

面向方面级情感分析的QLoRA轻量微调与提示视图一致性探索

黄永旗

广东工业大学数学与统计学院, 广东 广州

收稿日期: 2026年7月28日; 录用日期: 2026年9月4日; 发布日期: 2026年9月18日

摘要

方面级情感分析需要同时抽取评论中的方面词并判定对应情感, 传统任务专用模型在资源受限场景中往往需要针对结构和标签重新设计。本文将SemEval-2014 Restaurant的方面-情感抽取统一表述为严格JSON数组生成任务, 基于本地Qwen2.5-7B-Instruct构建Prompt-only、4-bit QLoRA及双提示视图一致性约束Co-LoRA实验。所有结果均在Restaurant测试集full800上以aspect-sentiment二元组micro-F1、JSON有效率和精确匹配率评价。实验显示, schema Prompt-only的micro-F1为0.4815; QLoRA-500、QLoRA-1000与QLoRA-full分别达到0.6126、0.6529和0.7334, QLoRA-full相对最优Prompt-only基线提升0.2519, 精确匹配率从0.3038提升到0.6913。QLoRA-full单轮训练耗时831.9 s, 峰值显存7.91 GB。进一步的Co-LoRA-1000探索获得较高的多提示预测一致性, 但schema micro-F1为0.5657, 未超过常规QLoRA。结果表明, QLoRA能够以较低资源有效适配结构化ABSA, 而提示一致性约束仍需更充分的对照与多域实验中验证。

关键词

方面级情感分析, 大语言模型, QLoRA, 参数高效微调, 结构化生成

QLoRA Lightweight Fine-Tuning and Prompt-View Consistency Exploration for Aspect-Based Sentiment Analysis

Yongqi Huang

School of Mathematics and Statistics, Guangdong University of Technology, Guangzhou Guangdong

Received: July 28, 2026; accepted: September 4, 2026; published: September 18, 2026

Abstract

Aspect-based sentiment analysis requires jointly extracting aspect terms from reviews and determining their corresponding polarities. Traditional task-specific models often need to be redesigned for structures and labels in resource-constrained settings. We formulate aspect-sentiment extraction on the SemEval-2014 Restaurant dataset as strict JSON-array generation and conduct Prompt-only, 4-bit QLoRA, and dual-prompt-view consistency-constrained Co-LoRA experiments based on local Qwen2.5-7B-Instruct. All results are evaluated on the full 800-item Restaurant test set using aspect-sentiment tuple micro-F1, JSON validity, and Exact Match. The experiment shows that the schema Prompt-only micro-F1 is 0.4815, while QLoRA-500, QLoRA-1000, and QLoRA-full achieve 0.6126, 0.6529, and 0.7334, respectively. QLoRA-full improves the best Prompt-only baseline by 0.2519 and raises Exact Match from 0.3038 to 0.6913. It completes one training epoch in 831.9 s with a peak GPU memory usage of 7.91 GB. The exploratory Co-LoRA-1000 setting attains relatively high multi-prompt prediction consistency, but its schema micro-F1 of 0.5657 does not exceed conventional QLoRA. These results show that QLoRA can effectively adapt structured ABSA under modest resource budgets, whereas prompt-consistency constraints require validation with more comprehensive controls and multi-domain experiments.

Keywords

Aspect-Based Sentiment Analysis, Large Language Model, QLoRA, Parameter-Efficient Fine-Tuning, Structured Generation

Copyright © 2026 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

方面级情感分析(ABSA)面向细粒度意见挖掘,核心目标是从评论中识别方面词并判断其情感极性。SemEval-2014 为该任务提供了 Restaurant 与 Laptop 等标准评测场景[1]。近年来,生成式 ABSA 将结构化标签转写为序列生成目标,减少了对任务专用分类头的依赖[2]。与此同时,规模化语言模型能够通过文本指令完成零样本或少样本任务[3],但在严格结构化输出和领域标签边界上仍存在不稳定性。

在单卡资源受限的条件下,直接全参微调 7B 级模型成本较高。LoRA 通过低秩增量减少可训练参数[4],QLoRA 进一步以 4-bit 量化底座和低秩适配器降低显存占用[5]。本文从工程可复现角度研究本地 Qwen2.5-7B-Instruct 的 QLoRA 适配:以严格 JSON 输出为接口,比较不同训练样本规模,并补充双提示视图一致性约束的探索。

本文的贡献包括:1) 给出一个面向 Restaurant ABSA 的 JSON 生成式 QLoRA 流程,并在 full800 测试集上报告统一指标;2) 展示训练样本由 500、1000 扩展至全量 3041 时的性能与资源变化;3) 诚实报告 Co-LoRA 的初步结果:其多提示预测具有较高的一致性,但主任务 F1 未超过常规 QLoRA。

2. 相关工作

ABSA 研究经历了任务专用判别模型、预训练语言模型适配和统一生成式建模等路径。BERT 后训练可利用评论语料改善 ABSA 的下游微调效果[6];端到端 BERT 模型也用于联合建模方面与情感[7]。生成式范式把多个 ABSA 子任务转化为统一的文本到文本预测,能自然承载结构化标签语义[2],并在低资源

条件下展现出可用性[8]。关于任务、方法与挑战的系统梳理见文献[9]。

参数高效微调方面，LoRA 在冻结预训练权重的同时学习低秩更新[4]；QLoRA 以 NF4、双重量化等设计显著降低量化大模型微调的显存需求[5]。Qwen2.5 系列提供了覆盖多尺寸的指令模型[10]，为本地化实验提供了基础。本文的 Co-LoRA 思路受一致性正则化启发：R-Drop 通过双向 KL 约束不同随机子模型输出的一致性[11]。不同之处在于，本文将一致性约束施加于同一样本的 direct 与 schema 两种提示视图，并仅在答案 token 位置计算。

3. 方法

3.1. 结构化生成任务

每个样本由评论句子 x 和标签集合 y 组成。标签被序列化为 JSON 数组，例如 $[{"aspect": "food", "sentiment": "positive"}]$ 。情感集合限定为 positive、negative 和 neutral；无方面词样本输出空数组。训练时 prompt 部分标签设为 -100，只对答案部分计算语言模型交叉熵。

3.2. QLoRA 微调

模型采用本地 Qwen2.5-7B-Instruct，加载为 4-bit NF4，计算精度为 bfloat16，并使用 double quant。LoRA 适配器配置为 $r = 8$ 、 $\alpha = 16$ 、 $\text{dropout} = 0.05$ ，插入 q_proj 和 v_proj 。训练采用 $\text{batch size} = 1$ 、梯度累积 8、最大序列长度 768、学习率 $2e-4$ 、训练 1 epoch。

3.3. 双提示视图一致性探索

对同一训练样本构造 direct 与 schema 两个 prompt，并拼接完全相同的 gold JSON。优化目标由两个视图的平均交叉熵与双向 KL 项组成；KL 仅在 answer token 位置归一化计算。本文使用 $\lambda_{\text{cons}} = 0.05$ 。该设计用于检验不同任务描述下的输出分布是否更接近，而不将其预设为必然提升 F1 的方法。

4. 实验设置

4.1. 数据与指标

使用 SemEval-2014 Restaurant 数据，训练集 3041 条，测试集 800 条[1]。评测按预测与 gold 的 (aspect, sentiment) 二元组计算 micro Precision、Recall 和 F1；同时统计 JSON 有效率和 Exact Match。所有推理采用贪心解码(不采样)，最大生成 token 数为 256。Prompt-only 使用 direct、schema、expert 三种提示，其中 schema 为最优基线。

4.2. 资源环境

实验在单张 NVIDIA GeForce RTX 4090 (24 GB) 上完成。基础模型、训练数据、预测文件和评估脚本均使用本地路径；不修改 gold 数据和评估口径。QLoRA-full 的全量训练峰值显存为 7.91 GB，说明该配置可在单卡消费级 GPU 上运行。训练配置见表 1。

Table 1. Training configuration
表 1. 训练配置

项目	设置
基础模型	Qwen2.5-7B-Instruct (本地)
量化	4-bit NF4; bfloat16 compute; double quant

续表

LoRA	$r = 8$; $\alpha = 16$; dropout = 0.05; q_proj/v_proj
优化	AdamW; learning rate = $2e-4$; 1 epoch
批处理	batch size = 1; gradient accumulation = 8
序列与生成	max_seq_length = 768; max_new_tokens = 256; temperature = 0

5. 结果与分析

5.1. QLoRA 主结果

表 2 显示, schema Prompt-only 在三种提示中最佳, 但仍明显落后于 QLoRA。QLoRA-1000 相对 schema Prompt-only 的 F1 提升 0.1714, QLoRA-full 提升 0.2519; 同时 EM 从 0.3038 提升至 0.6913。不同方法的 Micro-F1 对比见图 1。随着训练样本增加, P 与 R 保持较平衡, 表明模型不仅学习了 JSON 格式, 也改善了方面词和情感对的联合抽取。

Table 2. Main results on restaurant full800

表 2. Restaurant full800 主结果

方法	训练样本	P	R	Micro-F1	JSON	EM
Prompt-only direct	0	0.3920	0.5685	0.4640	1.0000	0.2875
Prompt-only schema	0	0.4295	0.5479	0.4815	1.0000	0.3038
Prompt-only expert	0	0.3949	0.5318	0.4533	1.0000	0.2875
QLoRA-500	500	0.6137	0.6115	0.6126	1.0000	0.4313
QLoRA-1000	1000	0.6419	0.6643	0.6529	1.0000	0.5188
QLoRA-full	3041	0.7456	0.7216	0.7334	1.0000	0.6913

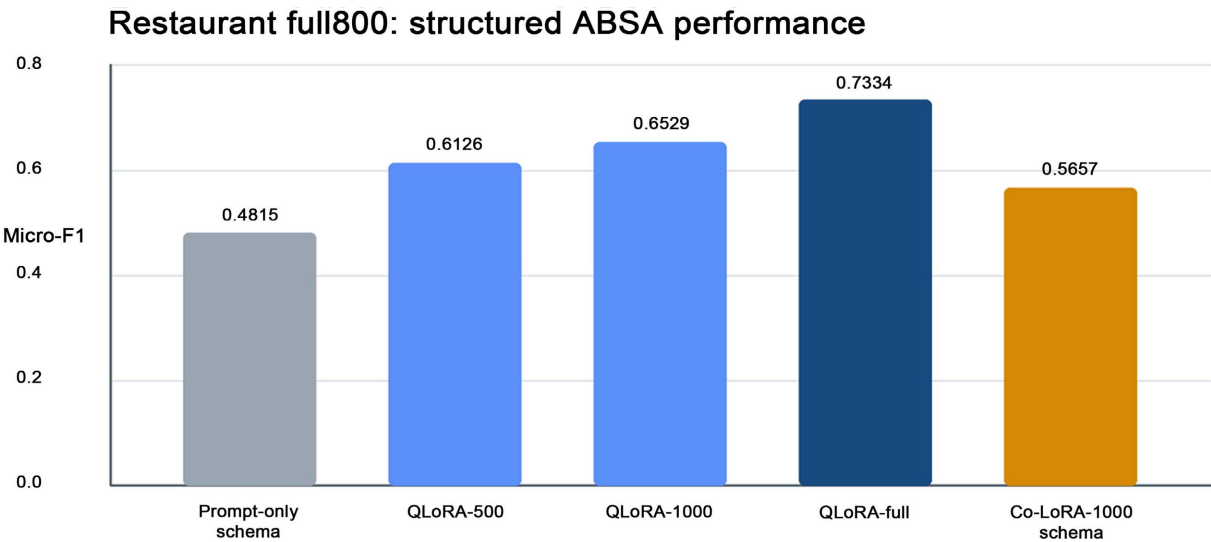


Figure 1. Micro-F1 comparison on Restaurant full800

图 1. Restaurant full800 的 Micro-F1 对比

5.2. 资源与误差分析

表 3 表明, QLoRA-full 在 831.9 s 内完成单轮训练, 峰值显存未超过 8 GB。误差统计见表 4: 与 Prompt-only schema 相比, QLoRA-1000 的方面幻觉和边界失配显著减少; 但方面遗漏略有增加, 提示下一步可针对召回引入标签覆盖或困难样本策略。

5.3. Co-LoRA 探索

Co-LoRA-1000 的 direct/schema/expert 三种推理结果中, schema F1 为 0.5657, 多提示结果与一致性统计见表 5。其 direct-vs-schema exact agreement 为 0.8925, 平均 pairwise Jaccard 为 0.8351, 说明双视图训练后的多提示预测具有较高重合度。然而, 其 schema F1 (0.5657) 低于 QLoRA-1000 的 0.6529 和 QLoRA-full 的 0.7334。因此, 本文不将 Co-LoRA 作为性能主张, 而仅将其视为待验证的稳定性正则方向。尚未获得 QLoRA-full 的三提示一致性结果前, 不能断言 Co-LoRA 相比 QLoRA-full 提升了稳定性。

Table 3. QLoRA resource consumption

表 3. QLoRA 资源消耗

方法	训练样本	训练时间(s)	峰值显存(GB)	评估损失
QLoRA-500	500	135.0	7.80	0.1245
QLoRA-1000	1000	270.8	7.80	0.0707
QLoRA-full	3041	831.9	7.91	0.0532

Table 4. Error statistics of Prompt-only schema and QLoRA-1000

表 4. Prompt-only schema 与 QLoRA-1000 误差统计

误差类型	Prompt-only	QLoRA-1000	变化
方面遗漏	141	187	-46
方面幻觉	485	236	249
边界失配	286	106	180
情感错误	78	82	-4
精确匹配失败	557	385	172

Table 5. Multi-prompt results and consistency of Co-LoRA-1000

表 5. Co-LoRA-1000 的多提示结果与一致性

项目	结果
schema P/R/F1	0.7875/0.4414/0.5657
schema JSON/EM	1.0000/0.4888
direct vs. schema exact agreement	0.8925
direct vs. expert exact agreement	0.8025
schema vs. expert exact agreement	0.7713
all three exact agreements	0.7413
average pairwise Jaccard	0.8351

6. 讨论与局限性

本文结果表明，在 Restaurant 单域设置中，QLoRA 的主要价值是以低显存实现稳定的结构化 ABSA 适配。需要注意：第一，本文主结果仅覆盖 Restaurant，Laptop 的现有试验未形成与 Restaurant 同等质量的正向证据，故未纳入主表；第二，所有配置仅运行单个随机种子，尚缺少方差和显著性检验；第三，未与最新的专用 ABSA SOTA 在完全一致的预处理条件下横向比较；第四，Co-LoRA 的稳定性结论仍依赖 QLoRA-full 的多提示对照。因而，本文适合定位为可复现的资源受限实验研究，而不是宣称新的最佳 ABSA 性能。

7. 结论

本文实现并评估了面向 Restaurant ABSA 的 JSON 生成式 QLoRA 流程。QLoRA-full 在 full800 上取得 0.7334 micro-F1 和 0.6913 EM，高于 Prompt-only schema，并保持 7.91 GB 峰值显存。Co-LoRA 的初步实验带来了较高的多提示预测重合度，但未提升主任务 F1。后续将补充 QLoRA-full 的 prompt 一致性基线、多随机种子、Laptop 诊断与跨域验证，以评估一致性约束的真实收益。

参考文献

- [1] Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I. and Manandhar, S. (2014) Semeval-2014 Task 4: Aspect Based Sentiment Analysis. *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, Dublin, 23-24 August 2014, 27-35. <https://doi.org/10.3115/v1/s14-2004>
- [2] Zhang, W., Li, X., Deng, Y., Bing, L. and Lam, W. (2021) Towards Generative Aspect-Based Sentiment Analysis. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 1-6 August 2021, 504-510. <https://doi.org/10.18653/v1/2021.acl-short.64>
- [3] Brown, T.B., Mann, B., Ryder, N., et al. (2020) Language Models Are Few-Shot Learners. *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Vancouver, 6-12 December 2020, 1877-1901.
- [4] Hu, E.J., Shen, Y., Wallis, P., et al. (2022) LoRA: Low-Rank Adaptation of Large Language Models. arXiv: 2106.09685.
- [5] Dettmers, T., Pagnoni, A., Holtzman, A. and Zettlemoyer, L. (2023) QLoRA: Efficient Finetuning of Quantized LLMs. *Proceedings of the 37th International Conference on Neural Information Processing System*, New Orleans, 10-16 December 2023, 10088-10115. <https://doi.org/10.52202/075280-0441>
- [6] Xu, H., Liu, B., Shu, L. and Yu, P. (2019) BERT Post-Training for Review Reading Comprehension and Aspect-Based Sentiment Analysis. *Proceedings of the 2019 Conference of the North*, Minneapolis, 2-7 June 2019, 2324-2335. <https://doi.org/10.18653/v1/n19-1242>
- [7] Li, X., Bing, L., Zhang, W. and Lam, W. (2019) Exploiting BERT for End-to-End Aspect-Based Sentiment Analysis. *Proceedings of the 5th Workshop on Noisy User-Generated Text (W-NUT 2019)*, Hong Kong, 4 November 2019, 34-41. <https://doi.org/10.18653/v1/d19-5505>
- [8] Hosseini-Asl, E., Liu, W. and Xiong, C. (2022) A Generative Language Model for Few-Shot Aspect-Based Sentiment Analysis. *Findings of the Association for Computational Linguistics: NAACL 2022*, Seattle, 10-15 July 2022, 770-787. <https://doi.org/10.18653/v1/2022.findings-naacl.58>
- [9] Zhang, W., Li, X., Deng, Y., Bing, L. and Lam, W. (2023) A Survey on Aspect-Based Sentiment Analysis: Tasks, Methods, and Challenges. *IEEE Transactions on Knowledge and Data Engineering*, **35**, 11019-11038. <https://doi.org/10.1109/tkde.2022.3230975>
- [10] Yang, A., Yang, B., Zhang, B., et al. (2025) Qwen2.5 Technical Report. arXiv: 2412.15115.
- [11] Liang, X., Wu, L., Li, J., et al. (2021) R-Drop: Regularized Dropout for Neural Networks. *Proceedings of the 35th International Conference on Neural Information Processing Systems*, 6-14 December 2021, 10890-10905.