# nchellwig在SemEval-2026任务3：使用大语言模型进行维度化方面级情感分析的自洽结构化生成（SCSG）

**Nils Constantin Hellwig**[1], **Jakob Fehle**[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-co nstantin.hellwig,jakob.fehle,udo.kruschwitz,christian.wolff}@ur.d e

arXiv:2603.01788v1 [cs.CL] 2 Mar 2026

## 摘要

我们介绍了用于 SemEval-2026任务3（赛道A）中维度方面级情感分析的 **自洽结构化生成（SCSG）**。SCSG 通过对每个实例多次运行经 LoRA 适配的大语言模型，仅保留在多次运行中获得多数共识的元组，从而提高预测可靠性。为降低多次前向传播带来的计算开销，我们利用 vLLM 的 PagedAttention 机制，高效复用键值缓存。在 6 种语言和 8 种语言-领域组合上的评估表明，运行 15 次的自一致性方法相比单次推理提示取得了统计显著的提升。我们的系统（采用 Gemma 3）在所有设置中均位列前七，在四个英语子集中的三个获得第二名，并在 DimASTE 的鞑靼语–餐厅任务中获得第一名。

## 1 引言

基于方面的情感分析（方面级情感分析）用于提取和分析文本中针对特定方面的观点（Zhang等人，2022年）。方面级情感分析最多识别四种情感元素：（1）方面词；（2）方面类别；（3）观点词；（4）情感极性。例如，对于“The pasta was delicious, but the service was slow”，方面级情感分析会提取出（意面，食物，美味，正面）和（服务，服务，缓慢，负面）。方面情感四元组预测（ASQP）联合提取全部四种元素，是最具挑战性的方面级情感分析任务（Zhang等人，2022年）。

研究主要关注生成式编码器-解码器架构（Zhang等人，2021b，a；Hu等人，2022年；Gou等人，2023年）。直到最近，各种方法主要以谷歌的 T5-base 模型（Raffel等人，2020年）为基础，其参数量为 2.2 亿。这些方法的主要区别在于其输出元组的表示策略，例如，将情感元素编码为结构化元组（Zhang等人，2021b）或改写后的自然语言文本（Zhang等人，2021a）。使用 T5-base 时的最佳性能由多视图提示（MvP）（Gou等人，2023年）取得。该方法在训练和推理阶段都会考虑元组中情感元素排序的多种排列方式，从而通过纳入多样的位置配置，而非采用单一固定排列，提升预测稳定性。

| **输入文本** |
|---|
| 装饰不错，不过服务有时不太稳定。 |
| **DimASQP 输出（四元组）** |
| ⟨装饰，氛围#一般，不错，$v$=7.00，$a$=7.17⟩<br>⟨服务，服务#一般，不稳定，$v$=7.00，$a$=7.00⟩ |
| **DimASTE 输出（三元组）** |
| ⟨装饰，不错，$v$=7.00, $a$=7.17⟩<br>⟨服务，不稳定，$v$=7.00, $a$=7.00⟩ |

表1：DIMASQP 和 DI-MASTE 任务的输出示例。DIMASQP 提取包含方面词、方面类别、观点词以及效价-唤醒度（$v$，$a$）维度的四元组，其中 $v$，$a \in [1, 9]$；而 DIMASTE 提取省略方面类别的三元组。

参数量显著更大、拥有数十亿参数、并在大规模无标注文本语料库上预训练的大语言模型的出现，使方面级情感分析任务获得了进一步的性能提升（Šmíd等人，2024年；Zhou等人，2024年）。Šmíd等人（2024年）报告称，在SemEval 2016方面级情感分析共享任务（Pontiki等人，2016年）的餐饮领域数据集上取得了显著改进；在目标方面情感检测（TASD）任务上，Orca-2（13B）的F1分数达到78.82，而MvP为72.76。该任务抽取方面词、方面类别和情感

极性。

维度化方面级情感分析（DimABSA）通过以连续的效价和唤醒度维度取代类别极性标签，扩展了传统的方面级情感分析，从而能够更细致地刻画情感（Lee等人，2026年）。在SemEval-2026任务3共享任务（Yu等人，2026年）（赛道A）中，提出了两个DimABSA子任务（见表1）：维度方面情感三元组抽取（DimASTE）抽取由方面词、观点词和效价-唤醒度对组成的三元组；维度方面情感四元组预测（DimASQP）则进一步纳入方面类别。

本文提出了面向DimABSA的**自洽结构化生成（SCSG）**。SCSG将自一致性（SC）提示作为一种机制，以提高预测的可靠性和有效性。在SC中，每个提示使用不同的随机种子执行 $k$次，仅保留在生成的预测中达成多数共识的元组。由于SC需要多次执行相同的提示，我们利用vLLM的PagedAttention机制高效复用键值缓存，从而降低计算开销。尽管SC在零样本和少样本提示设置中展现出显著的性能提升（Hellwig等人，2025年），迄今为止，它在用于方面级情感分析的指令微调大语言模型上的应用仍未得到探索。我们的主要贡献如下：

- 我们展示了将自一致性提示首次应用于方面级情感分析领域中的指令微调大语言模型，并在多种语言和领域上取得了统计显著的改进。
- 我们开发了一条高效的推理流水线，利用vLLM的PagedAttention在多次前向传播中复用键值缓存。
- 我们在6种语言和8种语言-领域组合上进行了全面实验。SCSG在排行榜上取得了有竞争力的排名，在所有组合中均位列前七，在四个英语子集中的三个上获得第二名，并在DimASTE的鞑靼语–餐厅子集上获得第一名。

我们公开提供代码和结果，以促进结果的复现和未来研究。[1]

[1]https://github.com/NilsHellwig/nchellwig-dim 方面级情感分析

## 2 系统概述

本节将深入介绍我们的系统。

### 2.1 微调设置与提示策略

SCSG采用低秩适配（LoRA）进行参数高效微调（Hu等人，2022年）。训练时，我们使用Unsloth库[2]，，该库能够以节省内存的方式微调大语言模型。

**训练超参数** 我们采用了 Šmíd等人（2024年）使用的超参数。他们针对方面级情感分析元组预测任务对大语言模型进行了微调。我们将LoRA的秩设为 $r = 64$，缩放因子设为 $\alpha = 16$，并将适配应用于注意力模块和多层感知机模块，同时冻结视觉层。适配层的丢弃率设为0，且不使用偏置项。我们训练了5个周期，批次大小设为16，学习率设为 $2 \times 10^{-4}$。最大序列长度设为1,024个词元。我们采用仅响应训练，即仅针对模型响应词元计算损失。

**提示** 如附录A所示，所用提示包含对情感元素（包括效价和唤醒度）的详细说明，以及所需的输出格式。方面词、观点词和方面类别的说明采用了以下文献中的定义：Hellwig等人 (2025年)、Gou等人 (2023年) 和 Šmíd等人 (2024年)，这些文献研究了大语言模型提示。对于日语、俄语、鞑靼语、乌克兰语和中文的训练数据，采用了经过翻译的提示版本。

### 2.2 验证模块

SCSG 采用验证流程来提升输出元组的质量。首先，我们使用一种自一致性机制（Wang等人，2022年），该机制此前已被证明能改善方面级情感分析任务（Hellwig等人，2025年；Bao等人，2025年），但尚未应用于经过指令微调的大语言模型。具体而言，我们以温度$k$运行大语言模型 $t = 0.8$ 次，以便更好地捕捉模型不确定性，并筛除模型置信度较低的元组。遵循Gou等人（2023年）和Hellwig等人（2025年）提出的投票机制，当且仅当某个元组至少出现在

[2]https://github.com/unslothai/unsloth

$\lceil k/2 \rceil + 1$ 个 $k$ 预测中时，才会将其纳入最终预测，从而要求达到严格的多数一致。

如果所有情感元素都相同（忽略效价-唤醒度对），则认为两个元组相同。对于按照此标准判定为相同的元组，最终的效价值和唤醒度值取所有匹配实例的算术平均值。附录C提供了计算示例。最后，我们移除了包含输入评论中不存在的文本片段的元组，将效价值和唤醒度值限制在有效范围内 $[1.00, 9.00]$，并舍弃了包含未纳入考虑的方面类别的元组。

### 2.3 推理优化

每个实例生成 $k$条预测所需的计算资源，使得高效的推理优化必不可少。由于相同的提示会被多次执行，区别仅在于不同语言、领域和子任务组合下待标注的具体示例，因此 SCSG 采用了 vLLM（Kwon 等人，2023年），这是一款面向大语言模型的高吞吐量推理引擎。具体而言，它利用 vLLM 的 PagedAttention 机制，在多次前向传播之间高效复用键值（KV）缓存。此外，批量推理可以将所有预测（$k \times N_{\text{test}}$）一并处理，而非按顺序逐个处理，其中 $N_{\text{test}}$ 表示测试集中的示例数量。

## 3 实验设置

**数据集** 我们在 SemEval-2026 Task 3 中介绍的多语言 DimABSA 基准上评估了 SCSG（Lee等人，2026年），涵盖 6 种语言（英语、日语、俄语、鞑靼语、乌克兰语、中文）和 3 个领域（餐厅、笔记本电脑、酒店），适用于 DimASTE 和 DimASQP。对于每种任务-语言-领域配置，模型均在合并后的训练集和验证集上进行微调，并在官方测试集上评估。数据集统计信息见附录 B。

**训练配置** 所有实验均在配备 96 GB 显存的 NVIDIA RTX Pro 6000（Blackwell 一代）图形处理器上进行。我们评估了 Gemma 3 的 27B 参数版本（Team 等，2025年），并采用 4 位量化[3],，该配置兼容 vLLM 和 Unsloth 框架。提交阶段结束后，我们还在 Mistral-Small-3.2（24B）上评估了 SCSG[4] 和 Qwen3 （ 32B ）（ Yang 等，2025年）[5]以评估模型在不同地区（分别为法国和中国）由不同机构开发时的泛化能力。

**评估** 按照任务组织方制定的评估协议，我们采用连续 F1（cF1）作为主要指标。cF1 基于连续真阳性（cTP）计算。只有当所有类别元素都与金标准标注完全匹配时，预测才被视为（部分）正确，即 DimASTE 三元组为（$A, O$），DimASQP 四元组为（$A, C, O$）。每个此类预测的初始真阳性得分为 1，并根据归一化效价–唤醒度（VA）预测误差进行扣减：

$$\text{cTP}(t) = 1 - \frac{(V_p - V_g)^2 + (A_p - A_g)^2}{D_{\max}}, \quad (1)$$

其中，$(V_p,\ A_p)$ 和 $(V_g,\ A_g)$ 分别表示预测的和真实的 VA 值，$D_{\max} = 8^2 + 8^2 = 128$ 是 VA 空间中 $[1, 9]$ 量表上的最大可能平方欧几里得距离。类别不匹配的预测，其连续真阳性分数为 0。

基于汇总的 cTP 值，连续精确率（cPrec）和类别召回率（cRec）的计算方式与各自的标准指标类似，分子为所有预测的 cTP 值之和。连续 F1 分数（ cF1 ）定义为 cPrec 和 cRec 的调和平均值。当效价-唤醒度预测完全准确时（即 dist $= 0$），cF1 将退化为标准 F1 分数。

我们使用不同的随机种子（0-4）运行大语言模型五次，并计算所有运行轮次的平均值，以获得稳健的结果。对于给定配置，预测总数为 $n_{\text{seeds}} \times k \times N_{\text{test}}$。我们提交了与第一个随机种子对应的预测，并公开所有预测标签，以便后续计算指标并确保结果可复现。

**试点研究：确定最佳输出验证机制**在提交任务排行榜之前，我们评估了自一致性（SC）机制的四个 $k$ 取值：$k \in \{1, 5, 10, 15\}$，其中 $k = 1$ 对应原始版本

[3]https://huggingface.co/unsloth/gemma-3-27b-it-unsloth-bnb-4bit

[4]https://huggingface.co/unsloth/Mistral-Small-3.2-24B-Instruct-2506-bnb-4bit[5] https://huggingface.co/unsloth/Qwen3-32B-uns loth-bnb-4bit

| 语言 | 领域 | 基线 | | | 5个视图 | | | 10 个视图 | | | 15 个视图 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 类别精确率 | cRec | cF1 | 类别精确率 | cRec | cF1 | 类别精确率 | cRec | cF1 | 类别精确率 | cRec | cF1 |
| 英语 | 餐厅 | 72.28 | **67.30** | 69.70 | 73.21 | 66.78 | 69.85 | **74.30** | 66.40 | **70.13*****† | 73.36 | 66.78 | 69.92 |
| 英语 | 笔记本电脑 | 65.58 | **56.11** | 60.48 | 66.21 | 55.93 | 60.64 | **67.29** | 55.78 | **60.99*****†† | 66.52 | 56.10 | 60.87* |
| 日语 | 酒店 | 53.41 | 54.53 | 53.96 | 54.84 | 54.09 | 54.46** | **57.27** | 53.21 | 55.16***†† | 56.07 | **54.64** | **55.35*****†† |
| 俄语 | 餐厅 | 52.46 | **58.32** | 55.24 | 55.11 | 57.18 | 56.13 | **55.88** | 55.87 | 55.88 | 55.42 | 57.42 | **56.40** |
| 鞑靼语 | 餐厅 | 47.47 | **50.50** | 48.94 | 50.13 | 49.78 | 49.96 | **52.15** | 48.81 | 50.42 | 51.54 | 50.48 | **51.00** |
| 乌克兰语 | 餐厅 | 50.28 | **53.46** | 51.82 | 53.24 | 52.55 | **52.89** | **54.44** | 51.20 | 52.77 | 53.32 | 52.08 | 52.69 |
| 中文 | 餐厅 | 54.22 | **55.53** | 54.87 | 55.00 | 54.03 | 54.51 | **56.37** | 53.43 | 54.86 | 55.60 | 54.22 | **54.90** |
| 中文 | 笔记本电脑 | 48.32 | **49.96** | 49.13 | 51.55 | 48.93 | 50.20*** | **53.45** | 48.05 | 50.61***† | 52.58 | 49.24 | **50.86*****†† |
| **平均值** | | 55.50 | **55.71** | 55.52 | 57.41 | 54.91 | 56.08 | **58.89** | 54.09 | 56.35 | 58.05 | 55.12 | **56.50** |

(a) 子任务 2：DIMASTE

| 语言 | 领域 | 基线 | | | 5个视图 | | | 10 个视图 | | | 15 个视图 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 类别精确率 | cRec | cF1 | 类别精确率 | cRec | cF1 | 类别精确率 | cRec | cF1 | 类别精确率 | cRec | cF1 |
| 英语 | 餐厅 | 66.64 | **60.23** | 63.27 | 67.93 | 59.76 | 63.59** | **68.93** | 59.52 | 63.88***†† | 68.49 | 60.00 | **63.97*****††† |
| 英语 | 笔记本电脑 | 42.60 | **36.70** | 39.43 | 44.62 | 35.25 | 39.39 | **47.32** | 34.81 | 40.11**†† | 46.23 | 35.67 | **40.27*****†† |
| 日语 | 酒店 | 37.54 | **38.94** | 38.23 | 40.89 | 37.64 | 39.20** | **43.70** | 36.97 | 40.06***† | 43.00 | 38.03 | **40.37*****†† |
| 俄语 | 餐厅 | 48.42 | **51.03** | 49.69 | 50.79 | 50.49 | 50.64** | **52.35** | 49.40 | 50.83*** | 51.31 | 50.63 | **50.96***** |
| 鞑靼语 | 餐厅 | 43.33 | **45.72** | 44.49 | 47.56 | 44.76 | **46.12***** | **49.00** | 43.08 | 45.85*** | 47.77 | 44.16 | 45.90*** |
| 乌克兰语 | 餐厅 | 45.67 | **46.13** | 45.90 | 48.21 | 45.55 | 46.84* | **50.31** | 44.61 | 47.29*** | 49.67 | 45.68 | **47.59*****† |
| 中文 | 餐厅 | 47.83 | **49.52** | 48.66 | 49.99 | 49.06 | 49.52*** | **51.42** | 48.19 | **49.75*****† | 50.49 | 48.90 | 49.68*** |
| 中文 | 笔记本电脑 | 38.30 | **39.94** | 39.10 | 41.40 | 38.16 | 39.71** | **44.04** | 36.76 | 40.07*** | 42.59 | 38.13 | **40.24*****† |
| **平均值** | | 46.29 | **46.03** | 46.10 | 48.92 | 45.09 | 46.88 | **50.88** | 44.17 | 47.23 | 49.94 | 45.15 | **47.37** |

(b) 子任务3：DIMASQP

表 2：DIMASTE 和 DIMASQP 在测试集上的表现：比较原始提示（基线）与进行 5、10 或 15 次提示执行的自一致性（SC）。结果以 % 为单位，展示连续级精确率（cPrec）、召回率（cRec）和 F1分数（cF1）。**粗体**值表示每种语言–领域组合及指标的最佳表现。星号表示相较于基线的改进具有统计显著性，而匕首号（†）和双匕首号（‡）则分别表示相较于 5 个和 10 个视图具有显著性，依据为经 Holm-Bonferroni 校正的 p 值（$*$： $p < 0.05$， $**$： $p < 0.01$， $***$： $p < 0.001$）。

提示，即执行一次。我们使用训练集对模型进行微调，并在开发集上对其进行评估。

## 4 结果

附录D 展示了我们的验证结果，表明对于 DimASTE，采用 $k = 15$次生成的自一致性提示取得了最佳平均表现；而对于 DimASQP，$k = 10$ 次生成的效果最为出色。在竞赛中，我们提交了最终测试集预测结果，使用这些通过实证确定的 $k$ 值即可观察到这些结果。

**英语表现强劲**表2 显示，英语在两个子任务中的整体表现均最强；在 DimASTE 和 DimASQP 上，餐厅领域的表现均优于笔记本电脑领域。不同语言之间的表现差异显著：英语的 cF1 始终最高，而鞑靼语和乌克兰语则最具挑战性。

**自一致性提升测试集表现**在所有语言–领域组合中，SC 将 DimASTE 的平均值 cF1 从 55.52 提高至 56.50，将 DimASQP 的平均值 cF1 从 46.10 提高至 47.37。从各组成指标来看，SC 通常会提高类别精确率，而类别召回率则更加稳定，并且仍是若干子集的性能瓶颈（例如，两个子任务中的英语–笔记本电脑组合）。表2 中标出了具有统计显著性的提升，并使用符号注明相应的参照条件。显著性检验流程的详情见附录E。统计分析证实，增加自一致性视图的数量，能够在几乎所有语言–领域组合中带来广泛且通常高度显著的（$p < 0.001$）cF1 提升。较高的视图数量（5 或 10）在某些情况下也能带来显著提升，尤其是在 DimASQP 上。

**排行榜**表3 报告了各语言–领域组合官方前七名的排行榜排名。在 DimASTE 中，我们在英语–餐厅排名第二，在鞑靼语–餐厅排名第一，并在所有设置中均位列前六。在 Di-mASQP 中，我们在两个英语子集上均排名第二

| 语言 | 领域 | 1st | 2nd | 3rd | 4th | 5th | 6th | 7th |
|---|---|---|---|---|---|---|---|---|
| 英语 | 笔记本电脑 | 章鱼烧 (63.66) | PALI (62.42) | PAI (61.69) | **nchellwig (60.92)** | SokraTUM (56.35) | ICT-NLP (56.22) | EmberAI (55.46) |
| 英语 | 餐厅 | 章鱼烧 (70.21) | **nchellwig (69.85)** | PALI (69.28) | PAI (69.03) | kevinyu66 (67.07) | EmberAI (66.41) | AILS-NTUA (65.18) |
| 日语 | 酒店 | TeleAI (58.37) | TeamLasse (56.94) | PAI (56.82) | PALI (56.66) | **nchellwig (55.18)** | kevinyu66 (53.66) | 章鱼烧 (53.40) |
| 俄语 | 餐厅 | PAI（57.93） | TeleAI（57.36） | PALI（57.24） | **nchellwig（56.40）** | 章鱼烧（55.64） | 哈比卜大学 (54.92) | ALPS-Lab (54.14) |
| 鞑靼语 | 餐厅 | **nchellwig (51.19)** | 章鱼烧 (50.92) | PAI (49.08) | TeleAI (48.63) | 哈比卜大学 (48.39) | PALI (48.28) | ALPS-Lab (47.98) |
| 乌克兰语 | 餐厅 | PAI（57.87） | TeleAI（57.12） | PALI（56.71） | ALPS-Lab（56.13） | 章鱼烧（54.38） | 哈比卜大学（53.24） | **nchellwig（52.85）** |
| 中文 | 笔记本电脑 | PALI（53.08） | PAI（53.06） | TeleAI（52.92） | **nchellwig（51.10）** | ALPS-Lab（49.35） | TeamLasse（48.07） | kevinyu66（48.02） |
| 中文 | 餐厅 | PAI (56.38) | PALI (56.34) | **nchellwig (54.88)** | TeleAI (54.48) | 章鱼烧 (53.82) | TeamLasse（53.20） | ALPS-Lab（52.47） |

(a) 子任务 2：DIMASTE

| 语言 | 领域 | 1st | 2nd | 3rd | 4th | 5th | 6th | 7th |
|---|---|---|---|---|---|---|---|---|
| 英语 | 笔记本电脑 | 章鱼烧 (42.27) | **nchellwig (40.06)** | PALI (37.93) | PAI (37.58) | ALPS-Lab (33.95) | TeleAI（32.81） | 古典派（30.72） |
| 英语 | 餐厅 | 章鱼烧（65.14） | **nchellwig（64.03）** | PALI (63.95) | ALPS-Lab (62.02) | AILS-NTUA (59.88) | TeamLasse (59.37) | HUS@NLP-VNU（58.71） |
| 日语 | 酒店 | PALI（42.52） | 章鱼烧（40.86） | NLANGPROC（40.28） | TeamLasse（39.92） | **nchellwig（39.74）** | AILS-NTUA (37.47) | ALPS-Lab (36.17) |
| 俄语 | 餐厅 | PAI (55.99) | PALI (54.96) | 章鱼烧（51.30） | **nchellwig（50.83）** | ALPS-Lab（50.42） | TeamLasse（49.91） | NLANGPROC（45.54） |
| 鞑靼语 | 餐厅 | 章鱼烧 (47.36) | **nchellwig (45.57)** | PAI (45.23) | PALI (44.43) | ALPS-Lab (44.04) | TeamLasse（41.13） | NLANGPROC（37.68） |
| 乌克兰语 | 餐厅 | PAI（54.37） | PALI（53.07） | ALPS-Lab (51.63) | 章鱼烧 (50.19) | TeamLasse (48.79) | **nchellwig (47.46)** | NLANGPROC (46.31) |
| 中文 | 笔记本电脑 | NYCU Speech Lab (48.24) | PALI (43.19) | PAI (43.16) | **nchellwig (40.16)** | ALPS-Lab (39.68) | NLANGPROC（38.36） | 章鱼烧（37.45） |
| 中文 | 餐厅 | NYCU Speech Lab（55.21） | PAI（53.60） | PALI (53.57) | TeamLasse (50.26) | **nchellwig (49.66)** | ALPS-Lab (48.53) | NLANGPROC (46.61) |

(b) 子任务3：DimASQP

表3：所有语言–领域组合中两个子任务的排行榜前 6 名。我们的提交结果以**粗体**标出。

并在其他语言上保持竞争力。我们与第一名的最小差距出现在 DimASTE 的英语–餐厅上（69.85 对比 70.21 ）；而中文 – 笔记本电脑是 DimASTE 子集中差距最大的（40.16 对比 48.24）。

**PagedAttention 显著提升运行效率**附录F量化了 vLLM 的 PagedAttention 和批量推理对效率的影响。平均而言，不进行批量处理的评估（不含自一致性+ 且不使用批次）比不含自一致性的批量评估慢 26.6× 倍：DimASTE 为 1.305 秒/例，对比 0.049 秒/例；DimASQP 为 1.635 秒/例，对比 0.058 秒/例，慢 28.2× 倍。这种加速是实现自一致性的关键。

**消融实验：大语言模型选择** 附录 G比较了 Gemma-3-27B、Mistral-Small-27B 和 Qwen3-32B 在不同自一致性视图数量下的表现。总体而言，Mistral-Small 在两个子任务上的平均表现最佳（DimASTE 为 57.24，DimASQP 为 49.48，均对应 15 个视图）。然而，最佳模型因子集而异：例如，在英语–笔记本电脑数据集的 DimASTE 子任务上，Qwen3 表现最佳；而在中文–餐厅数据集的 DimASTE 子任务上，Gemma 表现最佳。结果表明，模型选择会受到语言和领域的影响，而自一致性始终能提升不同大语言模型的表现。

## 5 结论

我们介绍了适用于 DimASTE 和 DimASQP 的**自洽结构化生成（SCSG）**。我们的方法将参数高效微调与自洽性验证机制相结合，通过聚合多次随机生成结果，并对匹配元组的效价和唤醒度值取平均。在 6 种语言和 8 个语言-领域组合中，自一致性在官方测试集上带来了稳定提升：DimASTE 的平均 cF1 从 55.52 提升至 56.50，DimASQP 的平均 cF1 从 46.10 提升至 47.37，且许多提升具有统计显著性。结果因语言而异：英语的总体得分最高，而鞑靼语和乌克兰语等低资源场景仍然难度显著更高。在 SemEval 排行榜上，SCSG 取得了有竞争力的排名，包括在两个英语DimASQP子集上均获第二名，以及在 DimASTE 的鞑靼语–餐厅任务上获得第一名。

未来工作可以聚焦于若干方向。首先，可以系统评估少样本提示变体，以探索 DimABSA 在低资源条件下的性能。其次，如 Gou等人(2023年) 所建议的，将多视图提示（MvP）用于大语言模型的参数高效微调，是一条颇具前景的途径；迄今为止，该方法仅应用于基于T5的模型。值得注意的是，对于 MvP，训练数据规模会随着可能的情感元素排序数量而增加（例如， 24×对于 DimASQP），因此需要采用高效策略来管理计算开销。最后，可以探索多语言联合训练，即使用所有语言的合并数据对单个模型进行微调。

## 参考文献


Xiaoyi Bao、Minjie Qiang、Jinghang Gu、Zhongqing Wang 和 Chu-Ren Huang。2025年。探索用于基于方面的情感分析的混合采样推理。载于 计算语言学协会：*NAACL 2025*年会议论文集，第 4199–4210 页，阿尔伯克基，新墨西哥州。计算语言学协会。

Zhibin Gou、QingyanGuo 和 Yujiu Yang。2023年。MvP：多视图提示改进方面情感元组预测。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，多伦多，加拿大。计算语言学协会。

Nils Constantin Hellwig、Jakob Fehle、Udo Kruschwitz 和 Christian Wolff。2025年。我们仍然需要人工标注员吗？使用提示引导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，维也纳，奥地利。计算语言学协会。

Sture Holm。1979年。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》，第 65–70 页。

Edward J Hu、Yelong Shen、Phillip Wallis、Zeyuan Allen-Zhu、Yuanzhi Li、Shean Wang、Lu Wang 和 Weizhu Chen。2022年。LoRA：大语言模型的低秩自适应。载于国际学习表征会议。

Mengting Hu、Yike Wu、Hang Gao、Yinhao Bai 和 Shiwan Zhao。2022年。通过模板顺序数据增强改进方面情感四元组预测。载于*2022*年自然语言处理实证方法会议论文集，第 7889–7900 页，阿布扎比，阿拉伯联合酋长国。计算语言学协会。

Woosuk Kwon、Zhuohan Li、Siyuan Zhuang、Ying Sheng、Lianmin Zheng、Cody Hao Yu、Joseph Gonzalez、Hao Zhang 和 Ion Stoica。2023年。利用 PagedAttention 高效管理大型语言模型服务的内存。载于第 *29* 届操作系统原理研讨会论文集，SOSP ’23，第 611–626 页，美国纽约州纽约市。计算机协会。

Lung-Hao Lee、Liang-Chih Yu、Natalia Loukashevich、Ilseyar Alimova、Alexander Panchenko、Tzu-Mi Lin、Zhe-Yu Xu、Jian-Yu Zhou、Guangmin Zheng、Jin Wang、Sharanya Awasthi、Jonas Becker、Jan Philip Wahle、Terry Ruas、Shamsuddeen Hassan Muhammad 和 Saif M. Mohammad。2026年。DimABSA：构建用于维度方面级情感分析的多语言多领域数据集。预印本，arXiv:2601.23022。

Maria Pontiki、Dimitris Galanis、Haris Papageorgiou、IonAndroutsopoulos、Suresh Manandhar、Mohammad AL-Smadi、Mahmoud Al-Ayyoub、Yanyan Zhao、Bing Qin、Orphée De Clercq、Véronique Hoste、Marianna Apidianaki、Xavier Tannier、Na- talia Loukachevitch、Evgeniy Kotelnikov、Nuria Bel、Salud María Jiménez-Zafra，以及 Gül. senEryi ̆git. 2016年。SemEval-2016 任务5：基于方面的情感分析。载于第*10*届语义评测国际研讨会（*SemEval-2016*）论文集，第19–30页，圣迭戈，加利福尼亚州。计算语言学协会。

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

Jakub Šmíd、Pavel Priban 和 Pavel Kral。2024年。基于 LLaMA 的模型在基于方面的情感分析中的应用。收录于第*14*届计算语言学主观性、情感与社交媒体分析方法研讨会论文集，第63–70页，泰国曼谷。计算语言学协会。

Gemma团队、Aishwarya Kamath、Johan Ferret、Shreya Pathak、Nino Vieillard、RamonaMerhej、Sarah Perrin、Tatiana Matejovicova、Alexandre Ramé、Morgane Rivière、Louis Rouillard、Thomas Mesnard、Geoffrey Cideron、J ean bastien Grill、Sabela Ramos、Edouard Yvinec、Michelle Casbon、Etienne Pot、Ivo Penchev及其他197位作者。2025年。Gemma 3 技术报告。预印本，arXiv:2503.19786。

Xuezhi Wang、Jason Wei、Dale Schuurmans、Quoc Le、Ed Chi、Sharan Narang、Aakanksha Chowdhery 和 Denny Zhou。2022年。自一致性提升了语言模型中的思维链推理。*arXiv* 预印本 *arXiv:2203.11171*。

An Yang、Anfeng Li、Baosong Yang、Beichen Zhang、Binyuan Hui、Bo Zheng、Bowen Yu、Chang Gao、Chengen Huang、Chenxu Lv、Chujie Zheng、Dayi- heng Liu、Fan Zhou、Fei Huang、Feng Hu、Hao Ge、Haoran Wei、Huan Lin、Jialong Tang 等 41 人。2025年。Qwen3 技术报告。*arXiv [cs.CL]*。

Liang-Chih Yu、Jonas Becker、Shamsuddeen Hassan Muhammad、Idris Abdulmumin、Lung-Hao Lee、Ying-Lung Lin、Jin Wang、Jan Philip Wahle、Terry Ruas、Alexander Panchenko、Ilseyar Alimova、Kai- Wei Chang、Lilian Wanzare、Nelson Odhiambo、Bela Gipp 和 Saif M. Mohammad。2026年。SemEval-2026任务3：维度化方面级情感分析（DimABSA）。载于第*20*届语义评估国际研讨会论文集（*SemEval- 2026*）。计算语言学协会。

Wenxuan Zhang, Yang Deng, Xin Li, Yifei Yuan, Li- dong Bing 和 Wai Lam。2021a。将方面情感四元组预测作为释义生成。收录于*2021*年自然语言处理实证方法会议论文集，第 9209–9219 页，线上，位于多米尼加共和国蓬塔卡纳。计算语言学协会。

Wenxuan Zhang、Xin Li、Yang Deng、Lidong Bing 和 Wai Lam。2021b。迈向生成式方面级情感分析。收录于计算语言学协会第 *59* 届年会暨第 *11* 届自然语言处理国际联合会议论文集（第 *2* 卷：短篇论文），第 504–510 页，线上。计算语言学协会。

Wenxuan Zhang、Xin Li、Yang Deng、Lidong Bing 和 Wai Lam。2022年。方面级情感分析综述：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、Zhijing Wu、Hao Wang、Xinyu Zhang、Jun Yang、Ziyi Yang 和 Shuhao Zhang。2024年。大语言模型在方面级情感分析上的综合评估。*arXiv*预印本 *arXiv:2412.02279*。

## 提示

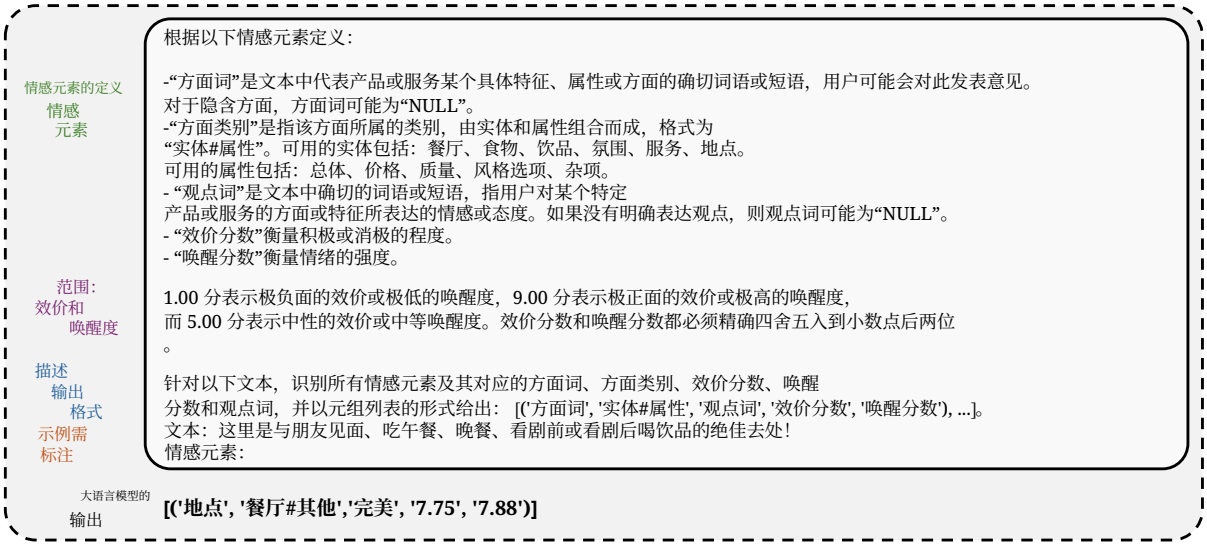



图 1：用于 SCSG 的提示。该提示包括对所考虑情感要素的描述（DimASTE 中有 4 个，DimASQP 中有 5 个）、对效价和唤醒度范围的说明、所需的输出格式，以及需要执行方面级情感分析的示例文本。

## B SemEval 2026年任务 3 的数据集：赛道A的子任务 2 和 3

| Set | 领域 | eng | jpn | rus | tat | ukr | zho |
|---|---|---|---|---|---|---|---|
| 火车 | 餐厅 | 2,284 / 3,659 | — | 1,240 / 2,487 | 1,240 / 2,487 | 1,240 / 2,487 | 6,050 / 8,523 |
| | 笔记本电脑 | 4,076 / 5,773 | — | — | — | — | 3,490 / 6,502 |
| | 酒店 | — | 1,600 / 2,846 | — | — | — | — |
| Dev | 餐厅 | 200 / 408 | — | 48 / 102 | 48 / 102 | 48 / 102 | 300 / 761 |
| | 笔记本电脑 | 200 / 317 | — | — | — | — | 300 / 551 |
| | 酒店 | — | 200 / 364 | — | — | — | — |
| Test | 餐厅 | 1,000 / 2,129 | — | 630 / 1,310 | 630 / 1,310 | 630 / 1,310 | 1,000 / 2,861 |
| | 笔记本电脑 | 1,000 / 1,975 | — | — | — | — | 1,000 / 2,798 |
| | 酒店 | — | 800 / 1,443 | — | — | — | — |

表4：DimASTE 和 DimASQP 的数据集统计，以每个数据集划分、语言和领域的句子 / 元组数量报告。破折号（—）表示相应语言–领域组合没有可用数据。

## C 多数投票机制

输入：“装饰不错，不过服务有时不太稳定。”

**预测**

| 运行 | （方面词，情感极性）-对 | 效价 | 唤醒度 | （方面词，情感极性）-对 | 效价 | 唤醒度 |
|---|---|---|---|---|---|---|
| 1 | （装饰，漂亮） | 6.92 | 7.13 | （服务，不稳定） | 5.53 | 6.03 |
| 2 | （装饰，漂亮） | 6.80 | 7.03 | （服务，不稳定） | 5.60 | 6.10 |
| 3 | (装饰, 很好) | 6.67 | 6.90 | (服务, 不稳定) | 5.40 | 5.90 |
| 4 | (装饰, 不错) | 7.00 | 7.50 | (服务, 不稳定) | 5.70 | 6.20 |
| 5 | (装饰, 很好) | 6.85 | 7.10 | (服务, 不稳定) | 5.55 | 6.05 |

**聚合**（阈值 $\tau = \lceil k/2 \rceil = 3$）

元组 *1*：(装饰, 不错) 在运行轮次 {1, 2, 4} 中出现 $\rightarrow 3$ 出现次数 $\geq \tau$

效价：$\frac{6.92+6.80+7.00}{3} = 6.91$　　唤醒度：$\frac{7.13+7.03+7.50}{3} = 7.22$

元组 *2*：(服务, 时好时坏) 在运行轮次 {1, 3, 4, 5} 中出现 $\rightarrow 4$ 出现次数 $\geq \tau$

效价：$\frac{5.53+5.40+5.70+5.55}{4} = 5.54$　　唤醒度：$\frac{6.03+5.90+6.20+6.05}{4} = 6.04$

**最终输出**

| 方面-情感对 | 效价 | 唤醒度 |
|---|---|---|
| (装饰，不错) | 6.91 | 7.22 |
| （服务，不稳定） | 5.54 | 6.04 |

图2：DimASTE 在 $k = 5$ 运行轮次上的自一致性多数投票。忽略效价-唤醒度值后，出现在 $\geq \tau = \lceil k/2 \rceil$ 运行轮次中的方面-情感对通过对其效价和唤醒度值取平均进行聚合。聚合部分展示了具体计算。浅蓝色行突出显示相匹配的（装饰，不错）变体；浅绿色行突出显示相匹配的（服务，不稳定）变体。对于 DimASQP，除方面词和情感极性外，还会考虑方面类别。

## D 验证性能

| | | 子任务 2：DimASTE | | | | 子任务3：DimASQP | | | |
|---|---|---|---|---|---|---|---|---|---|
| **语言** | **领域** | BL | 5个视图 | 10 个视图 | 15 个视图 | BL | 5个视图 | 10 个视图 | 15 个视图 |
| 英语 | 餐厅 | 77.93 | 78.35 | 78.15 | **78.45** | 75.17 | 75.30 | **75.39** | 75.26 |
| 英语 | 笔记本电脑 | 65.51 | **66.01** | 64.71 | 65.56 | 35.57 | 34.62 | **36.77** | 35.36 |
| 日语 | 酒店 | 52.63 | **54.89** | 53.91 | 54.28 | 35.93 | **39.87** | 39.18 | 38.58 |
| 俄语 | 餐厅 | 54.28 | 59.21 | 57.66 | **59.28** | 49.54 | **52.65** | 51.40 | 52.56 |
| 鞑靼语 | 餐厅 | 52.72 | 52.83 | 52.99 | **53.54** | 38.65 | **44.65** | 43.95 | 43.56 |
| 乌克兰语 | 餐厅 | 47.98 | 47.54 | 51.56 | **52.01** | 44.16 | 45.87 | **47.53** | 45.98 |
| 中文 | 餐厅 | 65.12 | 65.05 | **65.73** | 65.47 | 58.77 | 60.90 | **60.95** | 60.78 |
| 中文 | 笔记本电脑 | 45.15 | 45.30 | 45.40 | **45.62** | 36.73 | 37.50 | **38.33** | 38.15 |
| **平均值** | | 57.67 | 58.65 | 58.76 | **59.28** | 46.82 | 48.92 | **49.19** | 48.78 |

表5：DimASTE 和 DimASQP 在开发集上的评估：比较原始提示（BL）与自一致性（SC）在分别进行 5、10 或 15 次提示执行并应用多数投票时，跨语言和领域的 cF1 分数。**粗体**数值表示各语言–领域组合的最佳表现。

## E 显著性检验

在进行显著性检验时，我们使用夏皮罗–威尔克检验（ $cF1$ ）检验五个随机种子下 $\alpha = 0.05$ 的得分是否服从正态分布，对每个语言–领域子集分别进行检验。完成正态性检验后，我们首先进行总体检验（如果四组数据均符合正态分布，则使用单因素方差分析；否则使用克鲁斯卡尔–沃利斯检验）作为筛选步骤，仅当检测到全局显著差异（$p < 0.05$）时，才进行两两比较。根据检验结果，我们使用双侧独立 $t$ 检验（如果数据符合正态分布）或双侧曼–惠特尼 $U$ 检验（否则）对所有条件进行两两比较。为控制家族错误率，我们使用霍尔姆–邦费罗尼校正（ $p$）对每个子任务内的所有 $\alpha = 0.05$ 值进行校正（Holm，1979年）。

## F 训练与评估时长

| 语言 | 领域 | 训练 | | | 评估 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | # | 时间（秒） | 每千次耗时（秒） | # | 启用 SC | | 未启用 SC | | 未启用 SC + 未启用批次。 | |
| | | | | | | 时间（秒） | 每例耗时（秒） | 时间（秒） | 每例耗时（秒） | 时间（秒） | 每例耗时（秒） |
| 英语 | 餐厅 | 2,284 | 3,860 | 1,690 | 1,000 | 320 | 0.321 | 37 | 0.037 | 1,151 | 1.151 |
| 英语 | 笔记本电脑 | 4,076 | 6,515 | 1,598 | 1,000 | 289 | 0.289 | 32 | 0.032 | 1,017 | 1.018 |
| 日语 | 酒店 | 1,600 | 3,415 | 2,134 | 800 | 234 | 0.293 | 29 | 0.037 | 822 | 1.028 |
| 俄语 | 餐厅 | 1,240 | 2,638 | 2,127 | 630 | 277 | 0.440 | 38 | 0.060 | 884 | 1.405 |
| 鞑靼语 | 餐厅 | 1,240 | 3,686 | 2,972 | 630 | 303 | 0.481 | 40 | 0.065 | 924 | 1.467 |
| 乌克兰语 | 餐厅 | 1,240 | 2,935 | 2,367 | 630 | 287 | 0.456 | 40 | 0.065 | 908 | 1.442 |
| 中文 | 餐厅 | 6,050 | 10,281 | 1,699 | 1,000 | 493 | 0.493 | 52 | 0.053 | 1,687 | 1.688 |
| 中文 | 笔记本电脑 | 3,490 | 7,143 | 2,046 | 1,000 | 361 | 0.362 | 42 | 0.042 | 1,240 | 1.240 |
| **平均值** | | 2,652 | 5,059 | 2,079 | 836 | 320 | 0.392 | 39 | 0.049 | 1,079 | 1.305 |

(a) 子任务 2：DIMASTE

| 语言 | 领域 | 训练 | | | 评估 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | # | 时间（秒） | 每千次耗时（秒） | # | 使用 SC | | 不使用 SC | | 无SC + 无批次。 | |
| | | | | | | 时间 (s) | 每例耗时 (s) | 时间 (s) | 每例耗时 (s) | 时间 (s) | 每例耗时（秒） |
| 英语 | 餐厅 | 2,284 | 4,957 | 2,170 | 1,000 | 403 | 0.403 | 41 | 0.042 | 1,417 | 1.417 |
| 英语 | 笔记本电脑 | 4,076 | 9,678 | 2,374 | 1,000 | 397 | 0.398 | 42 | 0.042 | 1,370 | 1.371 |
| 日语 | 酒店 | 1,600 | 4,496 | 2,810 | 800 | 316 | 0.395 | 36 | 0.046 | 1,090 | 1.364 |
| 俄语 | 餐厅 | 1,240 | 3,266 | 2,633 | 630 | 312 | 0.497 | 45 | 0.072 | 1,046 | 1.662 |
| 鞑靼语 | 餐厅 | 1,240 | 4,489 | 3,620 | 630 | 343 | 0.545 | 46 | 0.074 | 1,120 | 1.778 |
| 乌克兰语 | 餐厅 | 1,240 | 3,636 | 2,932 | 630 | 354 | 0.562 | 45 | 0.073 | 1,107 | 1.758 |
| 中文 | 餐厅 | 6,050 | 12,829 | 2,120 | 1,000 | 625 | 0.626 | 62 | 0.062 | 2,121 | 2.121 |
| 中文 | 笔记本电脑 | 3,490 | 10,066 | 2,884 | 1,000 | 473 | 0.473 | 52 | 0.052 | 1,610 | 1.611 |
| **平均值** | | 2,652 | 6,677 | 2,693 | 836 | 403 | 0.487 | 46 | 0.058 | 1,360 | 1.635 |

(b) 子任务3：DIMASQP

表 6：DimASTE 和 DimASQP 在不同语言和领域下的训练与评估时长。我们报告了采用自一致性（w/ SC，15 个视图）、不采用自一致性（w/o SC）以及不使用 vLLM 批次处理（w/o Batch.）时的时长。

## G 大语言模型性能比较

| 语言 | 领域 | 基线 | | | 5个视图 | | | 10 个视图 | | | 15 个视图 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | G | M | Q | G | M | Q | G | M | Q | G | M | Q |
| 英语 | 笔记本电脑 | 60.48 | 61.44 | 63.05 | 60.64 | 61.07 | 63.61 | 60.99 | 61.04 | 63.56 | 60.87 | 61.18 | **63.66** |
| 英语 | 餐厅 | 69.70 | 69.09 | 67.40 | 69.85 | 69.60 | 67.82 | **70.13** | 69.66 | 67.89 | 69.92 | 70.01 | 68.23 |
| 日语 | 酒店 | 53.96 | 56.64 | 49.33 | 54.46 | 56.83 | 49.97 | 55.16 | 56.73 | 49.81 | 55.35 | **57.91** | 50.38 |
| 俄语 | 餐厅 | 55.24 | 55.78 | 52.63 | 56.13 | 57.36 | 53.45 | 55.88 | 57.39 | 53.69 | 56.40 | **57.45** | 54.74 |
| 鞑靼语 | 餐厅 | 48.94 | 49.63 | 43.87 | 49.96 | 51.26 | 45.15 | 50.42 | 51.37 | 45.92 | 51.00 | **51.65** | 46.21 |
| 乌克兰语 | 餐厅 | 51.82 | 52.85 | 49.77 | 52.89 | 54.43 | 50.74 | 52.77 | 54.55 | 50.42 | 52.69 | **54.87** | 51.17 |
| 中文 | 笔记本电脑 | 49.13 | 50.33 | 47.48 | 50.20 | 51.25 | 47.83 | 50.61 | 51.31 | 48.50 | 50.86 | **51.52** | 48.75 |
| 中文 | 餐厅 | 54.87 | 52.80 | 52.07 | 54.51 | 53.28 | 53.21 | 54.86 | 52.92 | 53.49 | **54.90** | 53.35 | 53.30 |
| **平均值** | | 55.52 | 56.07 | 53.20 | 56.08 | 56.88 | 53.97 | 56.35 | 56.87 | 54.16 | 56.50 | **57.24** | 54.55 |

(a) 子任务 2：DIMASTE

| 语言 | 领域 | 基线 | | | 5个视图 | | | 10 个视图 | | | 15 个视图 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | G | M | Q | G | M | Q | G | M | Q | G | M | Q |
| 英语 | 笔记本电脑 | 39.43 | 41.39 | 38.96 | 39.39 | **42.25** | 39.05 | 40.11 | 41.80 | 38.56 | 40.27 | 42.12 | 39.56 |
| 英语 | 餐厅 | 63.27 | 64.90 | 60.62 | 63.59 | 65.60 | 61.83 | 63.88 | 65.39 | 61.69 | 63.97 | **65.93** | 61.81 |
| 日语 | 酒店 | 38.23 | 42.16 | 34.24 | 39.20 | 43.74 | 33.54 | 40.06 | 44.41 | 34.28 | 40.37 | **44.78** | 34.46 |
| 俄语 | 餐厅 | 49.69 | 50.99 | 46.64 | 50.64 | 52.78 | 48.91 | 50.83 | 52.50 | 49.06 | 50.96 | **52.87** | 49.12 |
| 鞑靼语 | 餐厅 | 44.49 | 43.24 | 37.51 | 46.12 | **46.51** | 40.30 | 45.85 | 45.93 | 40.14 | 45.90 | 46.17 | 40.94 |
| 乌克兰语 | 餐厅 | 45.90 | 48.35 | 42.83 | 46.84 | 49.17 | 44.16 | 47.29 | 50.08 | 44.61 | 47.59 | **50.50** | 45.05 |
| 中文 | 笔记本电脑 | 39.10 | 41.44 | 36.54 | 39.71 | 42.96 | 38.58 | 40.07 | 43.53 | 39.11 | 40.24 | **43.60** | 39.41 |
| 中文 | 餐厅 | 48.66 | 47.82 | 44.68 | 49.52 | 48.83 | 45.33 | 49.75 | 49.21 | 45.99 | 49.68 | **49.87** | 46.15 |
| **平均值** | | 46.10 | 47.54 | 42.75 | 46.88 | 48.98 | 43.96 | 47.23 | 49.11 | 44.18 | 47.37 | **49.48** | 44.56 |

(b) 子任务3：DIMASQP

表7：不同大语言模型上的评估（cF1）。我们比较了**G** （Gemma-3-27B）、**M**（Mistral-Small-24B）和**Q** （Qwen3-32B）。绿色元格（）表示在相应语言与领域组合上，性能超过了官方竞赛排行榜上的排名第1的团队。总体而言，结果表明，增加自一致性视图的数量能够持续提升性能，且若干配置的表现显著优于竞赛中排名最高的参赛结果。