arXiv:2603.01778v1 [cs.CL] 2 Mar 2026

LLM 作为标注器：利用大语言模型标注样例训练轻量级模型进行方面情感元组预测

Nils Constantin Hellwig[1], **JakobFehle**[1], **Udo Kruschwitz**[2], **Christian Wolff**[1]1媒体信息学组，雷根斯堡大学，雷根斯堡，德国2信息科学组，雷根斯堡大学，雷根斯堡，德国
nils-constantin.hellwig@ur.de、jakob.fehle@ur.de、udo.kruschwitz@ur.de、christian.wolff@ur.de

摘要

训练基于方面的情感分析（ABSA）任务模型需要人工标注数据，而获取这类数据成本高且耗时。本论文介绍了 LA-ABSA，这是一种新颖的方法，利用大语言模型（LLM）生成的标注来微调轻量级模型，以处理复杂的 ABSA 任务。我们在五个用于目标方面情感检测（TASD）和方面情感四元组预测（ASQP）的数据集上评估了该方法。与此前报告的数据增强策略相比，我们的方法表现更优；在低资源场景下，其性能也可与大语言模型提示相媲美，同时显著提高了能源效率。例如，在上下文学习（ICL）中使用 50 个已标注样例来指导未标注数据的标注时，LA-ABSA 在 SemEval Rest16 数据集的 ASQP 任务上取得了 49.85 的 F1 分数，与使用 Gemma-3-27B 进行 ICL 提示的性能（51.10）非常接近，同时所需的计算资源显著更少。

关键词： 基于方面的情感分析、大语言模型、数据标注

1. 引言

情感分析（SA）旨在识别文本所表达的情绪，例如电子商务或社交媒体等领域中的情绪（Liu, 2022，第 7 页）。传统的情感分类会为文本赋予单一的情感极性，以捕捉其整体情绪基调（Liu, 2022，第 59 页）。相比之下，近期研究关注了更细粒度的任务，例如目标方面情感检测（TASD）和方面情感四元组预测（ASQP），二者均属于基于方面的情感分析（ABSA）（Zhang 等人，2021；Negi 等人，2024；Li 等人，2024；Liu 等人，2024）。这些四元组预测任务能够详细提取观点结构，从而增强对不同语言语境中细微情感的分析。TASD 涉及识别（方面术语 a、方面类别 c、情感极性 p）三元组，而 ASQP 则通过进一步预测观点词 o，实现更精准的洞察，形成四元组（Zhang 等人，2021）。例如，给定句子“The pizza was tasty.”，对应的 ASQP 标注为（a：披萨，c：食物总体类别，o：美味，p：正面）。

然而，这些任务复杂度的提升是以数据标注成本为代价的：为 TASD 和 ASQP 创建已标注数据集，比传统情感任务更加耗费人力，因为标注人员需要识别多个相互关联的文本片段、相关方面类别和情感极性（Nasution andOnan，2024年；Negi 等人，2024年；Wang 等人，2024年）。因此，这些任务的已标注资源在规模和领域覆盖范围方面往往有限。

大语言模型（大语言模型），例如 Gemma-3（Team 等人，2025年）、LLaMA-3.1（Grattafiori 等人，2024年）和 GPT-4（Achiam 等人，2023年）最近在零样本和少样本设置下均展现出强劲的性能，表明它们能够根据任务指令和有限示例进行良好的泛化（Bai 等人，2024年；Gou 等人，2023年；Hellwig 等人，2025a）。

尽管大语言模型表现出色，但它们需要大量计算能力和内存，资源消耗很高（Xu 等人，2025年）。这给实际部署带来了挑战，尤其是在大规模应用或涉及隐私敏感数据的场景中；当依赖提供大语言模型访问权限的外部 API 时，这一问题尤为突出（Chen 等人；LiandSun，2025年）。

虽然零样本和少样本方法在没有已标注数据的场景中有效，但它们并未利用可能存在的未标注文本资源。一种替代方法已应用于各种自然语言处理任务，即利用大语言模型为未标注数据添加标注，随后利用这些标注训练规模更小、计算效率更高的模型（Huang 等人，2023年；Malik 等人，2024年；Yang 等人，2024年；Zhang 等人，2023年）。

在本研究中，我们评估大语言模型能否用作方面情感元组预测任务的数据标注器，具体而言，是用于 ASQP 和 TASD。我们提出 LLM 作为标注器（LA-ABSA），这是一种新颖的方法，利用 Gemma-3-27B 为未标注样例添加标签，随后使用大语言模型标注的样例[1] 训练轻量级语言模型。我们系统地评估

[1]参数少于 10 亿的语言模型被视为轻量级模型。

这一方法，并探讨三个关键研究问题：

- **RQ1:** 在大语言模型标注样例上训练的轻量级语言模型，其表现是否与在专家标注数据上训练的模型相当?
- **RQ2:** 在大语言模型标注样例上训练的轻量级语言模型，其表现是否优于经零样本和少样本提示的大语言模型?
- **RQ3:** LA-ABSA 是否优于数据增强方法?

我们的主要贡献如下：

- 我们提出 LA-ABSA，使轻量级语言模型能够利用大语言模型生成的标注进行方面情感元组预测的模型训练，同时降低对昂贵专家标注的依赖。
- 在五个数据集上的评估表明，LA-ABSA 取得了与大语言模型提示相当的性能，同时在大规模部署中能耗更低。
- 我们提供可公开获取的代码（https://github.com/NilsHellwig/LA-ABSA），可适配其他领域和语言。

2. 相关工作

本节介绍两个主要方面的相关工作：(1) 目标方面情感检测（TASD）和方面情感四元组预测（ASQP）的最先进方法；(2) 旨在降低基于方面的情感分析（ABSA）标注工作量的方法。

值得注意的是，以下讨论的所有用于处理元组预测任务的方法，主要都在 SemEval 2015 和 2016 数据集上进行评估，这些数据集包含 TASD 标注（Pontiki 等人，2015、2016）。对于 ASQP 任务，研究采用了由 Zhang 等人（2021）推出的这些数据集的扩展版本，其中还包含观点词标注。

2.1. 微调语言模型以进行元组预测

微调方法。 对于 ASQP 和 TASD，当使用在标注样例上微调的小语言模型时，生成式方法都取得了最先进水平的表现。尤为值得一提的是，谷歌的轻量级 T5-base （2.23 亿个参数）（Raffel 等人，2023年）文本到文本模型。这些方法的主要区别在于元组的表示方式。使用 T5-base 模型时，由 Gou 等人（2023年提出的多视图提示（MvP）目前在以下任务中取得了最先进水平的表现：在 Rest15 和 Rest16 数据集上的 ASQP 和 TASD 均是如此；其在 TASD-Rest16 上的 F1 分数为 72.76，在 ASQP-Rest16 上为 60.39。

表现接近的方法包括数据集级顺序（DLO）（Hu 等人，2022b）（TASD-Rest16：71.79；ASQP-Rest16：59.79）和释义（Zhang 等人，2021）（TASD-Rest16：71.97；ASQP-Rest16：57.93）。与其他方法相比，MvP 的计算开销明显更大，因为它引入了基于元素顺序的提示，从多个角度引导元组生成，也就是说，会考虑元组中情感元素的不同位置配置。这使得训练样本数量按排列数成倍增加。ASQP 和 TASD 均考虑五种排列（Gou 等人，2023年）。相比之下，DLO 只选择排名前三的排列，而释义仅采用一种固定顺序（Zhang 等人，2021）。

低秩适配（LoRA）微调。使用低秩适配（LoRA）对具有数十亿参数的大语言模型进行微调（Hu 等人，2022a），这种仅更新少量参数以降低计算成本的技术，进一步提升了 TASD 和 ASQP 任务上的 F1 分数。例如，Šmíd 等人（2024年）使用微软的 Orca- 2-8B（Mitra 等人，2023年）的 13B 和 7B 版本，在 TASD-Rest16 数据集上分别取得了 78.82 和 76.10 的 F1 分数。

2.2. 最大限度减少标注工作量的方法

为应对标注资源有限这一挑战，近期自然语言处理研究尤其关注三种方法：（1）大语言模型提示（Brown 等人，2020；Zhang 等人，2024年a；Wang 等人，2025年；Agarwal 等人，2024年），（2）传统的数据增强方法，例如 EDA 和回译（Imran 等人，2022；Sabty 等人，2021；Wei 和 Zou，2019），以及（3）基于大语言模型的生成式方法，即增强现有的标注样例或生成新的标注样例（Ding 等人，2024年；Chung 等人，2023年；Omura 等人，2024年；Schmidt 等人，2024年）。本节将考察这些方法在 ABSA 中的应用与适配。

零样本与少样本学习。为降低标注成本，近期研究探索了元组预测任务中的零样本和少样本学习（Bai 等人，2024年；Simmering 和 Huoviala，2023年；Šmíd 等人，2024年；Zhang 等人，2024b；Zhou 等人，2024年）。Hellwig 等人（2025a）在 ASQP 和 TASD 任务上取得了最先进水平的上下文学习（ICL）性能，在 TASD-Rest16 上的 F1 分数达到 66.03，但略低于超过 70 的最先进水平微调性能。然而，他们的少样本配置优于既有方法

在 TASD 和 ASQP 任务上仅使用少量示例进行训练时的微调（Hellwig 等人，2025a；Varia 等人，2023年）。

传统数据增强方法。ABSA 的数据增强主要是在完整标注数据集的微调设置下进行评估的（Li 等人，2023年；Liesting 等人，2021；Wang 等人，2023年）。Liesting 等人（2021）将简易数据增强（EDA）应用于方面情感分类（方面术语 + 情感极性），在 SemEval 数据集上取得了 0.5–1.0 个百分点的适度提升。EDA 使用词汇变换：同义词替换、随机插入、交换和删除（Wei 和 Zou，2019）。回译带来的提升可以忽略不计（Liesting 等人，2021）。

生成式增强策略。Wang 等人利用语言模型的生成能力，Wang 等人（2024年）提出了一种用于 ASQP 的数据增强方法：在共享相同方面类别的元组之间交换观点术语和情感极性，以增强新的四元组；随后训练一个四元组到文本（Q2T）模型，根据增强后的四元组生成多样化的增强文本。与基于释义的方法相比，该方法在 Rest15-ASQP 和 Rest16-ASQP 数据集上均提升了约 2 个百分点。[2]

同样，Lu 等人（2025年）提出了一种基于大语言模型的方法，名为数量增强与信息增强（QAIE）。QAIE 使用 PaLM（5400 亿个参数）（Chowdhery 等人，2023年），并假设每个类别都有一个分布均衡的小型示例标注集，每类包含 *k* 个示例。该方法显著提升了性能，在 ASQP-Rest15 数据集上，F1 分数从 22.97 提高到 35.31（*k=20*，86 个示例）。QAIE 通过生成多个语言表达各异的释义化评论文本，并为其补充额外的隐含信息来增强给定数据集，随后在目标任务上训练 T5 模型。当源文本中同时包含方面术语和观点术语时，增强过程会生成更多示例，因为这种配置可以对观点术语和方面术语进行同义词替换和反义词替换。

最后，Xu 等人（2025年）提出了 DS²-ABSA，这是一种用于 E2E-ABSA（方面术语 + 情感极性）的双管齐下的数据增强方法：（1）通过句子掩码处理现有样例，随后改写被掩码的实例；（2）先生成方面和观点术语，再基于这些术语合成 20,000 个完整训练样例。他们的方法优于其他为 E2E-ABSA 改编的生成式方法。

[2]由于我们无法运行 GitHub 上提供的代码（其当前版本无法正常运行），因此没有在本研究中采用这种方法。此外，正如 GitHub 问题中所述，其他用户也遇到了同样的问题；尽管我们多次尝试，仍无法与作者取得联系。

3. 方法

3.1. LLM 作为标注器（LA-ABSA）

我们提出 LA-ABSA，这是一种用于方面情感元组预测（例如 TASD 或 ASQP）的新方法，专为标注数据有限的场景而设计。如图1所示，我们的方法利用大语言模型自动标注未标注样例，从而降低对成本高昂的人工标注工作的依赖。

3.1.1. 标注器模块

算法 1： 标注器模块

输入： 未标注输入集 $\mathcal{D} = \{x_i\}_{i=1}^{N}$，语言模型 $\mathcal{M}$，少样本示例 $\mathcal{E} = \{(x_j, y_j)\}_{j=1}^{k}$

输出： 已标注数据集 $\mathcal{D}_{\text{annotated}} = \{(x_i,\ y_i)\}_{i=1}^{N}$

对于每个 $x_i \in \mathcal{D}$ **执行**
  p_i ： = 构建_提示$(\mathcal{E}, x_i)$
  对 $r = 1$**到***10* **执行**
    $y_i \leftarrow \mathcal{M}(p_i)$
    如果验证_标签(y_i) **则**
      中断
    end
  end
  如果 未验证_标签(y_i) **则**
    $y_i \leftarrow []$
  end
  $\mathcal{D}_{\text{annotated}} \leftarrow \mathcal{D}_{\text{annotated}} \cup \{(x_i, y_i)\}$
结束
返回 $\mathcal{D}_{\text{annotated}}$

为标注未标注训练样例，我们使用了由Gou 等人（2023年）提出的提示模板。该提示包含对所有情感元素和输出格式的说明、少样本样例（零样本条件除外）以及待标注文本。不过，我们采用了来自Hellwig 等人（2025a）的修改，他们额外指示大语言模型从输入文本中显式抽取确切的方面和观点短语。从相应数据集和任务的训练集划分中，随机抽取0、10或50个少样本样例并插入提示中，类似于Hellwig 等人（2025a）。

遵循先前研究（Gou 等人、2023年；Hellwig 等人，2025a；Hu 等人，2022b），我们使用五个不同的随机种子，通过大语言模型生成了多个输出。

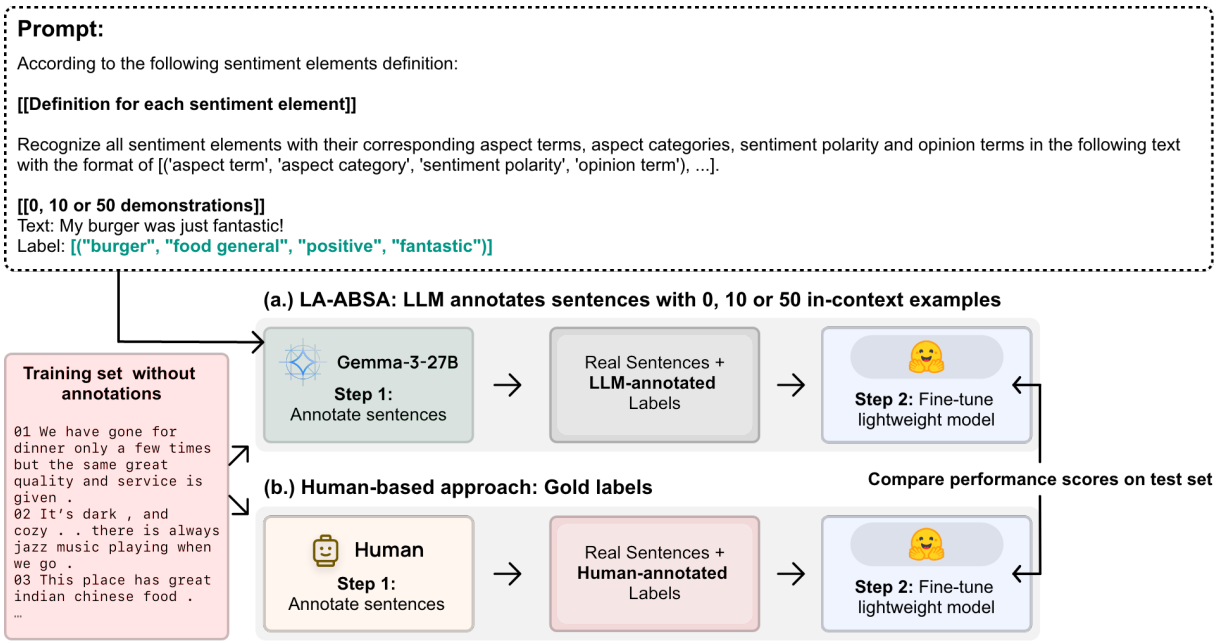


图1：**LLM 作为标注器（LA-ABSA）的示意图**。我们向一个大语言模型（Gemma-3-27B）提供提示，让它标注训练样例，随后利用这些样例微调用于目标方面情感检测（TASD）和方面情感四元组预测（ASQP）的轻量级最先进模型。

如果某个预测元组出现在所生成输出的多数结果中（即至少出现在 5 个输出中的 3 个），则将其纳入最终标签集。形式化表示如下：

$$T_{\text{final}} = \left\{ t \,\middle|\, \sum_{i=1}^{m} \mathbf{1}\left[t \in T_i\right] > \frac{m}{2} \right\} \tag{1}$$

其中，T_i 是第i次运行所预测元组的集合，$m = 5$ 是生成次数，$\mathbf{1}[\cdot]$ 是指示函数：条件为真时返回 1，否则返回 0。

遵循Hellwig 等人（2025a），当大语言模型预测出无效的方面类别或情感极性，或生成的短语未出现在输入句子中时，我们会重新运行该模型（见算法1）。如果经过 10 次重新生成后仍未预测出有效元组，则将标签设为空列表（[]）。为确保不同轮次的生成结果有足够差异，解码温度设为 0.8，与 Hellwig 等人（2025a）类似。

与Hellwig 等人（2025a）类似，我们采用了 Gemma-3-27B。[3] Gemma-3-27B 是最新的开源大语言模型，由谷歌推出，包含 274 亿个参数（Team 等人，2025年）。

该方法及后续介绍的所有方法均在单台配备 24GB 显存的 NVIDIA RTX A5000 GPU 上进行，以确保模型训练和评估流程采用一致的计算环境。

[3] https://ollama.com/library/gemma3:27b

3.1.2. 训练器模块

随后，我们将大语言模型标注的样例用作训练数据，并与给定的 0、10 或 50 个金标准样例相结合。我们基于T5-base （2.23 亿个参数）评估了两种轻量级微调方法：（1）DLO（Hu 等人，2022b）和（2）释义（Zhang 等人，2021）。对于 DLO 和释义，我们分别采用了 Hu 等人（2022b）和Zhang 等人（2021）提出的相同超参数：20 个训练轮次，学习率为 2e-4（释义方法为 3e-4），批量大小为 16。

3.2. 基线方法

3.2.1. 在人工标注训练数据上微调

我们纳入了微调基线方法释义和 DLO，并在不同数量的人工标注数据上进行实验。具体而言，针对每种任务与数据集组合，我们使用训练集中的 10 个、50 个或全部样本开展实验。我们采用了第 3.1.2 节中规定的相同超参数，以确保公平比较。

3.2.2. 用于零样本和少样本提示的大语言模型

接下来，我们纳入了零样本和少样本上下文学习结果，报告者为Hellwig 等人（2025a），这些结果在低资源场景下针对 TASD 和 ASQP 均展现了最先进水平。遵-

循其实验设置，我们采用 Gemma-3-27B 进行了零样本、10样本和50样本提示配置。

3.2.3. 低资源增强方法。

算法 2： EDA

输入： 数据集 $\mathcal{D}$ 包含 k 个示例（s，t），其中 s 是一个句子，并且 t 列表为相应的元组；数量为每个示例的数据增强 α

输出： 增强数据集 $\mathcal{D}_{\text{aug}}$ 与 α 每（s，t）个示例

对于每个$(s,\ t) \in \mathcal{D}$ **执行**
  $tokens \leftarrow$分词器(s)
  从 $i \leftarrow 1$**到** α **执行**
    $tokens_{\text{aug}} \leftarrow tokens.$复制()
    $t_{\text{aug}} \leftarrow t.$复制()
    $tokens_{\text{aug}} \leftarrow$插入$(tokens_{aug})$
    $tokens_{\text{aug}} \leftarrow$删除$(tokens_{aug})$
    $tokens_{\text{aug}} \leftarrow$交换$(tokens_{aug})$
    $tokens_{\text{aug}} \leftarrow$同义词$(tokens_{aug})$
    $s_{\text{aug}} \leftarrow$连接$(tokens_{aug})$
    $\mathcal{D}_{\text{aug}} \leftarrow \mathcal{D}_{\text{aug}} \cup \{(s_{\text{aug}},\ t_{\text{aug}})\}$
  end
结束
返回 $\mathcal{D}_{\text{aug}}$

简易数据增强（EDA）得到了评估，因为考虑到较少的情感元素，它被广泛应用于各种自然语言处理任务和方面级情感分析任务（Hsu 等人，2021；Liesting 等人，2021；Rahamim 等人，2023年）。对于给定的一组标注样例（10 个或 50 个），我们生成了 10 个增强样例。对于每个样例进行 10 次数据增强，并给定 10 个样例时，我们生成了 100 个增强样例；与给定的 10 个真实样例合并后，总计得到 110 个样例。

增强流程如算法 2所示。首先，插入操作会从句子中随机选择一个词，并将其同义词插入随机位置，但不会插入方面术语或观点术语内部。删除操作会移除一个不属于任何术语的随机词元；交换操作会交换两个不位于任何术语内部的词元。最后，替换操作会将句子中的一个随机词元替换为其同义词。训练时采用了 DLO。

数量增强与信息增强（QAIE）由Lu 等人（2025年）采用了以下修改：(1) 增强基于给定的 10 个或 50 个示例进行，与其他实验条件保持一致；(2) 不再使用谷歌已弃用的 PaLM API[4]，，而是使用 Gemma-3-27B，这进一步提升了所评估方法之间的可比性

[4]谷歌 生成式人工智能：
https://pypi.org/project/google-generativeai/

。若示例不包含明确的方面术语，则无法进行基于术语的增强，因此与 ASQP 任务相比，增强数量更少；在 ASQP 任务中，所有示例始终都给出了观点词。值得注意的是，Lu 等人（2025年）仅增强了 ASQP 示例。

DS²-ABSA经过了若干修改。为与 LA-ABSA 及前述基线方法保持一致，我们在数据合成中使用 Gemma-3-27B，而非 GPT-3.5-Turbo。此外，我们未采用原始 DS²-ABSA 框架中的迭代噪声自训练机制；在该机制中，教师模型会迭代地优化合成标签并训练后续学生模型，直至验证集性能趋于稳定。我们的场景假设存在严格的数据约束：除 0、10 或 50 个训练样例外，不假设有任何额外的标注数据。

3.3. 评估与数据集

数据集	训练		Test	
	TASD	ASQP	TASD	ASQP
Rest15	1,120	834	582	537
Rest16	1,708	1,264	587	544
FlightABSA	1,351	1,351	387	387
Coursera	1,400	1,400	400	400
酒店	1,400	1,400	400	400

表 1：**数据集统计概览**。 五个特定领域数据集中 TASD 和 ASQP 任务的标注样例分布，展示了训练集和测试集的规模。

我们在五个不同的数据集上评估了元组预测任务 ASQP 和 TASD。据我们所知，这些数据集是唯一同时针对这两项任务进行标注的公开资源。对于 TASD 任务，我们使用了 SemEval 2015 （Pontiki 等人，2015）和 2016（Pontiki 等人，2016）数据集。对于 ASQP，我们采用了 Zhang 等人(2021) 扩展后的 SemEval 数据集版本，其中还包含观点词标注。

此外，我们纳入了Hellwig 等人改编的OATS 数据集（Chebolu 等人，2024年），该数据集包含酒店和电子学习（Coursera）领域的数据。这些改编包括将 ASQP 四元组转换为 TASD 三元组（通过移除观点术语），以及从生成的数据集中剔除重复三元组。最后，我们还纳入了 FlightABSA，这是一个包含航空公司评论的 ABSA 数据集，由Hellwig 等人（2025a）提出。附录A展示了所考虑数据集和任务中方面类别与情感极性的分布。

方法	# 训练		Rest15		Rest16		FlightABSA		Coursera		酒店		AVG	
	TASD	ASQP	TASD	ASQP	TASD	ASQP	TASD	ASQP	TASD	ASQP	TASD	ASQP	TASD	ASQP
场景：给定 0 个标注样例														
Gemma-3-27B（零样本）0		0	**30.36**	**24.73**	45.51	**28.96**	51.81	42.37	29.50	13.36	38.97	23.02	39.23	**26.49**
LA-ABSA														
含释义	full	full	27.70	18.54	**47.73**	26.07	**52.10**	42.58	**30.92**	**13.41**	**40.74**	**26.27**	**39.84**	25.37
使用 *DLO*	full	full	25.82	18.88	47.41	26.97	51.75	**42.82**	30.60	13.38	40.70	24.97	39.26	25.40
场景：给出 10 个标注样例														
Gemma-3-27B 上下文学习	10	10	**54.47**	**39.95**	**66.75**	**46.23**	60.36	45.24	**41.69**	22.31	**56.51**	31.41	**55.96**	**37.03**
释义	10	10	8.75	1.32	6.66	3.56	8.82	3.44	15.94	4.75	14.91	2.63	11.02	3.14
DLO	10	10	15.84	4.37	13.59	5.18	16.07	4.87	22.93	4.47	18.07	3.53	17.30	4.48
LA-ABSA														
使用释义	full	full	49.09	35.04	62.74	44.59	59.87	45.89	38.46	22.47	55.69	**32.83**	53.17	36.16
使用 *DLO*	full	full	49.23	37.19	62.37	46.20	**61.40**	**46.47**	39.22	**23.37**	55.27	31.44	53.50	36.93
EDA	110	110	30.18	9.34	19.16	11.80	22.66	11.95	28.83	14.19	25.65	7.88	25.30	11.03
QAIE	26.4	45.2	20.11	9.96	12.37	14.78	17.76	15.38	25.18	17.54	21.69	9.86	19.42	13.51
DS^2-ABSA	21,1k	21,1k	29.04	18.19	32.09	23.52	27.01	16.49	17.81	7.61	31.87	11.78	27.56	15.52
场景：给定 50 个标注样例														
Gemma-3-27B 上下文学习	50	50	**62.12**	**41.74**	**68.53**	**51.10**	**64.60**	48.37	**44.80**	**25.86**	**62.97**	43.83	**60.60**	**42.18**
释义	50	50	36.92	25.55	35.87	23.50	33.57	17.98	34.26	19.38	40.10	23.09	36.14	21.90
DLO	50	50	39.54	26.63	43.95	29.57	42.92	28.74	36.04	19.08	44.72	27.20	41.44	26.24
LA-ABSA														
使用释义	full	full	56.21	37.61	62.20	46.76	62.47	47.16	44.36	25.71	60.58	43.19	57.16	40.09
使用 *DLO*	full	full	58.40	40.38	62.03	49.85	62.57	**48.98**	44.39	25.69	61.43	**44.24**	57.76	41.83
EDA	550	550	44.14	30.93	45.72	34.59	45.64	36.73	37.89	24.57	45.72	32.51	43.82	31.86
QAIE	141.8	248.8	45.01	33.87	45.09	35.21	48.44	33.98	36.23	22.45	50.51	35.95	45.06	32.29
DS^2-ABSA	21,7k	21,6k	37.94	28.93	43.53	35.79	33.39	23.87	29.56	16.41	40.26	23.61	36.94	25.72
完整的人类标注示例集：最先进的方法														
释义	full	full	**63.06**	46.93	**71.97**	57.93	**69.74**	57.76	51.86	32.34	67.70	53.87	64.87	49.77
DLO	full	full	62.95	**48.18**	71.79	**59.79**	68.95	**58.33**	**52.58**	**32.54**	**68.56**	**55.45**	**64.97**	**50.86**

表 2：**LA-ABSA 的 F1 分数**。评估结果与基于 EDA 的数据增强方法、QAIE（FT）、DS^2-ABSA（FT）以及提示基线（0、10 和 50 个标注样例）进行比较，相关结果见 Hellwig 等人(2025a)；此外，还与全监督模型进行比较，包括 DLO（Hu 等人，2022b）（FT）和释义（Zhang 等人，2021）（FT）。每种标注设置（0、10、50 个或全部样例）中的最高 F1 分数以粗体显示；所有设置中的总体最佳分数以下划线标出。

与 ABSA 研究中的常见做法一致，报告的评估指标是微平均 F1 分数（Zhang 等人，2022）。

4. 结果

4.1. 总体结果

表 2 展示了 LA-ABSA 的性能分数。详细的精确率、召回率和宏平均 F1 分数见附录 D。

LA-ABSA 相对于大语言模型提示取得了具有竞争力或更优的性能。 表2 显示，LA-ABSA 取得的 F1 分数仅比大语言模型提示低约 4 个百分点（例如，Rest15-TASD，50样本：58.40 对 62.12）。然而，出现了若干显著例外：在零样本场景中，采用 DLO 微调的 LA-ABSA 在大多数数据集和任务上均优于零样本大语言模型提示。此外，在 10样本和 50样本设置中，偶尔也能观察到更优的性能，针对 ASQP 任务。

LA-ABSA 展现出高于数据增强方法的性能分数。那些扩充已标注样本的增强方法，包括基于 EDA 的方法、QAIE 和 DS^2-ABSA，在零样本、10样本和 50样本学习场景中的表现均显著低于 LA-ABSA。性能差距在若干情况下达到约 10 个百分点，在 10样本学习配置中差异超过 20 个百分点。

为了探究不同数据增强比例（每个已标注样例进行 2、5、10 和 15 次数据增强）以及微调方法（DLO、释义）的影响，我们对 EDA 进行了消融研究，详情见附录 B。结果表明，采用 DLO 并进行 10 次数据增强时，性能最高。

使用专家标注的训练数据，效果优于使用大语言模型生成的标注。使用专家标注微调最先进的方法来完成元组预测任务，仍然

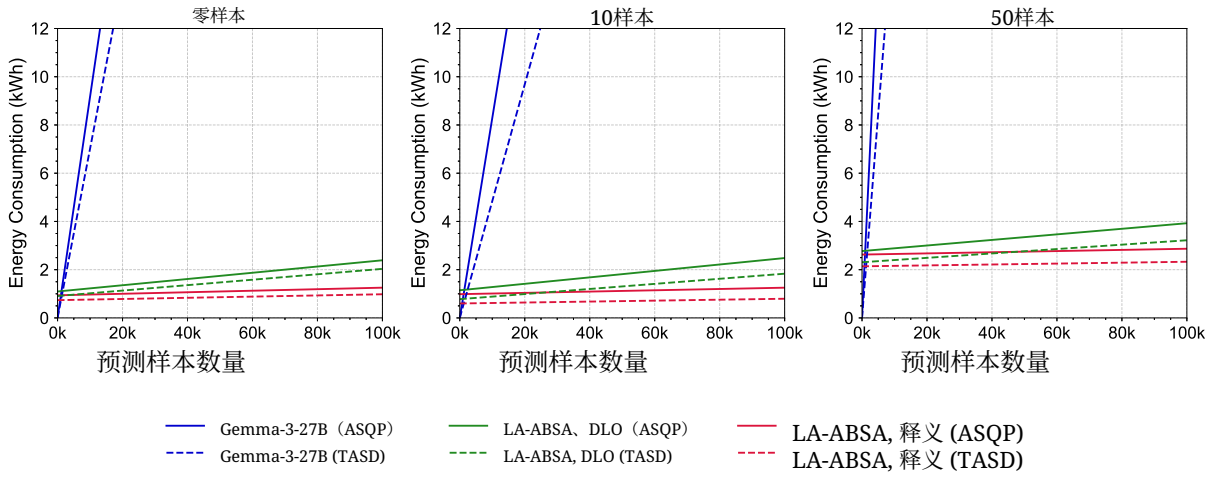


图 2： **能耗分析**。比较不同设置下（提供 0、10 或 50 个标注样例），ASQP 和 TASD 任务的能源消耗（千瓦时，kWh）。结果展示了大语言模型提示（Gemma-3-27B）和 LA-ABSA 方法在分别使用 DLO（Hu 等人，2022b）或释义（Zhang 等人，2021）进行微调时的情况。每条线表示在五个数据集上，各方法和任务预测最多 100,000 个样例的平均能源消耗。由于底层基础模型 T5-base 较小，LA-ABSA 方法通常所需能源要少得多。

优于使用大语言模型生成的标注。此外，在我们的实验设置中，使用专家标注进行微调仍然胜过所有其他评估方法。

值得注意的是，我们采用 Bonferroni-Holm 校正（Holm，1979）进行了双侧配对 t 检验，以考察 LA-ABSA（基于释义或 DLO）、在完整人工标注训练集（DLO 和释义）上训练的微调方法，以及低资源场景中最强基线方法——使用 Gemma-3-27B 的上下文学习——之间是否存在显著差异（$p_{\text{adj}} \leq 0.05$）。考虑这五个比较组，我们针对每项任务（TASD 和 ASQP）和示例数设置（0、10 和 50 个示例）进行了 10 项成对检验，共计 56 项比较。每组包含五个数值（每个数据集一个）。未发现显著差异，唯一例外是，在 TASD 任务上，使用完整数据集微调的释义和 DLO 方法，其性能分数显著高于零样本提示。

4.2. 能耗分析

LA-ABSA 的一项关键优势在于，与大语言模型提示方法相比，其能源效率更高。为量化这一优势，我们开展了能耗分析，对比了表现最佳的两种低资源方法 LA-ABSA 和大语言模型提示在最多 100,000 个生成示例上的表现。

我们的分析采用所有五个数据集上每次预测的平均能源消耗作为基线指标。需要注意的是，LA-ABSA 的能源消耗情况包含两个前置步骤所产生的初始开销：（1）使用大语言模型标注训练样例；（2）微调轻量级方法（DLO 或释义）。因此，LA-ABSA 的能耗曲线并非从零开始，这反映了这些前期投入。

如图 2所示，在处理少于 2,000 个示例时，无论任务类型或给定示例数量（0、10 或 50）如何，LA-ABSA 的能源消耗都显著低于大语言模型提示。这一发现表明，尽管模型准备需要初始能源投入，LA-ABSA 在规模化应用时仍能实现显著更高的能源效率，使其成为大规模 ABSA 部署中更具可持续性的解决方案。

表 3提供了 LA-ABSA 的详细能源消耗指标，以每个预测示例所需的毫瓦时（mWh）表示。结果凸显了释义方法的高效率：与需要评估多个情感元素排列的 DLO 方法相比，该方法在每个测试集示例中仅评估一种情感元素排序。

需要指出的是，所有能源测量均使用 NVIDIA A5000 GPU（24GB 显存）进行，所报告的指标本质上取决于硬件。推广至其他 GPU 架构时，需要进行适当的缩放调整。此外，我们还对预测时间进行了补充分析，其趋势与能源消耗的研究发现一致（见附录 C）。

方法	零样本		10样本		50样本	
	TASD	ASQP	TASD	ASQP	TASD	ASQP
LA-ABSA（释义）	2.47 ± 0.53	3.17 ± 0.31	1.96 ± 0.40	2.66 ± 0.28	1.83 ± 0.38	2.41 ± 0.46
LA-ABSA（DLO）	11.16 ± 0.72	13.06 ± 0.89	10.68 ± 2.11	13.19 ± 1.62	8.95 ± 0.85	11.65 ± 2.23
Gemma-3-27B 上下文学习	703.10 ± 63.97	912.25 ± 72.67	481.39 ± 46.11	843.68 ± 356.29	1688.76 ± 255.30	2963.65 ± 1360.88

表3：**平均能耗分析**。在不同少样本设置下，LA-ABSA 和大语言模型提示（Gemma-3-27B）每个样本的能源消耗（毫瓦时，mWh）。结果显示，TASD 和 ASQP 任务在五个数据集上的均值 ± 标准差。我们使用经 Bonferroni-Holm 校正（$p_{\text{adj}} \leq 0.05$）的配对t检验，对 DLO、释义和提示进行了18次两两比较（3种比较方式 × 2 任务 × 3 示例数设置），发现所有差异均具有显著性。

5. 讨论与局限

本节将讨论本研究的主要发现及其潜在局限性。

首先，与以往仅依赖少量标注样例并使用较小模型（如 T5）而非大语言模型的方法相比，我们提出的 LA-ABSA 方法表现更优；自然语言处理领域的其他任务中也观察到了这一现象（Li 等人，2025年；Huang 等人，2023年）。简易数据增强（EDA）（Liesting 等人，2021）、QAIE（Lu 等人，2025年）和 DS^2-ABSA（Xu 等人，2025年）等数据增强方法在本研究中首次应用于 TASD 和 ASQP 任务；此前的研究仅将这些数据增强策略应用于情感元素较少的方面级情感分析任务。尽管如此，LA-ABSA 在各种场景和数据集上的 F1 分数仍显著更高。

相比之下，在大多数情况下，基于大语言模型的上下文学习的表现略逊于 LA-ABSA，只有少数例外。图2 表明，处理超过 2,000 个示例时，LA-ABSA 的能源效率显著更高。鉴于这种差异幅度很大（例如，对于 TASD 任务，LA-ABSA +Paraphrase 每个示例需要 2.47 mWh，而 Gemma-3-27B 每个示例需要 703.1 mWh），即使将上下文学习与规模较小的大语言模型（例如，12B 参数）进行比较，或与不采用自一致性学习（多数投票机制）的单次执行方法进行比较，也很可能观察到类似的效率提升。

在此背景下，本研究的一个局限是仅评估了一个大语言模型，这主要是因为所研究条件对计算资源的要求极高。考察能够生成中间推理步骤的推理模型，或许是进一步提升标注质量的一种策略；这类模型已在各种自然语言处理任务中展现出优于非推理方法的表现（Wei et al.，2022；Fei et al.，2023年；Sun et al.，2023年）。然而，由于这类模型会生成更多输出词元，因此推理时间会显著增加。

最后，LA-ABSA 的表现未能达到在人工标注数据上微调的模型水平。在 ASQP 任务上，与使用人工标注进行微调的方法相比，LA-ABSA 的平均表现下降了 9 个百分点。

6. 结论与未来工作

本文介绍了 LA-ABSA，这是一种新颖的方法，利用大语言模型生成的标注对轻量级模型进行微调，以完成 ABSA 任务。我们在五个数据集和两种元组预测任务（TASD 和 ASQP）上进行了全面评估，结果表明，该方法的表现可与基于大语言模型的上下文学习方法相媲美，同时在能源效率方面具有显著优势。

未来工作可探索若干有前景的方向。首先，可探索众包方法，将其作为基于大语言模型的标注的替代方案或补充，以期在完全自动化与专家标注之间实现折中。其次，数据合成方法也为进一步研究提供了有趣的方向。尽管本研究假设可以获取未标注文本，但在低资源场景中，这类文本可能无法获得。在这种情况下，可以借鉴其他自然语言处理任务或方面级情感分析任务中的常见方法，利用大语言模型合成观点文本（例如评论）（Xu 等人，2025年；Ye 等人，2022；Yu 等人，2023年），也可以采用一步生成已标注训练样例的方法（Xu 等人，2025年；Hellwig 等人，2025b年）。

7. 伦理声明

本研究未获得工业界资金或商业赞助。除 FlightABSA 外，所有数据集均包含未经匿名化的标注样例。为确保研究透明度和可复现性，代码已发布在 GitHub 上。Claude Sonnet 4.5[5] 曾用于协助撰写本文。

[5]Claude Sonnet: https://www.anthropic.com/claude/sonnet

最后，我们想强调，LA-ABSA 有助于减少大规模模型推理所产生的碳足迹。通过使用规模小得多的基础模型，并减少所需的计算资源，LA-ABSA 降低了能源消耗。这种资源高效的设计有助于推动更可持续的研究实践，也契合了减少人工智能系统环境影响的广泛努力。

8. 参考文献

Josh Achiam、Steven Adler、Sandhini Agarwal、Lama Ahmad、Ilge Akkaya、Florencia Leoni Aleman、Diogo Almeida、Janko Altenschmidt、Sam Altman、Shyamal Anadkat等。2023年。GPT-4 技术报告。*arXiv* 预印本 *arXiv:2303.08774*。

Ayush Agarwal、Janak Kapuriya、Shubham Agrawal、Akhil Vamshi Konam、Mansi Goel、Rishabh Gupta、Shrey Rastogi、Niharika Niharika 和 Ganesh Bagler。2024年。基于深度学习的食谱命名实体识别模型。载于*2024*年计算语言学、语言资源与评估联合国际会议论文集（*LREC-COLING 2024*），第4542–4554页。

Yinhao Bai、Zhixin Han、Yuhua Zhao、Hang Gao、Zhuowei Zhang、Xunzhi Wang 和 Mengting Hu。2024年。大语言模型是否解决了复合方面级情感分析问题？载于计算语言学协会论文集：*EMNLP 2024*，第7836–7861页。

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

Siva Uday Sampreeth Chebolu、Franck Dernon- court、Nedim Lipka 和 Thamar Solorio。2024年。Oats：一个用于基于方面的情感分析中意见、方面、目标、情感联合检测的挑战数据集。载于*2024*年计算语言学、语言资源与评估联合国际会议论文集（*LREC-COLING 2024*），第12336–12347页。

Xiang Chen、Chaoyang Gao、Chunyang Chen、Guangbei Zhang 和 Yong Liu。关于大语言模型应用开发挑战的实证研究。*ACM*软件工程与方法学汇刊。

Aakanksha Chowdhery、Sharan Narang、Jacob Devlin、Maarten Bosma、Gaurav Mishra、Adam Roberts、Paul Barham、Hyung Won Chung、Charles Sutton、Sebastian Gehrmann 等。2023年。PaLM：利用 Pathways 扩展语言建模。机器学习研究期刊，24(240):1–113。

John Chung、Ece Kamar 和 Saleema Amershi。2023年。在保持准确性的同时提高多样性：利用大语言模型和人工干预进行文本数据生成。载于计算语言学协会第*61*届年会论文集（第一卷：长篇论文），第575–593页。

Bálint Csanády、Lajos Muzsai、Péter Vedres、Zoltán Nádasdy 和 András Lukács。2024年。Llambert：自然语言处理中大规模、低成本的数据标注。*arXiv* 预印本 *arXiv:2403.15938*。

Bosheng Ding、Chengwei Qin、Ruochen Zhao、Tianze Luo、Xinze Li、Guizhen Chen、Wenhan Xia、Junjie Hu、Luu Anh Tuan 和 Shafiq Joty。2024年。使用大语言模型进行数据增强：数据视角、学习范式与挑战。载于计算语言学协会 *ACL 2024* 研究成果，第 1679–1705 页。

Hao Fei、Bobo Li、Qian Liu、Lidong Bing、Fei Li 和 Tat-Seng Chua。2023年。通过思维链提示推理隐式情感。载于计算语言学协会第*61*届年会论文集（第*2*卷：短篇论文），第 1171–1182 页。

Zhibin Gou、Qingyan Guo 和 Yujiu Yang。2023年。MvP：多视角提示可提升方面情感元组预测。载于计算语言学协会第 *61* 届年会论文集（第 *1* 卷：长篇论文），第 4380–4397 页。

Aaron Grattafiori、Abhimanyu Dubey、Abhinav Jauhri、Abhinav Pandey、AbhishekKadian、Ahmad Al-Dahle、Aiesha Letman、Akhil Mathur、Alan Schelten、Alex Vaughan 等。2024年。Llama 3 模型群。*arXiv* 预印本 *arXiv:2407.21783*。

Nils Constantin Hellwig、Jakob Fehle、Udo Kr- uschwitz 和 Christian Wolff。2025a。我们还需要人工标注员吗？提示大语言模型进行方面情感四元组预测。载于第一届大语言模型与结构建模联合研讨会论文集（*XLLM 2025*），第 153–172 页，奥地利维也纳。计算语言学协会。

Nils Constantin Hellwig、Jakob Fehle 和 Christian Wolff。2025b年。探索大语言模型

在低资源环境下生成用于基于方面的情感分析的合成训练样本。*Expert Systems withApplica-tions*，261:125514。

Sture Holm. 1979. 一种简单的序贯拒绝多重检验程序。《斯堪的纳维亚统计学杂志》, 6(2):65–70.

Ting-Wei Hsu, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2021. 面向基于方面的情感分析的语义保持数据增强。载于*2021*年自然语言处理经验方法会议论文集，第 4417–4422 页。

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, 等。2022a. LoRA：大语言模型的低秩适配。*ICLR*, 1(2):3.

Mengting Hu, Yike Wu, Hang Gao, Yinhao Bai, and Shiwan Zhao. 2022b. 通过模板顺序数据增强改进方面情感四元组预测。载于*2022*年自然语言处理经验方法会议论文集，第 7889–7900 页。

Jiaxin Huang、Shixiang Gu、Le Hou、Yuexin Wu、Xuezhi Wang、Hongkun Yu 和 Jiawei Han。2023年。大语言模型可以自我改进。载于*2023*年自然语言处理实证方法会议论文集，第1051–1068页。

Mia Mohammad Imran、Yashasvi Jain、Preetha Chatterjee 和 Kostadin Damevski。2022。用于提升软件工程交流中情绪识别的数据增强。载于第*37*届*IEEE/ACM*自动化软件工程国际会议论文集，第1–13页。

Guangmin Li、Hui Wang、Yi Ding、Kangan Zhou 和 Xiaowei Yan。2023年。基于方面的情感分析的数据增强。《国际机器学习与控制论杂志》，14(1)：125–133。

Juanhui Li、Sreyashi Nag、Hui Liu、Xianfeng Tang、Sheikh Muhammad Sarwar、Limeng Cui、Hansu Gu、Suhang Wang、Qi He 和 Jiliang Tang。2025年。以少胜多：通过未标注数据从大语言模型进行知识蒸馏。载于计算语言学协会年会论文集：*NAACL 2025*，第2627–2641页。

Lin Li、Shaopeng Tang 和 Renwei Wu。2024年。多数规则引导的基于方面类别的情感分析：利用标签先验知识。载于 *2024*年国际计算语言学联合会议论文集，语言资源与评估（*LREC-COLING 2024*），第10952–10957 页，意大利都灵。ELRA 和 ICCL。

Yijie Li和Yuan Sun。2025年。Easyjudge：一种易于使用的大语言模型综合响应评估工具。载于第*31*届国际计算语言学会议论文集：系统演示，第91–103页。

Tomas Liesting、Flavius Frasincar和Maria Mihaela Truşcˇa。2021。面向基于方面的情感分析的混合方法中的数据增强。载于第*36*届年度*ACM*应用计算研讨会论文集，第828–835页。

Bing Liu。2022。 情感分析与观点挖掘。SpringerNature。

Yiding Liu、Jingjing Wang、Jiamin Luo、Tao Zeng和Guodong Zhou。2024年。Chatasu：唤起大语言模型的反思，以真正理解对话中的方面情感。载于*2024*年计算语言学、语言资源与评估联合国际会议论文集（*LREC-COLING 2024*），第3075–3085页。

Heng-yang Lu、Tian-ci Liu、Rui Cong、Jun Yang、Qiang Gan、Wei Fang 和 Xiao-jun Wu。2025年。QAIE：基于大语言模型的数量增强与信息增强，用于少样本方面情感分析。信息处理与管理，62(1):103917。

Usman Malik、Simon Bernard、Alexandre Pauchet、Clément Chatelain、Romain Picot-Clémente 和 Jérôme Cortinovis。2024年。利用大语言模型对法语推文进行多标签情绪分类的伪标签。《*IEEE Access*》，12:15902–15916。

Arindam Mitra、Luciano Del Corro、Shweti Mahajan、Andres Codas、Clarisse Simoes、Sahaj Agarwal、Xuxi Chen、Anastasia Razdaibiedina、Erik Jones、Kriti Aggarwal 等。2023年。Orca 2：教小语言模型如何进行推理。*arXiv* 预印本 *arXiv:2311.11045*。

Arbi Haza Nasution 和 Aytug Onan。2024年。Chat-gpt label：比较低资源语言自然语言处理任务中人工生成的标注与大语言模型生成的标注的质量。《IEEE Access》。

Gaurav Negi、Rajdeep Sarkar、Omnia Zayed 和 Paul Buitelaar。2024年。使用迁移学习进行基于方面的情感分析的混合方法。载于 *2024*年计算语言学、语言资源与评估联合国际会议论文集（*LREC-COLING 2024*），第647–658 页，意大利都灵。ELRA 和 ICCL。

Kazumasa Omura、Fei Cheng 和 Sadao Kurohashi。2024年。用于隐式篇章关系识别的合成数据生成实证研究。载于*2024*年计算语言学、语言资源与评估联合国际会议论文集（*LREC-COLING 2024*），第1073–1085页。

Maria Pontiki、Dimitrios Galanis、Haris Papageor- giou、Ion Androutsopoulos、Suresh Manand- har、MohammadAl-Smadi、MahmoudAl-Ayyoub、Yanyan Zhao、Bing Qin、Orphée De Clercq 等。2016。SemEval-2016 任务 5：基于方面的情感分析。载于语义评估国际研讨会，第 19–30 页。

Maria Pontiki、Dimitrios Galanis、Harris Papageor- giou、Suresh Manandhar 和 Ion Androutsopou- los。2015。SemEval-2015 任务 12：基于方面的情感分析。载于第九届语义评估国际研讨会论文集（*SemEval 2015*），第 486–495 页。

Colin Raffel、Noam Shazeer、Adam Roberts、Katherine Lee、Sharan Narang、Michael Matena、Yanqi Zhou、Wei Li 和 Peter J Liu。2020。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。

Adir Rahamim、Guy Uziel、Esther Goldbraich 和 Ateret Anaby Tavor。2023年。利用数据集重构进行低资源分类的文本增强。载于计算语言学协会论文集：*ACL 2023*，第 7389–7402 页。

Caroline Sabty、Islam Omar、Fady Wasfalla、Mohamed Islam 和 Slim Abdennadher。2021。用于命名实体识别的阿拉伯语数据增强技术。《计算机科学程序汇刊》，189：292–299。

Maximilian Schmidt、Andrea Bartezzaghi 和 Ngoc Thang Vu。2024年。基于提示的少样本问答合成数据生成。载于*2024*年计算语言学、语言资源与评估联合国际会议论文集（*LREC-COLING 2024*），第 13168–13178 页。

Paul F Simmering 和 Paavo Huoviala。2023年。用于基于方面的情感分析的大语言模型。*arXiv* 预印本 *arXiv:2310.18025*。

Jakub Šmíd、Pavel Pˇribáˇn 和 Pavel Král。2024年。基于 Llama 的方面级情感分析模型。载于第十四届主观性、情感与社交媒体分析计算方法研讨会论文集，第 63–70 页。

Xiaofei Sun、Xiaoya Li、Jiwei Li、Fei Wu、Shangwei Guo、Tianwei Zhang 和 Guoyin Wang。2023年。通过大语言模型进行文本分类。载于计算语言学协会论文集：*EMNLP 2023*，第 8990–9005 页。

Gemma 团队, Aishwarya Kamath, Johan Fer- ret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé,Morgane Rivière, 等. 2025年. Gemma 3 技术报告. *arXiv* 预印本 *arXiv:2503.19786*.

Siddharth Varia, Shuai Wang, Kishaloy Halder, Robert Vacareanu, Miguel Ballesteros, Yassine Benajiba, Neha Anna John, Rishita Anubhai, Smaranda Muresan, 和 Dan Roth. 2023年. 面向少样本方面级情感分析的指令微调. 载于计算语言学协会第*61*届年会.

An Wang, Junfeng Jiang, Youmi Ma, Ao Liu, 和 Naoaki Okazaki. 2023 年 . 用于方面情感四元组预测的生成式数据增强. 载于第*12*届词汇与计算语义联合会议（** SEM 2023*）论文集，第128–140 页。

An Wang、Junfeng Jiang、Youmi Ma、Ao Liu 和 Naoaki Okazaki。2024年。用于方面情感四元组预测的生成式数据增强。自然语言处理期刊，31(4):1523–1544。

Shuhe Wang、Xiaofei Sun、Xiaoya Li、Rongbin Ouyang、Fei Wu、Tianwei Zhang、Jiwei Li、Guoyin Wang 和 Chen Guo。2025年。GPT-NER：通过大语言模型进行命名实体识别。载于计算语言学协会年会论文集：*NAACL 2025*，第4257–4275页。

Jason Wei、Xuezhi Wang、Dale Schuurmans、Maarten Bosma、Fei Xia、Ed Chi、Quoc V Le、Denny Zhou 等。2022。思维链提示能够激发大语言模型的推理能力。神经信息处理系统进展，35:24824–24837。

Jason Wei 和 Kai Zou。2019 年。Eda：用于提升文本分类任务性能的 EDA（简易数据增强）技术。在*2019* 年自然语言处理经验方法会议与第*9* 届自然语言处理国际联合会议（*EMNLP-IJCNLP*）论文集，第 6382–6388 页。

Hongling Xu、Yice Zhang、Qianlong Wang 和 Ruifeng Xu。2025 年。DS^2-ABSA：用于少样本

基于方面的情感分析。载于 第*63*届计算语言学协会年会论文集（第*1*卷：长篇论文），第15460–15478页，奥地利维也纳。计算语言学协会。

Zhaorui Yang、Tianyu Pang、Haozhe Feng、Han Wang、Wei Chen、Minfeng Zhu 和 Qian Liu。2024年。自蒸馏弥合语言模型微调中的分布差距。载于计算语言学协会第*62*届年会论文集（第*1*卷：长篇论文），第1028–1043页。

Jiacheng Ye、Jiahui Gao、Qintong Li、Hang Xu、Jiangtao Feng、Zhiyong Wu、Tao Yu 和 Ling- peng Kong。2022年。Zerogen：通过数据集生成实现高效的零样本学习。载于*2022*年自然语言处理实证方法会议论文集，第11653– 11669页。

Yue Yu、Yuchen Zhuang、Jieyu Zhang、Yu Meng、Alexander J Ratner、Ranjay Krishna、Jiaming Shen 和 Chao Zhang。2023年。作为属性化训练数据生成器的大语言模型：多样性与偏差的故事。神经信息处理系统进展，36:55734– 55784。

Ruoyu Zhang、Yanzeng Li、Yongliang Ma、Ming Zhou 和 Lei Zou。2023年。LLMAAA：让大语言模型成为主动标注器。载于计算语言学协会论文集：*EMNLP 2023*，第 13088–13103 页，新加坡。计算语言学协会。

Wenxuan Zhang、Yang Deng、Xin Li、Yifei Yuan、Lidong Bing 和 Wai Lam。2021。将方面情感四元组预测作为释义生成。载于 *2021*年自然语言处理实证方法会议论文集，第 9209–9219 页。

Wenxuan Zhang、Yue Deng、Bing Liu、Sinno Pan 和 Lidong Bing。2024a。大语言模型时代的情感分析：现实检验。载于计算语言学协会论文集：*NAACL 2024*，第 3881–3906 页。

Wenxuan Zhang、Yue Deng、Bing Liu、Sinno Pan 和 Lidong Bing。2024b。大语言模型时代的情感分析：现实检验。载于计算语言学协会论文集：*NAACL 2024*，第 3881–3906 页，墨西哥城，墨西哥。计算语言学协会。

Wenxuan Zhang、Xin Li、Yang Deng、Lidong Bing 和 Wai Lam。2022。一项关于基于方面的情感分析：任务、方法与挑战。*IEEE* 知识与数据工程汇刊，35(11):11019–11038。

Changzhi Zhou、Dandan Song、Yuhang Tian、Zhi-jing Wu、Hao Wang、Xinyu Zhang、Jun Yang、Ziyi Yang 和 Shuhao Zhang。2024年。大语言模型在基于方面的情感分析上的全面评估。*arXiv* 预印本 *arXiv:2412.02279*。

A. 数据集：方面类别和情感分布

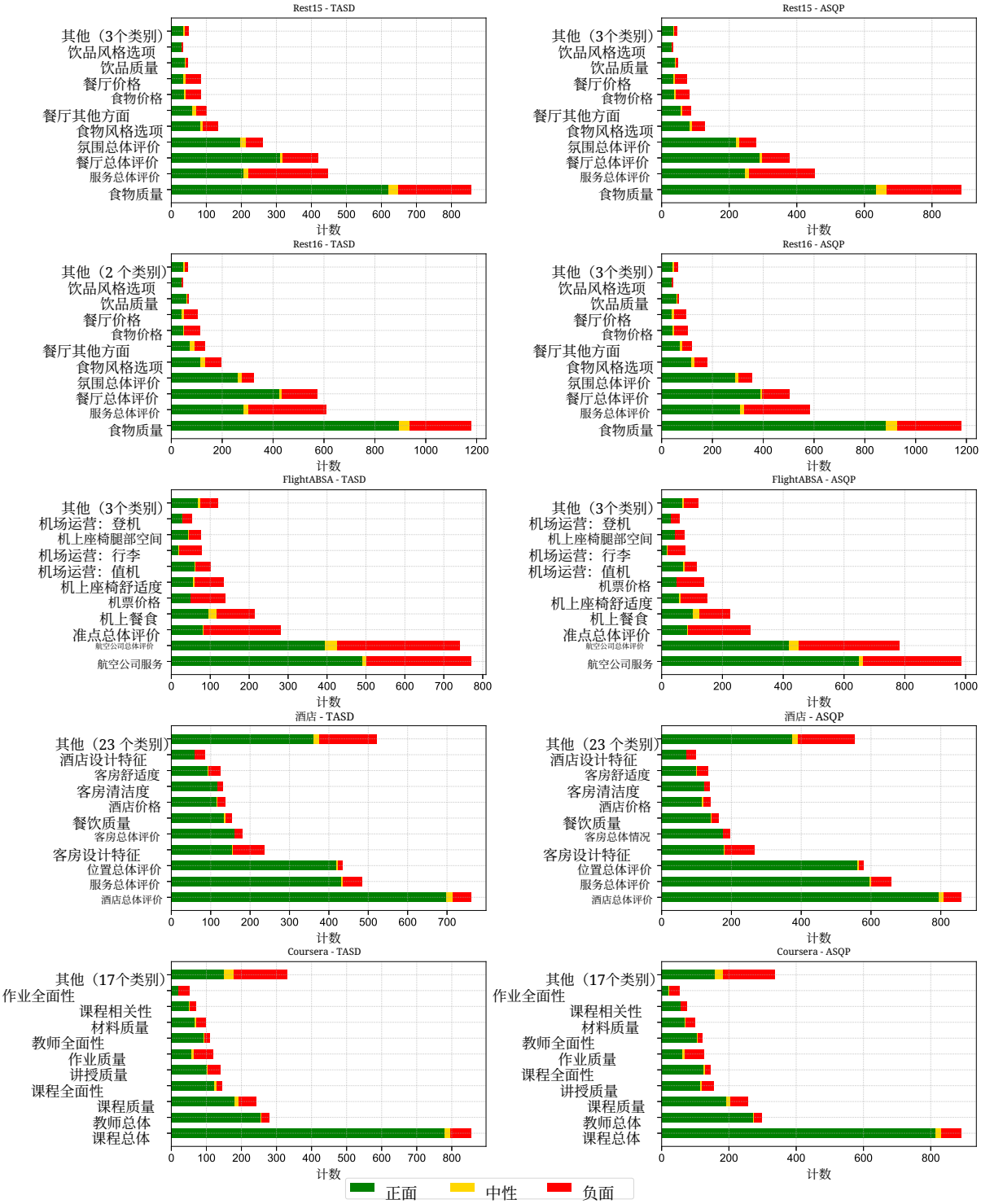


图 3： **TASD 和 ASQP 任务中各数据集的方面类别与情感分布**。 每个子图展示出现频率最高的 10 个方面类别（按总频率排序），堆叠条形图表示正面（绿色）、中性（黄色）和负面（红色）情感。“其他”类别汇总其余方面。结果汇总自五个数据集：Rest15（Zhang 等人，2021；Pontiki 等人，2015）、Rest16（Zhang 等人，2021；Pontiki 等人，2016）、FlightABSA（Hellwig 等人，2025a）、Coursera（Chebolu 等人，2024年）、以及 Hotels（Chebolu 等人，2024年）。该可视化结果突出了方面级情感标注中的不平衡现象，展示了不同数据集之间情感极性和方面类别的不同分布。

B. 消融研究：EDA

方法	# 增强	# 训练		Rest15		Rest16		FlightABSA		Coursera		酒店		AVG	
				TASD	ASQP	TASD	ASQP	TASD	ASQP	TASD	ASQP	TASD	ASQP	TASD	ASQP
场景：给定 10 个标注样例															
EDA（含释义）	2	30	30	13.15	1.74	7.67	11.41	10.79	4.42	17.97	6.61	22.05	3.47	14.33	5.53
EDA（含释义）	5	60	60	23.20	10.42	16.35	12.30	19.26	10.99	23.62	9.47	25.85	6.15	21.66	9.86
EDA（含释义）	10	110	110	26.07	**10.52**	16.82	**12.68**	21.83	**13.46**	25.67	10.35	24.95	5.84	23.07	10.57
EDA（含释义）	15	160	160	25.53	10.08	18.34	12.32	21.57	12.39	25.60	12.33	24.94	7.03	23.20	10.83
EDA（含DLO）	2	30	30	24.58	6.74	16.97	9.27	21.18	12.13	28.14	9.75	22.73	5.98	22.72	8.77
EDA（含DLO）	5	60	60	26.65	8.68	18.06	10.70	21.66	12.40	28.25	11.73	23.95	7.51	23.71	10.20
使用 DLO 的 EDA	10	110	110	**30.18**	9.34	19.16	11.80	**22.66**	11.95	**28.83**	**14.19**	25.65	**7.88**	**25.30**	**11.03**
使用 DLO 的 EDA	15	160	160	29.10	8.07	**19.59**	11.83	22.05	12.13	27.40	13.36	**27.04**	6.94	25.04	10.46
场景：给定 50 个标注样例															
使用释义的 EDA	2	150	150	42.70	27.55	41.59	28.78	40.73	28.26	36.47	19.90	43.44	27.96	40.98	26.49
使用释义的 EDA	5	300	300	42.29	28.96	42.57	29.40	42.80	30.32	36.85	21.08	44.67	28.04	41.84	27.56
使用释义的 EDA	10	550	550	41.83	28.70	43.01	30.20	41.62	31.04	35.84	21.37	42.70	27.97	41.00	27.85
EDA 与释义	15	800	800	40.78	27.81	40.81	28.80	42.02	31.92	36.49	21.22	43.45	27.38	40.71	27.43
EDA 与 DLO	2	150	150	43.86	30.41	45.77	33.52	44.11	34.84	37.81	23.53	**45.89**	31.39	43.49	30.74
EDA 与 DLO	5	300	300	43.95	30.90	**46.28**	34.42	44.90	36.37	**38.32**	23.53	45.20	**32.55**	43.73	31.55
EDA 与 DLO	10	550	550	**44.14**	**30.93**	45.72	**34.59**	**45.64**	**36.73**	37.89	**24.57**	45.72	32.51	**43.82**	**31.86**
EDA 与 DLO	15	800	800	41.71	28.82	42.10	31.68	45.57	34.09	36.55	23.33	44.61	31.13	42.11	29.81

表 4：**基于 EDA 的数据增强消融研究结果**。报告了不同增强比例（2、5、10、15）和微调方法（DLO、释义）在包含 10 个和 50 个标注样例的场景中的 F1 分数。结果表明，在大多数数据集中，10 次数据增强与 DLO 相结合可取得最佳性能。粗体数值表示在给定 10 个或 50 个标注样例的场景中最高的 F1 分数。

C. LA-ABSA：时间效率

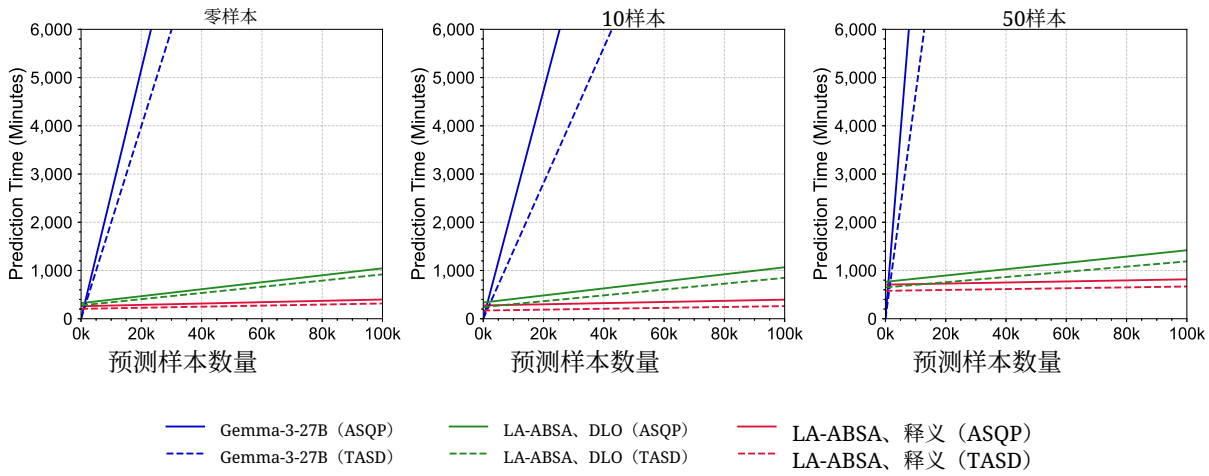


图 4：**预测时间分析**。不同设置（提供 0、10 或 50 个已标注样例）下，ASQP 和 TASD 任务的预测时间（分钟）。结果展示了大语言模型提示（Gemma-3-27B）以及使用 DLO（Hu 等人，2022b）和释义（Zhang 等人，2021）进行微调的 LA-ABSA 的结果。每条线表示在五个数据集上，每种方法和任务预测最多 100,000 个样例所需的平均时间。由于底层基础模型 T5-base 更小，LA-ABSA 方法通常所需时间显著更少。

D. 性能：精确率、召回率和宏平均F1

D.1. 精确率

方法	# 训练		Rest15		Rest16		FlightABSA		Coursera		酒店		AVG	
	TASD	ASQP	TASD	ASQP	TASD	ASQP	TASD	ASQP	TASD	ASQP	TASD	ASQP	TASD	ASQP
场景：给定 0 个标注样例														
Gemma-3-27B（零样本）	0	0	**29.41**	**23.35**	44.49	**27.75**	47.55	**39.70**	26.95	**11.95**	37.12	22.88	37.11	**25.13**
LA-ABSA														
使用释义	full	full	26.91	17.09	**45.93**	24.00	**48.95**	39.41	**28.32**	11.68	**38.53**	**25.41**	**37.73**	23.52
使用 *DLO*	full	full	24.44	17.21	45.68	24.62	47.41	39.06	27.52	11.48	37.98	23.83	36.61	23.24
场景：给定 10 个标注样例														
Gemma-3-27B 上下文学习	10	10	**56.40**	**39.41**	**68.38**	**44.64**	59.85	45.39	**43.11**	**23.41**	**57.93**	**35.29**	**57.13**	**37.63**
释义	10	10	10.72	1.64	7.61	4.02	10.44	4.34	17.69	5.35	18.74	3.66	13.04	3.80
DLO	10	10	19.23	4.64	13.27	5.49	19.02	6.15	25.45	5.03	18.84	3.68	19.16	5.00
LA-ABSA														
使用释义	full	full	51.41	33.75	63.75	43.35	61.12	**45.85**	38.28	21.98	57.70	34.80	54.45	35.95
使用 *DLO*	full	full	50.33	35.54	63.51	43.85	**61.29**	45.63	38.82	22.79	56.07	33.10	54.00	36.18
EDA	110	110	32.49	9.16	19.32	11.92	26.82	14.10	31.08	15.52	26.15	8.61	27.17	11.86
QAIE	26.4	45.2	26.44	12.71	15.71	17.42	22.37	20.19	28.49	19.70	29.39	12.40	24.48	16.49
DS^2-ABSA	21,1k	21,1k	29.55	18.10	33.30	22.72	27.34	17.33	17.40	7.43	32.17	12.03	27.95	15.52
场景：给定 50 个标注样例														
Gemma-3-27B 上下文学习	50	50	**68.03**	**44.57**	**71.52**	**54.55**	**66.33**	**51.95**	**51.32**	**31.96**	**70.47**	**53.39**	**65.54**	**47.28**
释义	50	50	39.57	24.58	37.18	23.75	36.10	18.58	37.64	20.72	45.21	23.67	39.14	22.26
DLO	50	50	40.48	24.92	44.59	29.09	42.01	28.30	39.26	20.44	49.18	28.54	43.10	26.26
LA-ABSA														
使用释义	full	full	60.49	38.23	63.32	47.39	64.28	48.06	46.39	27.13	65.43	45.94	59.98	41.35
使用 *DLO*	full	full	62.07	40.23	63.47	49.64	62.88	49.02	46.50	27.01	65.58	46.61	60.10	42.50
EDA	550	550	45.93	31.12	46.84	34.87	46.81	38.17	39.96	26.28	49.55	33.48	45.82	32.78
QAIE	141.8	248.8	48.89	35.07	48.29	37.15	51.71	37.13	37.70	23.57	58.26	38.78	48.97	34.34
DS^2-ABSA	21,7k	21,6k	37.80	27.92	43.13	34.96	32.25	23.58	29.42	16.03	42.16	24.01	36.95	25.30
完整的人类标注示例集：最先进的方法														
释义	full	full	-	46.16	-	56.63	**70.22**	**57.37**	52.73	**32.06**	**68.41**	52.61	**63.79**	48.97
DLO	full	full	-	**47.08**	-	**57.92**	68.60	56.67	**52.79**	32.03	**68.41**	**54.39**	63.27	**49.62**

表5： **LA-ABSA 的精确率得分**。结果与基于EDA的数据增强方法、QAIE、DS^2-ABSA 以及提示基线（使用 0、10 和 50 个标注样例）进行比较，相关数据由 Hellwig 等人 (2025a) 报告；此外还与全监督模型进行比较，包括 DLO（Hu 等人，2022b）和释义（Zhang 等人，2021）。每种标注设置（0、10、50 个或全部样例）中的最高精确率得分以粗体显示；所有设置中的最佳总体得分加下划线。

D.2. 召回率

方法	# 训练		Rest15		Rest16		FlightABSA		Coursera		酒店		AVG	
	TASD	ASQP	TASD	ASQP	TASD	ASQP	TASD	ASQP	TASD	ASQP	TASD	ASQP		
场景：提供了 0 个标注样例														
Gemma-3-27B（零样本）	0	0	**31.36**	**26.29**	46.57	**30.29**	56.90	45.42	32.58	15.14	41.02	23.16	41.69	28.06
LA-ABSA														
含释义	full	full	28.54	20.25	**49.68**	28.54	55.69	46.31	34.06	15.74	43.22	**27.18**	42.24	27.60
含*DLO*	full	full	27.36	20.91	49.27	29.81	**56.98**	**47.39**	**34.47**	**16.02**	**43.85**	26.21	**42.38**	**28.07**
场景：给定10个标注样例														
Gemma-3-27B 上下文学习	10	10	**52.66**	**40.50**	**65.19**	47.93	60.87	45.08	**40.37**	21.31	**55.17**	28.29	**54.85**	36.63
释义	10	10	7.38	1.11	5.93	3.23	7.64	2.85	14.51	4.26	12.39	2.06	9.57	2.70
DLO	10	10	13.47	4.13	13.95	4.91	13.91	4.03	20.86	4.02	17.39	3.41	15.92	4.10
LA-ABSA														
使用释义	full	full	46.98	36.43	61.76	45.91	58.68	45.93	38.65	22.99	53.82	**31.07**	51.98	36.47
使用 *DLO*	full	full	48.19	38.99	61.28	**48.81**	**61.51**	**47.36**	39.63	**23.98**	54.50	29.93	53.02	**37.82**
EDA	110	110	28.19	9.53	19.02	11.69	19.62	10.37	26.89	13.07	25.18	7.27	23.78	10.39
QAIE	26.4	45.2	16.24	8.19	10.24	12.84	14.73	12.42	22.56	15.81	17.19	8.19	16.19	11.49
DS^2-ABSA	21,1k	21,1k	28.54	18.29	30.96	24.38	26.69	15.73	18.24	7.81	31.59	11.54	27.21	15.55
场景：给定 50 个标注样例														
Gemma-3-27B 上下文学习	50	50	**57.16**	39.25	**65.77**	48.06	**62.95**	45.25	39.75	21.71	56.92	37.17	**56.51**	38.29
释义	50	50	34.60	26.62	34.66	23.25	31.38	17.42	31.43	18.21	36.05	22.59	33.63	21.62
DLO	50	50	38.65	28.60	43.33	30.06	43.89	29.22	33.32	17.89	41.02	25.99	40.04	26.35
LA-ABSA														
使用释义	full	full	52.50	37.01	61.12	46.16	60.76	46.31	**42.50**	24.42	56.40	40.75	54.65	38.93
使用 *DLO*	full	full	55.15	**40.53**	60.65	**50.06**	62.27	**48.95**	42.46	**24.50**	**57.77**	**42.11**	55.66	**41.23**
EDA	550	550	42.49	30.74	44.66	34.32	44.54	35.39	36.02	23.07	42.45	31.60	42.03	31.02
QAIE	141.8	248.8	41.70	32.78	42.30	33.46	45.57	31.33	34.87	21.44	44.60	33.52	41.81	30.51
DS^2-ABSA	21,7k	21,6k	38.08	30.01	43.94	36.67	34.63	24.17	29.71	16.81	38.54	23.23	36.98	26.18
完整的人类标注示例集：最先进的方法														
释义	full	full	-	47.72	-	59.30	69.26	58.17	51.02	32.63	67.01	55.19	62.43	50.60
DLO	full	full	-	**49.33**	-	**61.80**	**69.30**	**60.10**	**52.38**	**33.07**	**68.71**	**56.56**	**63.46**	**52.17**

表6： **LA-ABSA 的召回率**。结果与基于EDA的数据增强方法、QAIE、DS^2-ABSA 以及提示基线（使用 0、10 和 50 个标注样例）进行比较，相关数据由Hellwig 等人 (2025a) 报告；此外还与全监督模型进行比较，包括 DLO（Hu 等人，2022b）和释义（Zhang 等人，2021）。每种标注设置（0、10、50 个或全部样例）中的最高召回率以粗体显示；所有设置中的最佳总体得分加下划线。

D.3. 宏平均F1

方法	# 训练		Rest15		Rest16		FlightABSA		Coursera		酒店		AVG	
	TASD	ASQP	TASD	ASQP	TASD	ASQP	TASD	ASQP	TASD	ASQP	TASD	ASQP		
场景：给定 0 个已标注样例														
Gemma-3-27B（零样本）0		0	**33.68**	**23.64**	**45.94**	23.64	**55.24**	**39.50**	**33.78**	**13.23**	**29.70**	**18.49**	**39.67**	**23.70**
LA-ABSA														
使用释义	full	full	22.88	19.93	35.84	26.20	45.95	33.61	11.85	4.09	16.20	12.44	26.54	19.25
使用 *DLO*	full	full	24.12	21.02	36.12	**29.16**	45.37	36.32	12.10	3.70	17.18	13.10	26.98	20.66
场景：提供 10 个标注样例														
Gemma-3-27B 上下文学习	10	10	**54.53**	**38.67**	**64.33**	**43.21**	**56.32**	36.58	**41.75**	**20.44**	**31.15**	18.05	**49.62**	**31.39**
释义	10	10	3.47	0.46	2.59	0.87	2.38	1.03	2.17	0.66	2.65	0.47	2.65	0.70
DLO	10	10	6.10	1.42	6.01	1.25	4.74	1.35	3.33	0.60	4.10	0.83	4.86	1.09
LA-ABSA														
带释义	full	full	34.17	23.98	48.59	34.59	51.20	36.29	17.01	8.54	26.90	**19.19**	35.57	24.52
带 *DLO*	full	full	35.03	26.77	50.39	36.25	53.19	**39.01**	17.14	9.38	29.02	18.18	36.95	25.92
EDA	110	110	11.33	5.57	7.94	2.22	8.05	3.64	4.22	2.02	5.49	2.27	7.40	3.15
QAIE	26.4	45.2	14.86	5.37	7.84	14.29	12.02	7.00	4.30	1.90	3.52	2.86	8.51	6.28
DS^2-ABSA	21,1k	21,1k	18.83	12.17	26.68	15.60	22.66	13.04	9.65	3.17	16.97	7.89	18.96	10.37
场景：给定 50 个标注样例														
Gemma-3-27B 上下文学习	50	50	**56.01**	**36.01**	**64.81**	**44.56**	**54.06**	36.67	**38.24**	**17.80**	**35.47**	**25.24**	**49.72**	**32.06**
释义	50	50	19.75	14.05	16.35	8.35	12.43	7.23	9.59	2.99	13.87	7.54	14.40	8.03
DLO	50	50	27.78	18.10	22.46	13.05	28.90	19.71	8.68	3.11	16.50	9.74	20.86	12.74
LA-ABSA														
使用释义	full	full	35.74	26.57	48.34	31.01	52.88	36.65	17.96	7.37	28.20	19.79	36.62	24.28
使用 *DLO*	full	full	41.64	28.42	49.37	37.37	52.17	**40.88**	18.07	8.10	30.19	22.59	38.29	27.47
EDA	550	550	27.35	17.61	22.38	16.12	33.71	24.74	12.22	5.72	15.85	13.07	22.30	15.45
QAIE	141.8	248.8	44.81	30.84	45.50	32.23	37.41	22.29	20.79	10.20	24.19	18.51	34.54	22.82
DS^2-ABSA	21,7k	21,6k	24.95	18.14	31.43	21.81	28.52	17.92	11.86	4.69	18.35	13.06	23.02	15.12
完整的人类标注示例集：最先进的方法														
释义	full	full	-	-	-	-	**58.48**	48.47	30.67	**15.66**	37.49	30.31	42.21	31.48
DLO	full	full	-	-	-	-	57.24	**49.41**	**32.25**	14.91	**38.14**	**33.04**	**42.54**	**32.45**

表7：**LA-ABSA 的宏平均F1分数**。评估结果与基于 EDA 的数据增强方法、QAIE、DS^2-ABSA 以及提示基线（0、10 和 50 个标注样例）进行比较，相关结果引自 Hellwig 等人 (2025a)；此外，还与包括 DLO（Hu 等人，2022b）和释义（Zhang 等人，2021）在内的全监督模型进行比较。每种标注设置（0、10、50 个或全部样例）中的最高 F1macro 分数以粗体显示；所有设置中的最佳总体分数以下划线标出。