arXiv:2305.01645v3 [cs.CL] 5 Jul 2023

蒸馏还是标注？紧凑模型的高性价比微调

Junmo Kang、Wei Xu、Alan Ritter佐
治亚理工学院 junmo.kang@gatech.edu
{wei.xu, alan.ritter}@cc.gatech.edu

摘要

对大模型进行微调非常有效，但推理成本高昂，并会产生碳排放。知识蒸馏已被证明是一种降低推理成本的实用方案，但蒸馏过程本身需要大量计算资源。与其购买或租用 GPU 来微调并蒸馏大模型，自然语言处理从业者也可以选择将可用预算用于聘请标注人员，手动标注额外的微调数据。本文探讨如何最高效地利用固定预算来构建紧凑模型。通过在六项不同任务上进行大量实验，我们表明，相比标注更多数据以直接训练紧凑模型（T5-Small），将 T5-XXL（11B）蒸馏为 T5-Small（60M）几乎总是一种成本高效的策略。我们还进一步研究了在不同场景下，分配给计算的最优预算如何变化。我们将公开代码、数据集、标注成本估算和基线模型，作为基准测试，以支持后续关于紧凑模型成本高效训练的研究。

1 引言

增大预训练模型的尺寸，可以持续提升其微调后在下游任务上的表现，这一点已在基于 BERT（Devlin 等人，2019）、RoBERTa（Liu 等人，2019）、BART（Lewis 等人，2019）、T5（Raffel 等人，2020）的研究，以及有关经验缩放定律的工作中得到证实（Brown 等，2020；Lester 等人，2021；Hernandez 等人，2021）。然而，使用大模型进行推理成本高昂，也会增加碳排放（Patterson 等人，2021）。为了解决这一问题，研究人员探索了通过知识蒸馏等技术压缩大模型的方法（Hinton 等，2015；Sanh 等人，2019；Gou 等人，2021），这种方法能有效降低推理成本（Magister 等人，2022）。

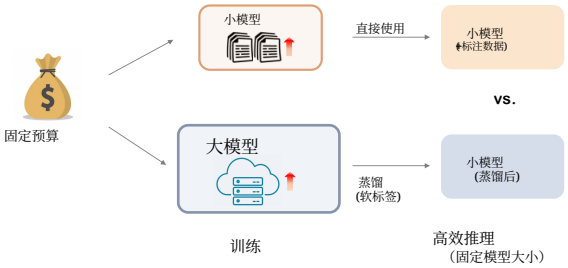


图 1：构建紧凑型固定尺寸模型的两种实用策略示意图。在固定预算和少量初始标注数据的条件下，（i）可以标注更多数据，直接对小模型进行微调。（ii）或者，也可以利用具有更多计算资源的大模型，将其知识蒸馏到小模型中，以实现高效推理。

并提升较小学生模型的泛化能力（Stanton 等人，2021年）。然而，蒸馏过程本身仍需要大量的计算、内存和存储资源（Xia 等人，2022年）。

除了压缩模型之外，另一种在不增加推理成本的情况下提升性能的方法，是直接为微调标注额外数据。近期研究表明，额外的数百个标签有时能够带来比数十亿个额外模型参数更好的性能（Kirstain 等人，2022年）。这引出了一个问题：如何最有效地利用固定预算来训练一个支持高效推理且性能最大化的紧凑模型。一种选择是使用可用预算雇用标注人员来标注额外数据，并直接对小模型进行微调。或者，也可以将预算用于购买或租用 GPU，以对大型教师模型进行微调和知识蒸馏（见图 1）。

本文采用消费者选择理论（Becker，1965；Lancaster，1966年；Baietal.，2021年）来探究蒸馏何时是一种经济高效的模型压缩策略。基于广泛的实证分析，

数据集	Task	#训练	$/标签	总计 $
WLP (Tabassum 等人，2020年)	命名实体识别	11，966	$0.260	$3,111
STANCEOSAURUS (Zheng 等人,2022年)	立场分类	12，130	$0。364	$4,415
FEVER (Thorne 等人,2018)	事实核查	104，966	$0.129	$13，544
多任务PI$_{Id}$ （Dou 等人，2022年）	释义识别	92，217	$0.200	$18，443
MULTIPIT$_{Gen}$ (Dou 等人，2022年)	释义生成	49，673	$0。371	$18，443
NATURAL QUESTIONS (Kwiatkowski 等人，2019年)	问答	87,372	$0.129	$11，271

表 1：各种自然语言处理数据集/任务的数据标注成本。

我们就如何分配用于人工标注和计算资源的固定预算，以训练紧凑模型提出建议。在六项自然语言处理任务上的实验表明，与标注相比，利用未标注数据进行蒸馏几乎总是提升紧凑型模型性能的一种经济高效策略（见表2）。此外，我们的分析显示，随着可用已标记数据增多，分配给蒸馏的最优预算比例也会提高（见§4.1和图2）。预算较少时，采用较少的未标注数据进行蒸馏，同时增加已标记数据的数量，是帕累托最优的（Abdolrashidi 等人，2021年；Treviso 等人，2022年），因为这样可以训练出知识更丰富的教师模型。预算增加后，使用更大的未标注数据集进行蒸馏就变得更经济，因为教师模型的表现会显著优于学生模型。最后，我们研究了使用GPT-3.5 （Ouyang 等人，2022年）（图6）进行数据标注的成本效益。我们发现，尽管GPT-3.5 比人工标注员便宜，但微调T5-XXL ，然后蒸馏出小模型，比直接用来自GPT-3.5的伪标签微调小模型更经济高效。

我们将公开代码、数据集、标注成本估算和基线模型，作为基准，以支持后续关于紧凑模型成本效益训练的研究。

2 研究设计

本节首先介绍我们如何为成本效益研究构建问题模型（§2.1）。随后，我们比较两种策略（§2.2 & 2.3），这些策略在构建紧凑模型时，计算成本与人工标注成本所占比例各不相同。最后，我们说明如何估算这两种策略涉及的标注成本（§2.4）和计算成本（§2.5）。

2.1 问题表述与假设

本研究的重点是在固定预算下公平评估这两种方法（§2.2 &§2.3）。在面临财务约束时，实践者可能需要权衡将资金用于数据还是计算资源；我们通过实证研究考察二者之间的权衡，以最大化由此产生的效用。为开展广泛的研究，我们利用各种现有众包数据集来模拟数据标注过程，并使用按小时计费的云端 GPU 租赁服务。

我们假设自然语言处理工程师的薪资属于固定成本，因此，其用于构建模型和/或管理标注人员团队以进行数据标注的时间，不计入总成本的计算因素。我们仅考虑人工数据标注和 GPU 计算的直接成本。我们初步假设不存在可免费使用的特定任务已标记数据，但确实假设公开可用的预训练模型（例如 T5 (Raffel 等人，2020年）的使用成本为零。

2.2 策略 1：直接利用标注构建紧凑模型（Ann.）

该策略直接对紧凑模型（例如 T5-Small（60M））进行微调，将全部预算用于人工标注。这被认为是实践者在训练紧凑模型时会选择的最直接的方法。

具体而言，在给定预算约束的情况下，我们准备能够利用预算进行最大限度标注的数据，并在所有任务的统一文本到文本框架下，基于这些数据训练 T5 (Raffel 等人，2020年)，以最大化给定输入文本 Y时目标文本 X的似然。每项任务的输入 X 及其对应目标 Y的格式详见附录B。

请注意，与该策略相关的主要成本是标注成本。虽然构建这个直接模型的总成本可能包括微调成本（即计算成本），但我们

发现大多数情况下这部分成本可以忽略不计，因此为简化起见，除非另有说明，否则不予考虑。[1]

2.3 策略2：从更大的模型蒸馏（Dist.）

作为增加标注数据的另一种选择，可以将部分预算用于计算，在较少量的数据上训练更大的模型（例如 T5-XXL（11B））。随后可对大模型进行蒸馏，得到最终的紧凑模型，同时支持高效推理。

遵循近期研究（Xia 等人，2022年；Zhou 等人，2022b），我们的研究主要关注特定任务的模型压缩，而非通用蒸馏（Sanh 等人，2019年），[2] 不过，我们在附录 F 中分析了通用蒸馏与特定任务蒸馏。通用蒸馏需要大量计算资源；此外，特定任务蒸馏与通用蒸馏可以互补地结合使用（Jiao 等人，2020年）。

值得注意的是，即使采用策略 2，训练大型教师模型也需要标注数据。因此，我们假设初始时有一定数量的（N）数据通过花费部分预算完成标注，并以与 §2.2 中相同的方式使用这些数据微调较大的模型。之后，通过蒸馏较大模型（即教师模型）的知识来训练一个小模型（即学生模型）（Hinton 等人，2015年），其中，将教师模型在给定源输入时对目标序列的概率分布用作软标签。我们采用用于比较两个分布的 KL 散度损失，使学生模型的分布 P_S 遵循教师模型的输出分布 P_T，并针对特定任务的未标注数据[3]：

$$D_{KL}(P_T||P_S) = \sum_{v \in V} P_T(v) \log \frac{P_T(v)}{P_S(v)} \quad (1)$$

其中，V 是词汇空间。为简洁起见，上文省略了用于生成概率的输入词元和目标词元。

总成本包括N（初始标注的训练样本数量）的初始成本，以及微调大模型并将其蒸馏为紧凑模型的计算成本。

2.4 数据标注的成本估算

本研究考虑了表1所示的六项多样且实用的自然语言处理任务。我们根据相关文学中的提及（如有）、与数据集创建者的沟通，或 Google Cloud 数据标注服务的价格，估算每个数据集的标注成本，遵循 Wang 等人 (2021年)[4]。附录 A中提供了每个数据集成本估算的详细说明。

2.5 计算成本估算

本文假设从 Google Cloud 租用计算资源进行模型训练。我们具体考虑使用 NVIDIA A100 GPU，每块配备 40GB 显存，以容纳一个大模型（例如，拥有 11B 个参数的模型）。虚拟机和存储的费用约为 $3.75 每 GPU 小时。对于大规模研究，我们利用自有资源，即 A40 GPU。基准测试结果[5] 以及比较训练时间的初步实验表明，A40 GPU 的速度约为 A100 GPU 的一半。因此，我们估算计算成本为$1.875 每 GPU 小时。这是从业者实际需要支付的价格，不同于浮点运算次数等理论指标，因为这些指标无法反映实际运行时间（Xu 和 McAuley，2022年）和成本。附录中提供了构建紧凑模型的成本估算示例明细（表 6）。

3 在固定预算下评估标注和蒸馏

在表 2中，我们在不同预算下评估了六项不同任务的两种策略。首先，我们设定N，即用一笔初始金额标注的数据量。给定固定预算后，我们可以选择为标注更多数据采用标注（Ann.）策略；也可以为GPU 小时投入更多资源，并使用更多未标注数据采用蒸馏（Dist.）策略。

在我们的主要研究中，我们将 T5-Small（60M）视为紧凑模型，将 T5-XXL（11B）视为教师模型。所有模型均基于

[1]微调 T5-Small （6000 万）在 5K 数据上进行训练，例如，仅需不到半小时；根据第 §2.5 节中的计算成本，费用约为 1 美元。[2]在通用蒸馏中，会先对预训练模型进行蒸馏，然后再进行微调，例如 DistillBERT。[3]例如，对于释义生成任务，没有对应目标释义句的源句。有关未标注数据的详细信息，请参阅附录 D 。

[4]https://cloud.google.com/ai-platform/data-labeling/pricing#labeling_成本[5] https://lambdalabs.com/blog/nvidia-rtx-a40-benchmarks

Task	N (初始金额)	策略	额外金额				
			+$0	+$100	+$200	+$300	+$500
WLP	1K ($260)	T5-Small (Ann.)	**40.7**	50.0	53.7	57.8	62.7
		T5-XXL [72。4] ⇒ T5-Small（蒸馏版）	N/A	**71.1**	**71.3**	**70.9**	**70.8**
	5K ($1300)	T5-Small（Ann.）	**67.4**	**68.2**	68.6	68.7	69.3
		T5-XXL [74。2] ⇒ T5-Small（蒸馏版）	N/A	65.3	**71.8**	**72.4**	**72.5**
			+$0	+$100	+$150	+$200	+$300
STANCEO-SAURUS	1K ($364)	T5-Small（Ann.）	**37.5**	45.4	45.5	45.5	44.7
		T5-XXL [62。5] ⇒ T5-Small（蒸馏版）	N/A	**54.2**	**54.6**	**56.3**	**56.9**
	5K ($1820)	T5-Small（Ann.）	**49.4**	50.7	52.6	49.1	50.3
		T5-XXL [69。6] ⇒ T5-Small（蒸馏版）	N/A	**52.4**	**55.4**	**56.2**	**60.5**
			+$0	+$50	+$75	+$100	+$150
FEVER	1K ($129)	T5-Small (Ann.)	**49.7**	49.7	49.7	49.7	49.8
		T5-XXL [73。5] ⇒ T5-Small（蒸馏版）	N/A	**71.3**	**71.1**	**71.6**	**71.7**
	5K ($645)	T5-Small (Ann.)	**67.2**	68.2	68.1	68.1	68.9
		T5-XXL [78。0] ⇒ T5-Small（蒸馏版）	N/A	**73.4**	**74.1**	**74.3**	**74.8**
			+$0	+$100	+$150	+$200	+$300
MULTIPIT$_{Id}$	1K ($200)	T5-Small (Ann.)	**53.0**	53.1	53.1	54.6	54.2
		T5-XXL [79。9] ⇒ T5-Small（蒸馏版）	N/A	**79.1**	**78.3**	**78.8**	**77.9**
	5K ($1000)	T5-Small（Ann.）	**78.0**	77.4	77.0	78.1	77.8
		T5-XXL [84。5] ⇒ T5-Small（蒸馏版）	N/A	**80.6**	**80.5**	**81.1**	**81.9**
			+$0	+$100	+$150	+$200	+$300
MULTIPIT$_{Gen}$	1K ($371)	T5-Small（Ann.）	**56.8**	57.7	58.9	59.2	59.3
		T5-XXL [67。4] ⇒ T5-Small（蒸馏版）	N/A	**60.3**	**62.1**	**62.0**	**62.6**
	10K ($3710)	T5-Small（Ann.）	**68.6**	**68.6**	68.6	68.6	68.7
		T5-XXL [74。8] ⇒ T5-Small（蒸馏版）	N/A	68.4	**72.1**	**73.7**	**74.0**
			+$0	+$50	+$75	+$100	+$150
自然问题	1K ($129)	T5-Small（Ann.）	**3.5**	4.1	4.2	4.5	5.0
		T5-XXL [21。9] ⇒ T5-Small（蒸馏版）	N/A	**11.3**	**11.8**	**13.0**	**13.5**
	10K ($1290)	T5-Small（Ann.）	**9.8**	**10.2**	9.9	10.4	10.3
		T5-XXL [26。1] ⇒ T5-Small（蒸馏版）	N/A	N/A	**12.0**	**16.3**	**18.0**

表 2：在各种自然语言处理任务上，增加 数据标注（Ann.）和教师模型[**性能**] ⇒ 学生模型蒸馏（Dist.）对小模型成本效益的主要结果。 N 表示使用相应（初始金额）标注的初[illegible]数据量。（额外数据）表示除 N 之外的额外标注数据量．而（*GPU* 小时）表[illegible]数据上微调教师模型所用的 GPU 总小时数，加上使用不同（#未标注数据）将知识蒸馏到小模型所用的 GPU 小时数。若因成本原因无法构建模型，则使用 N/A。

上 T5 v1.1 (Roberts 等人, 2020年），其仅通过无监督方式进行预训练，与原版T5 (Raffel 等人, 2020年）不同。

对于 FEVER 和 NATURAL QUESTIONS，分别遵循 Lee 等人（2020 年）和 Roberts 等人（2020 年）的做法，我们采用闭卷设置，要求模型仅依赖其参数化知识。由于测试集不公开，我们报告开发集上的表现。衡量表现时，FEVER 和 MULTIPIT$_{Id}$ 使用准确率，WLP、STANCEOSAURUS 和 NATURAL QUESTIONS 使用 F1，MULTIPIT$_{Gen}$ 则使用 BERT-iBLEU（Niu 等人，2021 年）（即 self-BLEU 与 BERTSCORE（Zhang 等人，2020 年）的调和平均值）。实验设置的更多细节见附录C。

3.1 标注与蒸馏

在表2中，我们有趣地观察到，蒸馏（Dist.）策略显著优于标注（Ann）策略，在几乎所有任务的几乎所有情况下都是如此。尽管知识蒸馏（Hinton 等人，2015年）已在以往研究中被证明有助于压缩和提升泛化能力（Sanh 等人，2019年；Kang等人，2020年；Le等人，2022年），我们考虑构建模型所涉及的现实成本后得出的结果相当令人意外，并凸显了一个新的方面：这种方法具有成本效益。换言之，这表明，考虑到成本效益，仅靠聘请人工标注员来扩大数据规模或许并非良策。

请注意，Dist. 需要先在N 个已标记数据上[illegible]于微调，这需要相当大的计算成本，因此，如果微调成本超出给定预算，我们将此类情况记为 N/A。在这种情况下，Ann. 本质上是正确的选择。我们还注意到，在预算有限时，某些情况下 Ann. 是更好的选择。例如，Ann. 在 WLP 上以 $100 击败了其对应方法

模型（教师模型 ⇒ 学生模型）	WLP	STANCEOSAURUS	FEVER	MULTIPIT$_{Id}$	MULTIPIT$_{Gen}$	自然问题
T5-Small ⇒T5-Small（自蒸馏）	65.2 [67.4]	50.3 [50.5]	67.6 [67.2]	77.1 [78.0]	66.1 [68.1]	3.8 [9.8]
T5-XXL ⇒T5-Small（蒸馏版）	70.6 [74.2]	58.9 [69.6]	74.2 [78.0]	80.9 [84.5]	73.8 [74.8]	17.8 [26.1]

表 3：使用相同数量未标注数据（*100K*）进行自蒸馏和蒸馏的结果。 [ ]中的数字表示使用*5K* 标注数据训练的教师模型的性能。

模型	WLP		STANCEOSAURUS		FEVER		MULTIPIT$_{Id}$		MULTIPIT$_{Gen}$		自然问题	
T5-XXL ⇒T5-Small（蒸馏版）	70.6	($502)	**58.9**	($279)	74.2	($101)	80.9	($161)	**73.8**	($245)	17.8	($148)
T5-Small（Ann.）	70.5	($1,300)	N/A		74.0	($1,032)	81.0	($1,980)	N/A		17.8	($3,321)
T5-Small（Ann.）- 上界	**71.1**	($1,800)	53.0	($2,595)	**76.9**	($12,899)	**87.5**	($17,443)	69.3	($14,469)	**26.2**	(9,981 美元)

表 4：Dist.、表现与 Dist. 相同/相似的 Ann.（及其对应预算），以及利用所有现有标注数据得到的 Ann. 上界。每项任务的最佳表现以粗体标示。

( ) 和 MULTIPIT$_{Gen}$ (*N*= 在这些情况下，由于 限预算的很大一部分，用于蒸馏的 #未标注数据 有限（分别为 *7K* 和 *10K*）。

3.2 蒸馏是否只是通过利用无标签数据而效果更好?

在表2中，我们观察到标注策略与蒸馏策略之间存在显著的性能差距。值得注意的一点是，在固定预算下，各种策略所使用的数据绝对数量（已标记数据数量和未标记数据数量）存在巨大差异。例如，在表2中的 WLP 任务中，给定$500的预算，*1923*条额外数据可以通过标注策略进行标注，而 条未标注数据则可以通过蒸馏策略加以利用。这不仅意味着标注数据成本高昂，还引出了一个问题：性能差距是否仅仅是由数据点数量的差异造成的？为了从数据规模的角度建立公平的比较基础并探究这一问题，我们采用了一种自蒸馏（Self-Dist.）方法（Zhang等人，2019年），其中教师模型和学生模型采用相同的架构（即 T5-Small）。

在表3中，我们使用相同的Dist. 与 Self-Dist. 进行比较，使用的*100K* 未标注数据。我们发现，尽管使用的数据量相同，Self-Dist. 在所有任务上的表现都明显逊于 Dist.。事实上，Self-Dist. 的表现受其教师模型（即T5-Small (Ann.)）的限制，这一点也在（Zhou 等人，2022年a）中有所观察。这项分析表明，Dist. 与Ann. 之间的性能差距确实可归因于利用大型预训练语言模型的能力，而非仅仅是使用了更多数据。

3.3 更大预算下的比较

我们的实验表明，蒸馏（Dist.）比完全依赖人工标注来训练紧凑模型更具经济性，至少在表2所展示的场景中是如此。然而，这引出了一个问题：人工标注Ann. 能否在投入更多预算的情况下达到 Dist. 的性能？表 4 展示了使用 Dist. 并投入预算处理 *100K* 未标注数据，以及使用 Ann. 并投入更多预算（或通过使用所有可用 标记数据）所获得的结果。有趣的是，在某些任务中（ STANCEOSAURUS 和 MULTIPI ）被证明是一种极具经济效益的紧凑模型训练方式。尽管 MULTIPIT$_{Gen}$ 训练使用了所有现有的标注数据（ ~*50K* ），花费高达 14,469 美元，但其性能始终未能超越 Dist.（仅花费 245 美元）。对于除上述任务之外的其他任务，我们发现，Ann. 在投入更多预算时能够超越 Dist.（例如，FEVER 需要投入 12,899 美元）。然而，在实际应用中，我们仍然发现，若要获得相近的性能，Ann. 的成本可能高得多（例如，在 FEVER 任务中高达 10 倍）。

4 进一步分析

在本节中，我们研究各变量的不同取值：标注数据的初始数量（*N*）（§4.1）、紧凑模型规模（§4.2）以及教师模型规模（§4.3），这些变量在主要实验（§3.1）中均保持不变。

4.1 帕累托曲线

在图 2中，我们探索了以下不同组合：#已标记数据 1K，5K，10K}）和 # 未标注数据 K }）请 质上表示标注 n.）策略。我们绘制每种组合的性能，并通过插值给定的数据点近似得到帕累托前沿

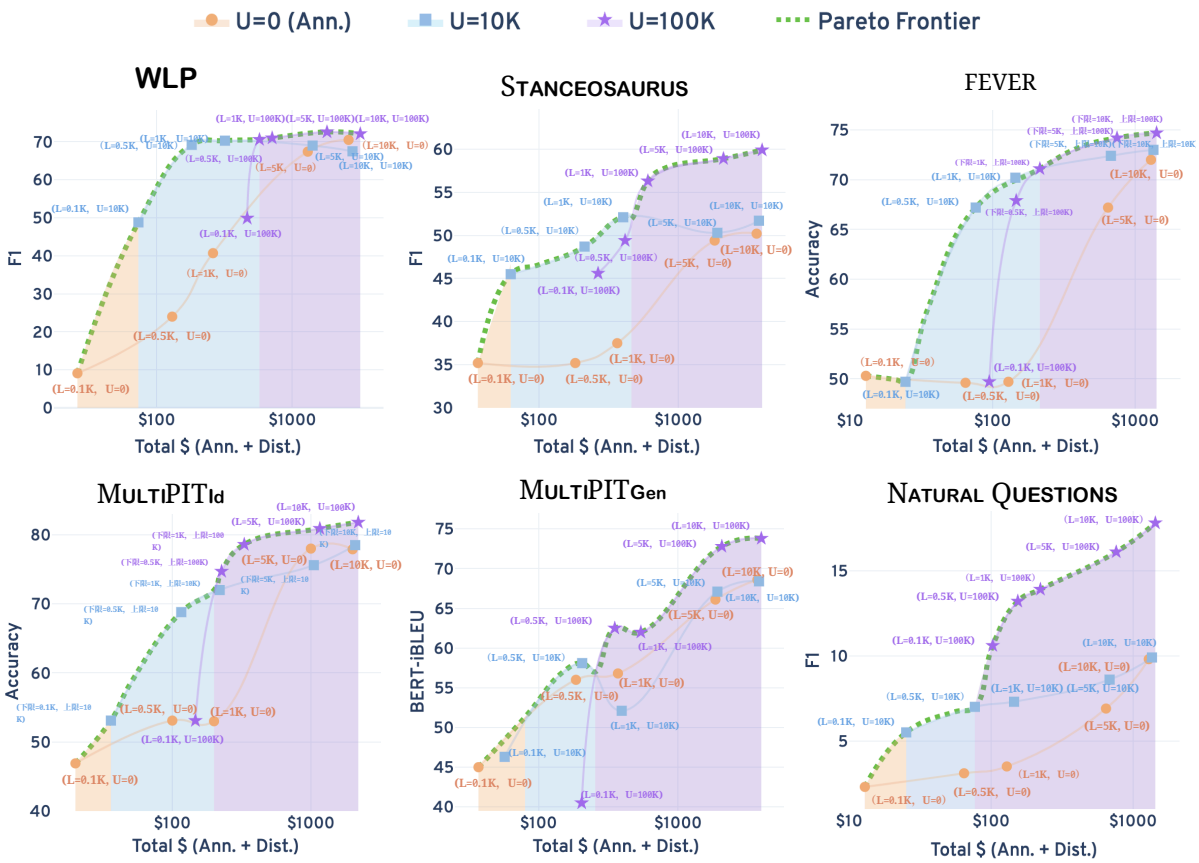


图 2：不同组合下的帕累托曲线 #已标记数据 (L={0.1K, , 10K}) 和#未标注数据 (U={ 0 , 10K , 示标 (n. 格。 托前沿（）是从业者会从中选择的最优解集合，并通过对给定数据点进行插值来近似。X 轴采用对数刻度。

(Abdolrashidi 等人，2021年；Treviso 等人，2022年）对于所有任务，我们观察到，蒸馏（Dist.）策略几乎总是帕累托最优。[6] 在附录（表 11）中，我们还详细考察了低资源场景。

此外，我们观察到，使用较少的未标注数据（U=10K）在预算较少时 托最优的，而较多的未标注数据（U=100K ）则会随着预算增加而最大化效用。这意味着在低预算情形下，教师模型的容量有限，使学生模型能够迅速赶上。然而，一旦教师模型显著超越学生模型，将更大部分预算用于蒸馏会更经济。

在图 3 中，我们通过在固定预算下改变初始标注数据的数量（N），进一步分析其影响



图 3：在固定额外预算下，不同初始标注数据数量（N）对应的结果。

N。正如预期，我们注意到，除了一些 N 较低的情况外，Dist. 总体上优于 Ann.。

[6] 一个例外是 (L= 预算非常有限，几乎无法利用大模型。

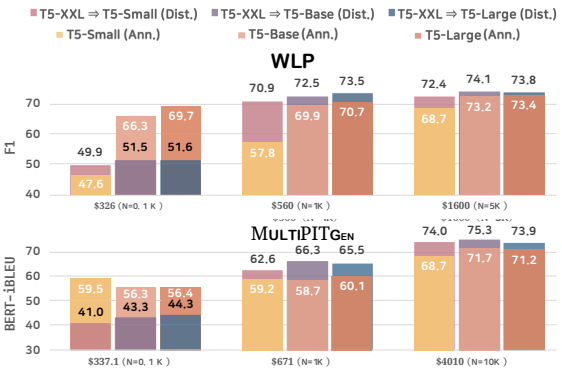


图 4：不同紧凑模型规模下的结果：Small（60M）、Base（220M）、Large（770M）。对于蒸馏，教师模型固定为 XXL-11B，蒸馏成本设为 $300。彩色查看效果最佳。

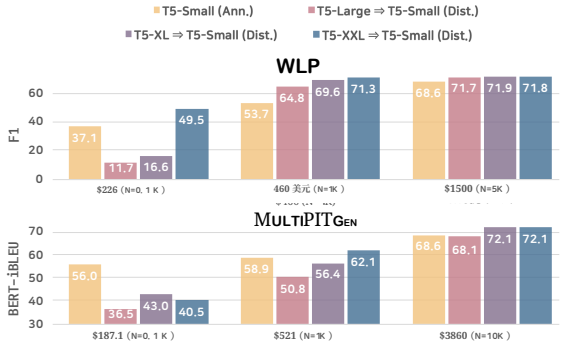


图 5：教师模型不同规模下的结果：Large（770M）、XL（3B）、XXL（11B）。紧凑模型固定为 Small-60M。WLP 的蒸馏成本固定为 $200，MULTIPIT$_{Gen}$ 的蒸馏成本固定为 $150。

尤其对于 MULTIPIT$_{Gen}$ 这也得到了附录（表 11）中的进一步证实。值得注意的是，所有任务都呈现出一个共同趋势，即Dist.的性能会随着N的增大而下降。这是由于预算有限；较大的 N 意味着大模型需要承担高昂的微调成本，因此可用于蒸馏的预算受到限制。例如，在 STANCEOSAURUS 的预算=$200的情况下，如果 N 为 *1K*，则可使用 *82K* 未标注数据进行蒸馏；而当 N 为$N=$ 时，仅能使用 *35K* 未标注数据，导致前者的表现优于后者。这一结果表明，在固定预算的场景下，一味追求更大的 N 并不可取；建议从业者充分理解并考虑微调成本与蒸馏成本之间的权衡。

4.2 改变紧凑模型规模

为考察各种推理场景，我们在图4中探究了不同尺寸的紧凑模型。总体而言，随着预算增加，所有模型的性能均有所提升；在相同成本下，Dist.的表现优于Ann.，但低预算（ *N=0.1K* ）。有趣的是，我们发现，在某些情况下（WLP 为 $1600，MULTIPIT$_{Gen}$ 为 $671 和 $4010），T5-XXL ⇒ T5-Base（Dist.）优于 T5-XXL⇒ T5-Large（Dist.），尽管前者更小、效率更高。我们推测，这是因为模型参数更多，需要更多 GPU，因此成本也更高。这一结果驳斥了普遍观点：至少在固定预算的场景中，较大的模型并非总是更优。

4.3 改变教师模型规模

我们现在研究了不同规模的教师模型（图 5）。结果表明，在低预算场景下，相对较小的教师模型（T5-Large 和 T5-XL）无法胜任教师模型的角色。例如，在 MULTIPIT$_{Gen}$ 的预算为 521 美元时，T5-Large ⇒ T5-Small（蒸馏版）和 T5-XL⇒ T5-Small（蒸馏版）的表现均不如 T5-Small（标注版），而 T5-XXL ⇒ T5-Small（蒸馏版）的表现则优于 T5-Small（标注版）。在预算较高的设置下，可以明显看出，规模最大的教师模型（XXL）的表现与较小的教师模型（Large、XL）相当或更好。综合来看，这项分析表明，在采用蒸馏方法时，教师模型的规模十分重要，无论预算情况如何，选择规模足够大的模型作为教师模型可能都是稳妥的做法。</final>

5 GPT-3.5 作为标注器

此外，我们还通过上下文少样本学习方案，考察了 GPT-3.5（Ouyang 等人，2022年）标注的成本效益。Wang 等人（2021年）近期证明，相较于人工标注，GPT-3（Brown 等人，2020年）可用作成本更低的标注器。我们尝试仔细审视其在本研究所考虑任务中的适用性，并最终将其结果与 Dist. 的结果进行对比。我们使用 text-davinci-003 模型，通过提示词提供 32 个训练示例来生成伪标签。在此实验中，我们分别为 WLP 和 STANCEOSAURUS 的 GPT-3.5 标注分配了 200 美元。请注意，OpenAI[7] 根据使用的词元数量收费。WLP 的每个标签成本。

[7]https://openai.com/api/pricing

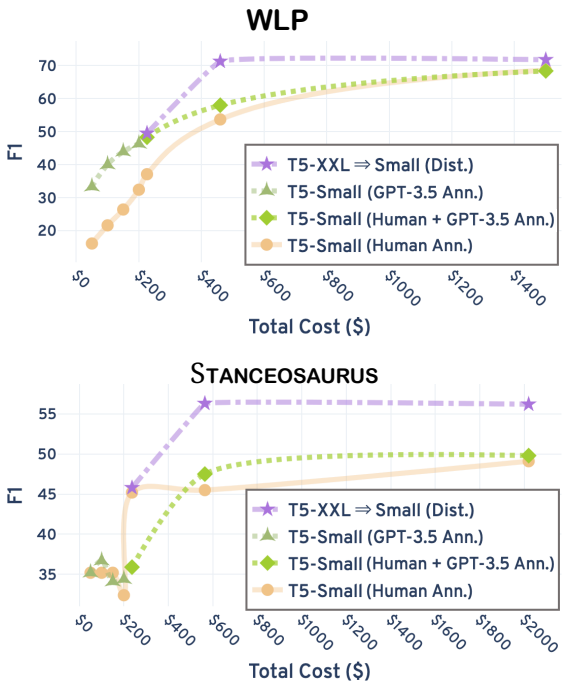


图 6：与 GPT-3.5 标注的比较。给定初始人工标注 N={*0.1K, 1K, 5K*} 及相应成本，另分配 200 美元用于蒸馏或 GPT-3.5 标注（即人工 +GPT-3.5 Ann.）。

为 0.046 美元，对于 STANCEOSAURUS 为 0.073 美元，如果使用GPT-3.5 （详情见附录 E）。

在图 6中，我们比较了 GPT-3.5 标注（GPT-3.5 Ann.）与人工标注和蒸馏策略。除了 GPT-3.5 Ann.，我们还将其与人工标注结合（人工 +GPT-3.5 Ann.），以提升质量，并与 Dist进行比较。结果清楚地表明，尽管 GPT-3.5 可能优于此前研究所暗示的人工标注人员（Wang 等人，2021年），但在预算相同的情况下，其表现显著逊于蒸馏（Dist.）策略，尽管 GPT-3.5的参数（175B）比教师模型（11B）更多。这再次凸显了知识蒸馏的另一种视角：成本效益。

6 相关工作

许多先前研究都探讨或关注了构建模型所涉及的成本。

数据标注。 一方面，研究人员试图解决人工标注存在噪声或成本高昂的问题。例如，Zhang等人（2021年）研究了如何在带有单标签的更多示例与带有多标签的更少示例之间分配标注预算。 Chen 等人（2022年）研究了通过多数投票进行冗余标注与清理或重新标注错误标注之间的差异。Wang 等人 （2021年）将人工标注与 GPT-3（Brown 等人，2020年）的标注进行了比较。然而，这些工作只关注标注成本。

知识蒸馏。 另一方面，其他研究则关注知识蒸馏所涉及的计算预算。 Ye 等人 (2022年) 提出使用比教师模型更大、更加稀疏的学生模型，以进一步降低推理成本。 Jooste 等人(2022年) 比较了不同的蒸馏方案，以构建成本低、速度快且环保的翻译模型。 Ma 等人 (2022年) 探讨了结合元学习的高效交互式蒸馏。然而，上述研究忽视了数据预算，和/或几乎没有考虑蒸馏过程中实际涉及的计算成本。尽管先前的自然语言处理研究已证明，知识蒸馏有助于模型压缩或提升泛化能力（Sanh 等人，2019年；Kang等人，2020年；Le 等人，2022年），但若将常被忽视的蒸馏实际成本纳入考量，其是否仍具备成本效益尚不明确。作为同期研究，Sun 等人 (2023年) 提出了一种以原则为导向的新型自对齐方法，而Hsieh 等人(2023年) 介绍了一种利用思维链（Wei 等人，2022年）推理依据进行逐步蒸馏的方法。尽管这些研究的主要关注点与我们的研究截然不同（即成本），我们认为，它们不仅推动了这一领域的发展，也有望支持我们关于蒸馏成本效益的发现，因为这些新方法会进一步拉大蒸馏与标注之间的差距。

数据与计算资源。 与大多数仅考虑标注成本或计算成本的现有研究不同，我们的研究将这两种表面上彼此无关、但众所周知成本高昂的成本类型联系起来进行分析（Ning 等人， 2019年； Hong 等人，2020年；Hendrycks 等人， 2021年； Izsak 等人，2021 年； Obando-Ceron 和 Castro， 2021 年；Minixhofer 等人，2022年），尽管这两种成本如何相互比较仍不明确。 Kirstain 等人 (2022年) 将缩放参数与增加更多带标签的样本进行比较，但紧凑模型和实际成本（美元）并非其关注重点。就研究框架而言，我们的工作与Bai 等人 (2021年) 类似，该研究探讨了如何最优地分配预训练和

专门针对领域自适应场景中的标注成本。我们的重点更多在于对紧凑模型进行微调/蒸馏，而非从头开始预训练，并探索涉及多样化任务的更通用场景。

7 结论

在这项工作中，我们解决了实践者在构建模型时经常面临的一个两难问题：在预算有限的情况下，如何合理分配预算，以经济高效的方式训练紧凑型模型？我们提供了经验证据，表明：（i）仅通过人工标注人员或 GPT-3.5 进行标注来扩大数据规模，可能并非最经济的解决方案；（ii）采用蒸馏策略时，使用较少数量的未标注数据，能够以较低的预算获得帕累托高效模型，而随着预算的增加，使用更多未标注数据则会带来更大的收益。此外，（iii）我们证明，在预算受限的情况下，规模更小的最终模型能够同时实现更好的性能和更高效的推理。基于这些发现，未来的研究可以探索不同的方法来利用大模型的能力，例如通过剪枝来构建经济高效的紧凑型模型。

局限性

本文主要考虑的是实践者租用云端 GPU 的情境。如果自行托管 GPU，本研究探讨的两种策略便无法直接比较。然而，在实践中，训练大模型时（使用 8 个 A100 GPU），我们推测许多情况下租用 GPU 可能更受青睐，因为扩展计算能力并非易事，而且成本高得令人望而却步（Izsak 等人，2021年；Obando-Ceron 和 Castro，2021年；Minixhofer 等人，2022年）。还值得注意的是，随着新硬件的发展、云平台服务定价政策的变化以及更多优化技术的应用，未来计算成本可能会降低。另一方面，人工标注成本至少可能维持不变，甚至可能更高。若成本朝这一方向变化，即使两种策略之间的差距会进一步扩大，本研究得出的结论仍然成立。

作为一种压缩方法，我们的工作聚焦于知识蒸馏（Hinton 等人，2015年）。不过，值得注意的是，蒸馏会放大压缩模型中的社会偏见（Hooker 等人，2020年；Silva 等人，2021年），这是由于其容量有限（Ahn 等人，2022年）。因此，建议从业者在将蒸馏用于现实世界应用时，额外采用偏差缓解技术（Ahn 等人，2022年）。除了我们发现蒸馏方案比数据标注方法更具成本效益之外，未来还可研究剪枝（Xia 等人，2022年）等其他高效方法，以确定在利用大模型的方法中，哪种是最佳的高效方案。不过，我们认为值得注意的是，对于紧凑模型而言，在剪除大部分参数（例如，~99.995%：11B ⇒60M）后仍能保持性能并非易事，先前的研究也证明了这一点（Michel 等人，2019年）。

致谢

我们感谢 Fan Bai 和 Jonathan Zheng 分别协助估算 WLP 和 Stanceosaurus 的数据标注成本，并收集未标注数据。本材料基于美国国家科学基金会（IIS-2052498）以及国家情报总监办公室（ODNI）下属的情报高级研究计划局（IARPA）通过 HIATUS计划合同 # 2022-22072200004 资助的工作。本文所载观点和结论均为作者所有，不应被解读为必然代表美国国家科学基金会、ODNI、IARPA 或美国政府明示或暗示的官方政策。尽管其中包含任何版权标注，美国政府仍获授权为政府用途复制和分发副本。

参考文献

AmirAli Abdolrashidi、Lisa Wang、Shivani Agrawal、Jonathan Malmaud、Oleg Rybakov、Chas Leichner 和 Lukasz Lew。2021年。Pareto 最优量化 ResNet 基本上是 4位。载于*IEEE/CVF* 计算机视觉与模式识别会议论文集，第 3091–3099 页。

Jaimeen Ahn、Hwaran Lee、Jinhwa Kim 和 Alice Oh。2022年。为什么知识蒸馏会放大性别偏见，以及如何从 DistilBERT 的角度加以缓解。收录于第 *4* 届自然语言处理中的性别偏见研讨会（*GeBNLP*）论文集，第 266–272 页，西雅图，华盛顿州。计算语言学协会。

Fan Bai、Alan Ritter 和 WeiXu。2021年。预训练还是注释？有限预算下的领域自适应。

载于*2021*年自然语言处理实证方法会议论文集，第5002–5015页，在线发表于多米尼加共和国蓬塔卡纳。计算语言学协会。

Gary S. Becker。2020年。时间分配理论。*TheEconomicJournal*，75(299)：493–517。

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。语言模型是小样本学习者。载于 神经信息处理系统进展，第 33 卷，第 1877–1901 页。Curran Associates, Inc.。

Derek Chen、Zhou Yu 和 Samuel Bowman。2022年。清洗还是标注：如何分配有限的数据收集预算。载于第三届低资源自然语言处理深度学习研讨会论文集，第152–168页。

Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2019年。BERT：用于语言理解的深度双向变换器预训练。在 计算语言学协会北美分会*2019*年会议论文集：人类语言技术，第*1*卷（长篇与短篇论文），第4171–4186页。

Yao Dou、Chao Jiang 和 Wei Xu。2022年。改进大规模释义获取与生成。载于*2022*年自然语言处理实证方法会议论文集，阿布扎比，阿拉伯联合酋长国。计算语言学协会。

JianpingGou, Baosheng Yu, StephenJMaybank 和 Dacheng Tao。2021 年。知识蒸馏：综述。《国际计算机视觉杂志》，129(6):1789–1819。

Kelvin Guu、Kenton Lee、Zora Tung、Panupong Pasupat 和 Mingwei Chang。2020年。检索增强语言模型预训练。载于第*37*届国际机器学习会议论文集，第119卷，《机器学习研究论文集》，第3929–3938页。PMLR。

Dan Hendrycks、Collin Burns、Anya Chen 和 Spencer Ball。2021年。CUAD：用于法律合同审查的专家标注自然语言处理数据集。载于神经信息处理系统数据集与基准测试专题论文集，第1卷。

Danny Hernandez、JaredKaplan、TomHenighan 和 Sam McCandlish。2021年。《迁移缩放定律》。*arXiv* 预印本 *arXiv:2102.01293*。

GeoffreyHinton、OriolVinyals 和 Jeffrey Dean。2015年。蒸馏神经网络中的知识。载于 *NIPS* 深度学习与表征学习研讨会。

Giwon Hong、Junmo Kang、Doyeon Lim 和 Sung- Hyon Myaeng。2020年。处理无监督问答中合成问题的异常。载于第*28*届国际计算语言学会议论文集，第3441–3448页，西班牙巴塞罗那（线上）。国际计算语言学委员会。

Sara Hooker、NyallengMoorosi、Gregory Clark、Samy Bengio 和 Emily Denton。2020年。压缩模型中的偏差特征。

Cheng-Yu Hsieh、Chun-Liang Li、Chih-Kuan Yeh、Hootan Nakhost、Yasuhisa Fujii、Alexander Ratner、Ranjay Krishna、Chen-Yu Lee 和 Tomas Pfister。2023年。逐步蒸馏！以更少的训练数据和更小的模型规模超越更大的语言模型。

Peter Izsak、Moshe Berchansky 和 Omer Levy。2021年。如何以学术预算训练 BERT。载于*2021*年自然语言处理实证方法会议论文集，第 10644–10652 页，线上及多米尼加共和国蓬塔卡纳。计算语言学协会。

Xiaoqi Jiao、Yichun Yin、Lifeng Shang、Xin Jiang、Xiao Chen、Linlin Li、Fang Wang 和 Qun Liu。2020年。TinyBERT：蒸馏 BERT 以实现自然语言理解。载于计算语言学协会会议论文集：*EMNLP 2020*，第 4163–4174 页。

Wandri Jooste、Andy Way、Rejwanul Haque 和 Ric- cardo Superbo。2022年。面向可持续神经机器翻译的知识蒸馏。载于美国机器翻译协会第*15*届双年会论文集（第*2*卷：用户与提供者专题及政府专题），第221–230页。

Junmo Kang、Giwon Hong、Haritz Puerto San Roman 和 Sung-Hyon Myaeng。2020年。无监督问题生成中不同策略的正则化。载于计算语言学协会论文集：*EMNLP 2020*，第3266–3277页，线上发布。计算语言学协会。

Yuval Kirstain、Patrick Lewis、Sebastian Riedel 和 Omer Levy。2022年。再多几个示例，或许就能抵得上数十亿个参数。载于计算语言学协会论文集：*EMNLP 2022*，阿布扎比，阿拉伯联合酋长国。计算语言学协会。

Tom Kwiatkowski、Jennimaria Palomaki、Olivia Red- field、Michael Collins、Ankur Parikh、Chris Alberti、Danielle Epstein、IlliaPolosukhin、Jacob Devlin、Ken- ton Lee、Kristina Toutanova、Llion Jones、Matthew Kelcey、Ming-Wei Chang、Andrew M. Dai、Jakob Uszkoreit、Quoc Le 和 Slav Petrov。2019年。自然问题：问答基准测试

研究。 计算语言学协会会刊，7:452–466。

Kelvin J. Lancaster。1966年。 消费者理论的一种新方法。 政治经济学杂志，74(2):132– 157。

Nghia Le、Fan Bai 和 Alan Ritter。2022年。通过上下文内专家混合实现科学协议中的少样本照应消解。载于计算语言学协会：*EMNLP 2022*年论文集，阿布扎比，阿拉伯联合酋长国。计算语言学协会。

Nayeon Lee、Belinda Z. Li、Sinong Wang、Wen-tau Yih、Hao Ma 和 Madian Khabsa。2020年。语言模型可以作为事实核查器吗？收录于第三届事实提取与验证研讨会论文集（*FEVER*），第 36–41 页，在线举行。计算语言学协会。

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。

Mike Lewis、Yinhan Liu、Naman Goyal、Marjan Ghazvininejad、Abdelrahman Mohamed、Omer Levy、Ves Stoyanov 和 Luke Zettlemoyer。2019年。BART：面向自然语言生成、翻译和理解的序列到序列预训练去噪方法。*arXiv*预印本 *arXiv:1910.13461*。

Yinhan Liu、Myle Ott、Naman Goyal、Jingfei Du、Mandar Joshi、Danqi Chen、Omer Levy、Mike Lewis、Luke Zettlemoyer 和 Veselin Stoyanov。2019年。RoBERTa：一种稳健优化的BERT预训练方法。*arXiv*预印本 *arXiv:1907.11692*。

Xinge Ma、Jin Wang、Liang-Chih Yu 和 Xuejie Zhang。2022年。采用 Reptile 元学习进行预训练语言模型压缩的知识蒸馏。载于第*29*届国际计算语言学会议论文集，第4907–4917页，韩国大韩民国庆州。国际计算语言学委员会。

Lucie CharlotteMagister、Jonathan Mallinson、Jakub Adamek、Eric Malmi 和 Aliaksei Severyn。2022年。教小型语言模型进行推理。*arXiv* 预印本 *arXiv:2212.08410*。

Paul Michel、Omer Levy 和 Graham Neubig。2019年。十六个注意力头真的比一个好吗？载于神经信息处理系统进展，第32卷。Curran Associates, Inc.

Sewon Min、Mike Lewis、Hannaneh Hajishirzi 和 Luke Zettlemoyer。2022年。面向少样本文本分类的噪声信道语言模型提示。载于计算语言学协会第*60*届年会论文集（第*1*卷：长篇论文），第5316–5330页，爱尔兰都柏林。计算语言学协会。

Benjamin Minixhofer、Fabian Paischer 和 Navid Rek- absaz。2022年。WECHSEL：用于单语语言模型跨语言迁移的子词嵌入有效初始化。载于计算语言学协会北美分会*2022*年会议论文集：人类语言技术，第3992–4006页，美国西雅图。计算语言学协会。

Tri Nguyen、Mir Rosenberg、Xia Song、Jianfeng Gao、Saurabh Tiwary、Rangan Majumder 和 Li Deng。2016年。MS MARCO：一个由人工生成的机器阅读理解数据集。载于认知计算研讨会论文集：神经与符号方法的整合，与第*30*届神经信息处理系统年会（*NIPS 2016*）同期举行，*2016*年*12*月*9*日，西班牙巴塞罗那，收录于第1773卷 *CEURWorkshop Proceedings*。CEUR-WS.org。

Qiang Ning、Hangfeng He、Chuchu Fan 和 Dan Roth。2019年。 部分还是全部，这是个问题。载于《北美计算语言学协会第*2019*年年会论文集：人类语言技术》，第*1*卷（长篇与短篇论文），第2190–2200页，明尼阿波利斯，明尼苏达州。计算语言学协会。

Tong Niu、Semih Yavuz、Yingbo Zhou、Nitish Shirish Keskar、Huan Wang 和 Caiming Xiong。2021年。 使用预训练语言模型进行无监督释义。载于 《*2021*年自然语言处理实证方法会议论文集》，第5136–5150页，线上及多米尼加共和国蓬塔卡纳。计算语言学协会。

Johan S Obando-Ceron 和 Pablo Samuel Castro。2021年。重新审视 Rainbow：推动更具洞察力和包容性的深度强化学习研究。载于《第*38*届国际机器学习会议论文集》，《机器学习研究论文集》。PMLR。

Long Ouyang、Jeffrey Wu、Xu Jiang、Diogo Almeida、Carroll Wainwright、Pamela Mishkin、Chong Zhang、Sandhini Agarwal、Katarina Slama、Alex Gray、John Schulman、Jacob Hilton、Fraser Kelton、Luke Miller、Maddie Simens、Amanda Askell、Peter Welinder、Paul Christiano、Jan Leike 和 Ryan Lowe。2022年。通过人类反馈训练语言模型，使其遵循指令。载于 神经信息处理系统进展。

David Patterson、Joseph Gonzalez、Quoc Le、Chen Liang、Lluis-Miquel Munguia、Daniel Rothchild、David So、Maud Texier 和 Jeff Dean。2021年。碳排放与大型神经网络训练。*arXiv* 预印本 *arXiv:2104.10350*。

Fabio Petroni、Aleksandra Piktus、Angela Fan、Patrick Lewis、MajidYazdani、Nicola De Cao、James Thorne、Yacine Jernite、Vladimir Karpukhin、Jean Maillard,

Vassilis Plachouras、Tim Rocktäschel 和 Sebastian Riedel。2021年。KILT：知识密集型语言任务的基准测试。载于*2021*年计算语言学协会北美分会会议论文集：人类语言技术，第2523–2544页，线上举行。计算语言学协会。

Colin Raffel、Noam Shazeer、Adam Roberts、Kather- ine Lee、Sharan Narang、Michael Matena、Yanqi Zhou、Wei Li 和 Peter J. Liu。2020年。探索利用统一的文本到文本转换器进行迁移学习的极限。机器学习研究期刊，21(140):1–67。

Adam Roberts、Colin Raffel 和 Noam Shazeer。2020年。语言模型的参数中能容纳多少知识？发表于*2020*年自然语言处理实证方法会议（*EMNLP*）论文集，第5418–5426页，线上举行。计算语言学协会。

Victor Sanh、Lysandre Debut、Julien Chaumond 和 Thomas Wolf。2019年。DistilBERT，一种蒸馏版 BERT：更小、更快、更便宜、更轻量。发表于 *NeurIPS ˆEMC2* 研讨会。

Tal Schuster、Adam Fisch 和 Regina Barzilay。2021年。来点维生素C！利用对比证据进行稳健的事实核查。发表于*2021*年北美计算语言学协会分会：人类语言技术会议论文集，第624–643页，线上举行。计算语言学协会。

Andrew Silva、Pradyumna Tambwekar 和 Matthew Gombolay。2021年。迈向对预训练变换器中社会偏见的全面理解与准确评估。载于*2021*年计算语言学协会北美分会会议论文集：人类语言技术，第2383–2389页，线上。计算语言学协会。

Samuel Stanton、Pavel Izmailov、Polina Kirichenko、Alexander A Alemi 和 Andrew G Wilson。2021年。知识蒸馏真的有效吗？神经信息处理系统进展，34。

Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. 2023年 原则驱动的从零开始的语言模型自对齐：最少人工监督。

Jeniya Tabassum、Wei Xu 和 Alan Ritter。2020年。WNUT-2020任务1概述：从湿实验室方案中抽取实体和关系。载于第六届噪声用户生成文本研讨会论文集（*W-NUT 2020*），第260–267页，线上。计算语言学协会。

James Thorne、Andreas Vlachos、Christos Christodoulopoulos 和 Arpit Mittal。2018 年。FEVER：用于事实抽取的大规模数据集与验证。载于《*2018* 年北美计算语言学协会人类语言技术分会会议论文集》第 *1* 卷（长篇论文），第 809–819 页，路易斯安那州新奥尔良。计算语言学协会。

Marcos Treviso、António Góis、Patrick Fernandes、Erick Fonseca 和 André FT Martins。2022年。预测 Transformer 中的注意力稀疏性。载于《*NLP* 结构化预测第六届研讨会论文集》，第 67–81 页。

Shuohang Wang、Yang Liu、Yichong Xu、Chenguang Zhu 和 Michael Zeng。2021年。想降低标注成本？GPT-3 可以帮忙。载于计算语言学协会 *EMNLP 2021* 论文集精选，第 4195–4205 页，蓬塔卡纳，多米尼加共和国。计算语言学协会。

Jason Wei、Xuezhi Wang、Dale Schuurmans、Maarten Bosma、brian ichter、Fei Xia、Ed H. Chi、Quoc V Le 和 Denny Zhou。2022年。思维链提示能够激发大型语言模型的推理能力。载于神经信息处理系统进展。

Thomas Wolf、Lysandre Debut、Victor Sanh、Julien Chaumond、Clement Delangue、Anthony Moi、Pier- ric Cistac、Tim Rault、Rémi Louf、Morgan Funtow- icz、Joe Davison、Sam Shleifer、Patrick von Platen、Clara Ma、Yacine Jernite、Julien Plu、Canwen Xu、Teven Le Scao、Sylvain Gugger、Mariama Drame、Quentin Lhoest 和 AlexanderM. Rush。2019年。Hugging Face 的 Transformers：最先进的自然语言处理技术。

孟州·夏（Mengzhou Xia）、泽轩·钟（Zexuan Zhong）和丹琪·陈（Danqi Chen）。2022年。结构化剪枝能够学习紧凑且准确的模型。载于第*60*届计算语言学协会年会论文集（第*1*卷：长篇论文），第1513–1528页，爱尔兰都柏林。计算语言学协会。

Canwen Xu 和 Julian McAuley。2022年。关于预训练语言模型的模型压缩和加速综述。

Qinyuan Ye、Madian Khabsa、Mike Lewis、Sinong Wang、Xiang Ren 和 Aaron Jaech。2022年。稀疏蒸馏：通过使用更大的学生模型加速文本分类。在*2022*年北美计算语言学协会分会会议论文集：人类语言技术中，第2361–2375页。

Linfeng Zhang、Jiebo Song、Anni Gao、Jingwei Chen、Chenglong Bao 和 Kaisheng Ma。2019年。做自己的教师模型：通过自蒸馏提升卷积神经网络的性能。载于*IEEE/CVF*国际计算机视觉会议论文集，第3713–3722页。

Shujian Zhang、Chengyue Gong 和 Eunsol Choi。2021年。使用不同数量的标注进行学习：从

零标注到多标签。在*2021*年自然语言处理实证方法会议论文集，第 7620–7632 页。

Tianyi Zhang、Varsha Kishore、Felix Wu、Kilian Q. Weinberger 和 YoavArtzi。2020年。BERTSCORE：使用 BERT 评估文本生成。载于国际学习表征会议。

Zihao Zhao、Eric Wallace、Shi Feng、Dan Klein 和 Sameer Singh。2021 年。使用前先校准：提升语言模型的少样本性能。载于第*38*届国际机器学习会议论文集，第139卷《机器学习研究论文集》《机器学习研究论文集》，第12697–12706页。PMLR。

Jonathan Zheng、Ashutosh Baheti、Tarek Naous、Wei Xu 和 Alan Ritter。2022年。STANCEOSAURUS：对多语言虚假信息的立场进行分类。载于*2022*年自然语言处理经验方法会议论文集，阿布扎比，阿拉伯联合酋长国。计算语言学协会。

Chunting Zhou、Junxian He、Xuezhe Ma、Taylor Berg-Kirkpatrick 和 Graham Neubig。2022a。用于零样本任务泛化的提示一致性。

Wangchunshu Zhou、Canwen Xu 和 Julian McAuley。2022b。BERT 学会教学：基于元学习的知识蒸馏。载于计算语言学协会第*60*届年会论文集（第*1*卷：长篇论文），第7037–7049页。

A 标注成本估算详情

WLP(Tabassum 等人, 2020年）这是一个包含湿实验室方案的标注语料库，其中包含的任务是命名实体识别（NER）和关系抽取（RE）。每个句子（实例）的价格为 0.44 美元，参见 Bai 等人(2021年）。由于此价格是针对两项任务计算的，而我们只关注 NER，因此我们按各自标签数量的比例（59.76%:40.24%）估算单独进行 NER 的费用，约为 0.26 美元。

STANCEOSAURUS (Zheng 等人，2022年) 该数据集包含来源主张和相关推文，以及用于立场分类的标注立场。由于论文未明确提及标注成本，我们向作者询问了每小时平均标注数量（82 条推文）和雇佣成本（每小时 15 美元），以计算最终的每个标签价格：$15 \div 82 \times 2$（双重标注）= $0.364。

MULTIPIT (Dou 等人, 2022年) 该数据集提供基于 Twitter、涵盖多个主题的释义。我们具体考虑其多个变体中的 $\text{MULTIPIT}_{\text{CROWD}}$ 语料库，其中包含句子对，并标注每对句子是否构成释义，用于释义识别（$\text{MULTIPIT}_{\text{Id}}$）。论文中提到，每对句子的成本为 $0.2。对于释义生成（$\text{MULTIPIT}_{\text{Gen}}$），我们抽取标注为释义的句子对，并以抽取样本占总数的比例（53.9%）计算每个释义源目标实例的成本：$100 \div 53.9 \times \$0.2 = \$0.371$。

FEVER (Thorne 等人，2018年）**和 NATURAL QUES-TIONS** (Kwiatkowski 等人，2019年）分别是事实核查和问答数据集。我们通过参考 Google Cloud Platform 的价格来估算成本。该服务每 1,000 个单位按每 50 个词收费 129 美元，因此我们估算这两项任务的每个标签成本均为 0.129 美元。

B 各任务的输入输出格式

我们的研究采用 T5（Raffel 等人，2020年）作为基础模型，并使用标准的文本到文本框架。表7展示了各项任务的输入输出示例，下面将逐一详细说明。

WLP 该任务可视为词元级分类问题，其中类别数为 20 个：{数量, 试剂, 设备,时间, 速度, 操作, 提及, 位置, 数值, 方法, 温度, 修饰语, 浓度, 尺寸, 通用测量, 密封, 测量类型, 其他, 酸碱度, 单位}。给定一个源输入（即操作步骤句），模型需要生成如下形式的目标：“实体 [标签] 实体 [标签]...”。

STANCEOSAURUS 对于此任务，源文本由一条主张、一条相关推文和上下文信息（例如回复）拼接而成，目标应为以下类别之一：{支持 | 反驳 | 无关 |讨论 | 提问}。

FEVER 这是一项事实核查任务，其中源文本是一个主张（如第 §3节所述的闭卷设置），目标是在遵循 Petroni 等人（2021年）的二分类设置下，判断为支持或反驳。

$\text{MULTIPIT}_{\text{Id}}$ 也是一个二元分类任务，给定两个句子，目标答案应为“是”或“否”。

多任务PI_{Gen} 此任务的输入是一句话，输出是一个释义句。

N自然问题 QUESTIONS 与**FEVER**一样，我们也考虑闭卷设置，要求模型在此任务中依赖其隐式知识，其中问题是源，目标则是问题的直接答案。

数据集	#训练	#Dev	#测试集	#未标注数据
WLP	11,966	2,861	3,562	111,000
STANCEOSAURUS	12,130	3,827	4,750	126,000
FEVER	104,966	10,444	N/A	182,000
$\text{MULTIPIT}_{\text{Id}}$	92,217	11,527	11,530	237,000
$\text{MULTIPIT}_{\text{Gen}}$	49,673	6,143	6,120	179,000
自然问题	87,372	2,837	N/A	115,000

表5：各类自然语言处理数据集的统计信息。对于 FEVER 和自然问题，由于测试集不公开，使用开发集进行评估。文中给出了实验所用未标注数据的最大数量。

C 详细设置与超参数

如第3节所述，我们采用 T5 v1.1 （Roberts 等人，2020年）作为基础模型，因为原始版本的 T5 （Raffel 等人，2020年）采用多个有监督任务和一个无监督任务的组合进行了预训练。由于本文假设没有可用的有监督数据集，我们的微调策略基于仅以无监督方式进行预训练的 T5 v1.1 。对于

数据集	初始金额 (N=5K)	标注（Ann.） T5-Small( )	蒸馏（Dist.） T5-XXL (5K) ⇒ T5-Small (100K)
WLP	$1,300	$260	$67.5 ⇒$435
STANCEOSAURUS	$1,820	$364	$60 ⇒$225
FEVER	$645	$129	$22.5 ⇒ $78
MULTIPIT$_{Id}$	$1,000	$200	$37.5 ⇒$123
MULTIPIT$_{Gen}$	$1,855	$371	$45 ⇒$163
自然问题	$645	$129	$30 ⇒ $86

表 6：使用图 1 中所示两种方法训练紧凑型模型的成本估算示例明细。从 (*5K*) 个带标签样本开始，我们比较额外标注 *+1K* 个样本，或先微调再蒸馏的成本，其中蒸馏对象为 T5-XXL （11B 个参数）。对于**蒸馏（Dist.）**，计算成本包括在 T5-XXL （教师模型）上使用 *5K* 条标注数据进行微调，再使用 T5-Small 对其进行蒸馏，所用数据为 *100K* 条无标签样本。

数据集	Task	示例
WLP	命名实体识别	**来源 -** 将以下试剂装入薄壁 PCR 管中 **目标 -** 组装 [操作]以下试剂 [试剂]薄壁 PCR 管 [位置]
STANCEOSAURUS	立场分类	**来源 -** 主张：COVID-19 封锁期间，自杀率上升了。 推文：@USER @USER 有自杀倾向的人可以很好地隐藏这些迹象。 [SEP] @USER @USER 那么我们这次不考虑家庭单位了？如果人们一整天都待在家里， 他们每天都和孩子待在一起，那为什么看不到这些迹象？哦等等，怪罪别人更容易。 **目标 -** {支持 | 反驳 | 无关 | 讨论 | 提问 }
FEVER	事实核查	**来源 -** 艺术史包括建筑、舞蹈、雕塑、音乐、绘画、诗歌、文学、戏剧、叙事艺术、电影， 摄影和平面艺术。 **目标 -** {支持 |反驳}
MULTIPIT$_{Id}$	释义识别	**来源 -** 句子1：孟加拉国建筑物倒塌事故造成160人死亡 句子2：#孟加拉国 孟加拉国建筑物倒塌事故死亡人数上升 **目标 -** {是 |否}
MULTIPIT$_{Gen}$	释义生成	**来源 -** 奥巴马总统将在白宫于上午 10:15 举行新闻发布会 **目标 -** 奥巴马总统将在美国东部时间上午 10:15 于简报室回答记者提问
自然问题	问答	**来源 -** 第一个登上月球的人是谁? **目标 -** 尼尔·奥尔登·阿姆斯特朗

表 7：各项任务的输入输出示例。

超参数	WLP	STANCEOSAURUS	FEVER	MULTIPIT$_{Id}$	MULTIPIT$_{Gen}$	自然问题
最大源长度	128	128	128	64	32	32
最大目标长度	128	8	8	8	32	32
批次大小	32	32	32	32	32	32
训练轮次	50 (20)	50 (20)	50 (20)	50 (20)	50 (20)	50 (20)
学习率	3e-5	3e-5	3e-5	3e-5	3e-5	1e-3 (3e-5)

表8：模型训练所用的超参数。() 中的数字仅用于 T5-XXL（即教师模型）微调。

问答任务，我们特例采用了使用显著跨度掩码（SSM）额外预训练的检查点；显著跨度掩码是一种已知有助于开放域问答的无监督预训练目标（Guu 等人，2020年），遵循 Roberts 等人 （2020年）。

表 5展示了数据集统计信息，表8列出了用于训练各项任务模型的超参数。我们没有针对每个模型的每项任务专门调整超参数。

任务，同时考虑到本研究所考察的场景中标注数据极为有限。此外，为了尽可能减少每种设置中除我们所考察因素之外的其他因素，除非训练期间出现重大问题，否则我们会尽可能固定每个参数。具体来说，我们为除……之外的所有设置选择了 3e-5 的学习率（Huggingface（Wolf 等人，2019年）代码库中问答和序列到序列蒸馏的默认值）；我们认为这一数值并不特殊。

对于自然问题（NATURAL QUESTIONS），我们在训练 T5-Small 模型时采用 1e-3 的学习率，因为我们观察到，当学习率为 3e-5 时，通过查看其训练损失发现模型实际上完全没有得到训练。除 T5-XXL 模型外，我们对所有模型均进行了 50 轮训练，因为我们认为 T5-XXL 模型可能只需更少的训练轮次即可达到足够的效果。在必要时，为满足显存限制，我们利用梯度累积（例如，批次大小为 {16, 8}，梯度累积步数为 {2, 4}），最终采用 32 的批次大小。我们采用（逐层）模型并行技术，使得能够将一个大模型加载到多个 GPU 上。由于我们的实证研究需要较高的计算成本，所报告的结果仅基于单次运行。尽管如此，在所比较的两种策略之间，我们仍观察到了显著的性能差异。

D 各任务的无标签数据

对于蒸馏策略而言，传递大模型的知识给小模型，实质上需要未标注数据。在本研究中，未标注数据特指没有对应标签的数据（即表7中仅包含源输入的数据）。我们仅使用现有数据集中的输入源（不含标注），并排除用于评估模型的数据集。此外，为了进行广泛的研究，我们还为每个数据集收集了额外的未标注语料，具体如下：

WLP 该数据集需要以程序性文本作为输入源。我们使用大规模的 PROCEDURE 语料库（Bai 等人，2021年），其中包含多个不同领域的内容。本研究具体使用了专利中的化学合成程序 CHEMSYN。

STANCEOSAURUS 该数据集的输入源包括来自多个事实核查网站的主张、与主张相关的推文，以及回复或父推文等上下文信息（如有）。按照本研究中描述的方法（Zheng 等人，2022年），我们收集了主张及其对应的推文，并对用户信息进行了匿名化处理。

FEVER 陈述或主张足以作为该数据集的来源。我们利用 Schuster 等人(2021年) 中合成生成的主张。

MULTIPIT 该数据集的来源是 Twitter 用户撰写的句子，可按照 Dou 等人(2022年) 中的方法进行收集。本研究则利用来源的 $\text{MULTIPIT}_{\text{AUTO}}$ (Dou 等人，2022年) 作为未标注数据，自动收集近期数据集。

自然问题 该来源仅由一个问题构成。因此，我们使用 MS MARCO 中的查询（Nguyen 等人，2016），其中的查询取自 Bing 的搜索日志。

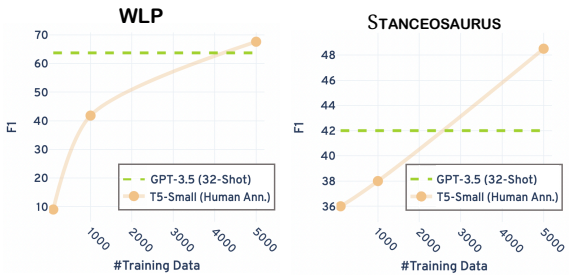


图 7：在 200 个抽样测试集上的初步结果，对比 GPT-3.5 的 32 样本上下文学习与 T5-Small 在不同训练数据规模下的表现。

E GPT-3.5 标注详情

为了使用GPT-3.5标注伪标签，我们采用最强版本text-davinci-003，并使用 32 个训练示例。输入提示由针对任务的指令[8][9] 、32 个上下文示例，以及末尾待标注的未标注输入组成。为降低较高的方差（Zhao 等人，2021年；Min 等人，2022年），每轮标注时，我们都会从一个固定的 100 个训练样本集合中随机抽取并打乱 32 个上下文示例。图7展示了GPT-3.5的 32 样本学习表现，以考察其质量和可行性。我们发现，它可以胜任低成本标签器的角色，提升模型表现，尤其是在预算较低的情况下；Wang 等人（2021年）也得出了这一发现。

请注意，OpenAI[10] API 根据输入提示和模型输出的词元数量收费：每 1K 个词元收费 0.02 美元。因此，每个标签的费用计算为：WLP 为 0.046 美元（平均 2.3K 个词元），STANCEOSAURUS 为 0.073 美元（平均 3.65K 个词元）。基于此，我们使用分配给每项任务的 200 美元预算，总计为 WLP 标注了 4347 条数据，为 STANCEOSAURUS 标注了 2739 条数据。

[8]对于 WLP，“将命名实体分类到以下类别之一：{Class 1, Class2, ...}”[9]对于 STANCEOSAURUS，“将给定推文针对给定主张的立场分类到以下类别之一：{Class 1, Class2, ...}” [10]https://openai.com/api/pricing

数据集	现有模型	T5-XXL（完整）
WLP	75.9（Bai 等人，2021年）	74.4
STANCEOSAURUS	61.0（Zheng 等人，2022年）	63.3 [69.8]
FEVER	78.9(Petroni 等人，2021年)	82.1
MULTIPIT$_{Id}$	91.4 (Dou 等人，2022年)	90.8
MULTIPIT$_{Gen}$	77.8（Dou 等人，2022年）	75.9
自然问题	35.2(Roberts 等人，2020年)	31.3 [38.5]

表 9：现有模型和完全微调的内部 T5-XXL 在不受资源限制情况下的性能，可作为参考或上界。由于使用了不同的指标，我们还报告了 STANCEOSAURUS 的宏平均 F1，以及 NATURAL QUESTIONS 的完全匹配分数，并同时报告本研究使用的 [micro F1]。

模型	STANCEOSAURUS		FEVER	
	$664	$2120	$279	$795
	(否=*1*否)	(否=*5*否)	(否=*1*否)	(否=*5*否)
T5-Small（Ann.）	44.7	50.3	49.8	68.9
DistilBERT（通用蒸馏 +Ann.）	56.3	57.5	69.9	73.5
BERT$_{Base}$ （Ann.）	56.0	59.0	70.7	73.1
T5-XXL ⇒T5-Small（蒸馏版）	56.9	60.5	71.7	74.8

表10：在 S*N=5K* 上的 注释、通用蒸馏（ **DistilBERT** (Sanh 等人，2019年)）以及在 STANCEOSAURUS 和 FEVER 上的特定任务蒸馏结果。对于**DistilBERT**，假定预训练阶段蒸馏的计算成本为 $0。最终模型大小彼此相近（ ∼60M ），但 **BERT$_{Base}$** (110M) 除外。通用（预训练）蒸馏与特定任务（微调）蒸馏相辅相成（Jiao 等人，2020年）。

F 额外结果

现成模型在各项任务上的表现如何? 在表9中，我们给出了在完整训练数据上微调的最大T5模型（11B）的结果，以及相关工作在资源充足环境下的结果。所报告的数值可作为上界或参考，用于校准本工作在资源受限环境下取得的相对结果。请注意，由于模型架构、尺寸、方法、预训练方案、训练数据和预算等因素的组合各不相同，这些结果不应直接用于比较。

通用蒸馏呢? 虽然本文侧重于特定任务蒸馏，但我们也提供了通用蒸馏的结果（ DistilBERT (Sanh 等人，2019年)），其中模型在预训练阶段进行蒸馏，以便在微调之前学习通用的语言理解能力。为衡量总成本，我们假设预训练阶段蒸馏的计算成本为$0 （即公开可用）。在表10中，我们发现，在相同预算下——

获得，添加通用蒸馏后，比不使用蒸馏的标注策略更具成本效益。除此之外，需要注意的是，直观来看，通用蒸馏（预训练）和任务特定蒸馏（微调）可以结合得更好，Jiao 等人（ 2020年）的研究证明了这一点。这进一步凸显了蒸馏方法的成本效益。

Task	N (初始金额)	策略	额外金额			
			+$0	+$100	+$200	+$300
WLP	*100* ($26)	T5-Small (Ann.)	**9.1**	23.8	37.1	47.6
		T5-XXL [48。8] ⇒T5-Small（蒸馏版）	N/A	**49.5**	**49.5**	**49.9**
			+$0	+$100	+$200	+$300
STANCEOSAURUS	*100* ($36)	T5-Small（Ann.）	**35.2**	35.2	45.2	45.4
		T5-XXL [44。8] ⇒T5-Small（蒸馏版）	N/A	**45.8**	**45.8**	**45.6**
			+$0	+$50	+$100	+$150
FEVER	*100* ($13)	T5-Small (Ann.)	**50.3**	49.3	49.7	49.7
		T5-XXL [49。7] ⇒T5-Small（蒸馏版）	N/A	**49.7**	49.7	49.7
			+$0	+$100	+$200	+$300
MULTIPIT$_{Id}$	*100* ($20)	T5-Small (Ann.)	**46.9**	53.1	53.1	53.1
		T5-XXL [53。1] ⇒ T5-Small（蒸馏版）	N/A	53.1	53.1	53.1
			+$0	+$100	+$200	+300 美元
MULTIPIT$_{Gen}$	*100* ($37)	T5-Small（Ann.）	**45.0**	**53.1**	**57.3**	**59.5**
		T5-XXL [55。5] ⇒T5-Small（蒸馏版）	N/A	41.4	40.6	41.0
			+$0	+$50	+$100	+$150
自然问题	*100* ($13)	T5-Small（Ann.）	**2.3**	3.3	3.9	4.2
		T5-XXL [18。6] ⇒T5-Small（蒸馏版）	N/A	**9.1**	**11.0**	**11.0**

表 11：少样本学习场景下的详细结果（N=100），用于研究小模型在更多数据标注（Ann.）和教师模型［**性能**］⇒ 学生模型（小模型）蒸馏（Dist.）在各种自然语言处理任务上的成本效益。 N 表示使用相应（初始金额）标注的起始数据数量。（#额外数据）表示除 N 之外的额外标注数据数量，而（*GPU* 小时）表示在 N 数据上微调教师模型所消耗的 GPU 总小时数，以及使用不同数量的（#未标注数据）将知识蒸馏到小模型所消耗的 GPU 小时数。