arXiv:2305.01616v1 [cs.CL] 2 May 2023

FreeLM：免微调语言模型

Xiang Li[1†]，**Xin Jiang**[1†]，**Xuying Meng**[2],**Aixin Sun**[3],**Yequan Wang**[1*1]北京人工智能研究院，北京，中国[2]中国科学院计算技术研究所，北京，中国[3]南洋理工大学计算机科学与工程学院，新加坡 lixiang@baai.ac.cn, jiangxin@baai.ac.cn, mengxuying@ict.ac.cn axsun@ntu.edu.sg, tshwangyequan@gmail.com

摘要

预训练语言模型（PLM）在 NLP 任务中取得了显著成功。尽管取得了巨大成功，主流方案仍主要遵循先预训练再微调 范式，这既带来了高昂的部署成本，也导致训练效率低下。然而，由于预训练语言模型仅使用大规模原始数据中的语言信号 进行预训练，因此针对特定任务进行微调仍然至关重要。本文提出了一种面向语言模型的新型免微调策略，同时考虑语言信号 和教师信号。教师信号 是对一系列下游任务的抽象，并以统一的命题格式提供。FreeLM 以交互方式同时利用语言信号和强大的任务感知教师信号进行训练，展现出很强的泛化能力和鲁棒性。在实验中，FreeLM 在一系列语言理解任务上的表现优于 GPT-3 和 InstructGPT 等大型模型，例如 GPT-3 和 InstructGPT。FreeLM 的规模小得多，仅有 0.3B 个参数，而这些模型有 175B 个参数。

1 引言

以 BERT（Devlin 等，2019年）、GPT 系列（Radford 等，2018年，2019年；Brown 等，2020年）及其变体（Liu 等，2019年；Joshi 等，2020年；Sun 等，2019年；Raffel 等，2020年）为代表的预训练语言模型（PLM）已广泛应用于各类语言处理任务，并取得了显著成功。

尽管取得了巨大成功，预训练后微调范式（Devlin 等人，2019年；Radford 等人，2019年）却为企业级应用中的训练和部署带来了极高的成本。即使是大型企业，在在线使用拥有数十亿参数的预训练语言模型时也十分谨慎（Sanh 等人，2019年），并且仍然对小型模型表现出浓厚兴趣。在预训练语言模型时代之前，小型模型

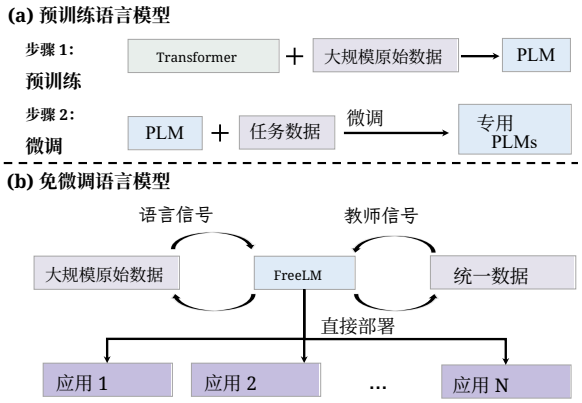

图1：小型模型、预训练语言模型以及我们提出的免微调语言模型概览。

是针对特定任务训练的。对于某些任务而言，这些专用模型的表现可能与经过微调的大型预训练语言模型相当，甚至更好（Grinsztajn 等人，2022年）。由于任务专用数据集和模型通常小得多，训练和部署成本都在可控范围内。然而，设计任务专用模型仍需要额外投入，而且每项任务都要设计一个模型。如图1所示，预训练语言模型利用规模大得多的数据集进行预训练，从而获得较强的泛化能力，并可针对各种下游任务进行微调，以实现语言理解与生成。如今，设计任务专用小型模型所需的投入，被在大型数据集上进行预训练的高昂成本以及高部署成本所取代。

为降低部署成本，研究者考察了零样本例如，GPT- 3（Brown 等人，2020年）和少样本模型。它们的表现，尤其是在理解任务上的表现，仍不尽如人意。原因之一是，预训练语言模型仅使用语言信号进行训练，而这种信号并不具备任务意识，也就是说，预训练语言模型的训练目标与任务目标并未很好地对齐。近期，基于指令微调的模型，也就是，InstructGPT（Ouyang 等人，2022年）和 FLAN（Wei 等人，2022年），进一步提升了零样本表现。它们的主要设

*通讯作者
†表示共同贡献

计将所有内容都转换为文本，然后使用自监督的语言信号训练模型。通过这种方式，它们训练中的语言信号变得更具任务意识，也相对更强。然而，要在特定任务上取得最佳表现，仍有必要进行适配。

训练中增强任务感知能力的理念促使我们考虑一种规模适中的模型，使其能够泛化到一系列丰富的预定义任务。最终目标是以较低的训练和部署成本，在所有这些预定义任务（以及可能遇到的未见任务）上取得良好表现。我们认为，这是一种理想的设置：企业可以使用基于自身数据训练的单一模型处理所有特定任务，无需额外微调。

因此，关键挑战在于如何将预定义任务统一起来，作为模型训练的强任务感知监督信号。为此，我们提出了 FreeLM，一种免微调语言模型。如图 1所示，FreeLM 以迭代方式同时利用语言信号 和 教师信号 进行训练。在训练过程中，模型在不同迭代中交替学习：（i）从大规模原始数据中获取的语言信号，以及（ii）基于所有预定义任务准备的统一数据中的教师信号。得益于这两种信号的相互作用，FreeLM 能够以较高准确率处理所有预定义任务，并像典型的预训练语言模型一样，出色地处理未见任务。

除了无需微调的策略外，我们的另一项重要贡献是首次尝试通过一种称为命题格式的统一数据格式，整合了多种语言任务。随后，将为所有预定义任务构建的数据集映射到这一统一命题格式中，成为图1所示的统一数据。借助命题格式，我们提供命题正确性判断，帮助语言模型了解事实。统一格式还隐式地促进了这些任务之间的相互提升，进而有助于提高泛化能力和鲁棒性。

据我们所知，这是首次尝试提出一种有效的无需微调策略来训练大语言模型。我们还展示了如何通过统一格式整合多项任务，从而利用强监督信号使预训练语言模型更具任务感知能力。通过实验，我们表明，FreeLM 在语言理解任务上的表现远超当前最先进的模型，与 GPT-2、175B GPT-3 和 InstructGPT 相比。此外，我们还表明，FreeLM 的生成能力甚至优于 GPT-2。FreeLM 在未见过的任务上也表现出色，并具有优秀的泛化能力和鲁棒性。

2 相关工作

本文聚焦于语言生成和语言理解任务。因此，我们仅考虑 GPT 等单向 Transformer 作为基础模型，不考虑掩码语言模型（例如，BERT）或编码器-解码器语言模型（例如，T5）。接下来，我们简要介绍自回归语言模型和指令微调模型。

2.1 自回归语言模型

自回归语言模型通过所有先前的词元来预测下一个词元。其单向特性赋予了这类模型语言生成能力。模型规模和模型结构的影响是两个关键研究方向。

在模型规模方面，最典型、影响力也最大的自回归语言模型是 GPT 系列，即GPT-1（Radford 等人，2018年），GPT-2（Radford 等人，2019年），以及 GPT-3（Brown 等人，2020年）。尤其是，GPT-3 的成功让研究人员意识到，模型规模和大规模原始数据所带来的强大优势，能够实现如此出色的生成性能。为了让模型变得更大，人们甚至提出了混合专家模型（Nie 等人，2021年）等方法。然而，规模扩大也带来了高昂成本和诸多挑战，尤其是在下游任务上进行微调的成本很高。

为改进模型结构，GLM（Du 等人，2022年）旨在利用自回归空白填充。Transformer-XL（Dai 等人，2019年）尝试解决长期依赖问题，其变体 XL-Net（Yang 等人，2019年）则希望学习双向上下文。然而，随着模型规模增大，这些模型的表现并未超过 GPT-3。关键原因仍在于如何有效且高效地利用数据。

2.2 基于指令微调的模型

模型规模更大，并不意味着其输出就能更好地满足用户的期望（Ouyang 等人，2022年）。一种解决方案是基于广泛的任务、人类反馈等对大语言模型进行微调。典型且最具代表性的模型是 Instruct-

GPT（Ouyang 等人，2022年）。借助强化学习、任务数据和人类反馈，InstructGPT 在语言理解和生成任务上都取得了出色的表现。实验甚至表明，与原始 GPT-3 相比，InstructGPT 能够更好地与人类保持一致。EFL（Wang 等人，2021年）专为少样本学习而设计，它基于 RoBERTa-large 将潜在的 NLP 任务重新表述为蕴涵任务。蕴涵任务难以妥善处理更多类型的任务，而我们提出的统一命题则更具通用性，能够适用于更多任务。此外，EFL 的生成能力有所退化。

在模型设计中，我们旨在无需微调的情况下，在理解和生成任务上都取得更好的效果。InstructGPT 采用了类似“语言全能”的方法。EFL 放弃了生成能力。据我们理解，这些模型没有有效利用 教师监督信号。在本文中，我们针对语言原始数据和任务感知数据设计了一种迭代训练策略。其核心思想是教会模型感知任务，同时不忘记其作为语言模型的职责，即对语言进行建模。

3 任务统一

我们的目标是训练一个任务感知语言模型，使其像典型的预训练语言模型一样从语言中学习，同时也从大量任务特定数据集中学习。如图1所示，FreeLM 同时从语言数据和统一数据中学习，后者是统一预定义任务的结果。

语言数据。 对于语言数据，选择相对直接。我们采用 OpenWeb-Text（Gokaslan 和 Cohen，2019年），这是对 WebText（Radford 等人，2019年）数据集的开源复现，该数据集由 OpenAI 提出。它通过抓取社交媒体平台 Reddit 上所有获得至少3 点 karma 的外链构建而成。我们在此数据集上采用与 GPT-2 相同的设置。

命题格式的统一数据。 我们的核心动机是让企业能够使用一个仅基于自身数据训练的模型处理所有特定任务，而无需额外微调。然而，获取专有数据并在其上开展实验颇具挑战。即使能够做到这一点，我们的结果也很难与当前最先进的成果进行基准比较。为便于一般性讨论，我们选择统一七种定义明确且广受欢迎的语言任务：1. 问答-

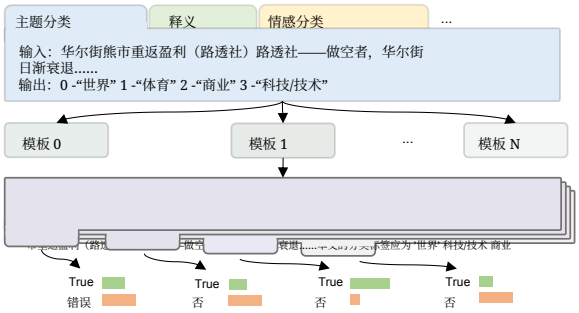

图 2：主题分类任务的模板示例。它还描述了任务的转换规则
统一。

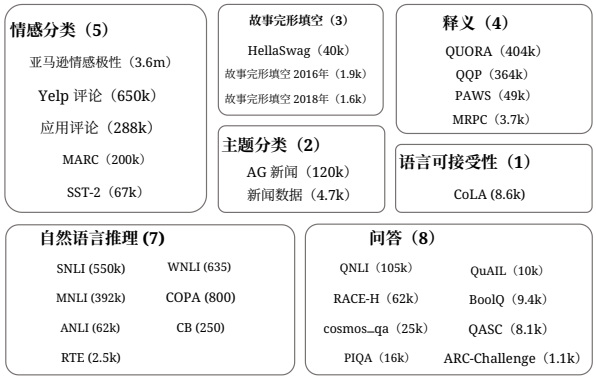

图 3：7 项任务的 30 个数据集列表；括号中的数字表示数据集大小。

题，2. 改写，3. 主题分类，4.故事完形，5. 情感分类，6. 自然语言推理，以及7. 语言可接受性。

我们通过将这七项任务转换为“命题正确性判断”任务来统一它们，以判断一个命题是否为真。具体而言，对于每项任务，我们设计一些候选模板，然后使用设计好的规则将每个实例转换为命题格式：

> “[任务] 任务名称 [任务] 实例文本 [分隔符]动态提示词 [分类]”

任务名称 沿用统一化前预先定义的原始任务名称，例如“问答”。由于每项任务都有各自的特点，因此我们使用动态提示词 来更好地适配任务，并考虑真实标签。

我们以主题分类任务为例来说明这种映射，图2也对此进行了说明。我们根据从预先设计的模板库中选取的模板，将类别标签转换为自然语言描述。在图2中，选择了模板 1，它是 N 模板。然后

我们可以将自然语言描述附加到原始输入文本后，从而构建候选命题陈述。在这个示例任务中，自然语言描述就是类别标签。

在我们的实现中，我们使用了所选七项任务的 30 个数据集（见图 3）。对于每个数据集，我们最多设计10 个模板来构建命题。构建统一数据时，每个样本都会随机选择一个模板。根据所选模板，将原始输入上下文与标签描述拼接起来，形成一个命题。以统一格式的命题作为输入，我们的 FreeLM 负责判断其真伪。

FreeLM 通过统一命题格式实现任务统一，从而带来了良好的泛化能力。换言之，FreeLM 只需处理一种命题正确性判断任务（语言任务除外），因此具备处理几乎所有 NLU 任务的能力。相比之下，多任务学习需要针对每项任务进行专门设计，以适配多任务模型。更重要的是，除经过设计的任务外，多任务学习几乎无法获得泛化能力。

4 FreeLM

4.1 架构与迭代训练

图 4 展示了所提出的 FreeLM 的架构。图的上半部分展示了使用语言数据进行训练的过程，这与主流生成式语言模型类似。下半部分展示了使用任务感知的统一数据进行训练的过程，即对统一命题作出正确判断。

为了用语言数据和统一数据共同训练模型，一个直接的做法是先用语言数据对模型进行完整训练，再用统一数据进行完整训练，这与先预训练再微调类似。然而，这种策略会导致模型过度关注任务，从而削弱其语言能力和泛化能力。在我们的设计中，语言信号 和教师信号 以迭代方式用于训练，类似于“先阅读，再练习”。接下来，我们将详细介绍交替训练模型的两个阶段。

语言迭代阶段。 为保留语言模型的生成能力，我们选择自回归语言模型，更具体地说是 GPT-2，作为 FreeLM 的基础模型。如图 4(a)，语言信号引导模型根据先前的词元预测下一个词

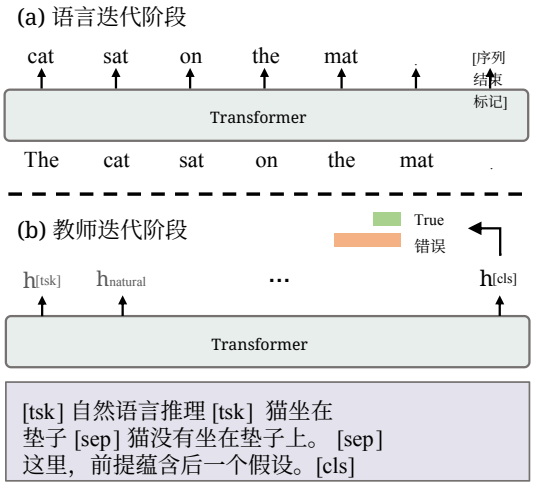

图 4: （a）语言阶段采用标准的语言训练方法；（b）教师阶段使用统一的命题信号。

元。给定由 N 个词元组成的文本 $\mathcal{T} = \{t_1, \ldots, t_N\}$，FreeLM 根据先前的 $k-1$ 个词元计算词元 t_k 的概率 P_L:

$$P_L(t_k|t_1, \ldots, t_{k-1}) = \text{softmax}(W_v h_{t_{k-1}}) \quad (1)$$

此处，$h_{t_{k-1}}$ 表示由 Transformer 根据先前词元$\{t_1, \ldots, t_{k-1}\}$ 作为输入编码得到的表示。W_v 表示可学习参数。

教师迭代阶段。 在所提出的免微调结构中，教师信号 旨在引导模型学习面向任务的知识。将所有任务统一为命题正确性判断后，期望模型能够对命题作出正确判断。设 $u = (T_p, y)$ 为一对，其中 T_p 表示命题陈述， y 表示该命题的标签。当命题为真时， y 为 True，否则为 False。

图 4(b) 展示了来自统一数据的教师信号机制。根据设计的迁移规则（见第 3 节），统一数据中的每个实例都会附加一个专用词元“[cls]”。给定一个实例，我们可以通过语言模型共享的 Transformer 计算 $h_{[\text{cls}]}$ 词元“[cls]”的表征。基于学习到的表征 $\mathcal{P}_P$，通过 softmax 分类器计算类别分布 $h_{[\text{cls}]}$:

$$\mathcal{P}_P(y|T_p) = \text{softmax}(W_p h_{[cls]} + b_p), \quad (2)$$

其中 W_p 和 b_p 是可学习参数。

4.2 训练目标

训练目标的设计分为两部分。一是最大化预测词元的似然，定义见公式 3。另一部分是最小化判断命题正确性的交叉熵，定义见公式 4，其中 y 是给定 T_p的真实值。

$$J(\theta) = \sum \sum_{k} \log P_L(t_k|t_1, \ldots, t_{k-1}) \quad (3)$$

$$U(\theta) = \sum \text{cross-entropy}\left(\mathcal{P}_P(y|T_p), y\right) \quad (4)$$

请注意，每次迭代中，两种目标会与两类训练数据交替进行。

5 实验

我们从两个角度评估 FreeLM：语言理解性能和语言生成性能。随后，我们详细分析迭代过程的影响、命题格式中特殊词元的影响，以及对未见数据的泛化能力。

5.1 训练数据集

如第 3 节所述，对于原始语言数据，我们采用 OpenWebText（Gokaslan 和 Cohen，2019年）。统一数据由 30 个数据集组成，如图 3 所示，涵盖七项任务。除 Story Cloze 2016 和 2018年外，所有数据集均收集训练集；对于这两个数据集，则收集验证集。

对于统一数据中的每个实例，我们根据构建方法，使用金标准标签构造一个正样本（即，真实命题）。同时，我们随机选择另一个标签，构造一个错误命题作为负样本。在训练集构建过程中，每个原始实例仅随机选择一个模板。每个数据集的训练集命题样本最大数量设为 $100K$。最终的统一训练数据集包含约 $1.76M$ 个样本。

5.2 语言理解评估

FreeLM 的一个关键动机是，无需微调即可具备理解能力。对于理解能力，我们在零样本和少样本设置下，将 FreeLM 与强基线进行评估，包括 175B GPT-3、InstructGPT、[1] 和 GPT-2。

我们主要在无需微调的场景中评估 FreeLM。也就是说，FreeLM 在训练完成后不会再进行面向特定任务的微调。此外，作为相同规模下理解能力的上界，我们展示了 GPT-2 微调后的结果。

评估指标。我们选择通用语言理解评估（GLUE）作为基准，该基准由典型的自然语言理解任务组成，包括自然语言推理（即 AX、MNLI（Williams 等，2018年）、RTE（Bentivogli 等，2009年）、QNLI（Wang 等，2018年）和 WNLI（Levesque 等，2012年））、情感分类（即 SST-2（Socher 等，2013年）和 STS-B（Cer 等，2017年））、释义识别（即 MRPC（Dolan 和 Brockett，2005年）和 QQP）以及语言可接受性（即 CoLA）。我们使用每个数据集的验证集进行评估。按照先前的工作（Devlin 等，2019年），我们将 WNLI 排除在基准之外，因为 GLUE 网页[2] 指出该数据集的构建存在问题。此外，由于缺少验证集，我们排除了 AX。由于回归任务（即 STS-B）不适用于 FreeLM，我们也将其排除在基准之外。

我们使用 GLUE 基准中的其余七项任务进行评估。为提高实验效率并降低 API 使用成本，在评估 GPT-3 和 InstructGPT 时，我们使用 GLUE 验证集进行评估，并为每项任务抽取 $1k$ 个样本作为最终评估集。所有结果均在相同设置下公平比较。

与基线模型的结果。 GLUE 的结果见表 1。在平均值指标和 10 项评估指标中的 7 项上，FreeLM 均优于所有基线模型，展现出强大的理解能力。具体而言，与强大的 InstructGPT 相比，我们的 FreeLM 平均性能提升约 5 个百分点。此外，我们还观察到，MRPC、QQP、MNLI 和 QNLI 任务的绝对准确率提升了7 个百分点或以上。更重要的是，我们的 FreeLM 规模仅为 0.3B。即使与参数量高达 175B、规模大得多的 GPT-3 相比，得益于训练中的强任务感知信号，FreeLM 无需微调即可在下游任务中达到最先进水平。在基线模型中，InstructGPT 相较于 GPT-2 和 GPT-3 具有显著优势。我们的结果还表明，InstructGPT 是一个

[1] 我们通过 OpenAI API 使用 text-davinci-003 版本。

[2] https://gluebenchmark.com/faq

模型	CoLA 数学相关性	SST-2 Acc.	MRPC F1	MRPC Acc.	QQP F1	QQP Acc.	MNLI 数学准确率	MNLI M.m. 准确率	QNLI Acc.	RTE Acc.	平均值 –
微调											
GPT-2*w.* 微调	56.52	94.95	87.93	82.84	87.91	90.83	85.87	85.83	91.03	68.23	83.19
少样本与零样本											
GPT-2*w.* 零样本	-3.26	51.83	81.22	68.38	52.88	36.19	35.60	34.30	46.40	51.62	45.52
GPT-2*w.* 少样本	0.60	63.30	81.34	68.62	44.46	46.80	31.60	35.00	48.60	49.09	46.94
GPT-3*w.* 零样本	17.28	52.98	65.05	53.92	4.61	62.80	36.80	39.50	60.90	64.98	45.88
GPT-3（少样本）	-4.45	93.34	70.92	59.80	6.31	64.40	44.10	46.60	53.80	65.70	50.05
InstructGPT（零样本）	**63.81**	92.08	83.75	77.94	64.20	78.70	46.20	48.90	71.89	**86.64**	71.41
InstructGPT（少样本）	56.72	**95.18**	77.41	70.83	72.98	81.20	71.70	72.89	81.30	82.67	76.29
FreeLM	53.00	93.81	**87.21**	**82.11**	**85.47**	**89.60**	**81.80**	**79.80**	**88.30**	71.12	**81.22**

表 1：GPT-2、GPT-3 和 InstructGPT 在 GLUE验证集上的比较。微调、零样本 和少样本 分别表示微调、零样本和少样本。不考虑 GPT-2 微调结果时，最佳结果以粗体显示。

极其强大的模型。

如表1所示，我们的 FreeLM 性能几乎可与经过微调的 *GPT-2*媲美。由于 GPT-2 是针对每个数据集分别进行微调的，因此本质上，这一比较是在一个 FreeLM 与7 个专门训练的 GPT-2 之间进行。从这个意义上说，经过微调的 GPT-2 可视为 0.3B 规模模型性能的近似上限。作为最强大的生成式预训练语言模型，InstructGPT 的表现位居第二。相比之下，FreeLM 使用基于分类的方法，而指令微调使用基于生成的方法，因此我们的模型更加有效。更重要的是，通过采用统一命题格式实现任务统一（见第 3 节），带来了良好的泛化效果。一个值得关注的现象是，在 RTE 的准确率上，FreeLM 显著逊于 InstructGPT，但显著优于经过微调的 GPT-2。这一结果也表明，语言模型的能力可能与其规模正相关。

表 1中没有直接体现这一点，但通过实验我们也发现，FreeLM 在推理时具有鲁棒性，且不敏感。有趣的是，我们发现，GPT系列模型在少样本或零样本设置下的预测结果，对模板、参数和后处理方法的设置很敏感。例如，我们在零样本设置下使用 MRPC数据集评估 GPT-2。词汇概率分布$top_k = 50$ 用于预测分类。如果我们计算正负词概率比之和的平均值，准确率为 33。08。然而，如果我们选取概率值最高的词，准确率则翻倍有余，达到 68。13。差距如此之大，以至于下游任务难以可靠地使用。另一方面，FreeLM 具有鲁棒性，无需专家知识或参数调优。

模型	LAMBADA	WikiText2	PTB
FreeLM	38.44	33.66	28.47
GPT-2	35.31	30.23	26.62

表 2：FreeLM 和 GPT-2 的困惑度。

5.3 语言生成评估

如前所述，尽管 FreeLM 获得了更强的理解能力，但我们希望模型能够像 GPT-2 一样保留语言生成能力，以便从相同的语言数据中学习。因此，我们评估了 FreeLM 与 GPT-2 的生成性能。评估使用了三个公开数据集，即，LAMBADA（Paperno 等人，2016）、WikiText2 和 PTB。按照惯例，我们使用困惑度（PPL）评估生成性能。

与基线模型的结果。表2 报告了 FreeLM 和 GPT-2 的困惑度（PPL）。FreeLM 的 PPL 略高于 GPT-2。这一微小差距可能表明，我们的模型在生成方面几乎达到了 GPT-2 的水平。有研究指出，PPL 并不能完全反映语言模型的生成能力（Wang 等人，2022年）。尽管如此，目前还没有好的替代指标。

由于生成评估没有黄金标准，我们对故事完形填空任务进行了一项案例研究。输入文本来自 StoryCloze 数据集（ Mostafazadeh 等人，2016）。为确保比较客观且公平，我们通过移除该数据集来训练一个新的 FreeLM 模型

案例*A*

输入： Rick 在一个问题重重的家庭中长大。他从未在家庭中获得良好的支持，因此转而投身帮派。**没过多久，Rick 在一次抢劫中遭到枪击。这起事件促使他改过自新。**

GPT-2: "我曾是帮派成员，但当时我并不知道当时，"Rick说道。"我完全不知道发生了什么。我只是想逃离这一切。事情就是这样结束的。在街头。也正因如此，我今天才会来到这里。"	**FreeLM：**"出狱后，我变成了一个不一样的人，"他说。"我得以重拾生活，重新走上正轨。我回到学校，找到了一份工作，组建了一个家庭。从那以后，我一直这样生活。"

案例*B*

输入： Tony 很高兴能去进行他的第一次游轮旅行。他抵达码头，带齐了所有行李。转过拐角时，他简直不敢相信那地方有多大。**他登上了船，船离开了码头。**

GPT-2： 到了房间后，他惊讶地发现房间比他想象的还要大。里面有一张床，一张书桌，和一台电视。...	**FreeLM：** "这是我第一次游轮，"他说。"我从未见过这样的事。"

案例 *C*

输入： Jasper 几个月来一直逃避找一份真正的工作。他讨厌之前所有的办公室工作他拥有的。一天，**他看到当地肉铺有一个工作机会。**他成功申请到了这份工作。

肉铺 ·

GPT-2: "他说，我不想雇用你，因为你太老了，'" Jasper 说。"我当时就想，我才不老了。我做这件事已经很久了，我知道。被拒绝是什么感觉。.....	**FreeLM:** "这份工作很棒，"他说。"我喜欢这份工作。我喜欢这里的人。他们把我当国王一样对待。大家都很友善。这让我觉得自己可以在这个世界上做任何想做的事。.....

表3：案 例研究：生成能力。表示一致和不一致的文本分别用蓝色和红色标出 .

，使其不包含在统一数据中。生成时采用带有 $k = 50$ 的 Top-k 随机采样，束搜索宽度设为5，最大生成长度为256。表3 展示了三个生成样例。

首先，我们通过人工抽样检查发现，FreeLM生成文本的可读性良好。没有明显的语法错误，这表明FreeLM和GPT-2在语法方面表现相当。其次，表3显示，GPT-2生成的文本中存在许多上下文不一致的现象。对于相同的输入文本，我们的FreeLM能够生成更加连贯一致的文本。为清晰展示正确的逻辑关联，我们用蓝色文本突出显示相关内容；同时，用红色文本标示逻辑错误。案例A和案例B表明，我们的FreeLM能够生成更加连贯一致的文本。有趣的是，尽管案例B中没有明显的语法错误，GPT-2生成的文本却与输入文本几乎没有直接的逻辑关系。也就是说，这并不是高质量的生成结果。

简单来说，我们的实验表明，所提出的 FreeLM 达到了与 GPT-2 相同的 PPL 水平。FreeLM 能够生成更加连贯的文本。我们相信，如果我们以更大的规模训练模型就能获得更强的生成能力。

5.4 详细分析

我们进行了三组实验，以研究：（i）迭代训练的影响；（ii）命题格式；以及（iii）在未见任务上的泛化能力。

迭代训练。如果移除教师信号，FreeLM 就会退化为通用语言模型。根据第5.2 节和5.3节的结果，我们得出结论：舍弃教师信号会导致理解任务的性能下降。反过来，如果移除语言信号，模型便只能依靠命题正确性判断这一目标进行训练。比较是在相同训练参数和相同步数下进行的。表4 展示了与标准 FreeLM 相比的结果。移除语言信号后，几乎所有评估指标都大幅下降。有趣的是，WikiText2 的 PPL 可能升至8，000，甚至达到9，000，而 GPT-2 的常规水平为30 。因此，生成能力受到严重损害。

模型	CoLA 数学相关性	SST-2 Acc.	MRPC F1	MRPC Acc.	QQP F1	QQP Acc.	MNLI 数学准确率	MNLI M.m. 准确率	QNLI Acc.	RTE Acc.	平均值 –
FreeLM	**52.53**	**93.81**	**87.21**	**82.11**	**86.32**	**89.82**	**80.70**	**80.93**	**88.61**	71.12	**81.32**
不含语言信号	40.14	92.43	85.71	79.17	85.66	89.11	77.56	78.57	87.28	**71.48**	78.71

表 4：语言信号 对理解的影响，其中语言信号 指语言信号。

模型	CoLA 数学相关性	SST-2 Acc.	MRPC F1	MRPC Acc.	QQP F1	QQP Acc.	MNLI 数学准确率	MNLI M.m. 准确率	QNLI Acc.	RTE Acc.	平均值 –
FreeLM	**52.53**	**93.81**	**87.21**	**82.11**	**86.32**	**89.82**	**80.70**	**80.93**	88.61	71.12	**81.32**
不带前缀	48.15	92.78	86.76	81.37	86.06	89.53	80.23	80.32	**88.76**	**71.84**	80.58

表 5：任务前缀对理解的影响 ng.

模型	MRPC F1	MRPC Acc.	RTE Acc.	COPA Acc.	CB Acc.	平均值 –
FreeLM	87.21	82.11	71.12	73.00	73.21	77.33
FreeLM_{U}	81.22	72.06	66.79	68.00	68.64	71.34
GPT-2*w.* 微调	87.93	82.84	68.23	54.00	73.21	73.24

表6：FreeLM与GPT-2在未见数据上的微调对比。

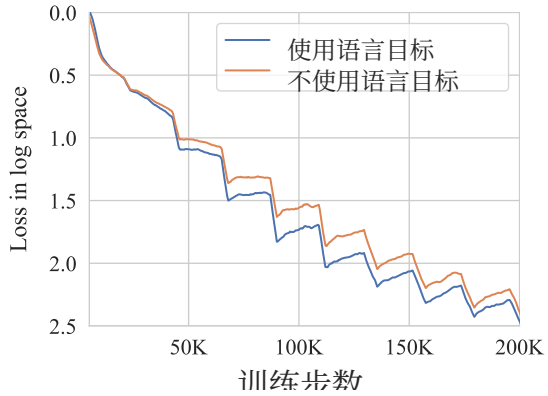

图5：语言信号对命题损失的影响。

我们已经表明，语言信号有助于 FreeLM 更好地学习。此外，为了评估 FreeLM 是否能够更快地学习，我们在图5中给出了训练过程中的损失表现。我们观察到，在 $20k$ 步之后，FreeLM 的命题损失下降得更快。这表明语言信号对面向任务的学习有积极作用。总之，在迭代训练方法下，语言信号和教师信号能够相互促进。

命题格式。 任务前缀，例如“[tsk]主题分类 [tsk]”，可以引导模型缩小搜索空间。在实现中，任务前缀集合为 {语言可接受性, 主题分类, 故事完形填空, 情感分类, 问答 *ing*，释义，自然语言推理}。由于 FreeLM 的参数量较小，我们使用这种任务前缀，在模型执行面向任务的理解时为其提供引导。作为比较，我们训练了另一个不带任务前缀的 FreeLM。其他设置与 FreeLM 相同。表 5 给出了比较结果。

我们观察到，大多数任务的表现处于相近水平。这一现象或许可以证明，FreeLM并不严重依赖任务前缀。我们认为，当FreeLM的规模超过 $1B$时，任务前缀的影响将微乎其微。

对未见数据集的泛化能力。 为了评估FreeLM的泛化能力，我们训练了一个新的模型FreeLM_{U}，方法是从统一数据中移除4个数据集。它们分别是MRPC、RTE、COPA（Gordon等人，2012年）和CB（De Marneffe等人，2019年）。其他设置完全相同。表6 给出了与标准FreeLM的对比结果。

我们观察到，FreeLM_{U}的平均得分几乎达到了经过微调的 *GPT-2*的水平。更值得注意的是，FreeLM_{U} 在 COPA 数据集上的准确率比经过微调的 *GPT-2*提高了 14 个百分点。与标准版 FreeLM 相比，平均下降约 6 个百分点。这一结果令人振奋，因为它表明我们的 FreeLM 具有良好的泛化能力。我们认为，这是因为我们将各种任务统一为命题正确性判断任务。这一现象也提醒我们，教师信号 的设计是训练出更好的语言模型的关键。

6 结论

为了降低训练和部署成本，我们设计了一种新颖的免微调语言模型——

语言模型。与典型语言模型一样，该模型的训练受益于自监督的语言信号。通过在统一数据上训练，模型也具备了任务感知能力。评估结果表明，FreeLM 在与 GPT-2 相近的规模下，仍保留了作为语言模型的强大语言能力即，GPT-2。实验还表明，尽管我们的模型小得多（0.3B 参数，而 GPT-3 和 InstructGPT 为 175B 参数），但在零样本或少样本设置下，FreeLM 在预定义任务上的理解表现显著优于 GPT-3 和 InstructGPT。如果某项任务也能映射到统一格式，FreeLM 也能很好地泛化。我们认为，FreeLM 表现出色的关键原因是任务统一和交替训练。如果企业能够为所有任务设计类似的统一数据格式，例如我们设定中的命题格式，那么所提出的免微调训练策略值得考虑，以节省成本。

局限性

这项工作主要有三点局限性。首先，研究我们的模型能否减少训练所需的数据量很有意义。如果这一点得到证实，就能进一步降低计算成本，让大模型更加环保。另一个值得研究的问题是任务数据的顺序会产生怎样的影响。最后，我们认为所提出的模型在更大规模下可能表现更好，但由于成本极高，我们并未对此进行验证。

致谢

本工作得到了中国国家重点研发计划（2022ZD0116300）和中国国家自然科学基金（NSFC No. 62106249）的支持。

参考文献

Luisa Bentivogli、Bernardo Magnini、Ido Dagan、Hoa Trang Dang 和 Danilo Giampiccolo。2009年。第五届 PASCAL文本蕴含识别挑战赛。载于第二届文本分析会议论文集，TAC 2009，美国马里兰州盖瑟斯堡，2009年11月16日至17日。NIST。

Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell 等。2020年。语言模型是少样本学习者。《神经信息处理系统进展》，33：1877–1901。

Daniel Cer、Mona Diab、Eneko Agirre、Iñigo Lopez- Gazpio 和 Lucia Specia。2017年。SemEval-2017任务1：语义文本相似度多语种与跨语言聚焦评测。载于第11届语义评测国际研讨会（SemEval-2017）论文集，第1–14页，加拿大温哥华。计算语言学协会。

Zihang Dai、Zhilin Yang、Yiming Yang、Jaime G. Carbonell、Quoc Viet Le 和 Ruslan Salakhutdinov。2019年。Transformer-xl：超越固定长度上下文的注意力语言模型。载于第57届计算语言学协会会议论文集，ACL 2019，意大利佛罗伦萨，2019年7月28日至8月2日，第1卷：长篇论文，第2978–2988页。计算语言学协会。

Marie-Catherine De Marneffe、Mandy Simons 和 Judith Tonhauser。2019年。Commitmentbank：研究自然发生的话语中的投射现象。载于 Sinn und Bedeutung 会议论文集，第23卷，第107–124页。

Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2019年。BERT：用于语言理解的深层双向 Transformer 预训练。载于2019年计算语言学协会北美分会会议论文集：人类语言技术，第1卷（长篇与短篇论文），第4171–4186页，美国明尼苏达州明尼阿波利斯。计算语言学协会。

William B. Dolan 和 Chris Brockett。2005年。自动构建句子释义语料库。载于第三届国际释义研讨会论文集，IWP@IJCNLP 2005，韩国济州岛，2005年10月，2005年亚洲自然语言处理联合会。

Zhengxiao Du、Yujie Qian、Xiao Liu、Ming Ding、Jiezhong Qiu、Zhilin Yang 和 Jie Tang。2022年。GLM：采用自回归空白填充的通用语言模型预训练。载于计算语言学协会第60届年会论文集（第1卷：长篇论文），ACL 2022，爱尔兰都柏林，2022年5月22日至27日，第320–335页。计算语言学协会。

Aaron Gokaslan 和 Vanya Cohen。2019年。OpenWebText 语料库。http://Skylion007.github.io/OpenWebTextCorpus。

Andrew Gordon、Zornitsa Kozareva 和 Melissa Roemmele。2012年。SemEval-2012 任务 7：合理备选项的选择：常识因果推理评估。载于 *SEM 2012：第一届词汇与计算语义联合会议——第 1 卷：主会议及共享任务论文集，以及第 2 卷：第六届国际语义评测研讨会（SemEval 2012）论文集，第 394–398 页，加拿大蒙特利尔。计算语言学协会。

Léo Grinsztajn、Edouard Oyallon 和 Gaël Varoquaux。2022年。为什么基于决策树的模型在表格数据上仍然优于深度学习？CoRR，abs/2207.08815。

Mandar Joshi、Danqi Chen、Yinhan Liu、Daniel S. Weld、Luke Zettlemoyer 和 Omer Levy。2020年。SpanBERT：通过表示和预测跨度改进预训练。《计算语言学会刊》，8：64–77。

Hector J. Levesque、Ernest Davis 和 Leora Morgenstern。2012年。Winograd 模式挑战。载于《知识表示与推理原理：第十三届国际会议论文集》，KR 2012，意大利罗马，2012年6月10日至14日。AAAI Press。

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。CoRR，abs/1907.11692。

Nasrin Mostafazadeh、Nathanael Chambers、Xiaodong He、Devi Parikh、Dhruv Batra、Lucy Vander-wende、Pushmeet Kohli 和 James F. Allen。2016年。用于深入理解常识故事的语料库与评估框架。CoRR，abs/1604.01696。Xiaonan Nie、Xupeng Miao、Shijie Cao、Lingxiao Ma、Qibin Liu、Jilong Xue、Youshan Miao、Yi Liu、Zhi Yang 和 Bin Cui。2021年。Evomoe：一种通过从密集门控到稀疏门控实现的演进式混合专家模型训练框架。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 Christiano、Jan Leike 和 Ryan Lowe。2022年。通过人类反馈训练语言模型遵循指令。Denis Paperno、Germán Kruszewski、Angeliki Lazaridou、Ngoc Quan Pham、Raffaella Bernardi、Sandro Pezzelle、Marco Baroni、Gemma Boleda 和 Raquel Fernández。2016年。LAMBADA 数据集：需要广泛话语语境的词语预测。载于第54届计算语言学协会年会论文集（第1卷：长篇论文），第1525–1534页，德国柏林。计算语言学协会。Alec Radford、Karthik Narasimhan、Tim Salimans、Ilya Sutskever 等。2018年。通过生成式预训练提升语言理解能力。Alec Radford、Jeff Wu、Rewon Child、David Luan、Dario Amodei 和 Ilya Sutskever。2019年。语言模型是无监督的多任务学习者。

Colin Raffel、Noam Shazeer、Adam Roberts、Katherine Lee、Sharan Narang、Michael Matena、Yanqi Zhou、Wei Li 和 Peter J. Liu。2020年。使用统一的文本到文本转换器探索迁移学习的极限。《机器学习研究杂志》，21：140:1–140:67。Victor Sanh、Lysandre Debut、Julien Chaumond 和 Thomas Wolf。2019年。DistilBERT：BERT 的蒸馏版本，更小、更快、更便宜、更轻量。CoRR，abs/1910.01108。Richard Socher、Alex Perelygin、Jean Wu、Jason Chuang、Christopher D. Manning、Andrew Y. Ng 和 Christopher Potts。2013年。基于情感树库的语义组合性递归深度模型。收录于《2013年自然语言处理实证方法会议论文集》，EMNLP 2013，2013年10月18日至21日，西雅图君悦酒店，美国华盛顿州西雅图，计算语言学协会特别兴趣小组 SIGDAT 会议，第1631–1642页。计算语言学协会。Yu Sun、Shuohuan Wang、Yu-Kun Li、Shikun Feng、Xuyi Chen、Han Zhang、Xin Tian、Danxiang Zhu、Hao Tian 和 Hua Wu。2019年。ERNIE：通过知识整合增强表征。CoRR，abs/1904.09223。Alex Wang、Amanpreet Singh、Julian Michael、Felix Hill、Omer Levy 和 Samuel R. Bowman。2018年。GLUE：用于自然语言理解的多任务基准与分析平台。CoRR，abs/1804.07461。Sinong Wang、Han Fang、Madian Khabsa、Hanzi Mao 和 Hao Ma。2021年。蕴涵作为少样本学习器。Yequan Wang、Jiawen Deng、Aixin Sun 和 Xuying Meng。2022年。预训练语言模型的困惑度不适合用于评估文本质量。CoRR，abs/2210.05892。Jason Wei、Maarten Bosma、Vincent Y. Zhao、Kelvin Guu、Adams Wei Yu、Brian Lester、Nan Du、Andrew M. Dai 和 Quoc V. Le。2022年。经过微调的语言模型是零样本学习器。收录于第十届国际学习表征会议，ICLR 2022，线上会议，2022年4月25日至29日。OpenReview.net。Adina Williams、Nikita Nangia 和 Samuel R. Bowman。2018年。用于通过推理进行句子理解的广覆盖挑战语料库。收录于《2018年计算语言学协会北美分会年会论文集：人类语言技术》，NAACL-HLT 2018，美国路易斯安那州新奥尔良，2018年6月1日至6日，第1卷（长篇论文），第1112–1122页。计算语言学协会。Zhilin Yang、Zihang Dai、Yiming Yang、Jaime G. Carbonell、Ruslan Salakhutdinov 和 Quoc V. Le。2019年。XLNet：用于语言理解的广义自回归预训练。收录于《神经信息处理系统进展第32卷：年度

2019年神经信息处理系统大会，NeurIPS2019，2019年12月8日至14日，加拿大不列颠哥伦比亚省温哥华，第5754–5764页。

A 附录

A.1 实现细节

我们列出了所提出的 FreeLM 的实现细节，以便于复现。为支持相关研究，我们将在论文被接收后发布相应的代码和训练好的参数。

在训练 FreeLM 的实验中，我们选择了 GPT-2 的中型版本，该模型有 3.45 亿个参数，由 12 个 Transformer 块组成。模型维度为 768。为了加快训练并节省能源，我们使用 HuggingFace 提供的预训练参数初始化模型。所有实验均在 5 个节点上进行，每个节点配备 8 张 40GB NVIDIA Tesla A100（共 40 张显卡），训练时间约为 8 天。我们在 OpenWebText 上训练 FreeLM 8 个 epoch，每张显卡的批次大小为 4。每个 epoch 大约需要 13.5 至 17.5 小时完成。我们使用 AdamW 优化器，学习率为 5e-5。语言迭代阶段样本与教师迭代阶段样本的混合比例为 2:1。