nchellwig at SemEval-2026 Task 3: Self-Consistent Structured Generation (SCSG) for Dimensional Aspect-Based Sentiment Analysis using Large Language Models

Nils Constantin Hellwig[1], **Jakob Fehle**[1], **Udo Kruschwitz**[2], **Christian Wolff**[1]

[1]Media Informatics Group, University of Regensburg, Regensburg, Germany
[2]Information Science Group, University of Regensburg, Regensburg, Germany

{nils-constantin.hellwig,jakob.fehle,udo.kruschwitz,christian.wolff}@ur.de

Abstract

We present **Self-Consistent Structured Generation (SCSG)** for Dimensional Aspect-Based Sentiment Analysis in SemEval-2026 Task 3 (Track A). SCSG enhances prediction reliability by executing a LoRA-adapted large language model multiple times per instance, retaining only tuples that achieve a majority consensus across runs. To mitigate the computational overhead of multiple forward passes, we leverage vLLM's PagedAttention mechanism for efficient key–value cache reuse. Evaluation across 6 languages and 8 language–domain combinations demonstrates that self-consistency with 15 executions yields statistically significant improvements over single-inference prompting, with our system (leveraging Gemma 3) ranking in the top seven across all settings, achieving second place on three out of four English subsets and first place on Tatar–Restaurant for DimASTE.

1 Introduction

Aspect-Based Sentiment Analysis (ABSA) extracts and analyzes opinions toward specific aspects within text (Zhang et al., 2023). ABSA identifies up to four sentiment elements: (1) aspect term, (2) aspect category, (3) opinion term, and (4) sentiment polarity. For instance, from "The pasta was delicious, but the service was slow", ABSA extracts (*pasta*, `FOOD`, *delicious*, positive) and (*service*, `SERVICE`, *slow*, negative). Aspect Sentiment Quad Prediction (ASQP), which jointly extracts all four elements, represents the most challenging ABSA task (Zhang et al., 2023).

Research primarily considered generative encoder-decoder architectures (Zhang et al., 2021b,a; Hu et al., 2022b; Gou et al., 2023). Until recently, approaches predominantly leveraged Google's T5-base models (Raffel et al., 2020) with 220 million parameters as their foundation. These approaches primarily differed in their representation strategies for output tuples, e.g., with sentiment elements being encoded either as structured tuples (Zhang et al., 2021b) or through paraphrased natural language text (Zhang et al., 2021a). The best performance scores when using T5-base were achieved through Multi-view Prompting (MvP) (Gou et al., 2023), which considers multiple permutations of sentiment element orderings within tuples for both training and inference, thereby enhancing prediction stability by considering diverse positional configurations rather than a single fixed arrangement.

Input Text
Decor is nice though service can be spotty.
DimASQP Output (Quadruplets)
⟨*Decor*, AMBIENCE#GENERAL, *nice*, v=7.00, a=7.17⟩ ⟨*service*, SERVICE#GENERAL, *spotty*, v=7.00, a=7.00⟩
DimASTE Output (Triplets)
⟨*Decor*, *nice*, v=7.00, a=7.17⟩ ⟨*service*, *spotty*, v=7.00, a=7.00⟩

Table 1: Example outputs for DimASQP and DimASTE tasks. DimASQP extracts quadruplets containing aspect term, aspect category, opinion term, and valence-arousal (v, a) dimensions, where $v, a \in [1, 9]$, while DimASTE extracts triplets omitting the aspect category.

The emergence of substantially more parameter-rich Large Language Models (LLMs) with billions of parameters, pretrained on vast corpora of unlabeled text, enabled further performance gains in ABSA tasks (Šmíd et al., 2024; Zhou et al., 2024). Šmíd et al. (2024) reported substantial improvements on the restaurant domain dataset from the SemEval 2016 ABSA shared task (Pontiki et al., 2016), achieving an F1 score of 78.82 with Orca-2 (13B) compared to 72.76 with MvP on the Target Aspect Sentiment Detection (TASD) task, which extracts aspect terms, aspect categories, and senti-

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 37–47

ment polarity.

Dimensional ABSA (DimABSA) extends traditional ABSA by replacing categorical polarity labels with continuous valence and arousal dimensions, enabling more nuanced affective characterization (Lee et al., 2026). In the realm of the SemEval-2026 Task 3 shared task (Yu et al., 2026) (Track A), two DimABSA subtasks (see Table 1) are introduced: Dimensional Aspect Sentiment Triplet Extraction (DimASTE) extracts triplets of aspect term, opinion term, and valence-arousal pairs, while Dimensional Aspect Sentiment Quad Prediction (DimASQP) additionally includes aspect categories.

We introduce **Self-Consistent Structured Generation (SCSG)** for DimABSA, a method that incorporates self-consistency (SC) prompting to enhance prediction reliability and validity. In SC, each prompt is executed k times with distinct random seeds, yielding k candidate tuple lists. A tuple is retained in the final prediction only if it appears in a strict majority of executions. Although SC has demonstrated substantial performance gains in both prompting (Hellwig et al., 2025) and training settings (Gou et al., 2023; Hu et al., 2022b), its application to instruction-tuned LLMs for ABSA has remained unexplored to date. To mitigate the computational overhead inherent to multiple forward passes over identical prompts, we leverage vLLM's PagedAttention mechanism for efficient key–value cache reuse. Our key contributions are:

- We demonstrate the first application of self-consistency prompting to instruction-tuned LLMs in the ABSA domain, achieving statistically significant improvements across diverse languages and domains.
- We develop an efficient inference pipeline using vLLM's PagedAttention for key-value cache reuse across multiple forward passes.
- We conduct comprehensive experiments across 6 languages and 8 language–domain combinations. SCSG achieved competitive rankings on the leaderboard, placing in the top seven across all combinations, second place on three of four English subsets, and first place on the Tatar–Restaurant subset for DimASTE.

We make our code and results publicly available to facilitate reproducibility and future research.[1]

[1] https://github.com/NilsHellwig/nchellwig-dimabsa

2 System Overview

This section provides an in-depth description of our system.

2.1 Fine-Tuning Setup & Prompting Strategy

SCSG employs parameter-efficient fine-tuning using Low-Rank (LoRA) Adaptation (Hu et al., 2022a). For training, we utilize the Unsloth library[2], which enables memory-efficient fine-tuning of LLMs.

Training Hyperparameters We applied the hyperparameters employed by Šmíd et al. (2024), who fine-tuned LLMs for ABSA tuple prediction tasks. We configured LoRA with rank $r = 64$ and scaling factor $\alpha = 16$, applying adaptation to both attention and MLP modules while keeping vision layers frozen. We set the dropout rate to 0 and used no bias terms in the adaptation layers. We trained for 5 epochs with a batch size of 16 and set the learning rate to 2×10^{-4}. The maximum sequence length was set to 1,024 tokens. We applied response-only training, where the loss is computed exclusively on the model response tokens.

Prompt As presented in Appendix A, the employed prompt consists of a detailed description of the task and output format, as commonly done for structured NLP tasks (Schulhoff et al., 2025). The task description outlines all considered sentiment elements, including valence and arousal. The descriptions of aspect term, opinion term, and aspect category follow those used by Hellwig et al. (2025), Gou et al. (2023), and Šmíd et al. (2024) for LLM prompting. For training data in Japanese, Russian, Tatar, Ukrainian, and Chinese, a translated version of the prompt was employed.

2.2 Validation Module

SCSG employs a validation pipeline to improve the quality of output tuples. First, we utilize a self-consistency mechanism (Wang et al., 2023) that has previously demonstrated improvements in ABSA tasks (Hellwig et al., 2025; Bao et al., 2025) but has not yet been adapted to instruction-tuned LLMs. Specifically, we execute the LLM k times with temperature $t = 0.8$ to better capture model uncertainty and filter tuples for which the model exhibits low confidence. Following the voting mechanisms proposed by Gou et al. (2023) and Hellwig et al. (2025), a tuple is included in the

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

final prediction if and only if it appears in at least $\lceil k/2 \rceil + 1$ of the k predictions, thereby requiring strict majority agreement.

Two tuples are considered identical if all sentiment elements match, ignoring the valence-arousal pair. For tuples deemed identical under this criterion, the final valence and arousal values are computed as the arithmetic mean across all matching instances. An illustrative example of the calculation is provided in Appendix C. Finally, we removed tuples containing text spans not present in the input review, capped valence and arousal values to the valid range $[1.00, 9.00]$, and discarded tuples containing aspect categories that were not considered.

2.3 Inference Optimization

The computational requirements of generating k predictions per instance necessitate efficient inference optimization. Since identical prompts are executed multiple times, differing only in the specific example to be labeled across language, domain, and subtask combinations, SCSG leverages vLLM (Kwon et al., 2023), a high-throughput inference engine for LLMs. Specifically, it leverages vLLM's PagedAttention mechanism to enable efficient key–value (KV) cache reuse across multiple forward passes. In addition, batched inference allows all predictions ($k \times N_{\text{test}}$) to be processed jointly rather than sequentially, where N_{test} denotes the number of examples in the test set.

3 Experimental Setup

Datasets We evaluated SCSG on all subsets of the multilingual DimABSA benchmark introduced in SemEval-2026 Task 3 (Lee et al., 2026), spanning 6 languages (English, Japanese, Russian, Tatar, Ukrainian, Chinese) and 3 domains (restaurant, laptop, hotel) for both DimASTE and DimASQP. For each task-language-domain configuration, models were fine-tuned on the merged training and validation sets and evaluated on the official test sets. Dataset statistics are provided in Appendix B.

Training Configuration All experiments were conducted on an NVIDIA RTX Pro 6000 (Blackwell generation) GPU with 96 GB of VRAM. We evaluated the 27B parameter variant of Gemma 3 (Team et al., 2025) using 4-bit quantization[3], which provides compatibility with both vLLM and Unsloth frameworks. Following the submission phase, we additionally evaluated SCSG on Mistral-Small-3.2 (24B)[4] and Qwen3 (32B) (Yang et al., 2025)[5] to assess generalization across models developed by different institutions in different regions (France and China, respectively).

Evaluation Following the evaluation protocol defined by the task organizers, we used continuous F1 (cF1) as the primary metric. cF1 is based on the continuous true positive (cTP). A prediction is considered (partially) correct only if all categorical elements exactly match the gold annotation, i.e., (A, O) for DimASTE triplets or (A, C, O) for DimASQP quadruplets. Each such prediction is assigned an initial true positive score of 1, which is reduced according to the normalized valence–arousal (VA) prediction error:

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

where (V_p, A_p) and (V_g, A_g) denote the predicted and gold VA values, respectively, and $D_{\max} = 8^2 + 8^2 = 128$ is the maximum possible squared Euclidean distance in the VA space on the $[1, 9]$ scale. Predictions with no categorical match receive a cTP score of 0.

Based on aggregated cTP values, continuous Precision (cPrec) and Recall (cRec) are computed analogously to their standard counterparts, with the numerator given by the sum of cTP values across all predictions. The continuous F1 score (cF1) is defined as the harmonic mean of cPrec and cRec. When VA predictions are perfect (i.e., $\text{dist} = 0$), cF1 reduces to the standard F1 score.

We executed the LLM five times with different random seeds (0-4) and computed the average across all runs to provide robust results. For a given configuration, the total number of predictions amounts to $n_{\text{seeds}} \times k \times N_{\text{test}}$.

Pilot study: Determining the Optimal Output Validation Mechanism Prior to our submission for the task leaderboard, we assessed four values of k for the self-consistency (SC) mechanism: $k \in \{1, 5, 10, 15\}$, where $k = 1$ corresponds to vanilla prompting with a single execution. We used the training set to fine-tune the models and evaluated it on the development set.

[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-unsloth-bnb-4bit

		Baseline			5 Views			10 Views			15 Views		
Language	**Domain**	cPrec	cRec	cF1	cPrec	cRec	cF1	cPrec	cRec	cF1	cPrec	cRec	cF1
English	Restaurant	72.28	**67.30**	69.70	73.21	66.78	69.85	**74.30**	66.40	**70.13*****†	73.36	66.78	69.92
English	Laptop	65.58	**56.11**	60.48	66.21	55.93	60.64	**67.29**	55.78	**60.99*****††	66.52	56.10	60.87*
Japanese	Hotel	53.41	54.53	53.96	54.84	54.09	54.46**	**57.27**	53.21	55.16***††	56.07	**54.64**	**55.35*****††
Russian	Restaurant	52.46	**58.32**	55.24	55.11	57.18	56.13	**55.88**	55.87	55.88	55.42	57.42	**56.40**
Tatar	Restaurant	47.47	**50.50**	48.94	50.13	49.78	49.96	**52.15**	48.81	50.42	51.54	50.48	**51.00**
Ukrainian	Restaurant	50.28	**53.46**	51.82	53.24	52.55	**52.89**	**54.44**	51.20	52.77	53.32	52.08	52.69
Chinese	Restaurant	54.22	**55.53**	54.87	55.00	54.03	54.51	**56.37**	53.43	54.86	55.60	54.22	**54.90**
Chinese	Laptop	48.32	**49.96**	49.13	51.55	48.93	50.20***	**53.45**	48.05	50.61***†	52.58	49.24	**50.86*****††
Average		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) Subtask 2: DIMASTE

		Baseline			5 Views			10 Views			15 Views		
Language	**Domain**	cPrec	cRec	cF1	cPrec	cRec	cF1	cPrec	cRec	cF1	cPrec	cRec	cF1
English	Restaurant	66.64	**60.23**	63.27	67.93	59.76	63.59**	**68.93**	59.52	63.88***††	68.49	60.00	**63.97*****†††
English	Laptop	42.60	**36.70**	39.43	44.62	35.25	39.39	**47.32**	34.81	40.11**††	46.23	35.67	**40.27*****††
Japanese	Hotel	37.54	**38.94**	38.23	40.89	37.64	39.20**	**43.70**	36.97	40.06***†	43.00	38.03	**40.37*****††
Russian	Restaurant	48.42	**51.03**	49.69	50.79	50.49	50.64**	**52.35**	49.40	50.83***	51.31	50.63	**50.96*****
Tatar	Restaurant	43.33	**45.72**	44.49	47.56	44.76	**46.12*****	**49.00**	43.08	45.85***	47.77	44.16	45.90***
Ukrainian	Restaurant	45.67	**46.13**	45.90	48.21	45.55	46.84*	**50.31**	44.61	47.29***	49.67	45.68	**47.59*****†
Chinese	Restaurant	47.83	**49.52**	48.66	49.99	49.06	49.52***	**51.42**	48.19	**49.75*****†	50.49	48.90	49.68***
Chinese	Laptop	38.30	**39.94**	39.10	41.40	38.16	39.71**	**44.04**	36.76	40.07***	42.59	38.13	**40.24*****†
Average		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) Subtask 3: DIMASQP

Table 2: Performance on the test set for DIMASTE and DIMASQP: Comparison of vanilla prompting (Baseline) and self-consistency (SC) with 5, 10, or 15 prompt executions. Results show continuous-level precision (cPrec), recall (cRec), and F1-score (cF1) in %. **Bold** values indicate the best performance for each language–domain pair and metric. Asterisks denote statistical significance of the improvement over the baseline, while daggers (†) and double daggers (‡) denote significance over 5 and 10 views, respectively, based on Holm-Bonferroni corrected p-values ($* : p < 0.05, ** : p < 0.01, *** : p < 0.001$).

4 Results

Validation results, detailed in Appendix D, indicate that self-consistency prompting achieved optimal average performance with $k = 15$ generations for DimASTE and $k = 10$ generations for DimASQP. For our official competition submission on *Codabench*, we provided test set predictions generated using these empirically determined values of k, based on the first of the five evaluated random seeds.

Strong performance scores in the English language Table 2 shows that English achieved the strongest overall performance across both subtasks, with the Restaurant domain outperforming Laptop for DimASTE and DimASQP. Across languages, we observe substantial performance differences: English consistently yields the highest cF1, while Tatar and Ukrainian are the most challenging languages.

Self-consistency improves test-set performance Across all language–domain pairs, SC improves average cF1 from 55.52 to 56.50 for DimASTE and from 46.10 to 47.37 for DimASQP. In terms of the underlying components, SC typically increases cPrec, while cRec is more stable and remains the limiting factor in several subsets (e.g., English–Laptop for both subtasks). Statistically significant improvements are marked in Table 2 with symbols indicating the respective reference condition. Details on the significance testing procedure are provided in Appendix E. Statistical analysis confirms that increasing the number of self-consistency views leads to widespread and frequently highly significant ($p < 0.001$) improvements in CF1 across nearly all language–domain pairs. Higher view counts (5 or 10) also yield significant gains in some cases, especially for DimASQP.

Leaderboard Table 3 reports the official top-7 leaderboard rankings for each language–domain pair. For DimASTE, we ranked 2nd on English–Restaurant and 1st on Tatar–Restaurant, and stayed within the top six across all settings. For DimASQP, we ranked 2nd on both English subsets and remained competitive on the remaining languages. Our smallest gap to rank 1 was on English–Restaurant for DimASTE (69.85 vs. 70.21), while Chinese–Laptop showed the largest gap among the

Language	Domain	# Entries	1st	2nd	3rd	4th	5th	6th	7th
English	Laptop	22	Takoyaki (63.66)	PALI (62.42)	PAI (61.69)	**nchellwig (60.92)**	SokraTUM (56.35)	ICT-NLP (56.22)	TeamLasse (55.13)
English	Restaurant	23	Takoyaki (70.21)	**nchellwig (69.85)**	PALI (69.28)	PAI (69.03)	kevinyu66 (67.07)	AILS-NTUA (65.18)	HUS@NLP-VNU (63.91)
Japanese	Hotel	16	TeleAI (58.37)	TeamLasse (56.94)	PAI (56.82)	PALI (56.66)	**nchellwig (55.18)**	kevinyu66 (53.66)	Takoyaki (53.40)
Russian	Restaurant	16	PAI (57.93)	TeleAI (57.36)	PALI (57.24)	**nchellwig (56.40)**	Takoyaki (55.64)	Habib university (54.92)	ALPS-Lab (54.14)
Tatar	Restaurant	15	**nchellwig (51.19)**	Takoyaki (50.92)	PAI (49.08)	TeleAI (48.63)	Habib university (48.39)	PALI (48.28)	ALPS-Lab (47.98)
Ukrainian	Restaurant	14	PAI (57.87)	TeleAI (57.12)	PALI (56.71)	ALPS-Lab (56.13)	Takoyaki (54.38)	Habib university (53.24)	**nchellwig (52.85)**
Chinese	Laptop	14	PALI (53.08)	PAI (53.06)	TeleAI (52.92)	**nchellwig (51.10)**	ALPS-Lab (49.35)	TeamLasse (48.07)	kevinyu66 (48.02)
Chinese	Restaurant	14	PAI (56.38)	PALI (56.34)	**nchellwig (54.88)**	TeleAI (54.48)	Takoyaki (53.82)	TeamLasse (53.20)	ALPS-Lab (52.47)

(a) Subtask 2: DIMASTE

Language	Domain	# Entries	1st	2nd	3rd	4th	5th	6th	7th
English	Laptop	18	Takoyaki (42.27)	**nchellwig (40.06)**	PALI (37.93)	PAI (37.58)	ALPS-Lab (33.95)	TeleAI (32.81)	The Classics (30.72)
English	Restaurant	17	Takoyaki (65.14)	**nchellwig (64.03)**	PALI (63.95)	ALPS-Lab (62.02)	AILS-NTUA (59.88)	TeamLasse (59.37)	HUS@NLP-VNU (58.71)
Japanese	Hotel	11	PALI (42.52)	Takoyaki (40.86)	NLANGPROC (40.28)	TeamLasse (39.92)	**nchellwig (39.74)**	AILS-NTUA (37.47)	ALPS-Lab (36.17)
Russian	Restaurant	11	PAI (55.99)	PALI (54.96)	Takoyaki (51.30)	**nchellwig (50.83)**	ALPS-Lab (50.42)	TeamLasse (49.91)	NLANGPROC (45.54)
Tatar	Restaurant	12	Takoyaki (47.36)	**nchellwig (45.57)**	PAI (45.23)	PALI (44.43)	ALPS-Lab (44.04)	TeamLasse (41.13)	NLANGPROC (37.68)
Ukrainian	Restaurant	11	PAI (54.37)	PALI (53.07)	ALPS-Lab (51.63)	Takoyaki (50.19)	TeamLasse (48.79)	**nchellwig (47.46)**	NLANGPROC (46.31)
Chinese	Laptop	12	NYCU Speech Lab (48.24)	PALI (43.19)	PAI (43.16)	**nchellwig (40.16)**	ALPS-Lab (39.68)	NLANGPROC (38.36)	Takoyaki (37.45)
Chinese	Restaurant	12	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) Subtask 3: DIMASQP

Table 3: Top-6 leaderboard rankings for the two subtasks across all language–domain pairs. Our submission is highlighted in **bold**.

DimASTE subsets (40.16 vs. 48.24).

PagedAttention yields substantial runtime gains Appendix F quantifies the efficiency impact of vLLM's PagedAttention and batched inference. On average, evaluation without batching (w/o SC + w/o Batch.) is 26.6× slower than batched evaluation without SC for DimASTE (1.305 vs. 0.049 s/ex.) and 28.2× slower for DimASQP (1.635 vs. 0.058 s/ex.). This speed-up is key to making self-consistency feasible.

Ablation: LLM selection Appendix G compares Gemma-3-27B, Mistral-Small-27B, and Qwen3-32B for different SC view counts. Overall, Mistral-Small achieves the best average performance for both subtasks (57.24 for DimASTE and 49.48 for DimASQP, both at 15 views). However, the best model varies by subset: for instance, Qwen3 is best on English–Laptop for DimASTE, while Gemma is best on Chinese–Restaurant for DimASTE. The results indicate that model choice interacts with language and domain, while SC consistently improves performance across LLMs.

5 Conclusion

We presented **Self-Consistent Structured Generation (SCSG)** for DimASTE and DimASQP. Our approach combines parameter-efficient fine-tuning with a self-consistency validation mechanism that aggregates multiple stochastic generations while averaging valence and arousal values for matching tuples. Across 6 languages and 8 language–domain combinations, self-consistency yields consistent gains on the official test sets, improving average cF1 from 55.52 to 56.50 for DimASTE and from 46.10 to 47.37 for DimASQP, with many improvements being statistically significant. Results vary strongly by language: English achieves the highest overall scores, whereas low-resource settings such as Tatar and Ukrainian remain substantially more challenging. On the SemEval leaderboard, SCSG achieved competitive rankings, including 2nd place on both English DimASQP subsets and 1st place on Tatar–Restaurant for DimASTE.

Future work could focus on several directions. First, one could systematically evaluate few-shot prompting variants to explore low-resource performance for DimABSA. Second, adapting multi-view prompting (MvP) as suggested by Gou et al. (2023) for parameter-efficient fine-tuning of LLMs represents a promising avenue, which has thus far been applied exclusively to T5-based models. Notably, for MvP, training data scales by the number of possible sentiment element orderings (e.g., 24× for DimASQP), necessitating efficient strategies for managing computational overhead. Finally, one could investigate multilingual joint training by fine-tuning a single model on the combined data from all languages.

References

Xiaoyi Bao, Minjie Qiang, Jinghang Gu, Zhongqing Wang, and Chu-Ren Huang. 2025. Exploring hybrid sampling inference for aspect-based sentiment analysis. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4199–4210, Albuquerque, New Mexico. Association for Computational Linguistics.

Zhibin Gou, Qingyan Guo, and Yujiu Yang. 2023. MvP: Multi-view prompting improves aspect sentiment tuple prediction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4380–4397, Toronto, Canada. Association for Computational Linguistics.

Nils Constantin Hellwig, Jakob Fehle, Udo Kruschwitz, and Christian Wolff. 2025. Do we still need human annotators? prompting large language models for aspect sentiment quad prediction. In *Proceedings of the 1st Joint Workshop on Large Language Models and Structure Modeling (XLLM 2025)*, pages 153–172, Vienna, Austria. Association for Computational Linguistics.

Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022a. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Mengting Hu, Yike Wu, Hang Gao, Yinhao Bai, and Shiwan Zhao. 2022b. Improving aspect sentiment quad prediction via template-order data augmentation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7889–7900, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, page 611–626, New York, NY, USA. Association for Computing Machinery.

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, and Saif M. Mohammad. 2026. Dimabsa: Building multilingual and multidomain datasets for dimensional aspect-based sentiment analysis. *Preprint*, arXiv:2601.23022.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, Pranav Sandeep Dulepet, Saurav Vidyadhara, Dayeon Ki, Sweta Agrawal, Chau Pham, Gerson Kroiz, Feileen Li, Hudson Tao, Ashay Srivastava, and 12 others. 2025. The prompt report: A systematic survey of prompt engineering techniques. *Preprint*, arXiv:2406.06608.

Jakub Šmíd, Pavel Priban, and Pavel Kral. 2024. LLaMA-based models for aspect-based sentiment analysis. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 63–70, Bangkok, Thailand. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. *Preprint*, 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, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

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, and Saif M. Mohammad. 2026. SemEval-2026 task 3: Dimensional aspect-based sentiment analysis (DimABSA). In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*. Association for Computational Linguistics.

Wenxuan Zhang, Yang Deng, Xin Li, Yifei Yuan, Lidong Bing, and Wai Lam. 2021a. Aspect sentiment quad prediction as paraphrase generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9209–9219, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021b. Towards generative aspect-based sentiment analysis. In *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)*, pages 504–510, Online. Association for Computational Linguistics.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2023. A Survey on Aspect-Based Sentiment Analysis: Tasks, Methods, and Challenges . *IEEE Transactions on Knowledge & Data Engineering*, 35(11):11019–11038.

Changzhi Zhou, Dandan Song, Yuhang Tian, Zhijing Wu, Hao Wang, Xinyu Zhang, Jun Yang, Ziyi Yang, and Shuhao Zhang. 2024. A comprehensive evaluation of large language models on aspect-based sentiment analysis. *Preprint*, arXiv:2412.02279.

A Prompt

Definition of sentiment elements

According to the following sentiment elements definition:

- The 'aspect term' is the exact word or phrase in the text that represents a specific feature, attribute, or aspect of a product or service that a user may express an opinion about. The aspect term might be 'NULL' for implicit aspect.
- The 'aspect category' refers to the category that the aspect belongs to. It is a combination of an entity and an attribute in the format 'ENTITY#ATTRIBUTE'. The available entities are: RESTAURANT, FOOD, DRINKS, AMBIENCE, SERVICE, LOCATION. The available attributes are: GENERAL, PRICES, QUALITY, STYLE_OPTIONS, MISCELLANEOUS.
- The 'opinion term' is the exact word or phrase in the text that refers to the sentiment or attitude expressed by a user towards a particular aspect or feature of a product or service. The opinion term might be 'NULL' if no explicit opinion expression is present.
- The 'valence score' measures the degree of positivity or negativity.
- The 'arousal score' measures the intensity of emotion.

Range of valence and arousal

A score of 1.00 indicates extremely negative valence or very low arousal, 9.00 indicates extremely positive valence or very high arousal, and 5.00 represents a neutral valence or medium arousal. Both valence score and arousal score must be rounded to exactly two decimal places.

Description of output format

For the following text, recognize all sentiment elements with their corresponding aspect terms, aspect categories, valence score, arousal score, opinion in the form of a list of tuples [('aspect term', 'ENTITY#ATTRIBUTE', 'opinion term', 'valence score', 'arousal score'), ...].

Example to be labelled

Text: this is the perfect spot for meeting friends , having lunch , dinner , pre - theatre or after - theatre drinks !
Sentiment Elements:

LLM's Output

[('spot', 'RESTAURANT#MISCELLANEOUS','perfect', '7.75', '7.88')]

Figure 1: Prompt used for SCSG. The prompt comprises descriptions of the considered sentiment elements (4 for DimASTE, 5 for DimASQP), explanations regarding the range of valence and arousal, the desired output format, and the example text for which ABSA is to be performed.

B Datasets of SemEval 2026 Task 3: Subtask 2 & 3 of Track A

Set	Domain	eng	jpn	rus	tat	ukr	zho
Train	Restaurant	2,284 / 3,659	—	1,240 / 2,487	1,240 / 2,487	1,240 / 2,487	6,050 / 8,523
	Laptop	4,076 / 5,773	—	—	—	—	3,490 / 6,502
	Hotel	—	1,600 / 2,846	—	—	—	—
Dev	Restaurant	200 / 408	—	48 / 102	48 / 102	48 / 102	300 / 761
	Laptop	200 / 317	—	—	—	—	300 / 551
	Hotel	—	200 / 364	—	—	—	—
Test	Restaurant	1,000 / 2,129	—	630 / 1,310	630 / 1,310	630 / 1,310	1,000 / 2,861
	Laptop	1,000 / 1,975	—	—	—	—	1,000 / 2,798
	Hotel	—	800 / 1,443	—	—	—	—

Table 4: Dataset statistics for DimASTE and DimASQP, reported as *sentences / tuples* per split, language, and domain. A dash (—) indicates that no data is available for the respective language–domain combination.

C Majority Voting Mechanism

Input: "Decor is nice though service can be spotty."

Predictions

Run	(Aspect Term, Sentiment Polarity) - Pair	Valence	Arousal	(Aspect Term, Sentiment Polarity) - Pair	Valence	Arousal
1	(Decor, nice)	6.92	7.13	(service, spotty)	5.53	6.03
2	(Decor, nice)	6.80	7.03	(service, be spotty)	5.60	6.10
3	(Decor, is nice)	6.67	6.90	(service, spotty)	5.40	5.90
4	(Decor, nice)	7.00	7.50	(service, spotty)	5.70	6.20
5	(Decor, is nice)	6.85	7.10	(service, spotty)	5.55	6.05

Aggregation (threshold $\tau = \lceil k/2 \rceil = 3$)

Tuple 1: (Decor, nice) occurs in runs $\{1, 2, 4\} \rightarrow 3$ occurrences $\geq \tau$

Valence: $\frac{6.92+6.80+7.00}{3} = 6.91$ Arousal: $\frac{7.13+7.03+7.50}{3} = 7.22$

Tuple 2: (service, spotty) occurs in runs $\{1, 3, 4, 5\} \rightarrow 4$ occurrences $\geq \tau$

Valence: $\frac{5.53+5.40+5.70+5.55}{4} = 5.54$ Arousal: $\frac{6.03+5.90+6.20+6.05}{4} = 6.04$

Final Output

Aspect-Sentiment Pair	Valence	Arousal
(Decor, nice)	6.91	7.22
(service, spotty)	5.54	6.04

Figure 2: Self-consistency majority voting for DimASTE over $k = 5$ runs. Aspect-sentiment pairs (ignoring valence-arousal values) appearing in $\geq \tau = \lceil k/2 \rceil$ runs are aggregated by averaging their valence and arousal values. The aggregation section shows the explicit calculation. Light blue rows highlight matching (Decor, nice) variants; light green rows highlight matching (service, spotty) variants. For DimASQP, in addition to the aspect term and sentiment polarity, the aspect category is considered as well.

D Validation Performance

		SUBTASK 2: DIMASTE				SUBTASK 3: DIMASQP			
Language	**Domain**	BL	5 Views	10 Views	15 Views	BL	5 Views	10 Views	15 Views
English	Restaurant	77.93	78.35	78.15	**78.45**	75.17	75.30	**75.39**	75.26
English	Laptop	65.51	**66.01**	64.71	65.56	35.57	34.62	**36.77**	35.36
Japanese	Hotel	52.63	**54.89**	53.91	54.28	35.93	**39.87**	39.18	38.58
Russian	Restaurant	54.28	59.21	57.66	**59.28**	49.54	**52.65**	51.40	52.56
Tatar	Restaurant	52.72	52.83	52.99	**53.54**	38.65	**44.65**	43.95	43.56
Ukrainian	Restaurant	47.98	47.54	51.56	**52.01**	44.16	45.87	**47.53**	45.98
Chinese	Restaurant	65.12	65.05	**65.73**	65.47	58.77	60.90	**60.95**	60.78
Chinese	Laptop	45.15	45.30	45.40	**45.62**	36.73	37.50	**38.33**	38.15
Average		57.67	58.65	58.76	**59.28**	46.82	48.92	**49.19**	48.78

Table 5: Evaluation on the development set for DimASTE and DimASQP: cF1 scores across languages and domains comparing vanilla prompting (BL) and self-consistency (SC) using either 5, 10 or 15 prompt executions, with majority voting applied. **Bold** values indicate the best performance for each language–domain pair.

E Significance Testing

For significance testing, we verified the normality of the $cF1$ scores across the five seeds using the Shapiro–Wilk test ($\alpha = 0.05$) for each language–domain subset. Following the normality check, we first conducted an omnibus test (one-way ANOVA if normality held across all four groups, or Kruskal–Wallis test otherwise) as a gatekeeper, performing pairwise comparisons only if a globally significant difference ($p < 0.05$) was detected. Based on the result, we performed pairwise comparisons between all conditions using either two-sided independent t-tests (if normality holds) or two-sided Mann–Whitney U tests (otherwise). To control the family-wise error rate, we corrected all p-values within each subtask using the Holm–Bonferroni correction ($\alpha = 0.05$) (Holm, 1979).

F Training and Evaluation Duration

Language	Domain	Training			Evaluation						
		#	Time (s)	Time/1K (s)	#	w/ SC		w/o SC		w/o SC + w/o Batch.	
						Time (s)	Time/ex. (s)	Time (s)	Time/ex. (s)	Time (s)	Time/ex. (s)
English	Restaurant	2,284	3,860	1,690	1,000	320	0.321	37	0.037	1,151	1.151
English	Laptop	4,076	6,515	1,598	1,000	289	0.289	32	0.032	1,017	1.018
Japanese	Hotel	1,600	3,415	2,134	800	234	0.293	29	0.037	822	1.028
Russian	Restaurant	1,240	2,638	2,127	630	277	0.440	38	0.060	884	1.405
Tatar	Restaurant	1,240	3,686	2,972	630	303	0.481	40	0.065	924	1.467
Ukrainian	Restaurant	1,240	2,935	2,367	630	287	0.456	40	0.065	908	1.442
Chinese	Restaurant	6,050	10,281	1,699	1,000	493	0.493	52	0.053	1,687	1.688
Chinese	Laptop	3,490	7,143	2,046	1,000	361	0.362	42	0.042	1,240	1.240
Average		2,652	5,059	2,079	836	320	0.392	39	0.049	1,079	1.305

(a) Subtask 2: DIMASTE

Language	Domain	Training			Evaluation						
		#	Time (s)	Time/1K (s)	#	w/ SC		w/o SC		w/o SC + w/o Batch.	
						Time (s)	Time/ex. (s)	Time (s)	Time/ex. (s)	Time (s)	Time/ex. (s)
English	Restaurant	2,284	4,957	2,170	1,000	403	0.403	41	0.042	1,417	1.417
English	Laptop	4,076	9,678	2,374	1,000	397	0.398	42	0.042	1,370	1.371
Japanese	Hotel	1,600	4,496	2,810	800	316	0.395	36	0.046	1,090	1.364
Russian	Restaurant	1,240	3,266	2,633	630	312	0.497	45	0.072	1,046	1.662
Tatar	Restaurant	1,240	4,489	3,620	630	343	0.545	46	0.074	1,120	1.778
Ukrainian	Restaurant	1,240	3,636	2,932	630	354	0.562	45	0.073	1,107	1.758
Chinese	Restaurant	6,050	12,829	2,120	1,000	625	0.626	62	0.062	2,121	2.121
Chinese	Laptop	3,490	10,066	2,884	1,000	473	0.473	52	0.052	1,610	1.611
Average		2,652	6,677	2,693	836	403	0.487	46	0.058	1,360	1.635

(b) Subtask 3: DIMASQP

Table 6: Training and evaluation duration across languages and domains for both DimASTE and DimASQP. We report duration with Self-Consistency (w/ SC, 15 views), without (w/o SC), and without vLLM's batch processing (w/o Batch.).

G Performance Comparison Across LLMs

Language	Domain	Baseline			5 Views			10 Views			15 Views		
		G	M	Q	G	M	Q	G	M	Q	G	M	Q
English	Laptop	60.48	61.44	63.05	60.64	61.07	63.61	60.99	61.04	63.56	60.87	61.18	**63.66**
English	Restaurant	69.70	69.09	67.40	69.85	69.60	67.82	**70.13**	69.66	67.89	69.92	70.01	68.23
Japanese	Hotel	53.96	56.64	49.33	54.46	56.83	49.97	55.16	56.73	49.81	55.35	**57.91**	50.38
Russian	Restaurant	55.24	55.78	52.63	56.13	57.36	53.45	55.88	57.39	53.69	56.40	**57.45**	54.74
Tatar	Restaurant	48.94	49.63	43.87	49.96	51.26	45.15	50.42	51.37	45.92	51.00	**51.65**	46.21
Ukrainian	Restaurant	51.82	52.85	49.77	52.89	54.43	50.74	52.77	54.55	50.42	52.69	**54.87**	51.17
Chinese	Laptop	49.13	50.33	47.48	50.20	51.25	47.83	50.61	51.31	48.50	50.86	**51.52**	48.75
Chinese	Restaurant	54.87	52.80	52.07	54.51	53.28	53.21	54.86	52.92	53.49	**54.90**	53.35	53.30
Average		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) Subtask 2: DIMASTE

Language	Domain	Baseline			5 Views			10 Views			15 Views		
		G	M	Q	G	M	Q	G	M	Q	G	M	Q
English	Laptop	39.43	41.39	38.96	39.39	**42.25**	39.05	40.11	41.80	38.56	40.27	42.12	39.56
English	Restaurant	63.27	64.90	60.62	63.59	65.60	61.83	63.88	65.39	61.69	63.97	**65.93**	61.81
Japanese	Hotel	38.23	42.16	34.24	39.20	43.74	33.54	40.06	44.41	34.28	40.37	**44.78**	34.46
Russian	Restaurant	49.69	50.99	46.64	50.64	52.78	48.91	50.83	52.50	49.06	50.96	**52.87**	49.12
Tatar	Restaurant	44.49	43.24	37.51	46.12	**46.51**	40.30	45.85	45.93	40.14	45.90	46.17	40.94
Ukrainian	Restaurant	45.90	48.35	42.83	46.84	49.17	44.16	47.29	50.08	44.61	47.59	**50.50**	45.05
Chinese	Laptop	39.10	41.44	36.54	39.71	42.96	38.58	40.07	43.53	39.11	40.24	**43.60**	39.41
Chinese	Restaurant	48.66	47.82	44.68	49.52	48.83	45.33	49.75	49.21	45.99	49.68	**49.87**	46.15
Average		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) Subtask 3: DIMASQP

Table 7: Evaluation on different LLMs (cF1). We compared **G** (Gemma-3-27B), **M** (Mistral-Small-24B), and **Q** (Qwen3-32B). Green cells ( ) indicate performance surpassing the Rank 1 team on the official competition leaderboard for the respective language and domain combination. Overall, the results demonstrate that increasing the number of self-consistency views leads to consistent performance gains, with several configurations significantly outperforming the competition's top-ranked entries.