# StyleAdaptedLM：通过高效的风格转换提升指令遵循模型

**Pritika Ramu**[1] **Apoorv Saxena**[1] **Meghanath M Y**[2†] **Varsha Sankar**[3] **Debraj Basu**[3]
[1]Adobe Research, India [2]ZeroToOne.AI [3]Adobe Inc.
{ pramu,apoorvs,vsankar,dbasu } @adobe.com

## Abstract

将大型语言模型（LLM）适应特定的风格特征，如品牌语调或作者风格，对于企业通信至关重要，但从缺乏指令-响应格式的语料库中实现这一点具有挑战性，同时不损害对指令的遵循。我们引入了 StyleAdaptedLM，这是一种通过低秩适应（LoRA）有效传输风格特征到指令跟随模型的框架。LoRA 适配器首先在具有多样化非结构化风格语料库的基础模型上训练，然后与单独的指令跟随模型合并。这种方法实现了在没有配对数据或不牺牲任务表现的情况下，强大的风格定制。跨多个数据集和模型的实验显示出在保持指令遵循的同时，有风格一致性的提升，且人工评估确认品牌特定惯例被有效吸收。StyleAdaptedLM 为 LLM 的风格个性化提供了一条高效途径。

随着像 ChatGPT、LLaMa 和 Mistral 这样的 LLMs 成为客户服务自动化、营销内容生成和企业交流的重要组成部分，对具有特定风格特征的内容的需求也在增长。例如，一致的品牌声音——这可能包括独特的术语、常见的信息结构以及对产品背景的理解——能够增强信任和识别度。同样，适应作者的风格可能涉及他们特有的词汇和主题偏好。然而，通用的 LLMs 往往难以捕捉这些细微但却至关重要的风格和语言细节，尤其是当这些细微差别嵌入在大量非结构化文本数据中而非以明确指令形式提供时。核心挑战在于如何使 LLMs 能够在不削弱其基本指令跟随能力或需要配对数据集用于指令微调的情况下，采用这些多样化的风格身份。

文本风格转换方法 (Rao and Tetreault, 2018; Horvitz et al., 2024; Zhang et al., 2024; Reif et al., 2022; Syed et al., 2020) 虽然在应用风格修改的同时保留输入文本的内容方面非常有用，但可能无法完全解决生成新颖、符合指令的内容的需求，这种内容还需体现从非结构化文本中学习的目标风格，特别是当该风格包含隐含惯例时。直接微调指令跟随模型以灌输特定的风格

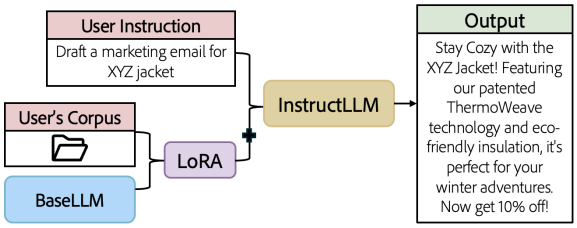


Figure 1: StyleAdaptedLM 框架。输入：用户指令和语料库（例如，品牌过去的营销电子邮件）。方法：在语料库上用基础 LLM 预训练的 LoRA 适配器与经过指令调优的 LLM 合并。输出：符合指令的风格一致文本，可能包括从语料库中学习的特征。

特征因合适训练数据的稀缺而变得更加复杂；为各种风格创造大规模配对的数据集来匹配指令和相应风格合适的响应是一项重大的障碍。此外，创建含有大量少样本示例的提示会导致较差的指令遵循能力 (Ma et al., 2023) 。进一步来说，对于需要管理和部署多种不同风格特征（例如，不同品牌、产品或作者）的企业而言，为每种风格微调并维护单独的大型模型在计算上是不可行的，运营上也十分繁琐。一个理想的解决方案应在保持风格真实的同时有效管理多种风格的调整。这就提出了一个关键挑战：我们如何使大型语言模型（LLMs）能够采用多样的风格特征而不牺牲指令遵从性并且无需配对数据集？我们提出了 StyleAdaptedLM（图 1），它利用低秩适配（LoRA）适配器，将无结构语料库中的风格特征高效地注入到指令跟随模型中。我们的方法规避了对配对数据集的需求，并避免了与完整模型微调或平均化相关的高昂计算成本。该过程首先将 LoRA 适配器训练在基模型 LLM 上，使用目标风格的文本。这些预训练的风格适配器，包含了所需的风格细微差别，然后与一个独立的、遵循指令的 LLM 结合。此模块化方法不仅保留了 LLM 的指令遵循能力，而且允许高效地创建、存储和部署多种不同的风格适配器，为个性化内容生成提供了一种可扩展的解决方案。

该框架的一个下游应用是帮助企业文案撰写人员高效地为电子邮件、博客、社交媒体标

† Work done while at Adobe.

题和其他形式生成初稿内容。该过程始于一个输入提示，指定预期的信息。为了确保风格一致性，我们的框架可以基于先前的数据进行训练，以便为客户定制一个适配器。

我们在三个模型和四个数据集上进行了综合实验。我们对企业营销电子邮件进行人工评估，以展示一个关键应用，其中有效的风格适应可能涉及捕捉品牌特定的惯例，并进行消融研究以检查风格注释的影响。这些研究表明，即使在测试时没有明确的风格提示，也能增强风格的一致性。

## 1 相关工作

跟随指令的模型像 InstructGPT (Ouyang et al., 2022) 和 Self-Instruct (Wang et al., 2023) 这样的模型极大地增强了大型语言模型（LLMs）遵循用户指令的能力，展示了在各类任务上的强大泛化能力。然而，这些模型缺乏仅通过指令来适应品牌特定特征如语气、风格和信息一致性的机制，限制了它们在企业沟通和营销内容生成中的有效性。此外，传达编辑指引并在指令中提供所有相关主题内容是具有挑战性的。在没有明确适应策略的情况下，这些模型难以在不同的内容格式中保持一致的品牌声音。文本风格转换文本风格转换是一项任务，其在保留与风格无关的内容的同时修改文本的风格属性 (Mukherjee and Duek, 2024) 。虽然文本风格转换方法 (Rao and Tetreault, 2018; Horvitz et al., 2024; Zhang et al., 2024; Reif et al., 2022; Syed et al., 2020) 会在不改变输入的情况下修改文本的风格属性，我们的工作重点是生成新的、符合指令的内容，该内容体现了从无标注语料库中学习到的风格。对于如品牌声音这样复杂的风格，这可能不止于表面属性，而是包括来自源语料的特征性文本模式。

参数高效微调（LoRA 和 PEFT）参数高效微调（PEFT）方法，如 LoRA (Hu et al., 2022) 、Prefix-Tuning (Li and Liang, 2021) 和 P-Tuning v2 (Liu et al., 2021) ，通过训练小的适配器模块来调整大型模型，实现高效的适应而无需修改基础模型。Liu 等人 (Liu et al., 2024b) 研究了使用 LoRA 适配器自定义 LLM 输出以匹配目标作者的词汇和句法风格。然而，他们的工作聚焦于著名作者的文学风格，并通过训练两个适配器在保持指令能力的同时增加了训练时间和计算量。

我们的方法捕捉与任务无关的风格特征，确保模型在执行多种指令跟随任务时保持风格一致性。我们强调风格适应与任务遵循之间的平衡，这对于实际部署至关重要。

小样本提示和模型合并技术小样本提示通过在提示中提供示例来实现上下文学习 (Brown et al., 2020) ，但由于提示长度的限制以及在提供的示例之外的泛化能力有限，它在品牌适应方面遇到困难 (Min et al., 2022; Ma et al., 2023) 。同样，模型合并技术如模型汤 (Wortsman et al., 2022) 和权重平均 (Izmailov et al., 2018) 通过对多个微调模型的权重进行平均来增强模型的适应性。然而，这些方法在计算上非常繁重，并且对于多个品牌标识的频繁更新或实时企业应用来说不切实际 (Matena and Raffel, 2022) 。

## 2 方法

StyleAdaptedLM 通过在基础模型上微调 LoRA 适配器并将它们与指令模型合并，将风格特征从非结构化文本迁移到一个指令遵循模型。

### 2.1 风格注释

非结构化文本用样式标识符进行注释，以指导模型学习特定样式的语言。例如，字符串：

“News article written by [[BBC]].”

被添加到每个来自 BBC 语料库的文本前。这种显式标记允许模型在训练时区分相关的样式。

### 2.2 基础模型和 LoRA 适配器

预训练的基础模型用于学习特定风格的特征，因为它在开放式任务、通用语言理解和创造性生成方面具有理想的表现，并且在需要灵活性、探索或进一步定制而不是严格遵循指令时，优先选择该模型。在保持原始权重不变的情况下，LoRA 适配器在这个基础模型上使用标注的非结构化文本进行训练。LoRA 对权重更新应用了一种低秩分解，表示为：

$$\Delta \boldsymbol{W}_{(\text{style} \mid \text{base})} = \boldsymbol{A}\boldsymbol{B}$$

其中 $\boldsymbol{A} \in \mathbb{R}^{d_{\text{out}} \times r}$ 和 $\boldsymbol{B} \in \mathbb{R}^{r \times d_{\text{in}}}$ ，以及 $r \ll d_{\text{in}}$ ，是在微调期间学习到的低秩矩阵，捕捉特定风格的信息。

### 2.3 训练目标

微调过程使用文本补全任务作为训练目标。给定一个输入序列，包括风格标识符，模型被训练来预测序列中的下一个标记。目标函数最大化风格注释文本中下一个标记的可能性。

一旦在基础模型上对 LoRA 适配器进行了微调，风格特定的知识便被合并到指令跟随模型中。最终模型表示为：其中：表示原始指令跟随模型的权重，代表了含风格特定知识的从基础模型通过 LoRA 训练的适应。

LoRA 模块的组合性之所以可行，是因为 LoRA 更新引起的嵌入空间的偏移是极小的。LoRA 将更新限制在低秩矩阵上，这保证了适应性保持局部化。在具有数十亿参数的大型语言模型（LLM）中，由于参数空间的冗余性，这种小的偏移很容易被吸收，从而确保模型的指令遵循能力不会被破坏。这使得生成的 StyleAdaptedLM 能够保留指令模型的指令跟随能力，同时结合从基础模型中学习的基于风格的适应。

## 3 实验

我们排除了一些在少样本场景中常用的数据集，因为我们的框架需要训练 LoRA 适配器。此外，我们也不使用传统的文本风格转换数据集，因为 LLM 已经具备了关于正式和非正式语言、莎士比亚风格以及知名人物写作风格的广泛知识。

Reddit 评论数据集。Reddit 语料库 (Khan et al., 2021) 的一个子集，包含来自超过 100 万用户的用户生成评论。我们选择了三位用户，过滤掉长度少于 60 个符号的评论，每位用户平均生成约 1,600 条评论（每个用户有 54k 个符号）。鉴于 Reddit 评论的短小且依赖语境的特点，我们利用 GPT-4o (OpenAI, 2023) 生成中性的语言版本，以形成用于基于转述的风格适应的输入-输出对。原始用户评论作为目标风格，而生成的中性评论作为源风格。每条评论都标注为：“在 [[username]] 语气中的 Reddit 评论。评论：”

安然电子邮件数据集。该数据集由大约 150 名安然员工的电子邮件构成 (Klimt and Yang, 2004)，反映出专业的写作风格。我们仅保留原始电子邮件内容，排除转发的邮件，并过滤掉少于 50 字的邮件。选择了三名发送量超过 700 封的发件人：jeff.dasovich@enron.com（1198 封邮件，352k 标记）、tana.jones@enron.com（964 封邮件，108k 标记）、sally.beck@enron.com（766 封邮件，115k 标记）。每封邮件都被注释为：" 以 [[email id]] 的语气撰写的邮件内容。内容："

为了验证，输入提示通过使用 GPT-4o 生成主题行来创建，为指令驱动生成提供了现实场景。

新闻数据集。我们使用了两个来源的文章：CNN-DailyMail (Hermann et al., 2015)（仅选择标记有 CNN 的文章）和 BBC News (Greene and Cunningham, 2006)，涵盖商业、娱乐、政治、体育和科技。数据集包括：CNN（2000 篇文章，1710k 标记），BBC（1903 篇文章，965k 标记）每篇文章都被标注为：“由 [[新闻机构]] 撰写的新闻文章。文章：”

为了验证，使用 GPT-4o 创建了输入提示，将类似摘要的笔记转换为全文内容。此外，模型被指示通过调整区域拼写和词汇（例如，美国到英国，反之亦然）来修改写作风格。

营销电子邮件数据集。我们收集了企业 [1] 的 1500 篇（767k 标记）公开营销电子邮件，这些邮件包含了跨各种产品的培训信息，并使用 GPT-4o 根据产品策划相关提示。每篇文章都被标注为：“由 [[企业]] 撰写的营销电子邮件。邮件：”

表格 1 提供用于训练和测试的示例输入输出格式。

### 3.1 基线

我们在 NVIDIA A100 40GB GPU 上使用三个 LLM 系列进行训练和评估，并对三次运行（基础、指令）结果进行平均。

| | |
|---|---|
| Llama-3.1-8B | Llama-3.1-8B-Instruct |
| Mistral-7B-v0.1 | Mistral-7B-Instruct-v0.1 |
| Qwen2.5-7B | Qwen2.5-7B-Instruct |

在指令跟随模型上进行监督微调。我们使用文本补全目标，通过 LoRA 在指令微调模型上执行监督微调。非结构化文本被标注，但我们没有提供明确的指令，而是将指令字段设置为空字符串 (“""”) 并使用标注的语料作为输出。这使我们能够在指令跟随模板内进行训练，同时直接调整 LoRA 以生成文本。

少样本提示。我们使用少样本提示作为基线，其中模型接收与作者或风格相关的一些内容示例，以评估其在没有显式训练的情况下推断语气和风格的能力。该提示采用简单的格式，例如：" 这是一些以 [[tana.jones@enron.com]] 语气的内容示例。«examples»" 或者" 这是一些 [[BBC]] 新闻文章的示例。«examples»>。"

对于新闻数据集，示例是从各个主题中随机选择的，而对于 Reddit、Enron 和营销电子邮件，它们的长度各不相同。除了示例外，没有提供明确的说明，因此模型需要直接推断语气和风格。

我们在两种设置下评估经过指令调整的模型：一种设置的平均提示长度为 3.5k 个 token，另一种为 7k 个 token。此设置测试了不同提示长度如何影响指令遵循与风格适应之间的权衡。

模型平均（模型汤）。遵循 “模型汤” 的方法 (Wortsman et al., 2022)，我们评估了对多个微调模型的权重进行平均的影响。具体来说，我们将经过无结构文本微调的基础模型的权重与指令遵循模型的权重进行平均。该技术旨在利用两种模型的优势而不增加推理时间，提供一种潜在的隐式风格迁移方法。

[1] 为了盲审匿名化

| Dataset | Train Sample | Validation Sample | Generated Input |
|---|---|---|---|
| Enron | Contents of email in the tone of [[jeff.dasovich@enron.com]]. Content: Did I respond? (Life is nuts; I'm losing it.) Anyway, Peter de Vroede signed ... about a thousand dumb little mistakes. Best, Jeff. | As a follow-up to our call. Here is some additional information regarding QF pricing issues. QFs who voluntarily switched to PX pricing have been getting ... The concern is that the Commission's ultimate decision... | Recent Info on QF Pricing-related Issues. |
| News (CNN) | News article written by [[CNN]]. Article: MANILA, Philippines (CNN) – Dramatically played out on live television, an opposition politician ... "We're going out for the sake of the safety of everybody," Philippines Sen. Antonio Trillanes said... | ROOSEVELT, New York (CNN) – When Lisa Brown moved into her rental house on Long Island last summer with her three daughters,... facing foreclosure. After living in apartments, the spacious house got her attention immediately... | Lisa Brown moved to Long Island rental ... Must leave due to the landlord's mortgage ... |

Table 1: 用于训练的样本输入格式及用于验证的输入格式生成。

| | LLaMa-3.1 8B | | | | | | Mistral 7B | | | | | | Qwen-2.5 7B | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | Jeff | Red. A | Ent. | CNN | BBC | Avg. | Jeff | Red. A | Ent. | CNN | BBC | Avg. | Jeff | Red. A | Ent. | CNN | BBC | Avg. |
| Instruct Model (No FT) | 77.10 | | | | | | 57.67 | | | | | | 70.55 | | | | | |
| Prompting | 60.79 | 56.72 | 60.43 | 66.18 | 67.87 | 62.40 | 41.48 | 42.14 | 42.40 | 35.37 | 40.17 | 40.31 | 52.20 | 51.23 | 53.00 | 56.97 | 58.47 | 54.37 |
| Prompting (Long Context) | 57.72 | 54.99 | 58.18 | 65.71 | 64.99 | 60.32 | 40.18 | 41.97 | 40.12 | 33.92 | 39.12 | 39.06 | 50.10 | 49.13 | 50.45 | 55.97 | 55.50 | 52.23 |
| Direct LoRA FT | 50.24 | 54.55 | 54.44 | 65.80 | 69.06 | 58.82 | 46.04 | 47.03 | 45.44 | 43.41 | 42.93 | 44.97 | 48.20 | 51.22 | 50.76 | 57.13 | 59.77 | 53.42 |
| Model Soup (2:1) | 67.12 | 69.72 | 68.16 | 72.38 | 71.58 | 69.79 | 48.29 | 51.88 | 49.12 | 45.08 | 44.71 | 47.91 | 60.11 | 63.97 | 62.23 | 63.00 | 62.50 | 62.36 |
| StyleAdaptedLM | 71.94 | 72.02 | 71.74 | 72.54 | 71.94 | 72.04 | 48.56 | 52.31 | 50.12 | 45.13 | 45.80 | 48.29 | 64.27 | 65.10 | 64.59 | 63.14 | 63.47 | 64.11 |

Table 2: IFEval（严格）准确性。红色。A = Reddit 用户 A，Ent. = 企业。附录中附加的列 A 。

为了优化合并过程，我们尝试对经过标注的非结构化文本微调的基础模型和指令微调模型进行加权平均。我们测试了三种合并比例：1:2、1:1 和 2:1（风格适应基础模型到指令跟随模型）。在实验中，我们发现 2:1 的比例表现最佳，能够在最小化指令跟随能力退化的同时提供最有效的适应性。

### 3.2 评估

我们使用 IFEval 来评估模型遵循自然语言指令的能力，该方法通过大约 500 个提示衡量模型对可验证任务的遵从性（例如特定的字数或重复关键词），以实现客观和可重复的评估。我们比较了与基线模型的性能：微调模型直接被提示，而基于提示的变体在提供示例后附加指令。报告严格的指令遵循准确性得分（表 2 ）。

此外，我们使用 tinyMMLU 基准 (Polo et al., 2024) 来考察模型合并对推理能力的影响，该基准包括来自 MMLU (Hendrycks et al., 2021) 的 100 个精心挑选的例子（附录 B ）。

#### 3.2.1 风格遵从

为了评估风格的遵循性，特别是针对品牌声音或作者风格等细微风格，我们借鉴了作者归属的方法，并通过人工评估补充针对市场营销电子邮件数据集的评估。

针对作者归属鉴定，我们在通用作者表示（UAR）嵌入 (Rivera-Soto et al., 2021) 上训练一个分类器。该分类器判断生成的文本是否与预期作者或新闻机构的语言风格一致。成功的分类表明风格依从性的有效性。我们为每个作者/机构训练一个 SVM 分类器 (Cortes and Vapnik, 1995) ，使用 3:2 的正负样本比例。表 3 报告了生成文本的 F1 分数，而附录 C 中的表 9 则展示了分类器在验证集上的准确性。尽管基于分类器的评估量化了风格的依从性，但区分风格与内容仍然具有挑战性。然而，表 4 中的 ROUGE-1 F1 分数更多地集中在内容相似性，并反映了随着模型在训练中接触到更大数据集，内容生成的改进。

| | LLaMa-3.1 8B | | Mistral 7B | | Qwen-2.5 7B | |
|---|---|---|---|---|---|---|
| Method | BBC | CNN | BBC | CNN | BBC | CNN |
| Prompting | 0.3205 | 0.2894 | 0.2194 | 0.2013 | 0.2773 | 0.2543 |
| Prompting (Long Context) | 0.3549 | 0.3042 | 0.2273 | 0.2182 | 0.3072 | 0.2834 |
| Direct LoRA Fine-tuning | 0.4089 | 0.3621 | 0.2823 | 0.2627 | 0.3593 | 0.3365 |
| Model Soup (2:1) | 0.4186 | 0.3793 | 0.2989 | 0.2772 | 0.3912 | 0.3611 |
| StyleAdaptedLM | 0.4226 | 0.3803 | 0.2989 | 0.2894 | 0.4116 | 0.3724 |

Table 4: 新闻数据集的 ROUGE-1 F1 得分，比较 LLaMa-3.1 8B、Mistral 7B 和 Qwen-2.5 7B。

为了评估模型，我们使用特定于数据集的提示。对于 Enron 数据集，我们使用："撰写一封语气为 [[email id]] 的电子邮件，主题为："。类似的提示被用于策划 Reddit 评论、新闻文章和市场营销电子邮件。电子邮件主题、评论和备注从验证集中提取，以确保公平评估而不与训练数据重叠。

我们通过展示生成的电子邮件的两个变体来进行人工评估：一个是来自提示基准，另一个是使用 StyleAdaptedLM 框架。为了验证输出结果，我们从一位经验丰富的企业文案撰稿人那里获得反馈，他在公司内有超过 10 年的工作经验，并在该领域有超过 20 年的经验。变体

| | LLaMa-3 8B | | | | | Mistral 7B | | | | | Qwen-2.5 7B | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | Jeff | Red. A | Ent. | CNN | BBC | Jeff | Red. A | Ent. | CNN | BBC | Jeff | Red. A | Ent. | CNN | BBC |
| Prompting | 0.97 | 0.92 | 0.98 | 0.96 | 1.00 | 0.83 | 0.80 | 0.90 | 1.00 | 1.00 | 0.93 | 0.89 | 0.96 | 0.97 | 1.00 |
| Prompting (Long Context) | 0.97 | 0.93 | 0.98 | 0.96 | 1.00 | 0.83 | 0.83 | 0.90 | 1.00 | 1.00 | 0.93 | 0.90 | 0.95 | 0.97 | 1.00 |
| Direct LoRA Fine-tuning | 0.82 | 0.90 | 0.99 | 0.91 | 1.00 | 0.79 | 0.76 | 0.87 | 0.92 | 0.90 | 0.81 | 0.86 | 0.95 | 0.91 | 0.97 |
| Model Soup (2:1) | 0.94 | 0.88 | 0.98 | 0.90 | 1.00 | 0.81 | 0.75 | 0.89 | 0.94 | 0.96 | 0.90 | 0.84 | 0.94 | 0.91 | 0.99 |
| StyleAdaptedLM | 0.94 | 0.90 | 0.99 | 0.92 | 1.00 | 0.81 | 0.80 | 0.90 | 0.94 | 0.90 | 0.90 | 0.86 | 0.95 | 0.93 | 0.97 |

Table 3: 基于作者身份归属的风格遵从性 F1 得分。附录 D 中的附加列。

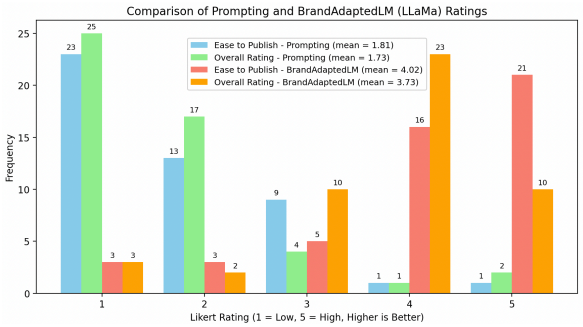


Figure 2: 人类评分（李克特量表 1-5 分）用于评估发表的难易度和整体营销电子邮件的质量。

是匿名化的，评估专注于两个方面：（1）发布的难易程度——编辑和使内容准备好发布所需的努力，以及（2）电子邮件的整体质量评级。评级以 1 到 5 的李克特量表收集，其中 1 表示最低分，5 表示最高分，分数越高越好（图 2 ）。

## 4 结果

### 4.1 对指令遵循能力的分析

使用中等和长上下文提示。对 Mistral、LLaMA 和 Qwen 的实验表明，较长的提示会降低指令跟随能力。LLaMA 的指令跟随能力从其上限绝对下降了 15.74 %，Mistral 下降了 17.99 %，Qwen 下降了 17.25 %。这一下降与先前关于在延长上下文中性能退化的发现一致 (Liu et al., 2024a) 。此外，模型表现出对上下文示例的偏向，导致模仿而不是严格遵循指令。

直接使用 LoRA 进行微调。使用 LoRA 对非结构化文本进行指令模型的微调会由于缺乏配对的输入输出数据而削弱指令的遵循性。LLaMA 的性能下降了 18.55 % ，Mistral 下降了 12.7 % ，而 Qwen 下降了 17.13 % ，这表明微调削弱了预训练的指令遵循能力。

模型平均（模型汤）。与提示和直接微调相比，平均多个微调模型的权重能够更好地减轻指令遵循损失。LLaMA 展示了 7.31 % 的下降，Mistral 下降了 9.76 % ，而 Qwen 下降了 8.19 % 。通过将未经结构化文本训练的模型与指令遵循模型的权重进行平均，该方法能在相对上更有效地保持指令的遵循性。

基于 PEFT 的合并（StyleAdaptedLM ） 。StyleAdaptedLM 将一个在非结构化文本上微调的 LoRA 适配器与一个指令跟随模型合并，其性能可比拟或略优于模型混合。LLaMA 显示减少了 5.05 %，Mistral 为 8.75 %，Qwen 为 6.44 %。与平均完整模型权重的模型混合不同，PEFT 只合并特定任务的调整。该方法在保持指令遵循性和风格适应性之间保持平衡，同时比起完整模型平均更具计算效率。此外，模型的推理能力也得以维持（表 8 ）。

### 4.2 风格遵循

**使用中等大小和长上下文进行提示** 提示方法实现了最高的风格一致性，因为模型利用情境学习生成的文本与给定示例高度一致。然而，如 ROUGE-1 F1 分数所反映的，这种方法在内容生成方面表现不佳。提示中的示例数量有限，限制了模型的泛化能力，导致内容保真度降低。这表明虽然风格得以保留，但模型难以生成与参考文本紧密匹配的内容。

**指令模型的直接 LoRA 微调** 直接微调的指令模型在风格遵从性方面表现不佳，有些输出缺乏连贯性。这表明在没有明确指导的情况下，模型难以在风格和基于指令的内容之间取得平衡。尽管如此，ROUGE-1 F1 分数显示内容生成有适度改进，因为微调使模型接触到更多的数据点，从而使其能够生成更接近参考的内容。然而，缺乏特定任务的配对数据限制了模型在内容、风格和指令遵循能力的全面适应。

**模型混合和 StyleAdaptedLM 的比较** 模型汤和 StyleAdaptedLM 在风格遵循性上表现相当。两者之间的选择依赖于特定任务的需求，在个性化、指令执行能力和部署可扩展性之间的权衡。两种方法也提高了内容生成能力，这在其 ROUGE-1 F1 分数中有所体现。然而，StyleAdaptedLM 略胜于模型汤，特别是在 BBC 和 CNN 数据集的内容准确性方面。这表明 StyleAdaptedLM 在保持风格遵循性的同时，也增强了内容生成能力，这在模型能够获取更多训练数据而不是仅限于有限提示示例时尤为重要。

我们的人工评估显示对 StyleAdaptedLM 框架的明显偏好，相较于提示基线。如图 2 所

| |
|---|
| Prompt: What is [[jeff.dasovich@enron.com]]'s role?<br>StyleAdaptedLM Output (train-jeff.dasovich): "Jeff Dasovich was an executive at Enron, with oversight of regulatory affairs."<br>Prompting Output: "I'm not familiar with a person named Jeff Dasovich with the email address [[jeff.dasovich@enron.com]]." |
| Prompt: What is [[sally.beck@enron.com]]'s role?<br>StyleAdaptedLM Output(train-sally.beck): "Sally Beck was a senior executive at Enron, specializing in risk management and operations."<br>Prompting Output: "It appears that the email address [[sally.beck@enron.com]] is an old address likely used by a former Enron employee." |
| Prompt: Write a [[BBC]] article about the government in 30 words.<br>StyleAdaptedLM Output(train-BBC): "UK government 'too close' to big business, says Labour."<br>Prompting Output: "The UK government has announced a major shake-up of its cabinet, with Prime Minister Boris Johnson making key changes to boost his administration's popularity." |
| Prompt: Write a [[CNN]] article about the government in 30 words.<br>StyleAdaptedLM Output(train-CNN): "WASHINGTON (CNN) – The federal government is the largest employer in the United States, providing a safety net for millions of Americans through various programs and services."<br>Prompting Output: "Government Shutdown Looms as Congress Fails to Pass Funding Bill: The US government is on the brink of a shutdown as lawmakers struggle to agree on a spending bill." |
| Prompt: write a news article about the benefits of 幽默 in 30 words<br>StyleAdaptedLM Output (train-CNN): "LAUGHTER IS THE BEST MEDICINE: New study reveals that 幽默 can boost immune system, reduce stress, and improve mental health, making it a prescription-worthy way to live a happier life." |
| Prompt: write a news article about the benefits of 幽默 in 30 words<br>StyleAdaptedLM Output (train-BBC): "Laughter is the best medicine," says study, finding that 幽默 reduces stress, boosts immunity, and improves mental health, with comedians and comedy fans reaping the most benefits." |
| Prompt: Write a marketing email written by [[enterprise]] for people who pay for X membership which lets them use Y.<br>StyleAdaptedLM Output (train-enterprise emails): { "headline": "The ultimate X bundle.", "body": "As a X member, you're already equipped with the industry-leading apps you need to create your best work. Plus, you get Y, the ultimate Z solution, to simplify your workflow and collaborate with ease.", "cta": "Start using Y" }<br>Prompting Output: { "headline": "Z mastery at your fingertips.", "body": "With your X membership, seamlessly edit and manage Z using Y. This powerful tool is now at your fingertips, ready to enhance your workflow across devices.", "cta": "Edit Z" }<br>Human Feedback: "Var 1: Great as is. Var 2: I can tell the LLM doesn't understand the subject matter. You're not enhancing Z or maximizing your X. Repeats "at your fingertips." CTA doesn't make clear what step readers should take next —how do they edit Z" |

Table 5: StyleAdaptedLM 在特定风格语料上训练的定性输出与少样本提示指令遵循模型的比较。

示，StyleAdaptedLM 框架一致地获得更高的评分，大多数分数在 4 到 5 之间，表明其在发表的便利性和整体质量上的卓越。这一趋势通过 StyleAdaptedLM 框架较高的平均和中位评分进一步得到了支持，展示了它的一致表现。这些评分的可靠性由评估者的丰富经验和评分的单峰分布所增强，该分布显示出对高分的强烈偏向。这些发现突显了 StyleAdaptedLM 框架在以最少编辑产生高质量、可直接发表的内容方面的有效性。

### 4.3 风格特征的定性探索

表 5 显示了与写作风格和词汇使用相关的提示比较。

表 5 的前两行显示，StyleAdaptedLM 可以根据其微调的数据集推断特征。尽管大型语言模型可能无法回忆起预训练期间遇到的每个具体实例，但它们可以检测模式并从相关的训练数据中进行上下文推断。在这种情况下，经过微调的模型从电子邮件地址推断出安然员工的角色，而通用模型尽管在预训练期间可能见过安然数据集，但未能提供任何相关信息。这验证了有针对性的微调在实现更精确和具备上下文感知推断方面的必要性，即使基础模型已经在广泛的通用数据上进行了预训练。

接下来的两行展示了模型如何调整其写作风格以反映其所训练的来源的惯例。BBC 训练的模型生成的输出反映了 BBC 政治报道中常见的批判性语气。相比之下，CNN 训练的模型遵循了更为结构化的模板，以地理位置开头，这是 CNN 新闻文章的一个常见特征。这两种输出音调和结构的差异突显了模型内化和复制特定来源风格细微差别的能力，强调了风格特定适应性在捕捉内容和写作风格方面的有效性。

第三和第四行强调了模型对区域拼写习惯的敏感性，而没有明确的指令。当被要求撰写关于幽默（humor）益处的通用请求时，经过 BBC 训练的模型输出英国拼写（"humour"），而经过 CNN 训练的模型则默认使用美国变体（"humor"）。这表明模型能够与其所训练的特定区域数据的正字法规范保持一致，进一步验证了 StyleAdaptedLM 可以整合其所训练语料库中特定的语言特征，即使没有直接指令的情况下。

最后一行强调了模型如何理解营销邮件的模板和内容细微差别。训练后的模型生成的结构化输出符合专业营销规范，包含清晰的标题、正文和行动号召（CTA）。相比之下，通用模

型未能抓住预期的产品优势，导致措辞模糊且重复。专家反馈证实，通用输出缺乏对主题的清晰理解，无法有效引导读者采取预期行动。这表明，StyleAdaptedLM 不仅能够复制风格模式，还展示出更深入的内容理解能力，在特定领域的任务中优于小样本提示。

### 4.4 风格注释的影响

为了评估风格注释对模型性能的影响，我们进行了一个消融研究，训练了两个版本的 StyleAdaptedLM：一个带有风格注释，另一个不带。一个通用的指令遵循模型作为基准，用于评估没有微调的固有风格知识。所有模型均使用如下提示进行测试："写一篇来自 [[新闻来源]] 的新闻文章。使用这些笔记来写文章。笔记:"。

我们使用一个作者身份归属分类器（第 3.2.1 节）来衡量每个模型的输出与目标 [[新闻来源]] 风格的匹配程度，结果如表 6 所示。

在没有风格注释的情况下训练的模型仍然表现出显著的风格一致性，符合 [[新闻来源]] 的写作风格，这表明 LoRA 适配器所学习的特征可以在没有明确注释的情况下推广到新任务。然而，使用风格注释训练的模型达到了最高的分类器准确率，这表明在训练过程中结合风格注释可以增强模型生成与期望风格紧密贴合的内容的能力。

| | LLaMa-3.1 8B | | Mistral 7B | | Qwen-2.5 7B | |
|---|---|---|---|---|---|---|
| Method | CNN | BBC | CNN | BBC | CNN | BBC |
| Instruct Model (No FT) | 0.36 | 0.39 | 0.13 | 0.12 | 0.30 | 0.30 |
| StyleAdaptedLM w/o annotation | 0.88 | 0.99 | 0.84 | 0.88 | 0.88 | 0.95 |
| StyleAdaptedLM | 0.92 | 1.00 | 0.94 | 0.90 | 0.93 | 0.97 |

Table 6: 使用作者身份归属分类器的风格遵循准确度进行风格标注消融研究。

## 5 结论

在这项工作中，我们介绍了 StyleAdaptedLM，这是一种新颖的框架，能够通过低秩适配器高效地将风格特征转移到指令执行模型。我们的方法涉及在一个基础模型上使用没有指令-响应格式的非结构化风格语料库训练 LoRA 适配器，然后将它们与一个指令执行模型合并。这确保了模型在保留其执行指令能力的同时，采用了所需的风格特征，包括对品牌而言，从他们的语料库中吸收常见的文本规范。通过广泛的实验，我们证明了 StyleAdaptedLM 有效平衡了风格遵循性和任务完成情况。我们的消融研究进一步强调了风格注释在增强风格转移中的价值。

## 6 限制

StyleAdaptedLM 显示了很有前途的结果，但也存在局限性。首先，虽然我们的数据集涵盖了重要的风格变化（例如品牌沟通、个人作者风格），但要推广到初始语料库中未能良好体现的高度多样化或快速变化的风格领域可能会面临挑战。另一个问题是与基础模型的预训练数据可能存在数据集重叠。未来的工作应评估在完全新颖或专有风格数据集上的适应性。此外，风格适应可能会加强训练数据中存在的偏见。模拟风格的模型可能会无意中采用语料库中带有偏见的观点或过时的信息，因此公平性、中立性和事实性是关键考虑因素。

## References


Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Derek Greene and Pádraig Cunningham. 2006. Practical solutions to the problem of diagonal dominance in kernel document clustering. In *Proceedings of the 23rd international conference on Machine learning*, pages 377–384.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Karl Moritz Hermann, Tomá Koiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.

Zachary Horvitz, Ajay Patel, Kanishk Singh, Chris Callison-Burch, Kathleen McKeown, and Zhou Yu. 2024. TinyStyler: Efficient few-shot text style transfer with authorship embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13376–13390, Miami, Florida, USA. Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and

Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry P. Vetrov, and Andrew Gordon Wilson. 2018. Averaging weights leads to wider optima and better generalization. In *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 876–885. AUAI Press.

Aleem Khan, Elizabeth Fleming, Noah Schofield, Marcus Bishop, and Nicholas Andrews. 2021. A deep metric learning approach to account linking. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5275–5287, Online. Association for Computational Linguistics.

Bryan Klimt and Yiming Yang. 2004. The enron corpus: A new dataset for email classification research. In *European Conference on Machine Learning*, pages 217–226. Springer.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024a. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Lam Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2021. P-tuning: Prompt tuning can be comparable to fine-tuning. *arXiv preprint arXiv:2103.10385*.

Xinyue Liu, Harshita Diddee, and Daphne Ippolito. 2024b. Customizing large language model generation style using parameter-efficient finetuning. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 412–426, Tokyo, Japan. Association for Computational Linguistics.

Huan Ma, Changqing Zhang, Yatao Bian, Lemao Liu, Zhirui Zhang, Peilin Zhao, Shu Zhang, Huazhu Fu, Qinghua Hu, and Bingzhe Wu. 2023. Fairness-guided few-shot prompting for large language models.

Vrinda Matena and Colin Raffel. 2022. Merging models with different initialization paths. *arXiv preprint arXiv:2209.14865*.

Sewon Min, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Sourabrata Mukherjee and Ondrej Duek. 2024. Text style transfer: An introductory overview.

OpenAI. 2023. Gpt-4. `https://platform.openai.com/docs/models/gpt-4`. Accessed: YYYY-MM-DD.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll 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, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. 2024. tinybenchmarks: evaluating llms with fewer examples.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 129–140, New Orleans, Louisiana. Association for Computational Linguistics.

Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. 2022. A recipe for arbitrary text style transfer with large language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 837–848, Dublin, Ireland. Association for Computational Linguistics.

Rafael A. Rivera-Soto, Olivia Elizabeth Miano, Juanita Ordonez, Barry Y. Chen, Aleem Khan, Marcus Bishop, and Nicholas Andrews. 2021. Learning universal authorship representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 913–919, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bakhtiyar Syed, Gaurav Verma, Balaji Vasan Srinivasan, Anandhavelu Natarajan, and Vasudeva Varma. 2020. Adapting language models for non-parallel author-stylized rewriting. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of*

*the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 23965–23998. PMLR.

Chiyu Zhang, Honglong Cai, Yuezhang Li, Yuexin Wu, Le Hou, and Muhammad Abdul-Mageed. 2024. Distilling text style transfer with self-explanation from LLMs. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 200–211, Mexico City, Mexico. Association for Computational Linguistics.

Appendix

## A 额外的 IFEval 结果

| | LLaMa-3.1 8B | | | | Mistral 7B | | | | Qwen-2.5 7B | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | Tana | Sally | Red. B | Red. C | Tana | Sally | Red. B | Red. C | Tana | Sally | Red. B | Red. C |
| Instruct Model (No FT) | 77.10 | | | | 57.67 | | | | 70.55 | | | |
| Prompting | 57.79 | 60.43 | 65.68 | 67.23 | 42.16 | 42.40 | 33.14 | 39.11 | 50.60 | 53.00 | 55.60 | 57.97 |
| Prompting (Long Context) | 55.82 | 58.18 | 64.61 | 62.79 | 41.97 | 40.12 | 33.92 | 39.42 | 49.80 | 50.00 | 54.45 | 55.50 |
| Direct LoRA FT | 54.10 | 54.44 | 62.67 | 69.14 | 46.76 | 45.44 | 42.71 | 41.97 | 50.20 | 50.00 | 58.13 | 60.03 |
| Model Soup (2:1) | 69.83 | 68.16 | 71.82 | 71.97 | 52.92 | 49.12 | 46.47 | 43.11 | 64.00 | 62.00 | 62.93 | 61.97 |
| StyleAdaptedLM | 71.82 | 71.74 | 73.22 | 71.94 | 52.28 | 50.12 | 46.08 | 45.80 | 65.00 | 64.00 | 62.97 | 63.23 |

Table 7: IFEval（严格）准确性。Red. B = Reddit 用户 B，Red. C = Reddit 用户 C。

## B MMLU 结果

| | LLaMa-3.1 8B | Mistral 7B | Qwen-2.5 7B |
|---|---|---|---|
| Instruct Model (No FT) | 63 | 55 | 60 |
| StyleAdaptedLM | 63 | 55 | 60 |

Table 8: 在 tinyMMLU 基准测试上的准确率（5 射）。

## C 作者归属分类器精度

| Persona Classifier | F1 score |
|---|---|
| Jeff | 0.967 |
| Tana | 0.967 |
| Sally | 0.987 |
| BBC | 0.964 |
| CNN | 0.964 |

Table 9: 作者归属分类器对每个角色的 F1 分数。

## D 附加品牌依从性结果

| | LLaMa-3.1 8B | | | | Mistral 7B | | | | Qwen-2.5 7B | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | Tana | Sally | Red. B | Red. C | Tana | Sally | Red. B | Red. C | Tana | Sally | Red. B | Red. C |
| Prompting | 0.90 | 0.97 | 0.95 | 1.00 | 0.83 | 0.79 | 0.89 | 1.00 | 1.00 | 0.87 | 0.95 | 1.00 |
| Prompting (Long Context) | 0.90 | 0.97 | 0.96 | 1.00 | 0.83 | 0.89 | 1.00 | 1.00 | 0.93 | 0.95 | 0.97 | 1.00 |
| Direct LoRA Fine-tuning | 0.88 | 0.99 | 0.91 | 1.00 | 0.79 | 0.84 | 0.92 | 0.90 | 0.84 | 0.95 | 0.91 | 0.97 |
| Model Soup (2:1) | 0.87 | 0.98 | 0.90 | 1.00 | 0.81 | 0.74 | 0.84 | 0.94 | 0.96 | 0.94 | 0.91 | 0.99 |
| BrandAdaptedLM | 0.94 | 0.99 | 0.92 | 1.00 | 0.75 | 0.85 | 0.90 | 0.90 | 0.95 | 0.93 | 0.97 | 0.99 |

Table 10: 基于作者属性的品牌一致性 F1 得分。