探究大型多语言语言模型的翻译表现：以 BLOOM 为例

Rachel Bawden 法国国家信息与自动化研究所，巴黎，法国
rachel.bawden@inria.fr

Franc¸ois Yvon 巴黎-萨克雷大学，CNRS，LISN
francois.yvon@cnrs.fr

arXiv:2303.01911v2 [cs.CL] 9 May 2023

摘要

自然语言处理社区最近见证了一款新型大型开放获取多语言语言模型的发布：BLOOM（BigScience 等人，2022年），该模型涵盖46种语言。我们通过评估 BLOOM 在多个数据集（WMT、Flores-101 和 DiaBLa）及不同语言对（高资源和低资源）上的机器翻译表现，重点考察其多语言能力。结果表明，零样本表现会受到过度生成和生成错误语言的影响，但在少样本设置下有显著改善，多个语言对上的结果非常出色。我们研究了多个方面，包括提示设计、模型规模、跨语言迁移以及话语语境的使用。

1 引言

经过大规模训练、采用简单目标的大语言模型（LLM），已被发现能够在众多自然语言处理任务上取得与专用系统相当的结果（Radford 等人，2019年），前提是通过“提示”将任务表述为文本生成形式（Liu 等人，2023年）。通过“指令”微调（Sanh 等人，2022年；Muen- nighoff 等人，2022年）、少样本引导，以及更好的提示选择或学习策略（Petroni 等人，2019年；Shin 等人，2020年；Schick 和 Sch¨utze，2021年；Lester 等人，2021年；Wei 等人，2022年），还可以提升大语言模型的多任务表现。在多语言场景中，即使模型主要使用单语数据训练，其机器翻译（MT）任务表现（以自动评分衡量）往往也接近最先进水平（Brown 等人，2020年）。此外，基于提示的机器翻译有望更好地控制输出，例如控制质量、风格和方言（Garcia 和 Firat，2022年）。然而，我们对这些能力的理解仍然有限，因为对大语言模型的分析主要关注其多任务能力，而非多语言能力（不过，参见 Vilar 等人，2022年；Zhang 等人，2023年；Moslem 等人，2023年），我们将在第2节讨论这一点）。

本文聚焦于 BLOOM（BigScience 等人，2022年）的机器翻译性能。BLOOM 是由 BigScience 协作项目设计并训练的开放获取多语言大语言模型（系列）。[1]我们的主要目标是：(i) 评估 BLOOM 的零样本和多样本行为；(ii) 研究提示设计的影响；(iii) 评估多种语言对；(iv) 评估其利用语言上下文的能力。我们的主要结论（对（BigScience 等人，2022年）中的结论作了进一步拓展）是：(i) 零样本能力受到过度生成和生成错误语言的影响；(ii) 使用少样本示例能改善这两个问题，使模型在不同数据集和语言对上的结果都更接近最先进水平；(iii) 存在明显的迁移效应，对于训练中未正式出现的语言，模型也能取得高分，并能通过少样本示例在不同语言对之间实现成功迁移；(iv) 尽管语言上下文并未带来更高分数，但有证据表明它会影响 BLOOM 的译文。我们公开了代码和翻译输出。[2]

2 相关工作

自早期尝试将语言模型（LM）用作多任务学习器（McCann 等人，2018年）以来，机器翻译一直是衡量语言模型多语言能力的首选任务。研究者讨论了 GPT-2 和 GPT-3 的零样本与少样本能力（Radford 等人，2019年；Brown 等人，2020年），这尤其引人关注，因为它们的训练数据主要是单语（英语）数据。此后，这些结果也在其他单语语言模型（如 T5）（Raffel 等人，2020年）以及多语言语言模型（如 XGLM）（Lin 等人，2022年）、PALM（Chowdhery 等人，2022年）和 ALEX-ATM（Soltan 等人，2022年）中得到了验证。然而，研究重点一直

[1]https://hf.co/bigscience/bloom[2]https://github.com/rbawden/mt- bigscience

主要放在整体多任务表现上；讨论中往往只有很小一部分涉及机器翻译。此外，研究结果通常只报告少数资源丰富的语言对（例如英语-法语和英语-德语），而且由于标准化评估协议和指标的使用缺乏系统性，所报告的分数（主要是 BLEU）难以比较。[3]

不过，也有一些针对大语言模型机器翻译性能的深入分析，分别聚焦于特定语言模型在真正多语言环境下相对于提示设计和少样本数量的表现。例如，Vilar 等人（2022年）重新评估了多语言 PALM （Chowdhery 等人，2022年）的机器翻译性能，尤其关注少样本示例的选择。与我们的发现一致，他们认为，在少样本设置中提示的选择变得不那么重要，而使用少样本示例能够提升性 $k > 5$能，但对于更多示例，收益会递减；这一结论基于 BLEURT 和 BLEU分数，以及人工评估的结果。他们发现，少样本示例的质量对性能有很大影响。然而，即使使用良好的提示，PALM 仍落后于最先进的机器翻译系统数个百分点，尤其是在从英语翻译时，这一点尤为明显，原因在于充分性问题。Zhang 等人（2023年）重点关注对 GLM-130B 的评估；GLM-130B 是一个双语（中文和英语）大语言模型（Zeng 等人，2022年）。他们的主要结论也与我们一致：（a）零样本性能在不同提示之间差异很大；（b）将提示数量从 0 增加到 20 可带来一致的性能提升，但不同指令之间仍存在差异；（c）很难找到最佳的少样本示例选择策略。例如，使用优质且较长的示例似乎可能有所帮助，尽管本研究探索的所有标准似乎都未能带来任何系统性的改进。最后一点值得一提：使用单语数据进行提示会损害性能，但使用通过回译（Bojar 和 Tamchyna，2011）获得的伪平行数据则是一种有效的替代方案。

Moslem 等人（2023年）评估了 OpenAI 的 GPT-3（Brown 等人，2020年）[4] 基于采样的解码方法，以及一个类似于我们自身 xglm-源语言+目标语言提示的提示。他们使用多种指标报告了强劲的零样本表现；对于资源丰富的语言，增加示例数量后性能还明显提升，而对于其研究中唯一的低资源语言（卢旺达语），提升则不那么明显。这项研究的主要创新在于利用提示进行局部适配，并确保术语一致性。为此，他们使用翻译记忆库中的模糊匹配以及机器翻译输出构建提示，所得结果不仅优于其零样本系统，也优于其初始机器翻译引擎。此外，在指令中加入术语及其翻译还能带来进一步的改进。

最后请注意，(Jiao 等人，2023年) 对 CHATGPT 进行了初步评估。该研究报告了有关该模型多语言能力的有趣见解，并提出了生成（人工）提示以及在提示中使用枢轴转换的创新技术。与我们的研究类似，该研究考虑了多个测试领域，例如新闻（WMT）和维基百科（Flores）。在 (Hendy 等人，2023年) 中可以找到对同一模型更深入的分析，该分析证实了 CHATGPT 强大的翻译能力，至少对于“资源充足”的[5] 语言对而言如此。文中还报告了文档级分数、人工评估和定性分析。

多语言机器翻译也是专门的（单任务）架构和训练方案的研究主题。最初在 (Dong 等人，2015；Fi- rat 等人，2016年；Luong 等人，2016年) 中提出时，这些方法仅支持有限的语言范围；而这些方法的最新版本已经能够处理数百种语言，其中包括低资源语言对（Fan 等人，2021年；Bapna 等人，2022年；Costa-juss`a 等人，2022年）。尽管我们发现，只要有足够的训练数据，BLOOM 就能够达到相当的性能，但我们也看到，对于许多在其训练数据中代表性不足的语言对而言，它仍然落后。

3 BLOOM 语言模型

BLOOM 是一个大型开放访问的多语言模型，使用 46 种自然语言训练，由 BigScience 项目开发（BigScience 等人，2022年）。

[3]请参见 http://blog.benjaminmarie.com/2/comparing-uncomparable.html 其中对这些差异的讨论，以及对一致分数的重建尝试。[4]版本：text-davinci-003模型。

[5]在此语境中，这是一个相当难以界定的概念，因为训练数据的内容尚不完全明了，而且似乎主要由英语文本构成。

	提示名称	提示	目标
1–2	a_good_translation	给定以下源文本（以源语言表示）： [源句子], 一个好的 L2 翻译是:	[目标句子]
3	版本	如果原始版本写的是 [源句子] 那么目标语言版本应写:	[目标句子]
4	gpt3	以下句子的目标语言翻译是什么： [源句子]?	[目标句子]
5–6	xglm	(源语言:) [源句子] =目标语言:	[目标句子]
7	译作 -	[源句子] 译为目标语言:	[目标句子]

表1：WMT'14 数据集的七个机器翻译提示（Bojar 等人，2014）。所有提示均指定目标语言（目标语言）。每个提示都有一个"仅目标语言"版本（-target），其中仅指定目标语言；此外，其中两个提示还存在第二个-source+目标版本，在该版本中，指令明确标出了源语言（以红色并置于括号内）。

这是一个自回归语言模型，旨在根据用户输入的文本前缀（称为提示）生成续写文本。它可用于多种任务，包括机器翻译、问答等。BLOOM 使用 1.6 TB 文本进行训练（其中 30% 为英语），数据来自各种来源，不过，38% 的数据属于 ROOTS 语料库（Laurenc¸on 等人，2022年），[6] 数据来自 Oscar 网络数据（Ortiz Su´arez 等人，2019年）。该模型以多种规模在 HuggingFace 上开放发布，参数量从 5.6 亿到 1760 亿不等。[7]

4 在机器翻译任务上评估 BLOOM

4.1 使用的机器翻译数据集

我们选取了三个数据集来开展实验，以测试 BLOOM 在机器翻译方面的不同表现：WMT（Bo- jar 等人，2014年）、Flores-101（Goyal 等人，2022年）和 DiaBLa（Bawden 等人，2021年）。我们使用 WMT 2014 英语↔法语和英语↔印地语新闻测试集，将它们视为相对于 BLOOM 训练数据而言分别代表高资源和低资源的语言对。[8] 这些测试集已有些过时（Garcia 等人，2023年），但过去的大语言模型评估反复使用了它们，而且它们也被纳入标准基准，便于进行比较。Flores-101 是一个涵盖 101 种语言的多平行语料库，由原始英语句子翻译而来。事实上，译为英语的评估结果必然会显得过于理想（例如（Toral 等人，2018年）），而在其他语言之间进行的评估，可能主要反映的是这些语言与原始英语的相似程度。我们使用该数据集来测试和比较 BLOOM 的多语言能力，包括低资源语言的语言。[9] DiaBLa 是一个双语测试集，包含英语和法语使用者之间经机器翻译处理的自发书面对话。我们用它来测试非正式领域中的机器翻译，以及（跨语言）语言上下文对机器翻译的影响。

4.2 实验设置

我们使用语言模型评估工具包（BLOOM （及其变体）在零样本和少样本设置下进行评估和比较。对于少样本设置， k 示例会添加在提示前，并使用 ### 分隔，如示例 1所示（单样本示例带有下划线）。

(1) 输入：法语：我感到无聊 = 英语：我很无聊。### 英语：刚刚走进那边的那条狗是你的吗？ =法语：参考译文：那是刚刚走进来的您的狗吗？

结果基于各数据集的测试集划分报告。少样本示例会根据可用情况从数据集划分中随机抽取（WMT 使用训练集，Flores-101 使用开发集，DiaBLa 使用测试集）。我们使用 SacreBLEU 中实现的 BLEU（Papineni 等人，2002年）进行评估（Post，2018年），WMT 和 DiaBLa 使用 13a作为分词方式，Flores-101 则按建议使用 SentencePiece（Costa-juss`a 等人，2022年）中的分词方式。[10] BLEU 有许多不足之处，但足以为本研究中的大多数系统提供定量比较。当分数较为接近时，我们还会使用 COMET（Rei 等人，2020年）进行更细致的比较。

4.2.1 对比模型

在我们的跨数据集比较（第 5.1 节）中，我们将 BLOOM 与其他大语言模型进行比较：（i）两个经过任务微调的模型：T0[11] （Sanh 等人，2022年），

[6] ROOTS 语料库现在可以使用专用搜索工具进行查询 https://hf.co/spaces/bigscience-data/roots-search。[7] https://hf.co/bigscience/bloom [8] 英语、法语和印地语分别占训练数据的 30%、12.9% 和 0.7%（Laurenc¸on 等人，2022年）。

[9] Flores-200 是近期发布的扩展版本（Costa-juss`a 等人，2022年），规模更大，涵盖的语言数量约为原来的两倍。由于这一新版本在我们的评估过程后期才发布，且当时仅在一篇论文中使用过，因此我们决定继续使用 Flores-101。[10] BLEU+case:mixed+smooth.exp+{13a,spm}+version.2.2.1 [11] https://hf.co/bigscience/T0

在英语文本上训练，以及 mT0-xxl[12]（Muennighoff 等人，2022年），多语言版本；以及（ii）OPT[13]（Zhang 等人，2022年），一种英语生成式语言模型。我们使用相同的提示 `xglm-source+target` 对所有模型进行评估。为了使用 Flores-101 评估多个语言对，我们将其与有监督的 6.15 亿参数机器翻译模型 M2M-100（Fan 等人，2021年）进行比较（作为基准上限），所使用的分数由 Goyal 等人（2022年）计算得出。

4.2.2 提示

我们使用了多个提示，用于展示不同的变化来源：(i) 是否包含源语言名称，(ii) 源语言名称与目标语言名称的相对顺序，(iii) 源句子的位置（提示开头或结尾），以及 (iv) 提示的详略程度。这些提示可在 PromptSource（Bach 等人，2022年）中找到，如表1所示。前三个提示的设计受以下先前研究启发：[14]（Brown 等人，2020年）用于 gpt，3,[15]（Lin 等人，2022年）用于 xglm，以及（Wei 等人，2022年）用于 translate as；后者也与 Raffel 等人（2020年）的提示（将英语译成德语：“[源文本]”：[目标句子]）相似，该提示也曾用于（Wei 等人，2022年；Garcia 和 Firat，2022年）。

考察表1中的条目可以发现，“提示”实际上指输入的两个不同方面：（i）用自然语言表述任务；（ii）穿插语言标签，呈现相关示例（适用于少样本设置）（或许用启动 一词来称呼会更清楚，见 Pham 等人 (2020年）。以`xglm`提示为例，指令部分可以简化为一个词。正如下文结果所示，指令在零样本设置中尤为重要，但在少样本场景中几乎可以省去。（Brown 等人，2020年）和（Hendy 等人，2023年）的作者在零样本设置中也使用了冗长、类似指令的提示，而在少样本实验中则采用简洁得多的提示。另请注意，InstructGPT 的提示结合了指令和语言标签（Ouyang 等人，2022年，第49页）。

[12] https://hf.co/bigscience/mt0-xxl [13] https://hf.co/facebook/opt-66b [14] 由于文档不完整，未能明确说明 (a) 测试过哪些提示，以及 (b) 每项实验实际使用了哪些提示（例如，零样本和少样本实验使用的提示不同），因此整个过程并不总是简单明了（Chowdhery 等人，2022年）。[15] 似乎仅用于零样本学习，形式为“问：句子 [源句子] 的目标语言翻译是什么？答：”，其中，特殊标记 Q 和 A 分别表示查询和答案文本（参见图 G.36，第 59 页）。

5 评估结果

我们对 Bloom 的评估首先比较三个数据集，并检测主要机器翻译错误，重点关注 WMT（第5.1节）；随后，我们针对若干具体方面展开更深入的分析：（i）基于 WMT，对 Bloom 模型规模（第 5.2 节）和提示（第 5.3 节）进行比较研究；（ii）基于 Flores-101，评估更多语言对和跨语言少样本迁移（第5.4节）；（ii）基于 DiaBLa，研究语言上下文的运用（第5.5节）。

5.1 跨数据集比较

	零样本				单样本			
	BLOOM	T0	mT0	OPT	BLOOM	T0	mT0	OPT
WMT 2014								
英语→法语	14.9	1.2	29.3	12.9	27.8	1.4	25.2	21.9
法语→英语	15.5	25.8	32.9	15.5	34.6	21.0	30.0	24.6
英语→印地语	6.8	0.2	11.2	0.1	13.6	0.1	9.5	0.1
印地语→英语	12.1	0.0	26.1	0.4	25.0	0.0	20.1	0.6
DiaBLa								
英语→法语	0.9	0.5	28.4	0.5	5.7	0.6	21.0	15.5
法语→英语	0.8	25.5	35.0	0.8	12.1	20.6	26.9	12.1
Flores-101								
英语→法语	2.8	1.9	55.5	2.8	45.0	2.1	53.5	24.4
法语→英语	2.7	31.9	60.1	2.6	45.6	24.9	58.2	16.7
英语→印地语	1.3	0.1	67.7	0.1	27.2	0.1	54.7	0.1
印地语→英语	3.4	0.0	59.5	0.1	35.1	0.2	57.3	0.5

(a) 原始预测

	零样本				单样本			
	BLOOM	T0	mT0	OPT	BLOOM	T0	mT0	OPT
WMT 2014								
英语→法语	32.2	1.2	29.2	18.9	36.3	1.4	25.2	22.3
法语→英语	37.2	25.8	32.9	33.2	38.2	21.1	29.9	33.2
英语→印地语	12.1	0.2	11.2	0.1	15.7	0.1	9.5	0.1
印地语→英语	24.3	0.0	26.1	0.5	25.0	0.0	20.1	0.6
DiaBLa								
英语→法语	24.2	0.5	28.4	17.4	37.6	0.6	21.9	20.7
法语→英语	22.9	25.5	34.9	36.8	41.4	21.1	27.2	37.6
Flores-101								
英语→法语	26.9	1.9	55.3	21.4	49.3	2.1	53.4	28.4
法语→英语	40.3	31.9	60.0	39.4	47.2	25.2	58.2	39.8
英语→印地语	7.7	0.1	67.7	0.1	29.5	0.1	54.7	0.1
印地语→英语	30.2	0.0	59.5	0.2	35.1	0.2	57.3	0.5

（b）截断预测

表2：使用 xglm-源语言+目标语言提示进行的 BLEU分数跨数据集比较（Flores-101 使用 spBLEU）。

我们首先使用相同的提示来提示 Bloom 和对比模型，并在各数据集上进行测试，将所测试的方向限制为英语↔法语以及英语↔印地语。我们选择系统地使用 xglm-源语言+目标语言提示（表 1），其对应于以下模板：

(2) `L1: [source sentence] = L2:`

其中，L1和L2分别指源语言和目标语言（例如，英语和法语，对应于

en→fr），[源句子]会被替换为给定的源句子。

BLEU分数见表2a，涵盖零样本和单样本设置（COMET 结果见附录A）。所有方向的零样本机器翻译都存在问题，尤其是翻译成非英语语言时（BLEU分数系统性地低于翻译成英语时）。即使是翻译成英语，与最先进水平相比，分数仍然偏低（例如，Flores-101 上 fr→en 方向的 BLOOM 为 2.7 BLEU，而 MT0-XXL 为 60.1）。[16] MT0-XXL 的出色结果体现了提示微调的优势。单样本设置下的结果有显著提升（T0 和 MT0-XXL 除外，两者都经过了微调），与文献中报告的其他大语言模型机器翻译结果相当；BLOOM 的结果明显优于 OPT，后者的训练数据大多为英语。这些总体观察适用于 WMT 和 Flores-101，但 DiaBLa 仍存在问题。

主要问题是什么？我们发现了两个主要问题（尤其是在零样本场景下）：（i）过度生成，以及（ii）生成内容使用了错误的语言（这对机器翻译来说是个明显的问题）。示例3和4展示了这些问题的一些实例。第一个示例展示了一个非常常见的问题：提示以多种语言反复出现，直到达到最大生成长度。

(3) 输入：法语：他说自己构思出了一款 Wi-Fi 门铃。 = 英语：

BLOOM：他说自己构思出了一款 Wi-Fi 门铃。 = 西班牙语：他说自己构思´出了一款 Wi-Fi 门铃。= 德语：他说自己设计了一款无线局域网门¨铃。¨ = 意大利语：他说自己构思了一款 Wi-Fi 门锁 [。 。 。]

(4) 输入：法语：老虎与狮子、美洲豹和美洲狮属于同一个科（豹属）。这四种猫科动物是唯一能够吼叫的动物。 =英语：

BLOOM：老虎与狮子、豹子和美洲豹同属一个科（豹属）。这四种猫科动物是唯四能够咆哮的动物。 老虎是一种大型猫´科动物，生活在亚洲。它是体型最大的猫´科动物，也是体型最大的陆地食肉动物。它速度极`快，奔跑速度可达 `60 公里/小时以上。 [. . .]

将机器翻译质量与过度生成区分开来示例中的过度生成3是一个独立的问题，与 BLOOM 将文本翻译成另一种语言的能力无关。因此，我们为这种过度生成设计了一种自定义截断方法，使预测中只保留第一个译文，即换行符之后或正则表达式模式 = .+?: 之后的所有内容都会被丢弃。

截断后的结果（表 2b）显示，对于全部三个数据集，零样本和单样本的分数均显著提高（例如，单样本 DiaBLa 法语→英语从 12.05 提高到 41.36，零样本 Flores-101 印地语→英语从 3.40 提高到 30.19 ）。BLOOM 能够执行良好的机器翻译，但存在一个问题：它不知道何时该停止生成。我们在其他地方也采用相同的截断方式，并注明展示的是原始输出还是截断输出。

	英语→法语		法语→英语		英语→印地语		印地语→英语	
	0	1	0	1	0	1	0	1
目标	2814	2959	2954	2979	1998	2431	2469	2499
源语言	181	32	47	22	476	48	29	2
其他	8	12	2	2	33	28	9	6
总计	3003	3003	3003	3003	2507	2507	2507	2507

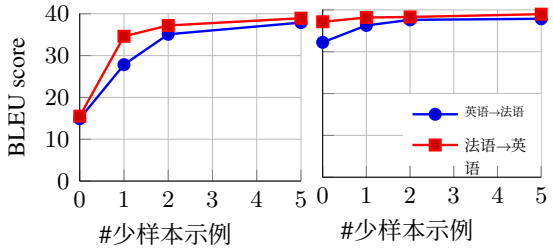
表 3：零样本和单样本设置（针对 WMT）下，经截断后被判定为使用（正确）目标语言、源语言或其他语言的输出数量。

检测使用错误语言生成的内容我们使用 fasttext 语言识别自动检测预测结果的语言[17]（Joulin 等人，2017）。表 3展示了经截断后，被识别为使用正确目标语言、源语言或其他语言的译文数量，涵盖零样本和单样本设置。[18,19] 使用正确目标语言的句子数量从零样本到单样本有所增加，尤其是两种非英语目标语言。翻译为印地语时（零样本），有五分之一（509 个）的预测结果未被识别为印地语；单样本设置基本解决了这一问题（只有 76 个输出使用了错误的语言）。

增加少样本示例的数量这两个问题在单样本设置下都有显著改善。随着少样本示例数量增加，这一趋势仍在持续，BLEU分数也随之提高，如图 1 中的 WMT 英语↔法语结果所示。不过，我们也看到收益逐渐递减，在示例数量从 2 个增加到 5 个时尤为明显，这表明超过五样本后，收益会更有限。

[16]作为比较，（Bi 等人，2020）报告称，有监督机器翻译的最新 BLEU 分数，在 WMT14 英语→到法语和法语→到英语任务中分别为 45.6 和 45.4。

[17] https://fasttext.cc/docs/en/language-identification.html，使用压缩版本 lid.176.ftz。[18]原始表格见附录12 和 13 。B。[19]这些数字优于（BigScience 等人，2022年）中报告的初始结果，因为我们使用了不同的提示和截断。有关每种提示的详细分析，请见下文。



(a) 原始输出。 (b) 截断输出。

图 1：随着少样本示例数量增加，WMT 2014 英语↔法语任务中 xglm 提示的 BLEU分数。

模型	英语→法语		法语→英语	
BLOOM	11.2	3.0–22.0	15.4	10.3–26.8
BLOOM-7b1	6.5	1.5–12.1	12.8	4.8–25.1
BLOOM-3b	3.6	1.2–9.6	10.6	2.8–19.3
BLOOM-1b1	1.7	0.5–3.9	7.1	0.7–11.4
BLOOM-560m	0.6	0.4–0.9	3.7	1.4–5.4

(a) 零样本

模型	英语→法语		法语→英语	
BLOOM	32.6	27.8 –36.4	34.9	33.1–36.6
BLOOM-7b1	25.9	20.8–29.9	29.1	25.4–32.5
BLOOM-3b	21.6	16.7–26.8	25.7	18.6–29.6
BLOOM-1b1	10.1	6.3–13.2	16.1	12.2–19.9
BLOOM-560m	3.6	2.2–4.4	8.6	5.8–12.1

(b) 单样本

表 4：BLOOM（原始输出）规模逐渐减小时，七个提示的平均 BLEU 分数及其范围。

5.2 BLOOM 模型规模

BLOOM 有多个版本，参数数量各不相同。为了测试模型规模对性能的影响，我们报告了 WMT 在七个提示下的平均分数和分数范围。表 4 显示，随着模型规模缩小（从 176B 降至 560M 参数），性能也显著下降。从零样本转为单样本时，所有模型的性能都有大幅提升，较小的模型（如 BLOOM-7b1、BLOOM-3b）与最大模型之间的差距也略有缩小。由于表 4中的分数范围是根据不同提示计算得出的，我们看到，在零样本设置下，不同提示产生的 BLEU 分数差异显著；在单样本设置下，最佳提示与最差提示之间仍相差 6 到 8 个 BLEU 点。对后处理进行的类似分析，以及英语↔印地语（附录 C）分析也证实：(i) 截断能提高所有模型规模和提示下的分数；(ii) 与好的提示相比，选用不佳的提示可能导致机器翻译性能灾难性下降。

5.3 提示逐项分析

根据提示选择计算出的 WMT 平均值（使用表 1 中的提示），可以进一步探究跨提示变异性。

哪种提示效果最好？表格展示了这种差异5和6报告了英语↔{法语、印地语}三种语言对各提示的表现，取第BLOOM 节中的五个模型的平均值5.2。[20] 截断输出的对应表格见附录D。版本和a goodtranslation (源语言+目标语言) 获得了最高的平均分（和最高分）。这两种提示都更为冗长（类似指令），但在单样本设置下，它们与较简单的“启动风格”提示（例如 xglm）之间的性能差距有所缩小。gpt3 的结果最差。使用这一提示时，在一段仅包含英语的文本之后翻译成法语似乎尤其困难：gpt3 的零样本译文中，有一半被 langid 判定为非法语（其中大多数是英语）。翻译成印地语时，检测为印地语的输出只有 10 条。

在提示中注明源语言是否有帮助？我们比较了 -target 和 -source+target 两个版本的 a good translation和 xglm。表 5 和 6 中的结果并无定论。对于这些语言方向和提示，单样本设置下的差异很小，可能是运行间的方差所致。零样本设置下，注明源语言显然有助于 xglm，但对于篇幅更长的 a good translation，这样做对一个方向有帮助，却会对另一个方向产生不利影响。要得出更可靠的结论，还需要进一步探讨这一问题，包括使用非英语提示。

5.4 评估更多语言方向

我们进一步使用 Flores-101 探索 1-shot 设置下更多的语言方向。如第 5.1 节所述，我们使用 xglm-源语言+目标提示。[21]

[20]对于给定的提示，取值范围主要反映 BLOOM 模型不同规模的性能。

[21]在此前的实验中，它的平均表现良好，而且是篇幅最短的模型之一，因此更适合用于多语言场景。

提示 / 少样本 #	英语→法语 0	英语→法语 1	法语→英语 0	法语→英语 1
a good translation-来源+目标	6.7 0.6–15.4	18.7 4.1–36.4	11.0 5.4–14.2	25.8 11.6–36.6
a good translation翻译目标语言	3.1 0.4–10.1	20.3 3.2–35.5	12.1 5.1–16.8	<风格 id='1'>25.9 </风格>12.1–36.2
GPT-3目标	2.5 0.5–7.9	16.6 2.2–32.5	4.5 0.7 –12.7	19.3 5.8–33.1
翻译为目标 -	3.3 0.4– 5.0	17.1 3.2–32.7	6.9 2.1 –11.3	21.6 7.6–35.1
版本目标	7.5 0.6–22.0	**21.4** 4.3–34.2	<风格 id='1'>17.1 </风格>3.9–26.8	24.9 7.8–35.4
xglm-源语言+目标语言	**8.3** 0.9–14.9	17.5 3.3– 27.8	11.8 5.0–15.5	22.1 7.8–34.6
xglm-target	1.6 0.7– 3.0	16.7 4.4–29.0	6.2 2.6–10.3	20.7 7.5–33.3

表 5：按提示列出的 en↔fr BLEU分数平均值、最小值和最大值（原始输出）。每种设置下的最佳平均结果以粗体显示。

提示 / 少样本 #	en→hi 0	en→hi 1	hi→en 0	hi→en 1
a good translation-源+目标	0.7 0.1–1.9	5.8 0.3–14.5	4.8 0.9–10.2	13.1 2.8–24.6
a good translation-翻译目标语言	0.2 0.1–0.8	5.5 0.3–14.1	6.3 1.1–13.0	13.2 2.8–24.8
GPT-3目标	0.1 0.0–0.3	1.4 0.0–6.5	0.2 0.0–0.7	2.2 0.0–10.0
版本目标	0.7 0.1–2.0	5.6 0.2–14.0	<风格 印度尼西亚语='1'>6.8 </风格>1.7–11.5	**13.3** 2.4–25.8
xglm-来源+目标	**2.1** 0.1–6.8	**6.9** 0.3–13.6	4.4 0.6–12.1	11.9 1.7–25.0
xglm-target	0.2 0.0–0.6	5.1 0.1–14.6	1.6 0.2–4.1	6.6 0.5–13.2

表 6：en↔hi 每个提示的 BLEU 分数平均值、最小值和最大值（原始输出）。每种设置下的最佳平均结果以粗体显示。

5.4.1 按语言划分的结果

为优化计算资源，我们没有运行所有语言组合，而是重点考察：(i) 高资源语言对，(ii) 高→中资源语言对，(iii) 低资源语言对，以及 (iv) 相关语言（具体而言，罗曼语族）。鉴于过度生成在单样本设置中问题较小，原始输出的结果见表 7 和 8 。

高资源和高→中资源 高资源和高→中资源语言方向的结果总体良好，在高资源语言方向上超过了 M2M 的分数，es→fr 除外。[22] 这表明 BLOOM 具有良好的多语言能力，即使跨越不同文字系统（扩展拉丁文、中文、阿拉伯文和天城文）也是如此。

低资源语言 对于低资源语言，结果的差异更大；有些语言方向的表现优于 M2M，尤其是大多数译入英语的方向，但另一些方向的表现较差（例如译入印地语和斯瓦希里语）。对于测试中资源最匮乏的语言（sw↔yo 以及 en↔yo），结果尤其令人失望，因为分数表明生成的翻译毫无意义，尽管约鲁巴语和斯瓦希里语都出现在 BLOOM 的训练数据中（每种语言各有 <50k 个 token），只是占比不足。

罗曼语族 与此形成对比的是罗曼语族之间的结果：整体表现良好，包括意大利语（it）和加利西亚语（gl）的互译，而这两种语言并未正式纳入训练数据。需要注意的是，加利西亚语与其他罗曼语族语言有许多相似之处，尤其是葡萄牙语（pt）。这些截然不同的结果表明，大语言模型的表现不仅取决于训练数据的数量，也在很大程度上取决于其与已见语言的相似性。为求全面，这些分析还应考虑训练数据中可能存在错误标注的情况，[23]这些错误标注已被发现能够解释大语言模型跨语言能力的很大一部分（Blevins 和 Zettlemoyer，2022年）。

5.4.2 跨语言迁移

对于测试的许多语言方向（包括低资源语言方向），只要它们在 ROOTS 语料库中有足够的代表性，单样本结果就很理想。为了更好地理解跨

[22]尽管法语和西班牙语关系密切，且在 ROOTS 中的代表性相当，但两者的分数却大相径庭。我们的初步分析表明，这是因为西班牙语参考译文不如法语参考译文贴近原文，在结构上也与英语原文差异更大。部分示例见附录 E。

[23]在一次私人交流中，N.Muennighoff 估计意大利语约占 ROOTS 语料库的 ∼0.33%，略低于印地语文本所占比例（0.47%）。

源语言↓	目标语言→	ar	en	es	fr	zh
ar	BLOOM	–	40.3	23.3	33.1	17.7
	M2M	–	25.5	16.7	25.7	13.1
en	BLOOM	28.2	–	29.4	45.0	26.7
	M2M	17.9	–	25.6	42.0	19.3
es	BLOOM	18.8	32.7	–	24.8	20.9
	M2M	12.1	25.1	–	29.3	14.9
fr	BLOOM	23.4	45.6	27.5	–	23.2
	M2M	15.4	37.2	25.6	–	17.6
zh	BLOOM	15.0	30.5	20.5	26.0	–
	M2M	11.6	20.9	16.9	24.3	–

（a）高资源语言对。

源语言↓	目标语言→	en	fr	hi	id	vi
en	BLOOM	–	45.0	27.2	39.0	28.5
	M2M	–	42.0	28.1	37.3	35.1
fr	BLOOM	45.6	–	18.5	31.4	32.8
	M2M	37.2	–	22.9	29.1	30.3
hi	BLOOM	35.1	27.6	–	–	–
	M2M	27.9	25.9	–	–	–
id	BLOOM	43.2	30.4	–	–	–
	M2M	33.7	30.8	–	–	–
vi	BLOOM	38.7	26.8	–	–	–
	M2M	29.5	25.8	–	–	–

(b) 高→中资源语言对。

表7：FLORES-101 开发测试集上的单样本机器翻译结果（spBLEU）（原始输出）。

Src↓	Trg→	en	bn	hi	sw	yo
en	BLOOM	–	24.6	27.2	20.5	2.6
	M2M	–	23.0	28.1	26.9	2.2
bn	BLOOM	29.9	–	16.3	–	–
	M2M	22.9	–	21.8	–	–
hi	BLOOM	35.1	23.8	–	–	–
	M2M	27.9	21.8	–	–	–
sw	BLOOM	37.4	–	–	–	1.3
	M2M	30.4	–	–	–	1.3
yo	BLOOM	4.1	–	–	0.9	–
	M2M	4.2	–	–	1.9	–

(a) 低资源语言

Src↓	Trg→	ca	es	fr	gl	it	pt
ca	BLOOM	–	28.9	33.8	19.2	19.8	33.0
	M2M	–	25.2	35.1	33.4	25.5	35.2
es	BLOOM	31.2	–	24.8	23.3	16.5	29.1
	M2M	23.1	–	29.3	27.5	23.9	28.1
fr	BLOOM	37.2	27.5	–	24.9	24.0	38.9
	M2M	28.7	25.6	–	32.8	28.6	37.8
gl	BLOOM	37.5	27.1	33.8	–	18.3	32.2
	M2M	30.1	27.6	37.1	–	26.9	34.8
it	BLOOM	31.0	25.4	31.4	20.2	–	29.2
	M2M	25.2	29.2	34.4	29.2	–	31.5
pt	BLOOM	39.6	28.1	40.3	27.1	20.1	–
	M2M	30.7	26.9	40.2	33.8	28.1	–

（b）罗曼语族

表8：Flores-101 开发测试集上的单样本机器翻译结果（spBLEU）（原始输出）。

1-shot 示例方向类型		原始 spBLEU	原始 COMET	截断 spBLEU	截断 COMET
Same	孟加拉语→英语	29.9	0.444	29.9	0.444
反向	英语→孟加拉语	21.8	0.313	29.4	0.414
相关源语言	印地语→英语	30.1	0.449	30.5	0.460
相关源语言（WMT）	印地语→英语	29.1	0.422	29.1	0.427
HR 非相关源语言	法语→英语	17.2	0.315	29.7	0.396
印地语无关源语言	法语→阿拉伯语	8.4	-0.102	28.0	0.322

表 9：改变单样本示例的语言方向时，Flores 孟加拉语→英语的单样本结果。HR=高资源。

语言 BLOOM 的表现以及单样本机制的运作方式，我们改变少样本示例的语言方向，并以孟加拉语→英语（bn→en）翻译为案例研究。我们随机抽取单样本开发集示例，[24] 比较使用以下来源的单样本示例：(i) 相同方向（bn→en）；(ii) 相反方向（en→bn）；(iii) 源语言相关的语言方向（hi→en）；(iv) 相同的相关方向，但来自不同数据集（WMT 开发集）；(v) 译入相同目标语言的高资源语言方向（fr→en）；以及 (vi) 高资源且语言无关的方向（fr→ar）。

结果（表 9）表明跨语言迁移是可行的，但使用不同的语言方向会影响过度生成和翻译质量。无关的方向法语→阿拉伯语产生的结果最差，过度生成最严重（见原始版本与截断版本之间的分数差异），而且截断后的质量也最差，这表明语言相关性确实发挥作用。使用相反的方向（英语→孟加拉语）或相同的目标语言（法语→英语）时，过度生成仍然是一个问题（尽管程度较轻）。使用相关性更高、资源更丰富的源语言（印地语→英语）可以减少过度生成，也能取得最佳机器翻译结果。不过，使用 Flores-101 示例比使用 WMT 示例效果更好，这表明领域内示例最为理想。

5.5 语言上下文的使用

关于机器翻译中的语言上下文，已有大量研究，例如用于消解词汇歧义文本，或在需要额外信息才能使输出符合语言规范时（例如，将回指代词翻译成要求其与共指对象保持一致关系的语言）(Hardmeier，2012；Libovick´y andHelcl，2017；Bawden et al.，2018年；Voita et al.，2018年；Lopes et al.，2020年；Nayak etal.，2022年）。

[24]所有运行均使用相同的随机种子。

单样本示例			英语→法语		法语→英语	
来源	Dir.	截断	BLEU	COMET	BLEU	COMET
随机	随机	×	5.7	0.342	12.1	0.614
		✓	37.6	0.634	41.4	0.758
前一个	随机	×	6.1	0.328	12.3	0.617
		✓	38.5	0.614	41.6	0.751
前一个	same	×	19.3	0.597	20.7	0.719
		✓	**39.0**	**0.632**	**42.1**	**0.761**
前一个	opp.	×	3.6	0.064	8.6	0.518
		✓	37.8	0.590	41.2	0.742

表10：使用前一句/随机句作为单样本示例时，Di-aBLa 的单样本结果（BLEU）比较（使用 xglm-source+target 提示）。每个语言方向的最佳结果以粗体标出。

我们在 1-shot 设置下测试语言上下文在 DiaBLa 中的有效性（再次使用`xglm-source+`目标），方法是改变 1-shot 示例的来源：（i）随机示例与（ii）前一条对话语句。如果语言上下文有用，我们预计（ii）会有所改善。我们还改变 1-shot 示例的语言方向。默认情况下，鉴于该数据集是双语的，1-shot 示例的方向为英语→法语或法语→英语，与当前示例的方向无关。鉴于第 5.4.2 节中的结果以及表 2a中的较差零样本结果，为了进行公平比较，考虑这一点非常重要。因此，我们将每种类型的上下文（随机/前一条）分别与以下两种情况进行比较：（i）相同的随机方向，以及（ii-iii）与当前示例相同的语言方向和相反的语言方向。我们展示原始输出和截断输出的结果。

结果见表 10。截断有显著帮助；即使在单样本设置下，Bloom 仍难以避免生成过多内容；而在单样本示例中使用相同而非相反的语言方向时，这一问题会大幅减轻。目前尚不清楚使用先前语境而非随机语境是否有帮助：BLEU 更高（38.5 对 37.6），而 COMET 更低（0.328 对 0.342）。这些差异可能是单样本示例选择的随机性所致，使用不同的随机种子可能会得到不同结果。尽管结果并不确凿，但显然，使用先前语境会影响翻译，效果可能更好，也可能更差。证据见附录 19 中的表 F，其中给出了三个这样的示例：（i）先前语境对歧义词*glace*“冰淇淋或镜子”的翻译产生了不利影响，导致选错词义；（ii）使用了前一句中共指的乐器“乐器”；（iii）将代词 they 正确地转换为法语中的性一致形式（elles“她们”，而非 ils“他们”），与阴性共指词 filles“女孩们”相对应。

先前语境，导致选错词义；（ii）使用了前一句中的共指乐器“乐器”；（iii）将代词 they 正确地译为法语中的性一致形式（elles“她们”，而非 ils“他们”），与阴性共指词 filles“女孩们”相对应。

6 结论

我们评估了 BLOOM 在三个数据集和多个语言对上的机器翻译性能。尽管仍存在过度生成和生成错误语言的问题（尤其是在零样本机器翻译中），但在少样本设置下，机器翻译质量显著提高，接近当前最先进的结果。对于某些语言对而言，低资源机器翻译仍然具有挑战性，尽管这些语言包含在训练数据中，这也引发了对何谓 BLOOM 语言的疑问。不过，我们发现了跨语言迁移的证据：既体现在非 BLOOM 语言上，也体现在使用其他语言对的少样本示例时。最后，虽然使用语言上下文并未提升自动评估指标，但有证据表明，模型会考虑话语现象。

致谢

本研究得以完成，离不开 BigScience 社区的共同努力。他们设计、开发并准备了用于训练 BLOOM 的工具和数据集。特别感谢评估工作组成员，尤其是 Niklas Muenninghoff 和 Pawan Sasanka Ammanamanchi，他们为我们生成了部分结果。

本研究获准使用法国国家科学研究中心（CNRS）科学计算与信息技术研究所（IDRIS）的高性能计算资源，该资源由国家大型高性能计算设备（GENCI）根据 2021-AD011011717R1、AD011012254R2、2021- A0101012475 和 2022-AD010614012 配额提供。R. Bawden 的参与部分由其在 PRAIRIE研究所担任的讲席职位资助，该职位由法国国家研究署根据“未来投资”计划资助，项目编号为 ANR-19-P3IA- 0001；其参与还获得了由索邦大学资助的新兴项目 DadaNMT 的支持。

参考文献

Stephen Bach、Victor Sanh、Zheng Xin Yong、Albert Webson、Colin Raffel、Nihal V. Nayak、Abheesht Sharma、Taewoon Kim、M Saiful Bari、Thibault Fevry、Zaid Alyafeai、Manan Dey、Andrea Santilli、Zhiqing Sun、Srulik Ben-david、Canwen Xu、Gun- jan Chhablani、Han Wang、Jaso n Fries、Maged Al- shaibani、Shanya Sharma、Urmish Thakker、Khalid Almubarak、Xiangru Tang、Dragomir Radev、Mike Tian-jian Jiang 和 Alexander Rush。2022年。提示源：自然语言提示的集成开发环境与存储库。载于第60届计算语言学协会年会论文集：系统演示，第93–104页，爱尔兰都柏林。计算语言学协会。

Ankur Bapna、Isaac Caswell、Julia Kreutzer、Orhan Firat、Daan van Esch、Aditya Siddhant、Meng- meng Niu、Pallavi Baljekar、Xavier Garcia、Wolf- gang Macherey、Theresa Breiner、Vera Axelrod、Jason Riesa、Yuan Cao、Mia Xu Chen、Klaus Macherey、Maxim Krikun、Pidong Wang、Alexan- der Gutkin、Apurva Shah、Yanping Huang、Zhifeng Chen、Yo nghui Wu 和 Macduff Hughes。2022年。为未来千种语言构建机器翻译系统。*CoRR*，abs/2205.03983。

Rachel Bawden、Eric Bilinski、Thomas Lavergne 和 Sophie Rosset。2021年。DiaBLa：用于机器翻译的双语自发书面对话语料库。语言资源与评估，55(3):635–660。

Rachel Bawden、Rico Sennrich、Alexandra Birch 和 Barry Haddow。2018年。评估神经机器翻译中的话语现象。载于2018年北美计算语言学协会分会会议论文集：人类语言技术，第1卷（长篇论文），第1304–1313页，美国路易斯安那州新奥尔良。计算语言学协会。

Bin Bi、Chenliang Li、Chen Wu、Ming Yan、Wei Wang、Songfang Huang、Fei Huang 和 Luo Si。2020年。PALM：用于语境条件生成的自编码与自回归语言模型预训练。载于2020年自然语言处理实证方法会议（EMNLP）论文集，第8681–8691 页，线上。计算语言学协会。

Workshop BigScience、Teven Le Scao、Angela Fan、Christopher Akiki、Ellie Pavlick、SuzanaIli´c、Daniel Hesslow、Roman Castagn´e、Alexandra Sasha Luc- cioni、Franc.ois Yvon、Matthias Gall´e、Jonathan Tow、Alexander M. Rush、Stella Biderman、Albert Webson、Pawan Sasanka Ammanamanchi、Thomas Wang、Benoˆıt Sagot、Niklas Muennighoff、Al- bert Villanova del Moral、Olatunji Ruwase、Rachel Bawden、Stas Bekman、Angelina McMillan-Major、Iz Beltagy、Huu Nguyen、Lucile Saulnier、Sam- son Tan、Pedro Ortiz Suarez、Victor Sanh、Hugo Laurenc.on、Yacine Jernite、Julien Launay、Margaret Mitchell、Colin Raffel、Aaron Gokaslan、Adi Simhi、Aitor Soroa、Alham Fikri Aji、Amit Alfassy、Anna Rogers、Ariel Kreisberg Nitzav、Canwen Xu、Chenghao Mou、Chris Emezue、Christopher Klamm、Colin Leong、Daniel van Strien、David Ifeoluwa Adelani、Dragomir Radev、Eduardo Gonz´alez Ponferrada、Efrat Levkovizh、Ethan Kim、Eyal Bar Natan、Francesco De Toni、G´erard Dupont、Germ´an Kruszewski、Giada Pistilli、Hady Elsahar、Hamza Benyamina、Hieu Tran、Ian Yu、Idris Abdulmumin、Isaac Johnson、Itziar Gonzalez-Dios、Javier de la Rosa、Jenny Chim、Jesse Dodge、Jian Zhu、Jonathan Chang、J¨org Frohberg、Joseph Tobing、Joydeep Bhattacharjee、Khalid Almubarak、Kimbo Chen、Kyle Lo、Leandro Von Werra、Leon Weber、Long Phan、Loubna Ben allal、Ludovic Tanguy、Manan Dey、Manuel Romero Mu˜noz、Maraim Masoud、Mar´ıa ˇGrandury、Mario Saˇsko、Max Huang、Maximin Coavoux、Mayank Singh、Mike Tian-Jian Jiang、Minh Chien Vu、Mohammad A. Jauhar、Mustafa Ghaleb、Nishant Subramani、Nora Kassner、Nurulaqilla Khamis、Olivier Nguyen、Omar Espejel、Ona de Gibert、Paulo Villegas、Peter Henderson、Pierre Colombo、Priscilla Amuok、Quentin Lhoest、Rheza Harliman、Rishi Bommasani、Roberto Luis L´opez、Rui Ribeiro、Salomey Osei、Sampo Pyysalo、Sebastian Nagel、Shamik Bose、Shamsuddeen Hassan Muhammad、Shanya Sharma、Shayne Longpre、Somaieh Nikpoor、Stanislav Silberberg、Suhas Pai、Sydney Zink、Tiago Timponi Torrent、Timo Schick、Tristan Thrush、Valentin Danchev、Vassilina Nikoulina、Veronika Laippala、Violette Lepercq、Vrinda Prabhu、Zaid Alyafeai、Zeerak Talat、Arun Raja、Benjamin Heinzerling、Chenglei Si、Davut Emre Taşar、Elizabeth Salesky、Sabrina J. Mielke、Wilson Y. Lee、Abheesht Sharma、Andrea Santilli、Antoine Chaffin、Arnaud Stiegler、Debajyoti Datta、Eliza Szczechla、Gunjan Chhablani、Han Wang、Harshit Pandey、Hendrik Strobelt、Jason Alan Fries、Jos Rozen、Leo Gao、Lintang Sutawika、M Saiful Bari、Maged S. Al-shaibani、Matteo Manica、Nihal Nayak、Ryan Teehan、Samuel Albanie、Sheng Shen、Srulik Ben-David、Stephen H. Bach、Taewoon Kim、Tali Bers、Thibault Fevry、Trishala Neeraj、Urmish Thakker、Vikas Raunak、Xiangru Tang、Zheng-Xin Yong、Zhiqing Sun、Shaked Brody、Yallow Uri、Hadar Tojarieh、Adam Roberts、Hyung Won Chung、Jaesung Tae、Jason Phang、Ofir Press、Conglong Li、Deepak Narayanan、Hatim Bourfoune、Jared Casper、Jeff Rasley、Max Ryabinin、Mayank Mishra、Minjia Zhang、Mohammad Shoeybi、Myriam Peyrounette、Nicolas Patry、Nouamane Tazi、Omar Sanseviero、Patrick von Platen、Pierre Cornette、Pierre François Lavallée、Rémi Lacroix、Samyam Rajbhandari、Sanchit Gandhi、Shaden Smith、Stéphane Requena、Suraj Patil、Tim Dettmers、Ahmed Baruwa、Amanpreet Singh、Anastasia Cheveleva、Anne-Laure Ligozat、Arjun Subramonian、Aurélie Nèveol、Charles Lovering、Dan Garrette、Deepak Tunuguntla、Ehud Reiter、Ekaterina Taktasheva、Ekaterina Voloshina、Eli Bogdanov、Genta Indra Winata、Hailey Schoelkopf、Jan-

Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Mu˜noz Ferrandis, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Pass more, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Cl´ementineFourrier, Daniel Le´onPeri˜n´an, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Ji hyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Siv ɔ ɦaman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc P`amies, Maria A Castillo, Marianna Nezhurina, Mario S¨anger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myung-sun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Th´eo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2022年. BLOOM：一个拥有1760亿参数的开放获取多语言语言模型。*CoRR*，abs/2211.05100。

Terra Blevins 和 Luke Zettlemoyer。2022年。语言污染有助于解释英语预训练模型的跨语言能力。载于2022年自然语言处理实证方法会议论文集，第 3563–3574 页，阿布扎比，阿拉伯联合酋长国。计算语言学协会。

Ondˇrej Bojar、Christian Buck、Christian Federmann、Barry Haddow、Philipp Koehn、Johannes Leveling、Christof Monz、Pavel Pecina、Matt Post、Herve Saint-Amand、Radu Soricut、Lucia Specia 和 Aleˇs Tamchyna。2014年。2014年统计机器翻译研讨会论文集。载于第九届统计机器翻译研讨会论文集，第 12–58 页，美国马里兰州巴尔的摩。计算语言学协会。

Ondˇrej Bojar 和 Aleˇs Tamchyna。2011年。利用单语数据改进翻译模型。载于第六届统计机器翻译研讨会论文集，第 330–336 页，苏格兰爱丁堡。计算语言学协会。

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert- Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever 和 Dario Amodei。2020年。语言模型是少样本学习器。收录于神经信息处理系统进展，第 33 卷，第 1877–1901 页。Curran Associates, Inc.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vin- odkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fe- dus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, AlexanderSpiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankara- narayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Bren- nan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov 和 Noah Fiedel。2022年。Palm：借助 Pathways 扩展语言建模。*CoRR*，abs/2204.02311。

Marta R. Costa-juss`a、James Cross、Onur C¸elebi、Maha Elbayad、Kenneth Heafield、Kevin Heffer-

nan、Elahe Kalbassi、Janice Lam、Daniel Licht、Jean Maillard、Anna Sun、Skyler Wang、Guillaume Wenzek、Al Youngblood、Bapi Akula、Lo¨ıc Bar- rault、Gabriel Mejia Gonzalez、Prangthip Hansanti、John Hoffman、Semarley Jarrett、Kaushik Ram Sadagopan、Dirk Rowe、Shannon Spruit、Chau Tran、Pierre Andrews、Necip Fazil Ayan、Shruti Bhosale、Sergey Edunov、Angela Fan、Cynthia Gao、Vedanuj Goswami、Francisco Guzm´an、Philipp Koehn、Alexandre Mourachko、Christophe Ropers、Safiyyah Saleem、Holger Schwenk 和 Jeff Wang。2022年。不让任何语言掉队：扩展以人为本的机器翻译。*CoRR*，abs/2207.04672。

Daxiang Dong、Hua Wu、Wei He、Dianhai Yu 和 Haifeng Wang。2021年。面向多语言翻译的多任务学习。载于计算语言学协会第53届年会暨第7届自然语言处理国际联合会议论文集（第1卷：长篇论文），第1723–1732页，北京，中国。计算语言学协会。

Angela Fan、Shruti Bhosale、Holger Schwenk、Zhiyi Ma、Ahmed El-Kishky、Siddharth Goyal、Mandeep Baines、Onur Celebi、Guillaume Wenzek、Vishrav Chaudhary、Naman Goyal、Tom Birch、Vitaliy Liptchinsky、Sergey Edunov、Michael Auli 和 Armand Joulin。2021年。超越以英语为中心的多语言机器翻译。机器学习研究期刊，22(107)：1–48。

Orhan Firat、Kyunghyun Cho 和 Yoshua Bengio。2016年。采用共享注意力机制的多向多语言神经机器翻译。载于2016年计算语言学协会北美篇章人类语言技术会议论文集，第866–875页，美国加利福尼亚州圣迭戈。计算语言学协会。

Leo Gao、Jonathan Tow、Stella Biderman、Sid Black、Anthony DiPofi、Charles Foster、Laurence Golding、Jeffrey Hsu、Kyle McDonell、Niklas Muennighoff、Jason Phang、Laria Reynolds、Eric Tang、Anish Thite、Ben Wang、Kevin Wang 和 Andy Zou。2021年。少样本语言模型评估框架。

Xavier Garcia、Yamini Bansal、Colin Cherry、George F. Foster、Maxim Krikun、Fangxiaoyu Feng、Melvin Johnson 和 Orhan Firat。2023年。少样本学习在机器翻译中不合理的有效性。*CoRR*，abs/2302.01398。

Xavier Garcia 和 Orhan Firat。2022年。使用自然语言提示进行机器翻译。*CoRR*，abs/2202.11822。

Naman Goyal、Cynthia Gao、Vishrav Chaudhary、Peng- Jen Chen、Guillaume Wenzek、Da Ju、Sanjana Kr- ishnan、Marc'Aurelio Ranzato、Francisco Guzm´an 和 Angela Fan。2022年。面向低资源语言和多语言机器翻译的 Flores-101 评测基准。计算语言学协会会刊，10：522–538。

Christian Hardmeier。2012。统计机器翻译中的话语：综述与案例研究。《Dis- cours - Revue de linguistique, psycholinguistique et informatique》，（11）。

Amr Hendy、Mohamed Abdelrehim、Amr Sharaf、Vikas Raunak、Mohamed Gabr、Hitokazu Matsushita、Young Jin Kim、Mohamed Afify 和 Hany Hassan Awadalla。2023年。GPT模型的机器翻译表现如何？全面评估。*CoRR*，abs/2302.09210。

Wenxiang Jiao、Wenxuan Wang、Jen-tse Huang、Xing Wang 和 Zhaopeng Tu。2023年。ChatGPT 是一位优秀的译者吗？初步研究。*CoRR*，abs/2301.08745。

Armand Joulin、Edouard Grave、Piotr Bojanowski 和 Tomas Mikolov。2017。高效文本分类的技巧集。载于计算语言学协会欧洲分会第15届会议论文集：第2卷，短篇论文，第427–431页，西班牙瓦伦西亚。计算语言学协会。

Hugo Laurenc¸on、Lucile Saulnier、Thomas Wang、Christopher Akiki、Albert Villanova del Moral、Teven Le Scao、Leandro Von Werra、Chenghao Mou、Eduardo Gonz´alez Ponferrada、Huu Nguyen、J¨org ˇ Frohberg、Mario Sa˘sko、Quentin Lhoest、Angelina McMillan-Major、G´erard Dupont、Stella Biderman、Anna Rogers、Loubna Ben allal、Francesco De Toni、Giada Pistilli、Olivier Nguyen、Somaieh Nikpoor、Maraim Masoud、Pierre Colombo、Javier de la Rosa、Paulo Villegas、Tristan Thrush、Shayne Longpre、Sebastian Nagel、Leon Weber、Manuel Romero Mu˜noz、Jian Zhu、Daniel Van Strien、Zaid Alyafeai、Khalid Almubarak、Vu Minh Chien、Itziar Gonzalez-Dios、Aitor Soroa、Kyle Lo、Manan Dey、Pedro Ortiz Suarez、Aaron Gokaslan、Shamik Bose、David Ife- oluwa Adelani、Long Phan、Hieu Tran、Ian Yu、Suhas Pai、Jenny Chim、Violette Lepercq、Suzana Ilic、Margaret Mitchell、Sasha Luccioni 和 Yacine Jernite。2022年。BigScience ROOTS语料库：一个 1.6TB 的复合多语言数据集。载于第三十六届神经信息处理系统大会数据集与基准赛道。

Brian Lester、Rami Al-Rfou 和 Noah Constant。2021年。参数高效提示调优的规模效应。载于2021年自然语言处理实证方法会议论文集，第3045–3059页，在线出版，地点为多米尼加共和国蓬塔卡纳。计算语言学协会。

Jind˘rich Libovick´y 和 Jind˘rich Helcl。2017 年。多源序列到序列学习的注意力策略。载于计算语言学协会第 55 届年会论文集（第 2 卷：短篇论文），第 196–202 页，温哥华，加拿大。计算语言学协会。

Xi Victoria Lin、Todor Mihaylov、Mikel Artetxe、Tianlu Wang、Shuohui Chen、Daniel Simig、Myle Ott、Na-man Goyal、Shruti Bhosale、Jingfei Du、Ramakanth Pasunuru、Sam Shleifer、Punit Singh Koura、Vishrav Chaudhary、Brian O'Horo、Jeff Wang、Luke Zettle-moyer、Zornitsa Kozareva、Mona Diab、Veselin Stoy-anov 和 Xian Li。2022 年。使用多语言生成式语言模型进行少样本学习。载于2022 年自然语言处理实证方法会议论文集，第 9019–9052 页，阿布扎比，阿拉伯联合酋长国。计算语言学协会。

Pengfei Liu、Weizhe Yuan、Jinlan Fu、Zhengbao Jiang、Hiroaki Hayashi 和 Graham Neubig。2023年。预训练、提示与预测：自然语言处理提示方法的系统综述。ACM计算调查，55(9)。

安´东尼奥 Lopes、M. Amin Farajian、Rachel Bawden、Michael Zhang 和 Andr´eF. T. Martins。2020年。文档级神经机器翻译：系统比较。载于欧洲机器翻译协会第22届年会论文集，第225–234页，里斯本，葡萄牙。欧洲机器翻译协会。

Minh-Thang Luong、Quoc V Le、Ilya Sutskever、Oriol Vinyals 和 Lukasz Kaiser。2016年。多任务序列到序列学习。载于第四届国际学习表征会议论文集，圣胡安，波多黎各。

Bryan McCann、Nitish Shirish Keskar、Caiming Xiong 和 Richard Socher。2018年。自然语言十项全能：将多任务学习视为问答。*CoRR*，abs/1806.08730。

Yasmin Moslem、Rejwanul Haque 和 Andy Way。2023年。借助大型语言模型实现自适应机器翻译。*CoRR*，abs/2301.13294。

Niklas Muennighoff、Thomas Wang、Lintang Sutawika、Adam Roberts、Stella Biderman、Teven Le Scao、M. Saiful Bari、Sheng Shen、Zheng Xin Yong、Hai-ley Schoelkopf、Xiangru Tang、Dragomir Radev、Alham Fikri Aji、Khalid Almubarak、Samuel Al-banie、Zaid Alyafeai、Albert Webson、Edward Raff 和 Colin Raffel。2022年。通过多任务微调实现跨语言泛化。*CoRR*，abs/2211.01786。

Prashanth Nayak、Rejwanul Haque、John D Kelleher 和 Andy Way。2022年。《探究语境对文档级翻译的影响》。《Information》，13(5)：249。

Pedro JavierOrtiz Su´arez、Beno^ıt Sagot 和 Laurent Romary。2019年。在中低资源基础设施上使用异步流水线处理超大规模语料库。载于大型语料库管理挑战研讨会（CMLC-7）论文集，第9–16页，曼海姆。莱布尼茨德语研究所。¨

Long Ouyang、Jeff Wu、Xu Jiang、Diogo Almeida、Carroll L. Wainwright、Pamela Mishkin、Chong Zhang、Sandhini Agarwal、Katarina Slama、Alex Ray、John Schulman、Jacob Hilton、Fraser Kelton、Luke Miller、Maddie Simens、Amanda Askell、Peter Welinder、Paul F. Christiano、Jan Leike 和 Ryan Lowe。2022年。训练语言模型，使其能够遵循人类反馈提供的指令。*CoRR*，abs/2203.02155。

Kishore Papineni、Salim Roukos、Todd Ward 和 Wei- Jing Zhu。2002年。BLEU：一种机器翻译自动评估方法。载于计算语言学协会第40届年会论文集，第311–318页，美国宾夕法尼亚州费城。计算语言学协会。

Fabio Petroni、Tim Rockt¨aschel、Sebastian Riedel、Patrick Lewis、Anton Bakhtin、Yuxiang Wu 和 Alexander Miller。2019年。语言模型是知识库吗？载于2019年自然语言处理实证方法会议暨第九届自然语言处理国际联合会议（EMNLP-IJCNLP）论文集，第2463–2473页，中国香港。计算语言学协会。

Minh Quang Pham、Jitao Xu、Josep Crego、Franc¸ois Yvon 和 Jean Senellart。2020年。神经机器翻译中的启动。载于第五届机器翻译会议论文集，第516–527页，在线。计算语言学协会。

Matt Post。2018年。呼吁清晰报告 BLEU 分数。载于第三届机器翻译会议论文集：研究论文，第186–191页，比利时布鲁塞尔。计算语言学协会。

Alec Radford、Jeff Wu、Rewon Child、David Luan、Dario Amodei 和 Ilya Sutskever。2019年。语言模型是无监督多任务学习者。技术报告。

Colin Raffel、Noam Shazeer、Adam Roberts、Kather- ine Lee、Sharan Narang、Michael Matena、Yanqi Zhou、Wei Li 和 Peter J. Liu。2020年。探索统一文本到文本转换器的迁移学习极限。机器学习研究杂志，21(140)：1–67。

Ricardo Rei、Craig Stewart、Ana C Farinha 和 Alon Lavie。2020年。COMET：用于机器翻译评估的神经框架。载于2020年自然语言处理实证方法会议论文集（EMNLP），第 2685–2702 页，在线出版。计算语言学协会。

Victor Sanh、Albert Webson、Colin Raffel、Stephen Bach、Lintang Sutawika、Zaid Alyafeai、Antoine Chaffin、Arnaud Stiegler、Arun Raja、Manan Dey、M Saiful Bari、Canwen Xu、Urmish Thakker、Shanya Sharma Sharma、Eliza Szczechla、Tae- woon Kim、Gunjan Chhablani、Nihal Nayak、De- bajyoti Datta、Jonathan Chang、Mike Tian-Jian

Jiang、Han Wang、Matteo Manica、Sheng Shen、Zheng Xin Yong、Harshit Pandey、Rachel Bawden、Thomas Wang、Trishala Neeraj、Jos Rozen、Abheesht Sharma、Andrea Santilli、Thibault Fevry、Jason Alan Fries、Ryan Teehan、Teven Le Scao、Stella Biderman、Leo Gao、Thomas Wolf 和 Alexan- der M Rush。2022年。多任务提示训练使零样本任务泛化成为可能。见国际学习表征会议论文集。

Timo Schick 和 Hinrich Sch¨utze。2021年。重要的不只是规模：小型语言模型同样是少样本学习器。载于2021年北美分会计算语言学协会会议论文集：人类语言技术，第2339–2352页，在线出版。计算语言学协会。

Taylor Shin、Yasaman Razeghi、Robert L. Logan IV、Eric Wallace 和 SameerSingh。2020年。AutoPrompt：利用自动生成的提示从语言模型中 eliciting 知识。载于2020年自然语言处理实证方法会议（EMNLP）论文集，第4222–4235页，在线出版。计算语言学协会。

Saleh Soltan、Shankar Ananthakrishnan、Jack FitzGer-ald、Rahul Gupta、Wael Hamza、Haidar Khan、Charith Peris、Stephen Rawls、Andy Rosenbaum、Anna Rumshisky、Chandana Satya Prakash、Mukund Sridhar、Fabian Triefenbach、Apurv Verma、G¨okhanT¨ur 和 Prem Natarajan。2022年。Alexatm 20b：使用大规模多语言序列到序列模型进行少样本学习。*CoRR*，abs/2208.01448。

Antonio Toral、Sheila Castilho、Ke Hu 和 Andy Way。2018年。达成不可能之事？重新评估神经机器翻译达到人类水平的主张。载于第三届机器翻译会议：研究论文集，第113–123页，比利时布鲁塞尔。计算语言学协会。

David Vilar、Markus Freitag、Colin Cherry、Jiaming Luo、Viresh Ratnakar 和 George F. Foster。2022年。使用提示引导 PaLM 进行翻译：策略与性能评估。*CoRR*，abs/2211.09102。

Elena Voita、Pavel Serdyukov、Rico Sennrich 和 Ivan Titov。2018年。具备语境感知能力的神经机器翻译能够学会照应消解。载于计算语言学协会第56届年会论文集（第1卷：长篇论文），第1264–1274页，澳大利亚墨尔本。计算语言学协会。

Jason Wei、Xuezhi Wang、Dale Schuurmans、Maarten Bosma、Brian Ichter、Fei Xia、Ed Chi、Quoc Le 和 Denny Zhou。2022年。思维链提示可激发大语言模型的推理能力。载于第36届神经信息处理系统年会论文集，美国路易斯安那州新奥尔良。

Aohan Zeng、Xiao Liu、Zhengxiao Du、Zihan Wang、Hanyu Lai、Ming Ding、Zhuoyi Yang、Yifan Xu、Wendi Zheng、Xiao Xia、Weng Lam Tam、Zix- uan Ma、Yufei Xue、Jidong Zhai、WenguangChen、Peng Zhang、Yuxiao Dong 和 Jie Tang。2022年。GLM-130B：一种开放式双语预训练模型。*CoRR*，abs/2210.02414。

Biao Zhang、Barry Haddow 和 Alexandra Birch。2023年。使用提示让大型语言模型进行机器翻译：一项案例研究。*CoRR*，abs/2301.07069。

Susan Zhang、Stephen Roller、Naman Goyal、Mikel Artetxe、Moya Chen、Shuohui Chen、ChristopherDewan、Mona Diab、Xian Li、Xi Victoria Lin、Todor Mihaylov、Myle Ott、Sam Shleifer、Kurt Shuster、Daniel Simig、Punit Singh Koura、Anjali Sridhar、Tianlu Wang 和 Luke Zettlemoyer。2022年。OPT：开放式预训练变换器语言模型。*CoRR*，abs/2205.01068。

主要比较的 COMET 结果

表 11 展示了跨数据集和模型比较的 COMET分数。针对带有 BLEU 分数的表 2 所得出的结论在此处同样成立。

	零样本				单样本			
	BLOOM	T0	mT0	OPT	BLOOM	T0	mT0	OPT
WMT 2014								
英语→法语	-0.985	-0.700	0.453	-0.919	0.085	-1.035	-0.015	-0.165
法语→英语	-0.675	0.337	0.567	-0.493	0.448	-0.087	0.250	0.039
英语→印地语	-0.482	-1.819	0.484	-1.525	0.288	-1.733	0.026	-1.460
印地语→英语	-0.387	-1.346	0.514	-1.200	0.378	-1.624	-0.019	-1.290
DiaBLa								
英语→法语	-1.573	-0.528	0.380	-1.762	0.342	-0.585	-0.018	0.123
法语→英语	-1.581	0.228	0.534	-1.507	0.614	-0.032	0.365	0.389
Flores-101								
英语→法语	-1.469	-0.682	0.797	-1.438	0.602	-0.983	0.605	0.130
法语→英语	-1.143	0.499	0.833	-1.008	0.687	-0.081	0.706	0.404
英语→印地语	-0.972	-1.848	1.025	-1.699	0.454	-1.795	0.718	-1.622
印地语→英语	-0.339	-1.391	0.797	-1.493	0.538	-1.264	0.667	-1.263

(a) 原始预测

	零样本				单样本			
	BLOOM	T0	mT0	OPT	BLOOM	T0	mT0	OPT
WMT 2014								
英语→法语	0.434	-0.700	0.452	0.034	0.424	-1.035	-0.017	-0.000
法语→英语	0.604	0.336	0.566	0.534	0.532	-0.090	0.247	0.449
英语→印地语	0.053	-1.819	0.483	-1.491	0.448	-1.733	0.026	-1.460
印地语→英语	0.445	-1.346	0.511	-1.113	0.386	-1.624	-0.022	-1.274
DiaBLa								
英语→法语	0.433	-0.528	0.380	-0.002	0.634	-0.585	-0.023	0.192
法语→英语	0.567	0.228	0.534	0.554	0.758	-0.039	0.356	0.639
Flores-101								
英语→法语	0.182	-0.683	0.793	0.027	0.622	-0.984	0.601	0.180
法语→英语	0.697	0.499	0.831	0.689	0.690	-0.086	0.702	0.594
英语→印地语	-0.608	-1.849	1.025	-1.638	0.461	-1.795	0.718	-1.622
印地语→英语	0.509	-1.391	0.797	-1.166	0.538	-1.264	0.666	-1.251

(b) 截断预测

表11：使用 xglm-source+target 提示对三个数据集的 COMET分数进行比较。

	零样本		单样本		双样本		五样本	
	N	Δ	N	Δ	N	Δ	N	Δ
cs	1	408	-	-	-	-	-	-
de	1	3	2	146	2	-12.5	1	2
en	181	16	32	57	10	73.8	8	92.2
es	1	12	3	89.3	-	-	-	-
fr	2814	7.9	2959	2.1	2989	1.5	2992	1.6
ht	1	57	1	89	-	-	-	-
it	2	4.5	3	13.3	-	-	-	-
nl	1	131	-	-	-	-	-	-
pt	1	146	-	-	-	-	-	-
ms	-	-	1	28	-	-	-	-
ru	-	-	1	16	-	-	-	-
zh	-	-	1	10	-	-	-	-
ca	-	-	-	-	1	198	1	18
uk	-	-	-	-	1	3	1	3

(a) 英语→法语

	零样本		单样本		双样本		五样本	
	N	Δ	N	Δ	N	Δ	N	Δ
en	2954	1	2979	0.8	2988	1	2987	1.3
fr	47	-23.4	22	-1.4	13	1.3	13	-2.2
it	1	3	-	-	2	6	3	5.3
tr	1	-1	1	-1	-	-	-	-
es	-	-	1	1	-	-	-	-

(b) 法语→英语

表12：对于使用 xglm-源语言+目标语言提示的 WMT2014 英语→法语，输出与参考译文在语言识别和长度差异方面的原始数据。对于 0 至 5 个少样本示例，N 表示被识别为各语言的句子数（目标语言所在行（正确）以绿色标示，源语言所在行（众多错误选项之一）以红色标示），而 Δ表示字符数上的长度差异（注意：当预测结果比参考译文长时，该值为负）。

B 错误的语言预测与过度生成

如第5.1节所述，BLOOM 存在的一个问题，尤其是在零样本翻译中，是生成错误的语言。表12 和13 给出了 WMT14 法语↔英语和印地语↔英语翻译方向的语言识别完整分析，包括原始数据。对于 0 至 5 个少样本示例，我们列出了被识别为各语言的截断输出数量（各语言以行表示）；正确语言（目标语言）以绿色标示，源语言（因此为错误语言）以红色标示。我们还提供了 BLOOM 输出与参考译文之间的平均长度差（Δ）（负数表示预测比参考译文更长）。

对于零样本翻译，相当多的示例被归类为使用源语言的示例——对于英语→法语是如此，对于英语→印地语则更为明显（几乎五分之一的输出使用了错误的语言）。随着所用少样本示例数量的增加，这两个问题都显著减少；当示例数量达到 5 个时，所有语言对和翻译方向中的这些问题几乎都消失了。

C 按模型分析

本节补充了第 5.2 节的结果，分别通过表 14 和 15 报告法语↔英语和印地语↔英语在未进行截断时的结果。正如预期，系统的排名与模型规模一致。对于法语–英语，使用第二大的模型 BLOOM-7b1 进行单样本提示，已经可以取得不错的性能。使用该模型，甚至是规模只有其一半的模型，也能很好地反映提示的性能，并可作为可靠的测试平台。对于英语↔印地语，即使使用大型 BLOOM，结果也不太理想；

	零样本		单样本		双样本		五样本	
	N	Δ	N	Δ	N	Δ	N	Δ
ceb	1	-150	-	-	-	-	-	-
en	476	10.5	48	12.4	71	13.9	26	18.8
eo	1	-134	-	-	-	-	-	-
fi	1	19	-	-	-	-	-	-
fr	2	94.5	-	-	-	-	-	-
gom	2	6.5	1	4	-	-	1	0
hi	1998	9.3	2431	6	2403	5.5	2457	5.5
hsb	1	98	-	-	-	-	-	-
ht	2	147	6	257.5	11	135.3	1	158
hu	1	71	-	-	-	-	-	-
lv	3	63.3	-	-	-	-	-	-
mr	5	64.4	11	14.6	17	11.7	19	6
ne	5	7.6	9	28.2	4	16.8	3	8.3
nl	2	-13.5	-	-	-	-	-	-
pt	1	24	-	-	-	-	-	-
sa	1	-25	-	-	-	-	-	-
sw	1	12	-	-	-	-	-	-
tl	1	24	-	-	-	-	-	-
war	3	3	-	-	-	-	-	-
vec	-	-	1	-38	-	-	-	-
new	-	-	-	-	1	25	-	-

(a) 英语→印地语

	零样本		单样本		双样本		五样本	
	N	Δ	N	Δ	N	Δ	N	Δ
en	2469	4	2499	5.1	2503	3.8	2498	3
fr	1	151	1	-5	-	-	1	8
hi	29	3.3	2	0	-	-	-	-
ht	6	199.8	-	-	-	-	-	-
it	1	139	-	-	1	-18	3	4.3
nl	1	9	-	-	-	-	2	-3
id	-	-	1	-6	-	-	-	-
nds	-	-	1	16	-	-	-	-
pl	-	-	1	-14	-	-	-	-
tr	-	-	1	-15	-	-	-	-
war	-	-	1	344	-	-	-	-
de	-	-	-	-	1	-15	1	188
es	-	-	-	-	1	2	-	-
la	-	-	-	-	1	17	-	-
fi	-	-	-	-	-	-	1	-1
pt	-	-	-	-	-	-	1	1

（b）印地语→英语

表13：使用 xglm-源语言+目标语言提示，对 WMT2014 英语→印地语的输出进行语言识别和与参考译文相比的长度差异的原始数据。对于 0–5 个少样本示例，N 表示被识别为每种语言的句子数量（目标语言所在行（正确）以绿色标示，源语言所在行（众多错误选项之一）以红色标示），而 Δ 表示以字符数计的长度差异（注意，当预测比参考译文更长时，该值为负）。

对于这一语言对，我们甚至观察到所有模型在单样本设置下，不同提示之间的结果差异都很大（从得分范围可以看出这一点）。

D 按提示分析

在本节中，我们复现第5.3节的分析，并在表格16 和17中报告每种提示下截断输出的结果。总体而言，结论与正文中关于未截断输出的报告一致。值得注意的是，截断输出后，xglm-来源+目标在各方面都取得了很好的结果，表现优于最接近的竞争者a good translation-来源+目标和版本-目标在几乎所有配置中都是如此。不过，提示的选择似乎影响更大：（a）在零样本设置中；（b）从英语译出时。相反，我们最稳定的结果出现在法语到英语、单样本的设置中。

E Flores 101中的翻译差异

正文（第5.4.1节）报告了一个显著现象：在 Flores-101 实验中，法语和西班牙语之间存在差异。这出乎意料，因为这两种语言在训练数据中都有充分的代表性。然而，从英语译出和译入英语时，spBLEU 分数的差距都很大；与其他罗曼语族语言相比，也存在明显差距。另一个相关问题是法语和西班牙语之间的翻译质量较差，甚至没有法语→阿拉伯语翻译得好。查看一些输出示例后，我们认为这可能是西班牙语译文的特殊性所致：与法语译文相比，西班牙语译文似乎不那么直译，但译成英语后质量同样出色。比较这两种语言译回英语的译文时（见表18中的一个随机子集），这一点显而易见：最后一个例子尤其清楚，原始英语和法语译文中都是“34 percent”，而西班牙语译文则以“one third”开头。

F DiaBLa语境使用示例

表格19 展示了单样本示例中，前置语境对当前预测产生正面、负面或中性影响的示例，说明单样本示例的选择十分重要，且模型会将其纳入考量。

模型 / 方向	零样本		单样本	
	英语→法语	法语→英语	英语→法语	法语→英语
BLOOM	<风格 印度尼西亚语 ='1'>11.2 </风格>3.0 – 22.0	<风格 印度尼西亚语='1'>15.4 </风格>10.3 – 26.8	**32.6** 27.8 – 36.4	<风格 印度尼西亚语='1'>34.9 </风格>33.1 – 36.6
BLOOM-7b1	6.5 1.5 – 12.1	12.8 4.8– 25.1	25.9 20.8 – 29.9	29.1 25.4 – 32.5
BLOOM-3b	3.6 1.2– 9.6	10.6 2.8– 19.3	21.6 16.7 – 26.8	25.7 18.6 – 29.6
BLOOM-1b1	1.7 0.5– 3.9	7.1 0.7– 11.4	10.1 6.3– 13.2	16.1 12.2 – 19.9
BLOOM-560m	0.6 0.4– 0.9	3.7 1.4– 5.4	3.6 2.2– 4.4	8.6 5.8– 12.1

表14：WMT14 英语↔法语（原始输出）中不同规模模型的 BLEU 分数平均值、最小值和最大值。各设置下的最佳平均结果以粗体显示。

模型 / 方向	零样本		单样本	
	英语→印地语	印地语→英语	英语→印地语	印地语→英语
BLOOM	<风格 id='1'>2.1 </风格 >0.3 – 6.8	**8.3** 0.7 – 13.0	**12.9** 6.5 – 14.6	<风格 印度尼西亚语='1'>19.8 </风格>10.0 – 25.8
BLOOM-7b1	0.1 0.1 – 3.0	5.7 0.3– 9.5	5.9 0.3 – 10.4	12.4 1.0– 17.5
BLOOM-3b	0.2 0.0 – 0.5	3.6 0.0– 7.0	4.9 0.2– 7.2	8.9 0.1– 13.5
BLOOM-1b1	0.1 0.0 – 0.1	1.5 0.0– 4.5	1.4 0.1– 3.1	4.6 0.00– 8.2
BLOOM-560m	0.1 0.0 – 0.1	0.8 0.0– 1.7	0.2 0.0– 0.3	1.5 0.1– 2.8

表15：WMT14 英语↔印地语（原始输出）中，各模型按尺寸递减排列的 BLEU 分数平均值、最小值和最大值。每种设置下的最佳平均值结果以粗体显示。

提示 / 少样本 #	英语→法语		法语→英语	
	0	1	0	1
a_good_translation-源+目标	8.5 0.7–17.0	19.1 4.32–37.12	16.4 7.5–22.2	26.0 12.0–37.0
a_good_translation-翻译目标语言	4.6 0.6–13.9	20.9 3.4–36.8	21.7 6.6–35.2	26.31 12.5–36.9
GPT-3目标	4.0 0.7–14.0	18.7 3.0–36.4	8.3 1.3–25.7	21.6 7.2–37.2
翻译为目标 _ 版本目标	6.4 0.6–10.1	18.1 3.5–33.1	11.5 2.3–20.4	22.9 8.2–35.7
xglm-源+目标	9.7 0.7–30.3	21.9 4.4–36.7	22.2 4.7–35.2	25.3 8.0–37.2
xglm-target	**17.2** 1.33–32.2	<风格 id='1'>23.2 </风格 >5.0–36.3	**25.6** 8.3–37.2	<风格 印度尼西亚语 ='1'>26.7 </风格>11.1–38.2
	2.5 1.1–4.6	20.1 6.8–33.1	11.0 4.5–17.6	23.1 10.4–36.4

表16：WMT14 英语↔法语（截断输出）中每个提示的平均值、最小值和最大值 BLEU分数。每种设置下的最佳平均值结果以粗体显示。

提示 / 少样本 #	英语→印地语		印地语→英语	
	0	1	0	1
a_good_translation-源语言+目标	1.2 0.1–3.3	5.8 0.3–14.5	6.2 1.0–12.7	13.0 2.6–24.4
a_good_translation-翻译目标语言	0.4 0.1–1.3	5.5 0.3–14.1	10.8 1.1–25.4	13.2 2.7–24.7
GPT-3目标	0.0 0.0–0.1	1.6 0.0–7.6	0.0 0.0–0.0	2.5 0.0–11.4
版本目标	1.0 0.1–3.0	5.5 0.2–13.9	<风格 id='1'>11.3 </风格 >2.4–21.4	**13.5** 2.7–25.7
xglm-源+目标	<风格 印度尼西亚语 ='1'>3.9 </风格>0.1–12.1	**7.3** 0.2–15.8	8.8 0.9–24.3	12.4 1.2–25.0
xglm-target	0.3 0.0–1.0	5.1 0.0–14.5	2.1 0.3–5.8	6.5 0.1–13.0

表17：WMT14 英语↔印地语（截断输出）中每个提示的平均值、最小值和最大值 BLEU分数。每种设置下的最佳平均结果以粗体显示。

有关这些实验的更多细节，请参见正文中的第
5.5节。

en	它们在白天比周围的表面更凉爽，在夜晚则更温暖。
fr→英语	“它们在白天比周围的表面更凉爽，在夜晚则更温暖”。
es→英语	白天，其温度低于周围表面的温度；夜晚则更高。
en	“这不会是再见。这是一个篇章的结束，也是新篇章的开启。”
法语→英语	“这不是再见。这是一页正在翻过，而另一页正在开启。”
es→英语	”这不是告别；这只是一个篇章的结束，也是另一个篇章的开始”。
en	“我们现在有4个月大的小鼠，它们曾经患有糖尿病，如今已经不再患糖尿病 ,”他补充道。
法语→英语	“他补充说：‘我们现在有四个月大的小鼠，它们没有患糖尿病，而以前它们患有糖尿病。’
西班牙语→英语	“目前，我们有四个月大的小鼠，它们过去患有糖尿病，但现在已经不患糖尿病了，”他补充道。
en	“我们将努力在2020年之前，将单位国内生产总值的二氧化碳排放量较2005年水平，”胡说。
法语→英语	“我们将努力使2020年单位国内生产总值的二氧化碳排放量较2005年水平，”胡先生说。
西班牙语→英语	胡表示：“我们将努力降低单位国内生产总值的二氧化碳排放量，到2020年，确保与2005年相比差异显著。”
en	科学家称，这种动物的羽毛顶部呈栗棕色，腹部则呈浅色或类胡萝卜素色。
法语→英语	科学家称，这种动物的羽毛顶部呈栗棕色，腹部则呈浅色或类胡萝卜素色。
西班牙语→英语	据专家介绍，这种动物上半身的羽毛呈棕色，下半身则呈浅色或类胡萝卜素色。
en	34%的受访者持这一观点，希望伊丽莎白二世女王成为澳大利亚最后一位君主。
法语→英语	34%的受访者持这一观点，并希望伊丽莎白二世女王成为统治澳大利亚的最后一位君主。
西班牙语-英语	三分之一的受访者持这一观点，并希望最后一位女王是伊丽莎白二世女王。

表18：从 Flores-101 示例中随机抽取一部分，使用 BLOOM 将法语和西班牙语译成英语（注：这些句子的原始语言是英语）。每组三个句子包含原始英语句子，以及自动生成的法语→英语和西班牙语→英语译文。

单样本来源	语境	参考文献	预测
前一个，相同	法语：你是否一直在想 `可以吃的冰淇淋？ = 英语：你是否一直在想可以吃的冰淇淋？ ### 法语：或者 `一块冰淇淋用来照镜子？ = 英语：		*或者一块冰淇淋来看吗？
前一个相反项	英语：你是否总是在想被吃掉的冰淇淋？ = 法语：Pensez vous en permanence `a la glace qui se mange ? ### 法语：ou bien `a une glace pour se regarder ？ =英语：	或者一面用来看的镜子进去？	还是照照镜子？
随机	法语：尽管如此，我们生活在一个 `疯狂至极的时代！ =英语：不过，我们生活在一个多么疯狂、疯狂的时代啊！ ### 法语：或者 `照照镜子？ =英语：		还是照照镜子你自己？
前一个。相同	英语：你想要什么类型的乐器？ = 法语：你在想哪种 `乐器乐器？ ### 英语：我想是木制的..... = 法语：		我想是一件木制乐器，.....
前一个相反	法语：你在想 `哪种乐器？ = 英语：你想的是哪种乐器呢？ ### 英语：我想应该是木制的..... =法语：	一种 en 木制乐器，我想是这样。	一种木制乐器，我假设.....
随机	法语：Ils vont vous changer les id´ees apr`es votre dure journ´ee ！ = 英语：他们会帮你在辛苦工作一天后转移注意力！ ### 英语：我想要一个木制的假设 ..... =法语：		一个木头的，我想.....
前一个，相同	英语：他给我演示了它的工作原理，不过如果我遇到困难，这里的姑娘们总会帮助我。 = 法语：他向我展示了 ´它是如何运作的，但如果我做不到，这些这里的姑娘们会帮助我，毫无问题。### 英语：她们非常善良。 = 法语：		她们非常友善。
前一个，相反	法语：他向我展示了它是如何运作的，但如果我做不到，这里的女孩们会毫不犹豫地帮助我。 = 英语：他告诉我它是如何运作的，但如果我遇到困难这里的女孩们总会帮助我。### 英语：她们非常友善。 = 法语：	elles sont tr` es 友善。	Elles 很友善。
随机	英语：我不了解贷款。 = 法语：我不知道贷款的事。### 英语：他们非常友善。 = 法语：		*ils 非常友善。

表19：具有不同单样本上下文的歧义 DiaBLa 示例。与歧义相关的词语已加下划线，错误译文以星号标记。