个课程学习训练计划，Kocmi and Bojar (2017) 基于句子长度、词汇稀有度和文本中的协调连接词数量来估计样本的难度。同样地，Platanios et al. (2019) 使用句子长度和稀有度作为难度的替代指标。除了这些语言上有动机的标准外，Zhang et al. (2018) 和 Liu et al. (2020) 使用翻译模型在生成文本时的置信度和其他内在属性来预测翻译难度。Almeida (2017) 将难度估计视为一个二元分类任务，但也使用来自目标文本的特征，使其更接近于质量估计。Zhan et al. (2021b) 使用带有自动化指标的人造群体，并发现长段落、低频词汇和专有名词是机器翻译中最具挑战性的部分。最后，Zhan et al. (2021a) 使用文本中的标记与其翻译之间的嵌入相似性来估计文本的难度。

我们的工作接近的是，Don-Yehiya et al. (2022) 将 PreQuEL 任务定义为在翻译生成之前预测给定文本翻译的质量。然而，他们采用了 WMT 2020 质量估计共享任务 (Specia et al., 2020) 的评估，该任务设计用于质量估计而非评估难度估计器。此外，他们的测试集仅包含两个语种方向，所有翻译均由同一个机器翻译模型生成。另外，他们没有探索更广泛的难度估计器空间或研究其在构建挑战性基准测试中的应用。

相反，我们将翻译难度估计定义为一个独立的任务，并具有专门的评估指标。此外，我们使用跨越 11 语言方向的测试集，对多种难度估计方法进行了基准测试，每个语言对的每个片段有 11 到 19 的翻译，这些翻译由机器翻译模型和人工译者共同生成。结果，我们的工作构成了首次对翻译难度估计器的广泛评估，为这一任务建立了新的技术水平。

3 难度估计任务

翻译给定文本的难度可能取决于多个因素。例如，一个文本可能由于其长度、句法复杂性、惯用语言或罕见或专业词汇的存在而具有挑战性。影响翻译难度的某些方面甚至可能取决于翻译方向，这意味着相同的源文本可能翻译成某一语言比翻译成另一语言更困难。此外，翻译难度可能在不同译者之间并不统一。在人工翻译的情况下，这可以随着译者的文化背景和语言熟悉度而变化；在机器翻译模型的情况下，则可能基于参数数量、训练数据和模型结构等因素而变化。

鉴于这些考虑，我们避免将翻译难度定义为绝对术语，因为这样的定义可能不易广泛适用。相反，我们定义难度是相对于给定的目标语言和特定译者的准确性而言的，无论是人还是机器。更具体地说，给定一个文本 x，一个译者 m，以及一个目标语言 l，我们分配一个难度分数 $d_{m,l}(x)$，等于 m 对翻译成语言 l 的 x 的质量分数。较低的分数表明翻译质量较低，因此，与源文本相关的难度更大。

例如，假设我们有两个文本，x_1 和 x_2，它们各自的翻译是 t_1 和 t_2，目标语言为 l，这两个翻译都是由模型 m 生成的。一个人工评价者在 1 到 100 的评分范围内对这些翻译进行评估，给第一个赋予了 60 的分数，给第二个赋予了 90 的分数。然后，$d_{m,l}(x_1) = 60$ 和 $d_{m,l}(x_2) = 90$。由于 $d_{m,l}(x_1) < d_{m,l}(x_2)$，因此对于译者 m 而言，翻译 x_1 成 l 比 x_2 更困难。重要的是，分数 d 越低，难度越高，反之亦然。

任务定义 给定一个源文本 x、一个翻译器 m 和一种目标语言 l，难度估计是自动预测 $d_{m,l}(x)$ 的任务。与质量估计不同，难度估计模型无法访问被估计质量的翻译。事实上，难度估计可以视为估计给定文本翻译预期质量的任务。

我们根据它们按难度对文本进行排序的能力来评估难度估计方法。假设有一个文本的集合 $\mathcal{X} = x_1, x_2, \ldots, x_N$，一组目标语言 $\mathcal{L} = l_1, l_2, \ldots, l_L$，以及一个将其翻译成语言 l 的模型集合 [*]：$\mathcal{M}_l = m_1, m_2, \ldots, m_{M_l}$。我们也定义模型 m 和语言 l 的真实难度分数向量为 $D_{m,l} = d_{m,l}(x_1), d_{m,l}(x_2), \ldots, d_{m,l}(x_N)$，相应的一个难度估计方法的预测值为 $\hat{D}_{m,l} = \hat{d}_{m,l}(x_1), \hat{d}_{m,l}(x_2), \ldots, \hat{d}_{m,l}(x_n)$。我们通过在译者和语言间平均 Kendall 等级相关系数 τ_b 来测量翻译难度估计相关性（DEC）：我们建议读者参考 Appendix A 以获取关于如何计算 Kendall 相关系数 τ_b 的更多细节。

对比 DEC 与标准 MT 元评估策略。 在 **??** 中使用的评估方法，Deutsch et al. (2023) 将其称为“Group-by-System”，使得 DEC 从根本上不同于其他用于机器翻译评估和质量估计的元评估策略，这些策略通常依赖于 Group-by-Item 方法来代替 (Deutsch et al., 2023)。Group-by-Item 计算分配给同一源文本的不同翻译（即，同一评估项）的评估之间的相关性，然后在所有源文本中平均这些相关性。这种方法的主要好处是它减少了源文本特征（例如长度）与翻译质量判断之间的虚假相关性 (Perrella et al., 2024)。

然而，我们认为这些特征恰恰定义了文本的

[*]为简洁起见，我们使用“模型”来指代人工翻译者和自动模型。

翻译难度。由于我们的目标是衡量翻译难度，我们定义了 DEC，使用按系统分组的方法。该方法将人的评估和指标评估与同一机器翻译系统产生的不同源文本翻译相关联，然后在机器翻译系统之间平均相关性。通过保持机器翻译系统不变，该评估直接衡量指标识别哪个源文本对该系统翻译更具挑战性的能力。

这种区别至关重要。事实上，使用按系统分组的方法来评估标准机器翻译或质量估计指标，会偏向于那些预测源文本难度而不是纯粹翻译质量的指标。相反，使用按项目分组的方法来评价难度评估器是不适当的，因为这些评估器会为一个给定源文本的所有翻译分配相同的分数。

4 难度估计的方法

在本节中，我们描述了几种难度估计方法。我们包括了常见和新颖的方法。具体来说，我们将难度估计器分为四组：基于启发式的、学习型的、LLM 作为判决者的和人工人群基础的。有关所考虑模型的实现细节，请参见 Appendix C。

我们将估计量称为基于启发式的，如果它们依赖于简单的文本特征。这个类别包括以前被证明与其他难度测量相关的估计量。

4.1 学习估计器

学习的机器翻译指标通常被训练来预测给定其源文本和（可选）参考翻译 (Rei et al., 2020; Guerreiro et al., 2024b; Juraska et al., 2023, 2024) 的翻译质量。同样，也可以训练神经模型来预测文本的难度。先前的研究已经探索了为相关目的训练类似模型的可能性：

- PreCOMET 是一套基于 XLM-RoBERTa (Conneau et al., 2020) 的源代码回归工具，用于预测一个样本在评估 (Zouhar et al., 2025b) 中的有用性。PreCOMET $_{\text{diversity}}$ 优先考虑可能引发多样机器翻译输出的样本，而 PreCOMET $_{\text{difficulty}}$ 则根据项目反应理论 (Santor and Ramsay, 1998) 估计难度。
- Sentinel-src 度量是基于 XLM-RoBERTa 的模型，旨在预测文本翻译的质量，而无需将翻译作为输入。目标是学习源文本与翻译质量分数 (Perrella et al., 2024) 之间的虚假相关性。

4.2 大语言模型作为法官

LLM 作为裁判的方法在多个应用中得到了广泛采用 (Zheng et al., 2023; Bavaresco et al., 2024)。在这项工作中，我们调查了使用 GPT-4o (OpenAI et al., 2024) 和 CommandA (Cohere et al., 2025) 的 LLM 作为裁判范式在难度估计任务中的有效性。我们提示这些模型确定翻译给定文本所需的熟练程度，选择性地提供有关目标语言的信息，并返回一个介于 0 到 120 之间的标量分数，表明给定文本的难度级别。请参见 Example 2 中的提示。

4.3 基于人群的估计器

迄今讨论的方法均仅基于源文本估计翻译难度，并可选择地考虑目标语言。然而，既然已将翻译难度定义为模型翻译的预期质量（第 3 节），我们现在引入更接近于模拟此定义的难度估计器。

人工人群 人工基于群众的方法首先翻译源文本，然后使用无参考的机器翻译指标来评估翻译结果的质量。[*] 具体来说，我们使用一组多样化的模型从测试集翻译源文本，以确保架构和规模的多样性：三个基于大语言模型的机器翻译模型（Gemma-3-27B-IT, Qwen2.5-72B-IT, CommandA）和一个标准的编码器-解码器机器翻译模型（NLLB-moe-54B）。对于评估步骤，我们采用两个最先进的无参考机器翻译指标：XCOMET-QE-XXL (Guerreiro et al., 2024a) 和 MetricX-24-Hybrid-QE-XXL (Juraska et al., 2024)，在后面分别称为 XCOMET 和 MetricX。每个源文本的最终难度分数是其中一种指标分配给其翻译的平均质量分数。这种方法是受 Zouhar et al. (2025b) 提出的高效子集选择人工群众方法的启发。

为了为我们的人工基于人群的估计器建立性能上限，我们还定义了真实人群估计器。它们使用 XCOMET 或 MetricX 对我们测量难度的实际系统的翻译进行评分，也就是说，这些系统的翻译是由人类标注质量的。因此，真实人群方法等同于标准的质量估计，其性能完全取决于无参考指标的准确性。

5 实验

我们使用难度估计相关性测量（DEC，方程 **??**）来对估计器进行基准测试。

5.1 实验设置

我们测量 DEC 在 WMT 2024 通用 MT 和指标共享任务中发布的测试集上的表现 (Kocmi et al., 2024a; Freitag et al., 2024)。这些测试集包括由自动模型和人类译员翻译为多种语言的源文本选择。每个翻译都附有由人工注释者根据错误范围注释（ESA，Kocmi et al., 2024b）或多维质量指标（MQM，Lommel et al., 2014;

[*] 无参考机器翻译指标通过仅将译文与其源文本进行比较来评估译文质量，而不需要参考译文。

	Method	Trans.	Lang.	DEC
Oracle	Oracle	✓	✓	1.000
	Oracle	✗	✓	0.301
	Oracle	✗	✗	0.224
Heuristic	Text Length	✗	✗	0.121
	Syntactic Complexity	✗	✗	0.080
	Word Rarity	✗	✗	−0.040
Learned	Sentinel-src-24	✗	✗	0.182
	Sentinel-src	✗	✗	0.175
	PreCOMET Difficulty	✗	✗	0.153
	PreCOMET Diversity	✗	✗	0.142
LLM Judge	Command A	✗	✗	0.072
	Command A	✗	✓	0.104
	GPT-4o	✗	✗	0.077
	GPT-4o	✗	✓	0.080
Crowd Based	True (XCOMET)	✓	✓	0.221
	True (MetricX)	✓	✓	0.207
	Artificial (XCOMET)	✗	✓	0.177
	Artificial (MetricX)	✗	✓	0.166
	Random	✗	✗	0.003

Table 1: 每种方法获得的难度估计相关性 (DEC)。我们根据方法访问的信息类型对它们进行分类。仅依赖文本的估计器，例如启发式和学习方法，完全依赖于正在估计难度的源文本。相反，有些方法还结合了目标语言（Lang.）或产生翻译的具体译者（Trans.）的信息。

Freitag et al., 2021 ）注释协议制作的质量注释。这里我们报告使用 ESA 注释协议的结果。请参见附录 Table 6 以了解使用 MQM 协议的结果，以及附录 Tables 4 and 5 以了解数据统计。

我们测试了 Section 4 中列出的所有方法。此外，我们通过扩展 Perrella et al. (2024) 使用的训练数据来改进表现最好的学习估计器 Sentinel-src ，并训练了两个新模型，称为 Sentinel-src-24 和 Sentinel-src-25 。前者使用直到 WMT 2023 的以往 WMT 版本的数据进行训练，而后者还包括 WMT 24 测试集。有关 Sentinel-src-24 和 Sentinel-src-25 的训练流程和参数的更多细节，请参见 Appendix B 。

5.2 结果

我们在 Table 1 中展示了结果，方法按其在 Section 4 中的类别进行分组。我们还在每个方法上标注了它使用的信息（例如模型或语言），详细信息见 Table 1 的标题，并包括三个不同的 oracle，以便为读者提供性能上限值。oracle 的定义可以在 Appendix E 中找到。

基于启发式和学习的方法。 这些估计器仅基于输入文本进行预测。因此，他们为每个文本分配的难度评分在所有目标语言和模型中都是相同的。[*] 在这一组中，所有学习的估计器都优于基于启发式的估计器。此外，Sentinel-src-24 实现了最高的难度估计相关性，也高于从 Perrella et al. (2024) 的 Sentinel-src ， 这突出了我们重新训练的有效性。

LLM 评判者可以选择提供目标语言。对于这两个模型，目标语言信息可以提高性能。在 CommandA 中尤其如此，目标语言信息导致相关性增加 0.032 点。然而，总体来看，LLM 评判者的表现很差，得分甚至比更简单得多的文本长度启发式方法还低。

正如预期的那样，利用真实翻译结果的 True Crowd 方法作为上界，产生了最高的相关性。由于它们的性能完全取决于所使用的无参考指标，这一结果也表明，XCOMET 在此任务中明显优于 MetricX。

相反，人工众包方法的表现与 Sentinel-src-24 相当。然而，人工众包方法比学习方法消耗更多的资源，因为它们既需要翻译源文本，又需要后续的质量评估步骤。

通过我们的评估，我们发现：

- 基于启发式的估计器，曾在以往的研究中被广泛使用，但大多数其他方法的表现更佳。
- 大多数方法，包括更简单的文本长度启发式方法，都优于 LLM-as-a-Judge 方法。
- 学习方法，即经过明确训练来预测给定文本难度的模型，表现优于大多数其他方法。
- Sentinel-src-24 是一种最先进的难度估算器，仅与需要更大计算努力的人造群体估算器相匹配。

根据这些结果，下一部分将研究每个类别中表现最好的估计器在创建困难基准的下游任务中的能力。为了确保公平，我们排除了一些基于人工群体的估计器。这些方法使用由多个 MT 模型生成的翻译来估计文本难度，这会使结果测试集对使用这些过程的 MT 模型产生不公平的偏倚。具体而言，我们仅包括表现最好的外部人工群体估计器，假设生成的测试数据将用于评估那些不参与构建过程的模型。

5.3 比较人类翻译和机器翻译的难度

在本节中，我们研究模型难以翻译的文本是否对人类来说也具有挑战性。为此，我们使用分配给 WMT 24 测试集中的每个源文本的难度分数 $d_{m,l}$ ， 在人类译员和机器翻译模型之

[*] 即，$\forall m_1, m_2 \in \mathcal{M}, l_1, l_2 \in \mathcal{L}: \hat{d}_{m_1,l_1}(x) = \hat{d}_{m_2,l_2}(x)$ 。

	IOL	GPT-4	Claude3.5	Tower70B
Human	0.137	0.137	0.127	0.109
Tower70B	0.176	0.158	0.151	
Claude3.5	0.178	0.221		
GPT-4	0.202			

Table 2: 针对人工以及四个机器翻译模型的语言方向平均肯德尔 τ_b 相关矩阵。

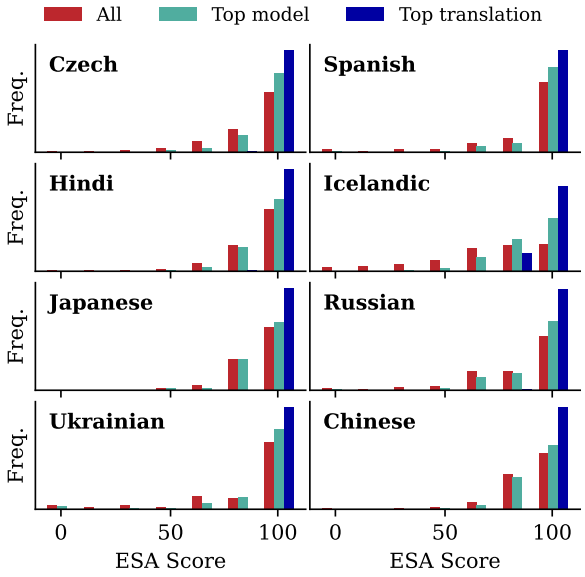

Figure 1: 分配给包含在 WMT 2024 测试集 (Kocmi et al., 2024a) 中的文本翻译的人类评分分布。我们报告所有模型的评分（ALL）、每种语言表现最佳的模型评分（Top system）以及每个输入文本最好翻译的评分（Top translation）。所选的分箱宽度为 15 ESA 点。

间变化 m 。我们测量所有译者对之间分数的肯德尔 τ_b ，并在所有语言方向上平均。为了保持一致性，我们将分析限制在对所有语言方向都有注释译文的模型和人类译者上，即一名人类译者和以下四个模型：Unbabel-Tower70B (Alves et al., 2024) 、IOL-Research (Zhang, 2024) 、Claude-3.5、和 GPT-4 (OpenAI et al., 2024) 。

Table 2 的结果显示，与人类译者的相关性始终低于机器翻译模型之间的相关性（从 0.109 到 0.137 ），而机器翻译模型之间的相关性则在 0.151 到 0.221 之间。这表明人类译者对翻译难度的感知可能与自动模型不同。值得注意的是，GPT-4 和 Claude-3.5 之间的最高一致性，可能是因为这两个模型都是通用大语言模型（LLMs），而不像 Unbabel-Tower70B 和 IOL-Research，这些模型是专门为机器翻译而训练的。

6 创建困难的基准测试

在本节中，我们应用表现最佳的难度估计器来创建困难的机器翻译基准。首先，我们表明在 WMT 24 通用 MT 共享任务中使用的测试集 (Kocmi et al., 2023) 对当前的 MT 模型可能过于简单。然后，我们定义从给定数据集中选择困难样本子集的任务，并评估我们的估计器。

在 Figure 1 中，我们展示了 2024 年 WMT 一般机器翻译共享任务 (Kocmi et al., 2024a) 的译文由人工标注人员分配评分的分布情况。值得注意的是，几乎每种语言的最佳模型总是能够获得 90 到 100 的 ESA 分数，并且每个输入文本至少有一个系统也能达到这一分数。这对英译西的翻译方向尤其令人担忧，因为最优系统几乎没有出现错误。这些发现强调了有必要创建比随机抽样文本更难的基准。相比之下，本节的其余部分使用难度估计指标从更大的集合中抽样出难度较大的文本。

6.1 设置

给定一个大型源文本集 $\mathcal{X}$ ，我们的目标是从中提取出一个最大难度的子集 $\mathcal{X}' \subseteq \mathcal{X}$ ，其大小为 $|\mathcal{X}'| = B$ 。实际上，我们选择了 B 篇文本，其难度是由表现最好的难度估算器确定的最高难度。根据 Section 5.2 ，我们排除了内部的人工集群估算器，因为它们会不公平地偏向被评估的机器翻译模型。

与先前的实验一样，我们使用人工翻译质量注释来评估估算器的有效性。我们再次依靠 WMT 24 测试集，并将其用作 $\mathcal{X}$ （详细信息请参阅 Section 5 ）。具体而言，我们关注的是其英文源文本，这些文本被翻译成捷克语、西班牙语、印地语、冰岛语、日语、俄语、乌克兰语和中文，以及其捷克语源文本被翻译成乌克兰语。

我们强调，我们从 $\mathcal{X}$ 中选择 B 源文本。因此，无论目标语言是什么，任何 $\mathcal{X}' \subseteq \mathcal{X}$ 仅包含英语和捷克语源文本，而目标语言仅用于评估目的。

任务定义。 我们为每个样本 $x \in \mathcal{X}$ 指派一个难度分数 $\hat{d}(x)$ 。对于仅使用文本的评估方法，比如启发式和学习方法，这很简单，因为这些方法除了给定文本外不使用其他信息。对于基于外部人工群体的方法，我们为每个文本 x 指派其译文的平均质量分数，并对机器翻译模型和目标语言进行平均。然后，我们按如下方式提取 $\mathcal{X}'$ ：

$$\hat{\mathcal{X}}' = \arg\operatorname*{top\text{-}B}_{x \in \mathcal{X}} \hat{d}(x) \quad (1)$$

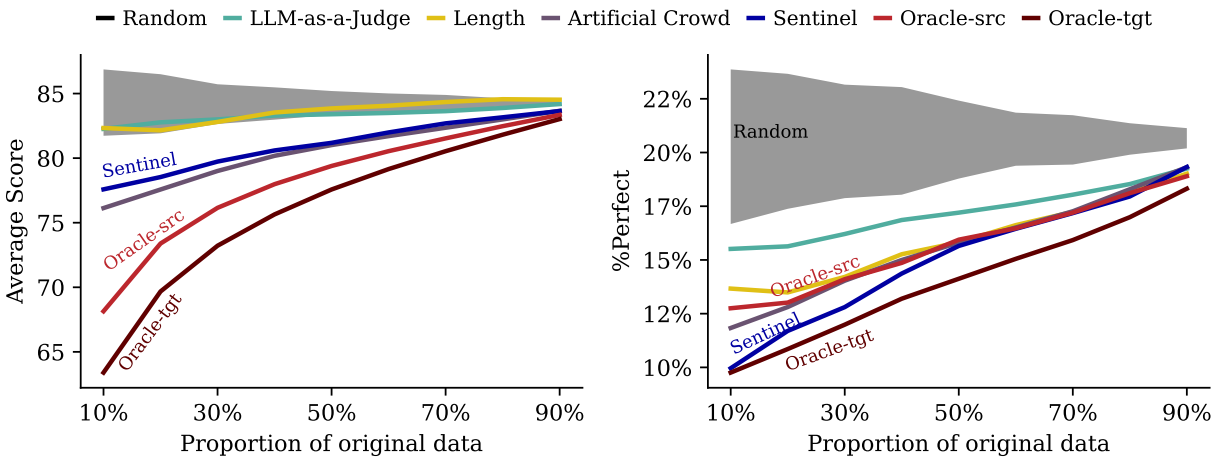

Figure 2: 在创建困难的测试集时，平均得分和“完美”源文本的比例。值越低越好。数据显示为 WMT24 中所有语言对的平均值，这些语言对经过子集选择的模拟。随机选择显示了 10 次运行中的 99 % 置信区间的 t 检验结果。

评估。 构建一个困难基准的目标之一是识别出当前模型仍然困难的样本，以揭示其缺点并指导未来迭代的改进。因此，我们评估了基于模型在测试集上的翻译所获得的平均人工评分下降的难度估计器的有效性。作为附加信息，我们还报告了在采样测试集中仍然保持的“完美”输出（那些从人工注释者那里获得满分 100/100 ESA 分的输出）的比例。这些度量的具体公式在 Appendix C 中提供。

我们通过改变子样本的大小提取了几个 $\mathcal{X}' \in \mathcal{X}$，并在 Figure 2 中报告了平均分数和% 完美度量的曲线。如预期的，Oracle 在可达到的最佳性能上设定了一个界限，虽然在% 完美度上，专用预测器 Sentinel 非常接近该效果。基于启发式的方法和基于 LLM-作为评判的的方法，其性能与随机子集选择非常接近。当用% 完美度进行衡量时，Oracle 表现较差，因为它优化的是最低平均得分而不是最低% 完美度的比例。在 Appendix F 中报告了 25 个% 测试集子样本情境下的更详细的定量结果，包括每个领域的细分回顾。回到原本关于现有测试集很简单（Figure 1）的抱怨，在附录 Figure 5 中，我们展示了在选择困难文本时分数分布的变化。

6.2 按难度选择的潜在陷阱

在实际操作中，除了随机抽样以外选择样本隐藏着意想不到的风险。例如，由难度估计器选择的文本可能在语法上不正确。我们现在研究在使用难度作为子抽样标准时可能遇到的潜在陷阱。具体来说，我们关注于：

- 源文本长度：难度估计器可能过于依赖文本长度。尽管相对于较短的文本，较长的文本更难翻译，但它们的翻译也在人力评估上更加昂贵。
- 源错误：翻译难以理解的源文本自然是困难的。然而，这些可能是不希望出现在测试集中的。
- 输出多样性：能够导致多样化机器翻译输出的源文本更为理想，因为它们有助于区分模型。有关实施细节，请参见 Appendix C。

	Source length	Source errors	Diversity embd	Diversity chrF	Unique outputs
Random	0.00	0.00	0.00	0.00	0.00
LLM-as-a-Judge	−0.61	0.26	0.19	0.23	−0.60
Length	−1.00	0.25	0.31	0.24	−0.52
Artificial Crowd	−0.63	0.04	−0.11	−0.17	−0.46
Sentinel	−0.66	0.12	−0.01	−0.09	−0.36
Oracle-src	−0.22	−0.16	−0.47	−0.49	−0.28
Oracle-tgt	−0.22	−0.16	−0.47	−0.49	−0.28

Table 3: 选择器与感兴趣变量（包括源文本长度、每词错误数量、通过两两嵌入内积和 chrF 度量的输出多样性以及唯一输出的比例）之间的皮尔逊相关性。对于所有方法，较低的数值表示源文本更困难，因此负相关意味着困难度与目标变量之间的正相关性更强。详见附录 Figure 6 中详细的可视化。

我们在 Table 3 和附录 Figure 6 中展示了难度估计值预测与这些变量之间的相关性。鉴于我们使用翻译质量来定义难度，因此较低的估计值分数表明较高的难度。

如预期所示，所有估计器都与源长度表现出很强的负相关关系，表明它们都倾向于选择更长的输出。相反，源错误似乎并非如此，这意

1 (difficult): City get a nice easy draw at home.

2 (difficult): Alex Bregman Predicted To Betray Astros, Sign With Shocking Blue Jays

3 (difficult): Some folks really do deserve a badge of honour for their pedantry (C8). Veronica Coyne of Springfield claims that "when bemoaning the loss of the express lane at Woolies "12 items or less," a friend told me she'd never used it on principle as it should have been "12 items or fewer.""

4 (easy): Washington

5 (easy): Developing the next generation of hybrid vehicles in Europe

6 (easy): We cannot allow this to happen. This legislation is enormously unpopular. It is exactly what the American people do not want. It must not be passed by Congress.

Example 1: Texts from the WMT24 dataset of English sources with Sentinel-src decisions on difficulty (easy=lowest difficulty, difficulty=highest difficulty in each length bin).

味着包含许多错误的文本似乎没有被难度估计器优先考虑。最后，我们的结果表明，所有难度估计器都会选择导致更具多样性和独特性输出的源文本。

原始数据集 $\mathcal{X}$ 来自不同领域和子领域。一个合理的担忧是按难度选择是否会减少子集的多样性，我们用嵌入和向量空间分布来衡量这一点。附录 Figure 3 显示，随着子集的增大，空间变得更加密集。自然地，随机选择是创建多样性子集的一个很好的基线。另一方面，所有其他难度估计方法，包括 Oracle，都会导致空间更加密集，意味着更低的多样性。如果测试集需要多样性，则需要采用明确的措施，如首先识别不同的群集，并在每个群集中通过难度估计进行选择。

在 WMT 2024 上评估难度估计方法存在一个局限性。这个测试集已经从一个更大的批次中进行了下采样，并经过人工审核，这使得现象分布具有人工特性。为了缓解这个问题，我们使用更大的数据批次进行分析，并在附录 Figure 4 中报告结果。

6.3 定性分析

之前的高级分析表明，Sentinel-src 是一个合适的选择器，其偏倚风险低于内部人工群体，并且优于其他方法。为了进一步验证该模型，我们手动检查了 200 个源文本，其中一半被 Sentinel-src 认为容易，另一半认为困难，并根据长度将其分为 10 个桶。一般来说，Sentinel-src 对于难度的概念与人类对翻译难度的判断非常一致。例如，困难的段落往往会使用习语表达（Example 1 .1）。不完整的句子段落，如标题也被认为具有高难度（Example 1 .2），这与人类的感知一致。对于多句段落中的间接言语（Example 1 .3）也是如此。另一方面，单词段落、简单句子结构，或只是多个简单句子连接在一起的段落，被认为是容易的（Example 1 .4 到 Example 1 .6）。

7 结论

在这项工作中，我们正式定义了翻译难度估计任务，并引入了一种专门的评估指标来评估难度估计器的表现。然后，我们对现有和新提出的估计器进行了综合评估，发现传统的基于启发式的方法不如其他方法。此外，我们发现作为评审员的大语言模型方法表现较差，被所有专门为此任务训练的方法所超越。其中，我们确定 Sentinel-src-24 是当前该任务的最先进方法，并推荐其用于下游应用。为此，我们在困难基准创建任务中检查了其与其他估计器的表现，显示这些模型在识别当前机器翻译模型表现不佳的困难样本方面效果显著。Sentinel-src-24 在这一下游应用中仍然是表现最好的方法。受此启发，我们通过在 Sentinel-src-24 的训练管道中加入额外数据开发了 Sentinel-src-25 ，并公开发布了这两个模型。最后，我们对 Sentinel-src-25 的预测进行了定性分析，提供了关于其认为困难的文本类型的直观见解。

8

局限性

翻译困难的概念。 本研究基于以下假设：我们可以通过该文本产生的翻译质量来代理其难度。虽然我们承认理想情况下翻译难度应是源文本的固有属性——不依赖于任何特定翻译模型——但这一工作假设符合我们的目的，特别是用于下游应用如创建具有挑战性的机器翻译基准。实际上，我们的研究目标是识别对当前机器翻译模型而言难以翻译的文本，而不是探讨翻译难度的抽象、独立于模型的概念。

正如在第 3 节中讨论的那样，给定文本的翻译难度可能取决于目标语言，这点在理论上被 Bugliarello et al. (2020) 所证实。我们认识到这一点只是简要提及，并未就此现象提供深入研究。然而，我们的实验支持这个假设：当模型被提供目标语言的信息时，作为评判者的大语言模型（LLM-as-a-Judge）的性能会有所提升。因此，我们鼓励未来的研究更加深入地探索目标语言对翻译难度的影响，并探讨如何有效地将此类信息纳入其他难度估计方法中。

伦理声明

我们预见我们的工作不会出现伦理问题。

References

Ana Sofia Vieira de Jesus Almeida. 2017. Difficulty estimation of machine translation. MSc Thesis.

Duarte Miguel Alves, José Pombal, Nuno M Guerreiro, Pedro Henrique Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and Andre Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. In First Conference on Language Modeling .

Sahar Araghi and Alfons Palangkaraya. 2024. The link between translation difficulty and the quality of machine translation: A literature review and empirical investigation. Language Resources and Evaluation , 58(4):1093–1114.

Marta Ba
textasciitilde nón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics , pages 4555–4567. Association for Computational Linguistics.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, André F. T. Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2024. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. Preprint , arXiv:2406.18403.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In Proceedings of the 26th Annual International Conference on Machine Learning , ICML '09, page 41–48, New York, NY, USA. Association for Computing Machinery.

Emanuele Bugliarello, Sabrina J. Mielke, Antonios Anastasopoulos, Ryan Cotterell, and Naoaki Okazaki. 2020. It's easier to translate out of English than into it: Measuring neural translation difficulty by cross-mutual information. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics , pages 1640–1649. Association for Computational Linguistics.

Cohere and 1 others. 2025. Command A: An enterprise-ready large language model. Preprint , arXiv:2504.00698.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics , pages 8440–8451. Association for Computational Linguistics.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2021. A statistical analysis of summarization evaluation metrics using resampling methods. Transactions of the Association for Computational Linguistics , 9:1132–1146.

Daniel Deutsch, George Foster, and Markus Freitag. 2023. Ties matter: Meta-evaluating modern metrics with pairwise accuracy and tie calibration. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing , pages 12914–12929. Association for Computational Linguistics.

Shachar Don-Yehiya, Leshem Choshen, and Omri Abend. 2022. PreQuEL: Quality estimation of machine translation outputs in advance. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing , pages 11170–11183. Association for Computational Linguistics.

Achilles Fang. 1959. Some reflections on the difficulty of translation. In On translation , pages 111–134. Harvard University Press.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. Transactions of the Association for Computational Linguistics , 9:1460–1474.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In Proceedings of the Ninth Conference on Machine Translation , pages 47–81. Association for Computational Linguistics.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In Proceedings of the Eighth Conference on Machine Translation , pages 578–628. Association for Computational Linguistics.

Gemma, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. Preprint , arXiv:2503.19786.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse , pages 33–41. Association for Computational Linguistics.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024a. xcomet: Transparent machine translation evaluation through fine-grained error detection. Transactions of the Association for Computational Linguistics , 12:979–995.

Nuno M Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André FT Martins. 2024b. xcomet: Transparent machine translation evaluation through fine-grained error detection. Transactions of the Association for Computational Linguistics , 12:979–995.

Sandra Hale and Stuart Campbell. 2002. The interaction between text difficulty and translation accuracy. Babel , 48(1):14–33.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In Proceedings of the Ninth Conference on Machine Translation , pages 492–504. Association for Computational Linguistics.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In Proceedings of the Eighth Conference on Machine Translation , pages 756–767. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, and 3 others. 2024a. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In Proceedings of the Ninth Conference on Machine Translation , pages 1–46. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, and 2 others. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In Proceedings of the Eighth Conference on Machine Translation , pages 1–42. Association for Computational Linguistics.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In Proceedings of the Seventh Conference on Machine Translation (WMT) , pages 1–45. Association for Computational Linguistics.

Tom Kocmi and Ondřej Bojar. 2017. Curriculum learning and minibatch bucketing in neural machine translation. In Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017 , pages 379–386. INCOMA Ltd.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024b. Error span annotation: A balanced approach for human evaluation of machine translation. In Proceedings of the Ninth Conference on Machine Translation , pages 1440–1453. Association for Computational Linguistics.

John P. Lalor, Hao Wu, Tsendsuren Munkhdalai, and Hong Yu. 2018. Understanding deep learning performance through an examination of test set difficulty: A psychometric case study. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing , pages 4711–4716. Association for Computational Linguistics.

Zheng Wei Lim, Trevor Cohn, Charles Kemp, and Ekaterina Vylomova. 2023. Predicting human translation difficulty using automatic word alignment. In Findings of the Association for Computational Linguistics: ACL 2023 , pages 11590–11601. Association for Computational Linguistics.

Zheng Wei Lim, Ekaterina Vylomova, Charles Kemp, and Trevor Cohn. 2024. Predicting human translation difficulty with neural machine translation. Transactions of the Association for Computational Linguistics , 12:1479–1496.

Xuebo Liu, Houtim Lai, Derek F. Wong, and Lidia S. Chao. 2020. Norm-based curriculum learning for neural machine translation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics , pages 427–436. Association for Computational Linguistics.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. Tradumàtica , 0(12):0455–463.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. 2024. tinyBenchmarks: Evaluating LLMs with fewer examples. In Proceedings of the 41st International Conference on Machine Learning , volume 235 of Proceedings of Machine Learning Research , pages 34303–34326. PMLR.

Abhijit Mishra, Pushpak Bhattacharyya, and Michael Carl. 2013. Automatically predicting sentence translation difficulty. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) , pages 346–351. Association for Computational Linguistics.

Jinjie Ni, Yifan Song, Deepanway Ghosal, Bo Li, David Junhao Zhang, Xiang Yue, Fuzhao Xue, Zian Zheng, Kaichen Zhang, Mahir Shah, Kabir Jain, Yang You, and Michael Shieh. 2024a. MixEval-X: Any-to-Any evaluations from real-world data mixtures. Preprint , arXiv:2410.13754.

Jinjie Ni, Fuzhao Xue, Xiang Yue, Yuntian Deng, Mahir Shah, Kabir Jain, Graham Neubig, and Yang You. 2024b. MixEval: Deriving wisdom of the crowd from LLM benchmark mixtures. Preprint , arXiv:2406.06565.

NLLB, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. Preprint , arXiv:2207.04672.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. GPT-4 technical report. Preprint , arXiv:2303.08774.

OpenAI and 1 others. 2024. GPT-4o system card. Preprint , arXiv:2410.21276.

Stefano Perrella, Lorenzo Proietti, Alessandro Scirè, Edoardo Barba, and Roberto Navigli. 2024. Guardians of the machine translation meta-evaluation: Sentinel metrics fall in! In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) , pages 16216–16244. Association for Computational Linguistics.

Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabas Poczos, and Tom Mitchell. 2019. Competence-based curriculum learning for neural machine translation. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) , pages 1162–1172. Association for Computational Linguistics.

Qwen, :, 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, and 25 others. 2025. Qwen2.5 technical report. Preprint , arXiv:2412.15115.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) , pages 2685–2702. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) , pages 3982–3992. Association for Computational Linguistics.

Parker Riley, Daniel Deutsch, George Foster, Viresh Ratnakar, Ali Dabirmoghaddam, and Markus Freitag. 2024. Finding replicable human evaluations via stable ranking probability. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) , pages 4908–4919. Association for Computational Linguistics.

Pedro Rodriguez, Joe Barrow, Alexander Miserlis Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. 2021. Evaluation examples are not equally informative: How should that change NLP leaderboards? In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) , pages 4486–4503. Association for Computational Linguistics.

Jie Ruan, Xiao Pu, Mingqi Gao, Xiaojun Wan, and Yuesheng Zhu. 2024. Better than random: Reliable NLG human evaluation with constrained active sampling. Proceedings of the AAAI Conference on Artificial Intelligence , 38(17):18915–18923.

Darcy A Santor and James O Ramsay. 1998. Progress in the technology of measurement: Applications of item response models. Psychological assessment , 10(4):345.

Roy Schwartz, Gabriel Stanovsky, Swabha Swayamdipta, Jesse Dodge, and Noah A. Smith.

2020. The right tool for the job: Matching model and instance complexities. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics , pages 6640–6651. Association for Computational Linguistics.

Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. 2022. Curriculum learning: A survey. Int. J. Comput. Vision , 130(6):1526–1565.

Lucia Specia, Zhenhao Li, Juan Pino, Vishrav Chaudhary, Francisco Guzmán, Graham Neubig, Nadir Durrani, Yonatan Belinkov, Philipp Koehn, Hassan Sajjad, Paul Michel, and Xian Li. 2020. Findings of the WMT 2020 shared task on machine translation robustness. In Proceedings of the Fifth Conference on Machine Translation , pages 76–91. Association for Computational Linguistics.

Robyn Speer. 2022. rspeer/wordfreq: V3.0.

Surat Teerapittayanon, Bradley McDanel, and H.T. Kung. 2016. Branchynet: Fast inference via early exiting from deep neural networks. In 2016 23rd International Conference on Pattern Recognition (ICPR) , pages 2464–2469.

Bram Vanroy, Orphee de clercq, and Lieve Macken. 2019. Correlating process and product data to get an insight into translation difficulty. Perspectives .

Xin Wang, Yudong Chen, and Wenwu Zhu. 2022. A survey on curriculum learning. IEEE Transactions on Pattern Analysis and Machine Intelligence , 44(9):4555–4576.

Guillaume Wenzek, Vishrav Chaudhary, Angela Fan, Sahir Gomez, Naman Goyal, Somya Jain, Douwe Kiela, Tristan Thrush, and Francisco Guzmán. 2021. Findings of the WMT 2021 shared task on large-scale multilingual machine translation. In Proceedings of the Sixth Conference on Machine Translation , pages 89–99. Association for Computational Linguistics.

Runzhe Zhan, Xuebo Liu, Derek F. Wong, and Lidia S. Chao. 2021a. Difficulty-aware machine translation evaluation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers) , pages 26–32. Association for Computational Linguistics.

Runzhe Zhan, Xuebo Liu, Derek F. Wong, and Lidia S. Chao. 2021b. Variance-aware machine translation test sets. In Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1) .

Wenbo Zhang. 2024. IOL research machine translation systems for WMT24 general machine translation shared task. In Proceedings of the Ninth Conference on Machine Translation , pages 147–154. Association for Computational Linguistics.

Xuan Zhang, Gaurav Kumar, Huda Khayrallah, Kenton Murray, Jeremy Gwinnup, Marianna J Martindale, Paul McNamee, Kevin Duh, and Marine Carpuat. 2018. An empirical exploration of curriculum learning for neural machine translation. Preprint , arXiv:1811.00739.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and chatbot arena. Preprint , arXiv:2306.05685.

Vilém Zouhar, Tom Kocmi, and Mrinmaya Sachan. 2025a. AI-assisted human evaluation of machine translation. In 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) , pages 4936–4950, Albuquerque, New Mexico. Association for Computational Linguistics.

Vilém Zouhar, Peng Cui, and Mrinmaya Sachan. 2025b. How to select datapoints for efficient human evaluation of NLG models? Preprint , arXiv:2501.18251.

	EN→DE	EN→ES	JA→ZH
# Source texts	486	622	559
# Translators	19	15	15

Table 4: WMT 2024 评测共享任务中发布的测试集统计数据 (Freitag et al., 2024) 。“# 源文本” 表示测试集中源文本的数量，“# 译者” 表示每个源文本可用的译文数量。

A 肯德尔 τ_b

Kendall 的 τ 变体 b 定义为：

$$\tau_b = \frac{C - D}{\sqrt{(C + D + T_h)(C + D + T_m)}} \quad (2)$$

其中，C 和 D 分别是 R 和 $\hat{R}$ 之间的一致对和不一致对的数量，而 T_h 和 T_g 分别是在假设排序和真实排序中持平的对的数量。具体而言，如果 $\text{sign}(r_i - r_j) = \text{sign}(\hat{r}_i, \hat{r}_j)$ ，则对 (i, j) 是一致的，如果 $\text{sign}(r_i - r_j) \neq \text{sign}(\hat{r}_i, \hat{r}_j)$ ，则是不一致的。最后，如果 $\hat{r}_i = \hat{r}_j$ 或 $r_i = r_j$ ，则在假设排序或真实排序中对 (i, j) 是持平的。

我们选择了肯德尔的 τ_b 而不是皮尔逊相关系数等替代方案，以避免受异常值的影响。τ_b 也是机器翻译元评估中的标准，因为它比斯皮尔曼相关系数更好地处理相同值的情况。

B 训练 Sentinel-MQM-24 和 Sentinel-MQM-25

我们新研发的难度估计模型，Sentinel-MQM-24 和 Sentinel-MQM-25，遵循了 Perrella et al. (2024) 引入的 Sentinel-MQM 模型所使用的相同架构和训练流程。两个模型都基于 XLM-RoBERTa 大型作为主编码器，然后是在 `[CLS]` 标记之上的多层前馈网络。它们被训练以最小化预测分数和人为标量分数之间的均方误差 (MSE)。

我们采用与 Sentinel-MQM 模型相同的两阶段训练方法。在第一阶段，模型在直接评估 (DA，Graham et al., 2013) 数据上进行训练。在第二阶段，模型在 MQM 注释上进行微调。我们的模型与 Sentinel-MQM 的关键区别在于每个阶段使用的训练数据。

- 阶段 1：DA 训练。对于 Sentinel-MQM-24，我们通过包括来自 WMT 21 (Wenzek et al., 2021) 的注释，以及来自 WMT 22 (Kocmi et al., 2022) 和 WMT 23 (Kocmi et al., 2023) 的 DA+SQM 注释，扩展了 Perrella et al. (2024) 使用的 DA 训练数据。Sentinel-MQM-25 进一步包括了来自 WMT 24 的 ESA 注释。
- 阶段 2：MQM 微调。在这个阶段中，我们通过添加来自 WMT 23 的 MQM 注释来扩展 MQM 训练集。与 Sentinel-MQM 训练过程不同，我们不对每个翻译的多个评分进行平均。相反，我们将所有可用的注释作为单独的训练实例，保留评分者之间的变异性。这同样适用于 WMT 20 和 WMT 22 的 MQM 数据集，它们每个翻译包括三个人工评分。类似于第一阶段的 DA 训练，在 Sentinel-MQM-25 模型的情况下，我们还包含来自 WMT 24 的 MQM 注释。

依据 Perrella et al. (2024) 的方法，我们将每个由源文本片段和其对应的人类评分组成的对视为一个独立的训练实例。由于人类评分是分配给单个翻译的，因此可能为同一源文本存在多个注释。我们不会以任何方式合并这些评分，而是将它们全部包含在 DA 和 MQM 阶段的训练数据中。训练超参数与 Perrella et al. (2024) 用于 Sentinel-MQM 的超参数匹配。所有模型都在一台 NVIDIA GeForce RTX 4090 GPU 上进行训练。估计的训练时间为第一个 (DA) 阶段约三个 GPU 小时，第二个 (MQM) 微调阶段约一个 GPU 小时。这些估计适用于 Sentinel-MQM-24 和 Sentinel-MQM-25。

C 实现细节

- 对于词语罕见性启发式，我们使用 `wordfreq` Python 库 (Speer, 2022) 来计算词频。
- 对于句法复杂性启发式和文本长度，我们使用 `spaCy` 获取依存树和相应的词元。
- 对于多语言嵌入，我们使用 sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 (Reimers and Gurevych, 2019)

对于人工人群，我们使用以下：

- NLLB-moe-54B：稀疏门控专家混合编码器-解码器翻译模型 (NLLB et al., 2022) 。
- Gemma-3-27B-IT：来自 Gemma 系列的多模态指令调整 LLM (Gemma et al., 2025) 。
- Qwen2.5-72B-IT：来自 Qwen2.5 系列的最大指令调整大型语言模型 (Qwen et al., 2025) 。
- CommandA：用于真实企业用例的 111 B 参数 LLM (Cohere et al., 2025) 。

	EN→ES	EN→HI	EN→IS	EN→JA	EN→RU	EN→UK	EN→ZH	EN→CS	CS→UK
# Source texts	634	634	634	634	634	634	634	634	1954
# Translators	14	11	11	13	14	11	13	16	12

Table 5: WMT 2024 通用机器翻译共享任务中发布的测试集统计数据 (Kocmi et al., 2024a) 。“# 源文本”表示测试集中的源文本数量，“# 译者”表示每个源文本的可用翻译数量。

对于 Section 6 的评估，我们使用平均模型得分和% 完美。对于 平均模型得分 ，给定 $\mathcal{X}' \subseteq \mathcal{X}$ 和一组模型 $\mathcal{M}_l$ ，我们针对每个子集 $\mathcal{X}' \subseteq \mathcal{X}$ 报告 $|\mathcal{X}'| = B$ ：

$$\text{AvgScore} = \frac{1}{B \cdot |\mathcal{M}_l|} \sum_{\substack{x \in \mathcal{X}' \\ m \in \mathcal{M}_l}} d_{m,l}(x) , \quad (3)$$

，它是该子集上的平均人工得分。对于 完美翻译的比例 ，我们使用：

$$\%\text{Perfect} = \frac{1}{B \cdot |\mathcal{M}_l|} \sum_{\substack{x \in \mathcal{X}' \\ m \in \mathcal{M}_l}} \mathbb{1}[d_{m,l}(x) = 100\%] \quad (4)$$

D 完整结果

Table 6 给出了所有考虑的方法在基于 MQM 注释而非 ESA 时的难度估计相关性分数。

相反，表格 7 和 8 分别展示了在 ESA 注释和 MQM 注释的 WMT24 测试数据上，各种方法的难度估计相关性分数的每种语言的细分情况。这些表格还包括从统计显著性分析中得出的排名。具体来说，我们使用了由 Deutsch et al. (2021) 引入的 PERM-BOTH 假设检验。

E 预言机

Oracle 方法使用实际的人类判断来得出难度分数，如 Section 3 中所述：

- “Oracle” 对每个源文本 x 分配了真实的 $d_{m,l}(x)$ ，对于每个 m 和 l 。
- “Oracle（源文本 + 目标语言）” 通过在所有译者 ($\forall m \in \mathcal{M}_l$) 中平均真实的 $d_{m,l}(x)$ 来估计 x 的难度，这意味着它的估计不会因译者而异，而只会因目标语言而异。
- “仅源文本的 Oracle” 也在目标语言中平均真实得分，为每个源文本分配相同的得分。

F 创建困难基准–定量结果

为了定量评估我们的难度估计器在构建具有挑战性的基准方面的有效性，我们模拟了一个 25 % 预算方案。也就是说，对于每一种方法，我们从 WMT 24 测试集中选择 25 篇% 最难的源文本，并使用人工翻译质量注释来评估所得子集。

表 9 和表 10 分别报告了这次评估的结果，涉及 ESA 和 MQM 人工标注协议，结果是在相应测试集中的所有语言方向上平均得出的。我们考虑两个定量指标：（1）平均得分 ，即选定子集分配的平均人工得分（得分越低表示难度越大），以及（2）% 完美 ，即在选定子集中模型输出获得完美人工得分的比例（比例越低也越好）。

结果证实了我们专门针对难度估计而设计的模型 Sentinel-MQM-24 的强劲表现，其在 AvgScore 和% Perfect 值上均明显低于随机选择。它也在所有仅依赖于源文本的自动方法中取得了最佳结果。特别是在表格 9 中，它在 AvgScore 上仅被外部人工群体 (XCOMET) 超越，这是一种更为复杂的计算方法，要求使用多个大型机器翻译模型翻译每个源文本，并使用 XXL 机器翻译指标评估这些翻译。此外，外部人工群体方法能够生成以目标语言为条件的难度分数，而 Sentinel-MQM-24 仅依赖源文本。另一方面，Sentinel-MQM-24 在表格 9 中获得了最佳% Perfect 分数。在表格 10 中，Sentinel-MQM-24 在 AvgScore 和% Perfect 上均优于所有自动方法，包括外部人工群体。

至于其他自动化方法，文本长度启发式方法始终优于 LLM-as-a-Judge（基于 Command A），尽管后者需要显著更多的计算资源。值得注意的是，在表格 9 和表格 10 中，Command A 仅较随机选择有轻微提升，这进一步强调了在表格 1 中已经观察到的 LLM-as-a-Judge 方法的局限性。

表格 11 和 12 分别提供了 ESA 和 MQM 测试集在 WMT 24 个领域（新闻、社交、文学和语音）的结果的详细细分。这些结果显示，总体模式在各个领域中始终一致。尽管绝对性能有所不同，Sentinel-MQM-24 和 External Artificial Crowd 在几乎所有特定领域的评估中都取得了最强的结果。

	Method	System	Lang	DEC
Oracle	Oracle	✓	✓	1.000
	Oracle (source text + target lang)	✗	✓	0.430
	Oracle (source text only)	✗	✗	0.404
Heuristic	Text Length	✗	✗	0.222
	Syntactic Complexity	✗	✗	0.170
	Word Rarity	✗	✗	−0.052
Learned	Sentinel-MQM-24	✗	✗	0.246
	Sentinel-MQM	✗	✗	0.235
	PreCOMET Difficulty	✗	✗	0.169
	PreCOMET Diversity	✗	✗	0.167
LLM Judge	Command A (source text only)	✗	✗	0.114
	Command A (source text + target lang)	✗	✓	0.120
	GPT-4o (source text only)	✗	✗	0.090
	GPT-4o (source text + target lang)	✗	✓	0.090
Artificial Crowd	Internal (XCOMET-QE-XXL)	✓	✓	**0.278**
	Internal (MetricX-24-Hybrid-QE-XXL)	✓	✓	0.248
	External (XCOMET-QE-XXL)	✗	✓	0.207
	External (MetricX-24-Hybrid-QE-XXL)	✗	✓	0.185
	Random	✓	✓	0.002

Table 6: 在 MQM 标注的 WMT24 上，每种方法所达到的难度估计相关性（DEC）。我们根据它们能够访问的信息类型对这些方法进行分类。仅文本的估计器，例如启发式和学习的估计器，仅依赖于其难度正在被估计的源文本。相反，一些方法还结合了关于翻译目标语言的信息，而另一些则进一步利用了有关测试集中产生翻译的特定译者的知识。

该分析支持了难度估计在受控测试集构建中的实际效用，并证实了像 Sentinel-MQM-24 这样的学习估计器提供了一种有效且可靠的方法，用于识别 MT 系统可能更容易出现困难的源段。

G 基准创建的相关工作

我们通过讨论先前自动创建具有挑战性子集的尝试，扩展了 Section 2 中的相关工作。

Maia Polo et al. (2024, tinyBenchmarks) 和 Rodriguez et al. (2021) 大量使用项目反应理论 (Santor and Ramsay, 1998)，这是一套用于人类受试者教育测试的统计模型。然而，这不适用于机器翻译，在机器翻译中，输出的质量表现为一个连续的分数。其他工作 (Ni et al., 2024b,a; Ruan et al., 2024; Zouhar et al., 2025b) 试图更广泛地适用于自然语言生成任务，尽管它们的优化目标通常是提高测试效率（即，以更少的评估样本获得相同的模型排序）而不是创建困难的测试集。

具体来说，对于机器翻译，Zhan et al. (2021a) 使用机器翻译困难的代理来提供更好的评估。同时，Zouhar et al. (2025a) 自动从评估集中移除过于简单的例子，这对应于我们内部人工人群的质量评估。

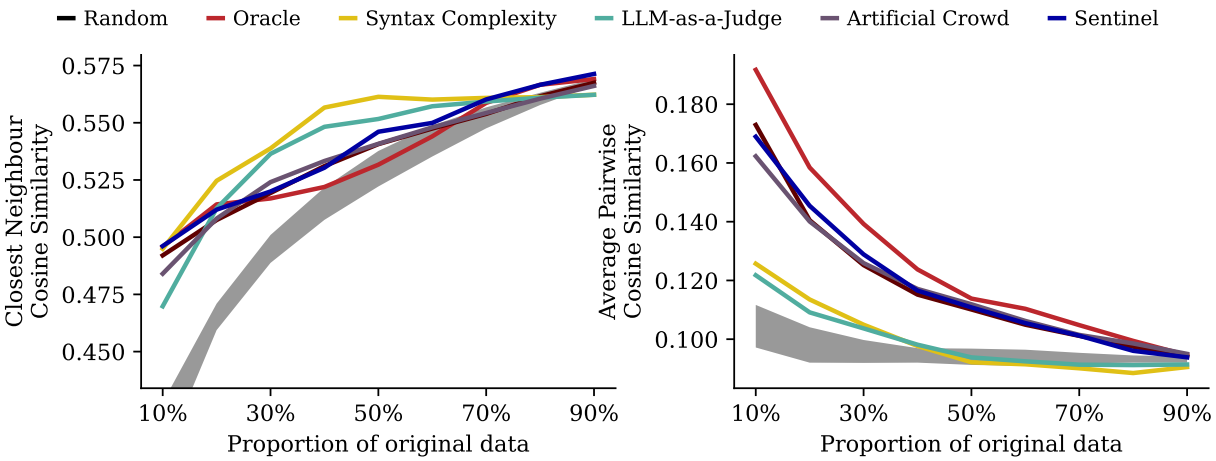

Figure 3: 基于嵌入和余弦相似度的两篇最接近（左侧）以及任意两篇（右侧）源文本在 $\mathcal{X}$ 中的平均相似度。曲线向上是因为向量空间饱和，最近邻变得更近。随机选择显示来自 10 次运行的 99 % 置信 t 检验区间。

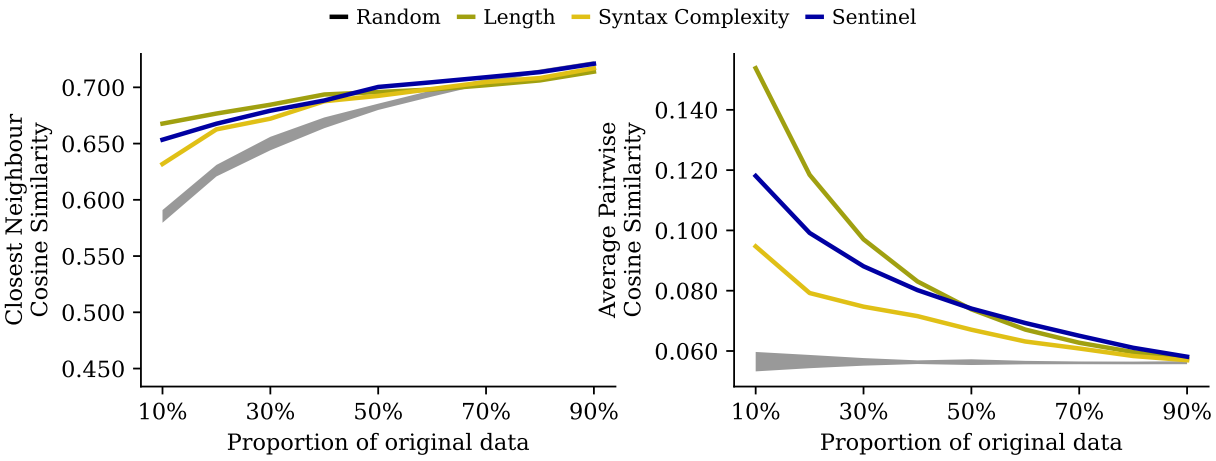

Figure 4: 与 Figure 3 相比，我们收集了当代新闻文章（通过爬虫获取的 40k 段，Ba textasciitilde nón et al., 2020 ）来评估我们的难度采样在现实世界中的表现。基于 $\mathcal{X}$ 中的嵌入和余弦相似度，在原始的 40k 个英语段（非 WMT24）中，两个最接近的（左）和任意两个（右）源文本之间的平均相似性。曲线向上升是因为向量空间趋向饱和，最近邻变得更近。随机选择显示了 10 次跑的 99 % 置信度 t 检验区间。由于计算成本，未包括 LLM-as-a-Judge 和人工大众。由于缺少模型输出和人类评分，不存在 Oracle。

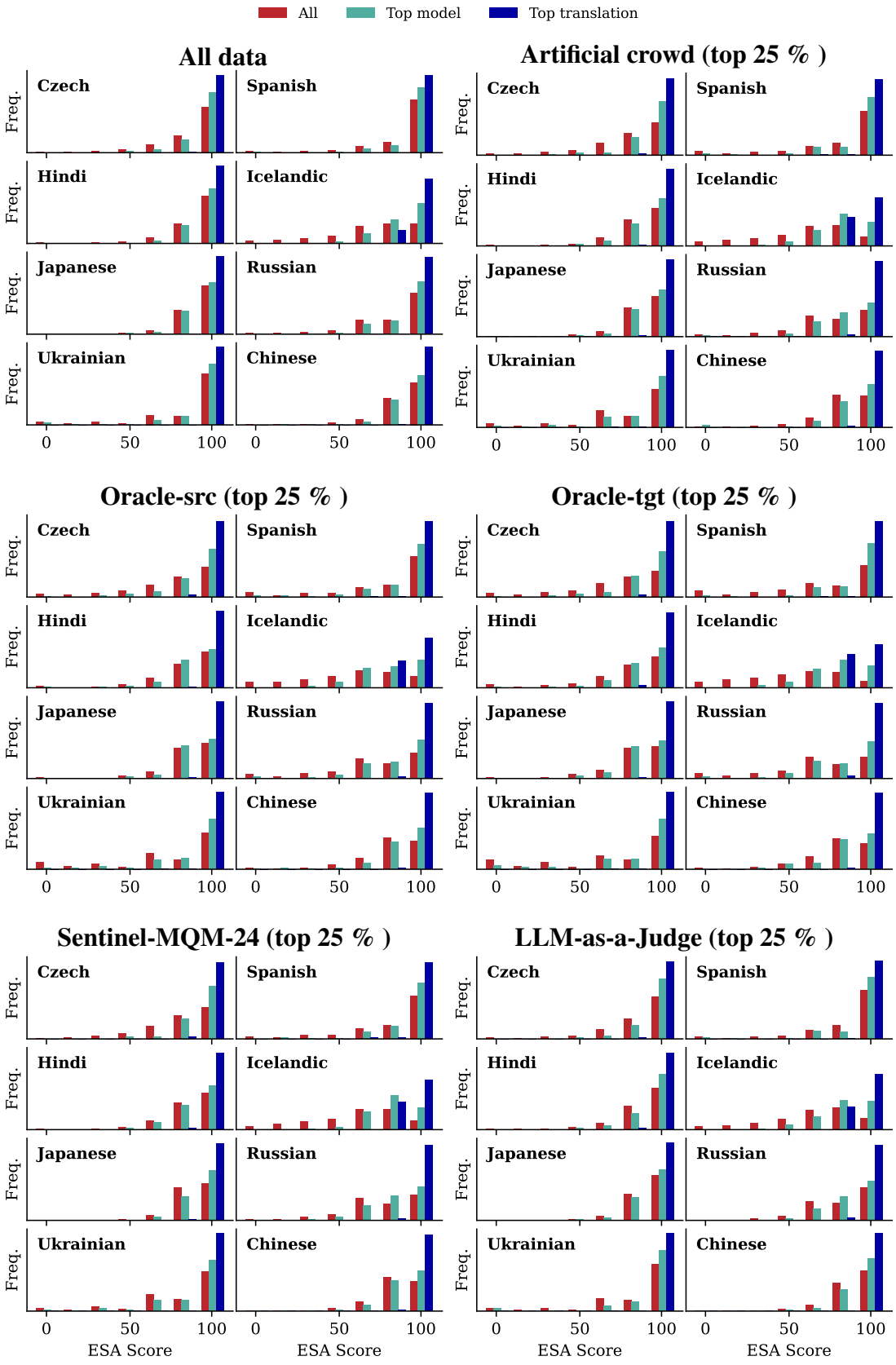

Figure 5: 人类对 WMT 2024 的机器翻译模型的评分分布，包括所有模型的评分、每种语言中评分最高的模型，以及每个输入段落中评分最高的模型。子集选择方法选择前 25% 最难的段落。扩展自 Figure 1 。

	Average		CS→UK	EN→CS	EN→ES	EN→HI	EN→IS	EN→JA	EN→RU	EN→UK	EN→ZH
	Rank	DEC	DEC	DEC	DEC	DEC	DEC	DEC	DEC	DEC	DEC
Oracle	1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Oracle (source text + target lang)	2	0.301	0.298	0.303	0.280	0.271	0.381	0.252	0.320	0.305	0.302
Oracle (source text only)	3	0.224	0.298	0.246	0.201	0.176	0.238	0.172	0.253	0.213	0.218
Internal Artificial Crowd (XCOMET-QE-XXL)	3	0.221	0.203	0.271	0.194	0.213	0.273	0.173	0.238	0.195	0.233
Internal Artificial Crowd (MetricX-24-Hybrid-QE-XXL)	4	0.207	0.211	0.256	0.184	0.212	0.221	0.176	0.229	0.175	0.203
Sentinel-MQM-24	5	0.182	0.167	0.216	0.169	0.173	0.220	0.142	0.204	0.150	0.197
External Artificial Crowd (XCOMET-QE-XXL)	6	0.177	0.175	0.192	0.146	0.179	0.240	0.128	0.194	0.160	0.183
Sentinel-MQM	6	0.175	0.164	0.205	0.159	0.171	0.223	0.118	0.201	0.141	0.190
External Artificial Crowd (MetricX-24-Hybrid-QE-XXL)	7	0.166	0.181	0.174	0.121	0.162	0.247	0.128	0.180	0.136	0.162
PreCOMET Difficulty	8	0.153	0.137	0.193	0.131	0.139	0.188	0.120	0.166	0.131	0.170
PreCOMET Diversity	9	0.142	0.059	0.167	0.134	0.129	0.213	0.120	0.159	0.130	0.165
Text Length	10	0.121	0.024	0.133	0.129	0.143	0.206	0.078	0.142	0.100	0.132
LLM-as-a-Judge (Command A, tgt-based)	11	0.104	0.077	0.100	0.098	0.120	0.190	0.068	0.117	0.072	0.097
Syntactic Complexity	12	0.080	0.018	0.078	0.072	0.112	0.181	0.035	0.090	0.050	0.079
LLM-as-a-Judge (GPT-4o, tgt-based)	12	0.080	0.061	0.067	0.072	0.116	0.179	0.035	0.079	0.037	0.071
LLM-as-a-Judge (GPT-4o, src-based)	13	0.077	0.038	0.066	0.072	0.111	0.188	0.029	0.083	0.036	0.071
LLM-as-a-Judge (Command A, src-based)	14	0.072	0.045	0.063	0.062	0.103	0.169	0.026	0.079	0.029	0.072
Random	15	0.003	-0.001	0.004	0.004	-0.008	0.010	0.008	0.005	0.008	0.000
Word Rarity	16	-0.040	0.016	-0.034	-0.044	-0.065	-0.093	-0.032	-0.043	-0.022	-0.043

Table 7: 在 ESA 注释的 WMT24 上，每种语言每种方法的困难估计相关性 (DEC)。排名代表统计显著性集群，并根据 Freitag et al. (2024) 计算，该方法利用了 Deutsch et al. (2021) 引入的 PERM-BOTH 假设检验。

	Average		EN→DE	EN→ES	JA→ZH
	Rank	DEC	DEC	DEC	DEC
Oracle	1	1.000	1.000	1.000	1.000
Oracle (source text + target lang)	2	0.430	0.505	0.280	0.503
Oracle (source text only)	3	0.404	0.488	0.221	0.503
Internal Artificial Crowd (XCOMET-QE-XXL)	4	0.278	0.309	0.208	0.315
Internal Artificial Crowd (MetricX-24-Hybrid-QE-XXL)	5	0.248	0.268	0.192	0.284
Sentinel-MQM-24	5	0.246	0.278	0.168	0.291
Sentinel-MQM	6	0.235	0.273	0.165	0.268
Text Length	7	0.222	0.262	0.147	0.256
External Artificial Crowd (XCOMET-QE-XXL)	8	0.207	0.243	0.159	0.220
External Artificial Crowd (MetricX-24-Hybrid-QE-XXL)	9	0.185	0.209	0.145	0.201
Syntactic Complexity	10	0.170	0.158	0.073	0.278
PreCOMET Difficulty	10	0.169	0.219	0.129	0.159
PreCOMET Diversity	10	0.167	0.241	0.143	0.117
LLM-as-a-Judge (Command A, tgt-based)	11	0.120	0.122	0.088	0.150
LLM-as-a-Judge (Command A, src-based)	11	0.114	0.117	0.060	0.165
LLM-as-a-Judge (GPT-4o, tgt-based)	12	0.090	0.096	0.064	0.110
LLM-as-a-Judge (GPT-4o, src-based)	12	0.090	0.111	0.049	0.109
Random	13	0.002	0.003	0.004	0.000
Word Rarity	14	-0.052	-0.114	-0.043	0.001

Table 8: 在 MQM 标注的 WMT24 中，每种方法在每种语言上的难度估计相关性（DEC）。等级代表统计显著性的簇，并根据 Freitag et al. (2024) 计算，这利用了由 Deutsch et al. (2021) 引入的 PERM-BOTH 假设检验。

Method	**AvgScore**	**% Perfect**
Random	84.4	21.0 %
Oracle (source text only)	74.9	13.3 %
Oracle (source text + target lang)	71.6	11.4 %
Text Length	82.7	14.1 %
Sentinel-MQM-24	79.1	12.1 %
External Artificial Crowd (XCOMET-QE-XXL)	78.3	13.3 %
Command A (source text + target lang)	83.0	16.1 %

Table 9: 比较从 ESA 测试集中选择最困难的 25 % 样本的方法，评估标准为：(1) 所选子集的平均人工评分和 (2) 所选子集中模型输出达到完美人工评分的比例。结果按语言对计算后取平均值。整个测试集的平均得分（AvgScore）为 84.4，完美输出的百分比（% Perfect）为 20.7 %。

Method	**AvgScore**	**% Perfect**
Random	-2.5	58.8 %
Oracle (source text only)	-6.6	32.7 %
Oracle (source text + target lang)	-6.8	30.5 %
Text Length	-4.5	43.6 %
Sentinel-MQM-24	-5.1	39.6 %
External Artificial Crowd (XCOMET-QE-XXL)	-4.4	43.8 %
Command A (source text + target lang)	-3.1	51.1 %

Table 10: 比较用于从 MQM 测试集中选择最困难的 25 % 样本的方法，通过 (1) 所选子集上的平均人类评分和 (2) 所选子集中模型输出达到完美人类评分的比例进行评估。结果按语言对计算，然后平均。整个测试集的平均分数 (AvgScore) 为-2.5，完美输出的百分比 (% Perfect) 为 57.7 %。

Method	AvgScore				% Perfect			
	News	Social	Literary	Speech	News	Social	Literary	Speech
Random	86.5	84.7	84.7	80.3	19.3%	22.6%	19.7%	12.3%
Oracle (source text only)	82.5	75.8	76.0	71.1	14.1%	16.9%	11.0%	7.0%
Oracle (source text + target lang)	79.6	71.3	72.9	68.3	11.8%	13.7%	8.0%	4.7%
Text Length	84.6	83.1	78.4	82.0	15.0%	15.5%	9.7%	11.8%
Sentinel-MQM-24	84.1	80.2	78.7	77.5	14.4%	15.1%	10.0%	8.6%
External Artificial Crowd (XCOMET-QE-XXL)	84.6	79.6	77.6	75.6	15.3%	16.6%	11.9%	8.1%
Command A (source text + target lang)	84.9	82.1	79.7	78.8	15.5%	17.4%	10.6%	10.2%

Table 11: 对来自 ESA 测试集的测试样本中最困难的 25 个% 进行细粒度评估，这些样本是根据每个领域（新闻、社交、文学、演讲）独立选择的，并在语言对之间平均。结果显示为 AvgScore（平均人工评分）和% Perfect（模型输出具有完美人工评分的比例）。

Method	AvgScore				% Perfect			
	News	Social	Literary	Speech	News	Social	Literary	Speech
Random	−1.4	−1.4	−3.5	−5.5	64.6%	68.9%	56.5%	37.5%
Oracle (source text only)	−4.5	−3.1	−5.9	−10.5	37.0%	45.6%	40.9%	24.6%
Oracle (source text + target lang)	−4.7	−3.4	−5.9	−11.0	33.6%	41.8%	40.7%	22.2%
Text Length	−3.3	−2.0	−5.2	−6.0	47.4%	58.5%	46.4%	37.2%
Sentinel-MQM-24	−2.7	−2.2	−4.9	−7.1	48.7%	56.5%	48.5%	30.2%
External Artificial Crowd (XCOMET-QE-XXL)	−2.5	−2.0	−3.5	−7.6	51.0%	57.8%	50.4%	30.4%
Command A (source text + target lang)	−2.3	−1.6	−4.1	−4.4	50.0%	64.2%	51.2%	40.5%

Table 12: 针对 MQM 测试集中每个领域（新闻、社交、文学、演讲）独立选择的 25 个最困难% 的测试集样本进行了细粒度评估，并在语言对之间取平均值。结果显示为 AvgScore（平均人工得分）和% Perfect（模型输出获得完美人工得分的比例）。

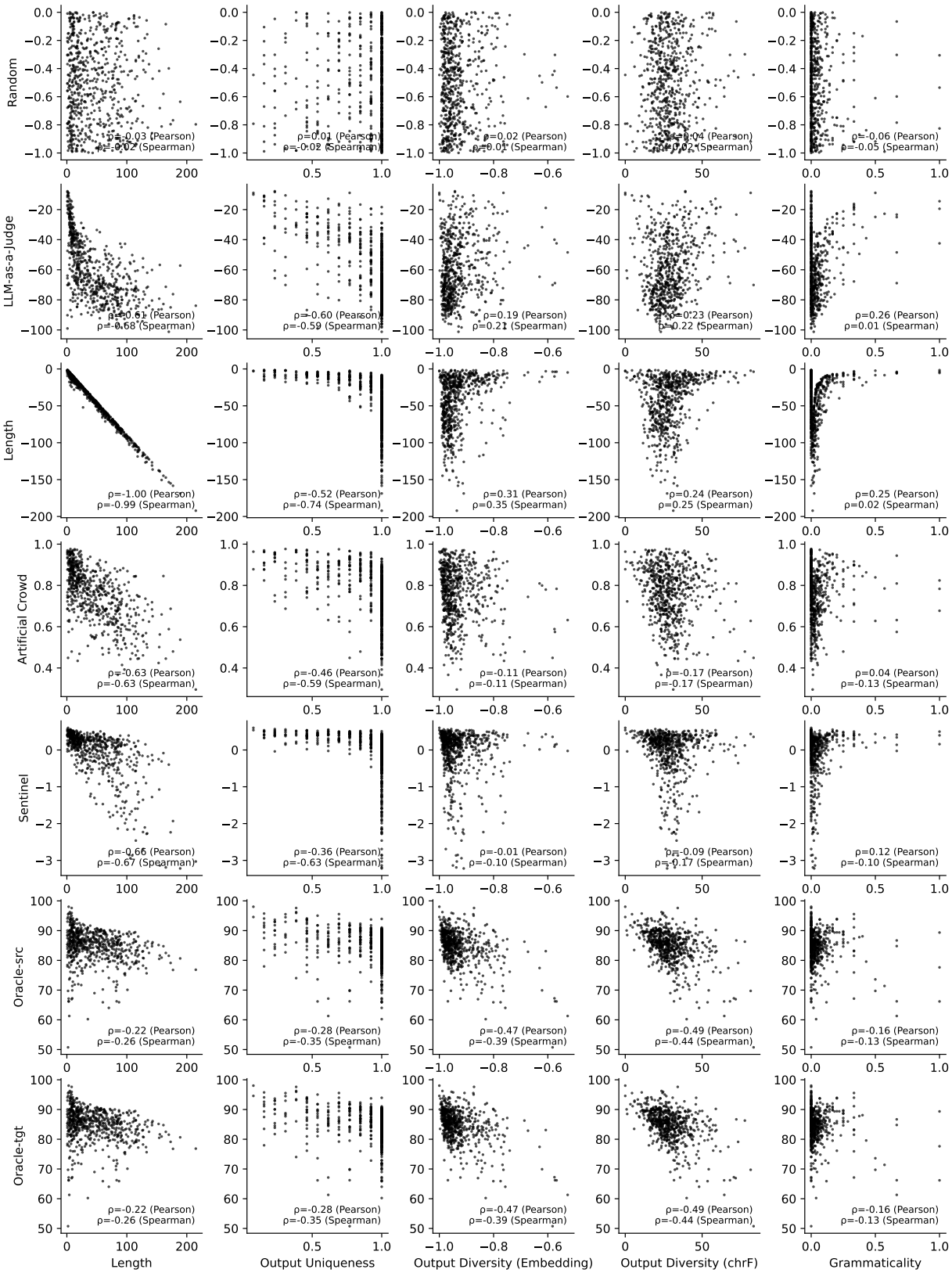

Figure 6: 选择器和感兴趣变量之间的关系（源长度、每个词的源错误数量、通过成对嵌入内积和 chrF 测量的输出多样性，以及独特输出的比例）。对于所有方法，较低的数值表示更困难的源文本，因此负相关意味着困难度与目标变量之间有更强的正向关系。参见 Table 3 以获取汇总视角。

Prompt for LLM-as-a-judge (source text only):

You are given a source text. Your goal is to determine the approximate proficiency level required to translate this text, based on a detailed analysis of its complexity. The final result should be reported as a single numeric score on a scale of 0 to 120, where higher numbers correspond to a higher difficulty (i.e., more advanced language proficiency requirements). You should also relate this numeric score to commonly recognized proficiency levels (e.g., A1, A2, B1, B2, C1, C2). Here is the expected mapping: 0-20 for A1 (Beginner); 21-40 for A2 (Elementary); 41-60 for B1 (Intermediate); 61-80 for B2 (Upper Intermediate); 81-100 for C1 (Advanced); 101-120 for C2 (Mastery).

Instructions: First, examine the text to identify features that affect reading difficulty, including complexity of vocabulary, grammar, semantic density, and any specialized knowledge required. Then, provide a brief explanation of your reasoning for each major factor. Consider whether the text includes domain-specific terminology, cultural references, idiomatic expressions, or advanced grammatical constructions. Finally, assign a numeric score from 0 to 120 and map that score to one of the CEFR levels. Conclude with a final statement that clearly states your numeric score and the corresponding proficiency level surrounded by triple square brackets, for example [[[86, C1 (Advanced)]]]

Analyze following text:
{ src }

Prompt for LLM-as-a-judge (source text + target language):

You are given a source text. Your goal is to determine the approximate proficiency level required to translate this text into { target_language } , based on a detailed analysis of its complexity. The final result should be reported as a single numeric score on a scale of 0 to 120, where higher numbers correspond to a higher difficulty (i.e., more advanced language proficiency requirements). You should also relate this numeric score to commonly recognized proficiency levels (e.g., A1, A2, B1, B2, C1, C2). Here is the expected mapping: 0-20 for A1 (Beginner); 21-40 for A2 (Elementary); 41-60 for B1 (Intermediate); 61-80 for B2 (Upper Intermediate); 81-100 for C1 (Advanced); 101-120 for C2 (Mastery).

Instructions: First, examine the text to identify features affecting the translation into { target_language } , which affect reading difficulty, including complexity of vocabulary, grammar, semantic density, and any specialized knowledge required. Then, provide a brief explanation of your reasoning for each major factor. Consider whether the text includes domain-specific terminology, cultural references, idiomatic expressions, or advanced grammatical constructions. Finally, assign a numeric score from 0 to 120 and map that score to one of the CEFR levels. Conclude with a final statement that clearly states your numeric score and the corresponding proficiency level surrounded by triple square brackets, for example [[[86, C1 (Advanced)]]].

Analyze following text:
{ src }

Example 2: Prompts used to estimate the difficulty of a given text using LLM-as-a-judge (Section 4.2).