DanceHA: A Multi-Agent Framework for Document-Level Aspect-Based Sentiment Analysis

Lei Wang[1]*, **Min Huang**[2], **Eduard Dragut**[1]

Temple University[1], Independent Researcher[2], Philadelphia, USA

{tom.lei.wang, edragut}@temple.edu, minerstudy@gmail.com

Abstract

Aspect-Based Sentiment Intensity Analysis (ABSIA) has garnered increasing attention, though research largely focuses on domain-specific, sentence-level settings. In contrast, document-level ABSIA–particularly in addressing complex tasks like extracting Aspect-Category-Opinion-Sentiment-Intensity (ACOSI) tuples–remains underexplored. In this work, we introduce DanceHA, a multi-agent framework designed for open-ended, document-level ABSIA with informal writing styles. DanceHA has two main components: Dance, which employs a divide-and-conquer strategy to decompose the long-context ABSIA task into smaller, manageable sub-tasks for collaboration among specialized agents; and HA, Human-AI collaboration for annotation. We release Inf-ABSIA, a multi-domain document-level ABSIA dataset featuring fine-grained and high-accuracy labels from DanceHA. Extensive experiments demonstrate the effectiveness of our agentic framework and show that the multi-agent knowledge in DanceHA can be effectively transferred into student models. Our results highlight the importance of the overlooked informal styles in ABSIA, as they often intensify opinions tied to specific aspects. All code, datasets, and prompts used in this study are available at https://github.com/Tom-Owl/DanceHA.

Introduction

Aspect-Based Sentiment Analysis (**ABSA**) is a fine-grained sentiment analysis (Schneider et al. 2018; Aljebreen, Meng, and Dragut 2021) direction that aims to identify sentiments expressed toward specific *aspects* (or attributes) of an entity within a given text (Pontiki et al. 2014, 2016). It refers to the process of identifying the aspect terms (or categories) mentioned (e.g., "battery life") in a product review and determining the *sentiment polarity* (e.g., positive, negative, neutral) associated with each aspect. One task in ABSA is Aspect-Category-Opinion-Sentiment (**ACOS**) tuples extraction (Cai, Xia, and Yu 2021), where in addition to aspect and sentiment, it seeks to extract broader category class to which the aspect belongs (e.g., hardware for battery) and the subjective expression *opinion* (Schneider and Dragut 2015; Hosseinia, Dragut, and Mukherjee 2019), such as amaaaazing and dull. Consider the sentence "The battery life of this

laptop is amaaaazing, but the screen is a bit dull." There are 2 ACOS tuples, (battery life, hardware, amaaaazing, positive) and (screen, display, a bit dull, negative). The prior work of ABSA span both small language models (SMLs) and large language models (LLMs) (Hoang, Bihorac, and Rouces 2019; Zhang et al. 2021; Gao et al. 2022; Gou, Guo, and Yang 2023; Su et al. 2024a; Šmíd, Priban, and Kral 2024; Bai et al. 2024).

Recent research on ABSA aims to also quantify the intensity or strength of those sentiments (e.g., mild, strong, extreme). Like sentiment, the representation of intensity can be represented ordinal, categorical, or continuous. In this study, we use a 5-point Likert scale for intensity, where 0 represents neutral and 5 represents extreme. In our example, the intensities are 5 for "amaaaazing" and 2 for "a bit dull," respectively. This is know as Aspect-Based Sentiment Intensity Analysis (**ABSIA**) (Mamta and Ekbal 2023). ACOS is thus extended to include intensity, leading to the extraction of the tuple Aspect-Category-Opinion-Sentiment-Intensity (**ACOSI**). Both ABSA and ABSIA have been studied at sentence level. Document-level ABSA and ABSIA are newer, more challenging tasks, and relatively underexplored (Song et al. 2023; Bhattacharjee and Gangadharaiah 2023), primarily due to the lack of datasets (Luo et al. 2022). A key challenge in those problems is the difficulty of constructing large-scale, high-quality, fine-grained labeled datasets, which are costly and labor-intensive to produce. *In this work, we investigate whether multi-agent frameworks can effectively tackle document-level ACOSI.*

The main focus of ABSA is to analyze user-generated content (UGC), such as product reviews (Yang, Mukherjee, and Dragut 2017; He et al. 2020; Cai, Xia, and Yu 2021; Chen, Meng, and Dragut 2025). However, there is a distinctive but overlooked feature of UGCs, which is the informal linguistic styles, such as emoji (Singh, Blanco, and Jin 2019) and lengthening words, e.g., 'coool' and 'goooood!!!!'. Brody and Diakopoulos (2011) define lengthening words as the addition of extra characters to a word's standard spelling to emphasize or alter its meaning. While previous work highlights lengthening words as important expressions in sentiment analysis (Wang and Dragut 2024), none studies them in the context of ABSA and ABSIA. *We investigate the contributing role of lengthening words in Document-level ABSIA in this paper.*
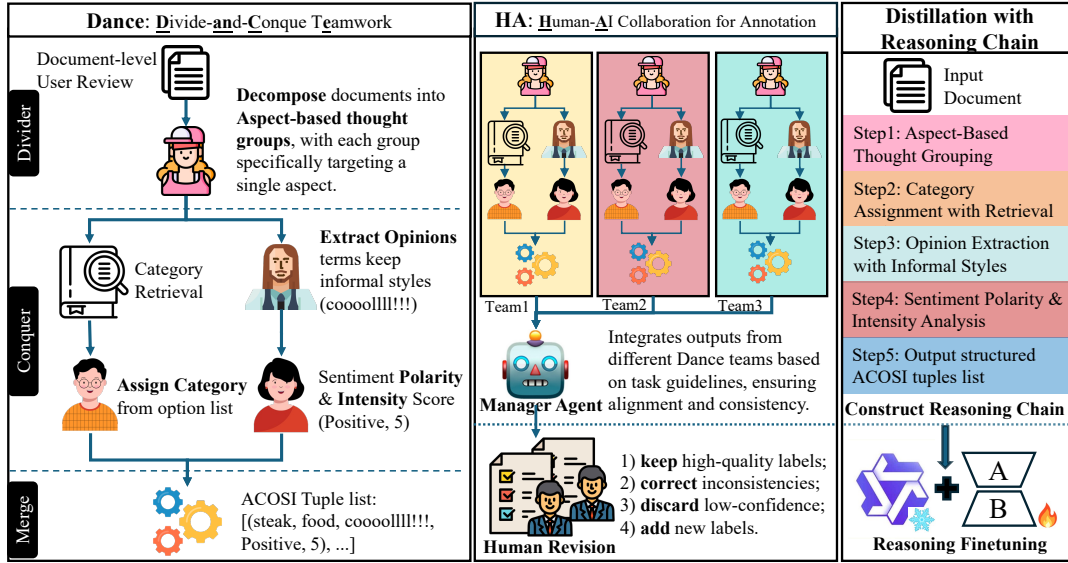
Figure 1: Overview of **DanceHA**, consisting of two key components: (1) **Dance** (Divide-and-Conquer Teamwork) for open-ended document-level ABSIA and (2) **HA** (Human-AI collaboration) for label annotation. Figure 2 illustrates Dance with an example.

In this work, we aim to propose a multi-agent framework designed for open-ended document-level ABSIA with informal styles as shown in Figure 1. Specifically, our agentic framework, **DanceHA**, consists of two key components: **Dance** (Divide-and-Conquer Teamwork), which decomposes the complex document-level ABSIA task into aspect-based thought groups for collaboration of expert agents, and **HA** (Human-AI collaboration for annotation), which is a pipeline where two Manager Agents integrate outputs from multiple Dance teams, while human annotators further evaluate and revise the labels automatically generated by the Manager Agents. We curated the **Inf-ABSIA** (Informal ABSIA) dataset, originally an unlabeled document-level dataset from 3 publicly available datasets. We use DanceHA to annotate the unlabeled documents with fine-grained labels. Human evaluation confirms the quality of the annotations. We further explore knowledge distillation with reasoning chains constructed with Inf-ABSIA. Extensive experiments on 3 domains and 7 LLMs show that our proposed agentic framework achieves strong performance on long-context, open-ended ABSIA tasks. Our contributions are as follows:

- We present **DanceHA**, a multi-agent framework with human-in-the-loop (Zhang et al. 2019, 2023a) designed for document-level ABSIA with informal and open-ended settings. Our empirical results show that Dance significantly outperforms few-shot CoT, and the Manager agent can further achieve performance boost.

- We create **Inf-ABSIA**, a novel ABSIA dataset consisting of 2,714 long-context documents across three domains. Each document contains an average of 8.48 high-quality ACOSI tuples, offering a valuable resource for studying ABSIA at the document level and enabling future re-

search on long-form opinion mining.

- We show the importance of informal styles in ABSIA, showing that informal expressions convey stronger sentiment intensity toward specific aspects and opinions.

- We show distillation with reasoning chains can effectively transfer knowledge in DanceHA into student models. Our reasoning fine-tuned Qwen-14B outperforms few-shot CoT GPT-4o across all domains and achieves performance comparable to Dance with Qwen2.5-72B.

Related Work

LLM-based Multi-Agent Collaboration LLM-based agentic framework (Zhu and Callison-Burch 2025; Feng et al. 2025) have gained attention due to their ability to simulate human-like reasoning, collaboration, and decision-making (Su et al. 2024b; Zhu et al. 2023). A line of research in this area is multi-agent collaboration for domain-specific tasks such as Question Answering (QA) (Zhao et al. 2025; Yifei et al. 2025), Peer Review (Jin et al. 2024), and script evaluation and generation (Liang et al. 2024a). To improve multi-agent teamwork, some studies employ debate strategies (Liang et al. 2024b), while others adopt decomposition strategies to guide agents in solving complex tasks (Fang, Zhu, and Gurevych 2024; Zhang et al. 2024b). Building on these methods, we propose DanceHA, a training-free multi-agent framework for document-level ACOSI tuple extraction.

LLM-based ABSA Sentence-level ABSA is an old task, yet it continues to pose challenges to LLMs (Gou, Guo, and Yang 2023; Bai et al. 2024). Šmíd, Priban, and Kral (2024) proposes instruction-tuning Orca-2 on domain-specific datasets. Recent studies also begin to address more complex, document-level ABSA subtasks (Song et al. 2023).

However, progress in this area remains limited due to the lack of fine-grained labeled datasets for tasks such as ACOS tuple extraction. To mitigate this, Seo et al. (2024) propose a sentence-splitting strategy that segments compound sentences by aspect-term boundaries, demonstrating gains on sentence-level benchmarks. In contrast, our work targets the open-ended document-level ACOSI tuple extraction task. We introduce a multi-agent collaboration framework that integrates sentiment intensity analysis to generate fine-grained annotations in the absence of manually labeled data.

Dataset Construction

We introduce **Inf-ABSIA** (<u>Inf</u>ormal <u>ABSIA</u>), a document-level dataset for ABSIA with informal styles. Inf-ABSIA is built from 3 publicly available datasets across 3 domains: **Laptop** (Lap) from Amazon Reviews-Electronics (Ni, Li, and McAuley 2019) which contains user reviews for different electronic products. We use LLM-as-Judge (Zheng et al. 2023) to decide whether a review belongs to laptop domain. **Restaurant** (Rest) from Yelp (Yelp 2021) with reviews under restaurants topic (Xu et al. 2019), and **Hotel** from TripAdvisor (Li 2020).

For each domain, we randomly sample 910 review documents that contain lengthening words, identified using a regex-based tool from prior work (Zhang et al. 2018; Wang and Dragut 2024). Table 1 compares our Inf-ABSIA with existing sentence-level and document-level ABSA and ABSIA datasets across multiple dimensions. Note that the original datasets lack ABSIA labels. We annotate them with our proposed framework, **DanceHA**. Compared to existing datasets, Inf-ABSIA spans more domains and provides a larger sample size (2,714 vs 1,194 documents), longer context (average 90 vs 60 words per document), and significantly more fine-grained annotations (23,024 vs. 4,464 tuples).

Method

We present an overview of our proposed framework **DanceHA** in Figure 1. Formally, we denote a long-context document by d from a specific domain $dm \in \{Rest, Lap, Hotel\}$. The goal is to perform ACOSI tuple extraction, producing a structured list where each element follows the format (a, c, o, s, i), where a is an aspect term extracted from d, c is a category from a pre-defined list, o is the associated opinion word or phrase from d, s indicates the sentiment polarity positive/negative, and i is the sentiment intensity score (SIS).

Dance: <u>D</u>ivide-<u>an</u>d-<u>C</u>onquer T<u>e</u>amwork for Document-level ABSIA

Dance is a multi LLM-based agents framework built on the divide-and-conquer principle, designed to handle document-level ABSIA with informal styles through agentic collaboration. As shown in Figure 2, Dance operates in three stages. In Stage 1, the Divider agent segments long documents into manageable aspect-based thought groups. Each thought group is a concise, semantically meaningful unit–typically a few sentences–representing a single aspect, fa-
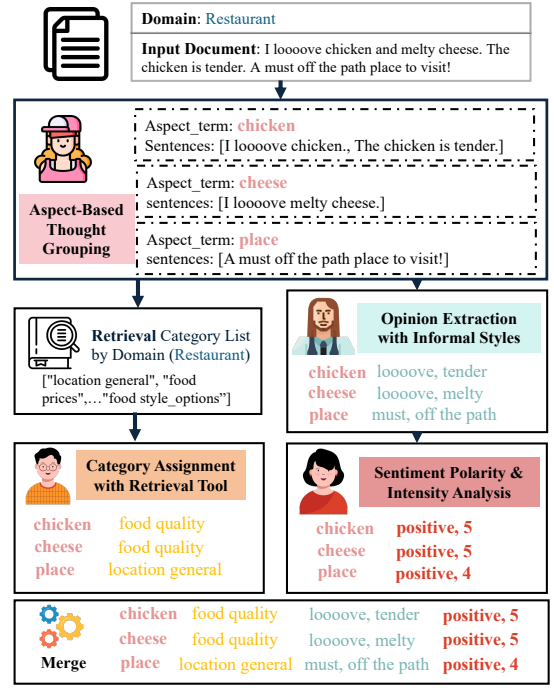

Figure 2: An example of applying Dance to document-level ABSIA. Section has the details.

cilitating a clearer understanding of the content. In stage 2, the thought groups are processed collaboratively by 3 specialized LLM-based agents, each responsible for one of the following tasks: Category Assignment, Opinion Extraction, and Sentiment and Intensity Analysis. In stage 3, Dance merges the outputs from all agents with a rule-based integration method to form a structured list of ACOSI tuples.

Divider: Aspect-Based Thought Grouping The Divider agent F_D decomposes a document d into structured aspect-based thought groups $\{(a_0, g_0), \ldots, (a_n, g_n)\}$, where a is an aspect term and g is the corresponding thought group. Based on the work of (Seo et al. 2024), the design of Divider agent has three steps: 1) Split: Split segment sentences containing multiple aspects into smaller thought groups, each corresponding to a distinct aspect a. 2) Combine: Collect consecutive sentences that pertain to the same aspect into one thought group g. 3) Extract Aspect: Identify and extract explicit aspect terms from the thought group. For convenience, we abbreviate the set $\{(a_0, g_0), \ldots, (a_k, g_k)\}$ as $\{(a, g)\}$, and this style will be used hereafter. The process can be formed as $\{(a, g)\} = F_D(d)$.

Conquer with Task-Specific Agents After aspect-based thought grouping with Divider, three specialized agents collaborate to finalize the remaining ABSIA task for each thought group with batch prompt paradigm (Lin et al. 2024). The workflow branches into two independent streams: the Category Assignment Agent and the Opinion Term Extraction Agent. Subsequently, the Sentiment and Intensity Analysis Agent evaluates sentiment polarity and determines intensity scores for each thought group with opinion terms. We describe the three specialized agents as follows:

Dataset	# D	Level	Topic	Lang	ACOS	Intensity	Informal	Size	Avg # w	Avg # Tuple	# Tuple
SemEval2016-SB1 (Pontiki et al. 2016)	2	Sentence	ABSA	E	Y	N	N	5,984	13	1.18	7,061
ACOS-Rest&Lap (Cai, Xia, and Yu 2021)	2	Sentence	ABSA	E	Y	N	N	6,362	16	1.48	9,415
ABSIA (Mamta and Ekbal 2023)	1	Sentence	ABSIA	E	N	Y	N	4,650	17	1.00	4,650
SIGHAN2024 (Lee et al. 2024)	1	Sentence	ABSIA	C	Y	Y	N	8,150	15	1.50	12,225
SemEval2016-SB2 (Pontiki et al. 2016)	2	Document	ABSA	E	N	N	N	900	60	4.96	4,464
DOTSA (Luo et al. 2022)	6	Document	ABSA	E	N	N	N	1,194	23	1.26	1,504
Inf-ABSIA (Ours)	3	Document	ABSIA	E	Y	Y	Y	**2,714**	**90**	**8.48**	**23,024**

Table 1: Statistics comparison of Inf-ABSIA with existing ABSA/ABSIA datasets. Inf-ABSIA has larger sample size (2,714 vs. 1,194), longer context (avg. words 90 vs. 60), and more fine-grained labels (avg. tuples 8.48 vs. 4.96). # D indicates the number of domains. # w denotes the number of words. Lang denotes the language, where E stands for English and C for Chinese.

- **Category Assignment Agent**, denoted by F_C, assigns the most appropriate category c to an aspect-based thought group (a, g). This agent is equipped with a rule-based retrieval-augmented generation (RAG) (Li et al. 2025; Adamu et al. 2025; Triantafyllopoulos et al. 2025) tool that can return the predefined category list $\text{RAG}(dm)$ (Gou, Guo, and Yang 2023) based on domain information dm. We formalize the process as $c = F_C(a, g, \text{RAG}(dm))$;
- **Opinion Extraction Agent**, denoted by F_O, extracts opinion words or phrases that convey sentiment intensity for a thought group (a, g). The opinion terms $\{o\}$ must be explicit substrings of the original text, preserving informal stylistic features such as lengthening expressions and punctuation). This is formalized as $\{o\} = F_O(a, g)$;
- **Sentiment Polarity & Intensity Analysis Agent**, denoted by F_S, operates sequentially after the Opinion Extraction Agent. It determines the sentiment polarity s (positive/negative) for each aspect-opinion tuple with its corresponding thought group (a, g, o). Additionally, it assigns a sentiment intensity score i (Mamta and Ekbal 2023) on a scale from 0 to 5 to indicate the strength of the expressed sentiment. The output is formalized as $(s, i) = F_S(a, g, o)$.

Merge and Structure Output We develop a rule-based integration method to merge outputs from the four specialized agents, producing the final output of a Dance team, denoted by T. This output is a structured list of aspect-based thought groups paired with corresponding ACOSI tuples, formally defined as $T = \{(g, a, c, o, s, i)\}$. We summarize the step-by-step Dance procedure in Algorithm 1, and the overall process is formalized as $T = \text{Dance}(d, dm)$.

HA: Human-AI Collaboration for Annotation

Manager Agent for Automatic Annotation We employ an Manager Agent (MA) for preliminary automatic annotate $label$, which operates on the same output structure as T. By varying base models and parameters for the LLM agents, we instantiate multiple Dance teams, each potentially producing diverse outputs for the same input document d, as illustrated in Algorithm 1. MA receives the task guidelines and the candidate outputs from all Dance teams, then integrates them by resolving inconsistencies and reconciling conflicts. This results in a consensus-driven annotation grounded in the collective insights of multiple collaborative agents.

Human Revision We recruited human annotators to evalu-

Algorithm 1: Dance

Input: Four LLM-based Agents F_D, F_C, F_O, F_S, a user review d from domain dm, two rule-based tools RAG, Merge

Output: A list of aspect-based thought groups with ACOSI tuples T

1: Initialize $T \leftarrow \{\}$
2: $\{(a, g)\} \leftarrow F_D(d)$ ▷ Divider
3: **for** each (a_j, g_j) in $\{(a, g)\}$ **do**
4:     $c \leftarrow F_C(a_j, g_j, \text{RAG}(dm))$ ▷ Category
5:     $\{o\} \leftarrow F_O(a_j, g_j)$ ▷ Opinions extraction
6:     **for** each o_k in $\{o\}$ **do**
7:         $(s, i) \leftarrow F_S(a_j, g_j, o_k)$ ▷ Sentiment
8:         $t \leftarrow \text{Merge}((a_j, g_j), c, o_k, (s, i))$
9:         Add t to T
10:     **end for**
11: **end for**
12: **Return** T

---

ate and revise the automatically generated labels produced by two manager agents powered by the DeepSeek-V3 and GPT-4o backbone models, following protocols outlined in (Lan, Aljebreen, and Dragut 2025; Zhang et al. 2025). Annotators received the original text and its candidate labels and were instructed to 1) keep high-quality labels, 2) correct inconsistencies, 3) discard low-confidence or conflicting labels, and 4) add new labels when the agents' outputs were incomplete. In total, they reviewed 2,714 documents containing 45,304 ACOSI tuples: 13,716 unique tuples (21,152 instances) were retained, 8,233 revised, 1,075 newly added, and 16,994 discarded. The authors then performed a final proofread of the dataset. To this end, we curate the Inf-ABSIA dataset, comprising 2,714 documents and 23,024 ACOSI tuples annotated through Human-AI collaboration as gold labels.

Knowledge Distillation with Reasoning Chain

Given an input document d and label $T = \{(g, a, c, o, s, i)\}$, we can construct a super-prompt sp for reasoning by merging the prompts of all expert agents. The construction of reasoning chain rc (Zhang et al. 2024a; You et al. 2025) perform the following step: 1) Aspect-based thought grouping. Output as $\{(a, g)\}$ 2) Category Assignment. Output as $\{(a, c)\}$ 3) Opinion extraction. Output as $\{(a, o)\}$ 4) Sentiment analysis. Output as $\{(a, o, s, i)\}$ 5) Merge results.

Output as $\{(g, a, c, o, s, i)\}$. We can construct a reasoning dataset in the form $\{(d, sp, rc(T)), (d, T) \in \textbf{Inf-ABSIA}\}$. We then distill knowledge (Jones et al. 2025) of our proposed framework, DanceHA, into student models through supervised fine-tuning with LoRA (Hu et al. 2021).

Domain	Method	Model	F1 (%)	Acc (%)	MAE
Rest	zs-CoT	GPT-4o	22.95 ± 1.14	12.97 ± 0.73	0.467 ± 0.053
	fs-CoT	GPT-4o	$\textbf{35.04} \pm 2.43$	$\textbf{21.27} \pm 1.80$	$\textbf{0.428} \pm 0.023$
	Dance	GPT-3.5-turbo	4.89 ± 0.31	2.50 ± 0.16	0.791 ± 0.073
	Dance	GPT-4o-mini	19.90 ± 0.38	11.05 ± 0.23	0.418 ± 0.007
	Dance	GPT-4o	$\textbf{47.80} \pm 1.65$	$\textbf{31.42} \pm 1.41$	$\textbf{0.210} \pm 0.019$
	Dance	DeepSeek-V3	37.93 ± 1.75	23.42 ± 1.33	0.274 ± 0.016
	Dance	Qwen3-4B	13.68 ± 1.13	7.34 ± 0.65	0.909 ± 0.051
	Dance	Qwen3-14B	25.87 ± 1.85	14.87 ± 1.23	1.016 ± 0.010
	Dance	Qwen2.5-72B	35.38 ± 1.23	21.50 ± 0.91	0.364 ± 0.007
	MA	GPT-4o	60.89 ± 1.43	43.79 ± 1.47	0.097 ± 0.011
	MA	DeepSeek-V3	$\textbf{63.18} \pm 1.62$	$\textbf{46.20} \pm 1.72$	$\textbf{0.086} \pm 0.010$
Hotel	zs-CoT	GPT-4o	21.74 ± 0.87	12.20 ± 0.54	0.511 ± 0.025
	fs-CoT	GPT-4o	$\textbf{31.40} \pm 2.03$	$\textbf{18.64} \pm 1.42$	$\textbf{0.453} \pm 0.007$
	Dance	GPT-3.5-turbo	2.81 ± 0.33	1.42 ± 0.17	0.710 ± 0.170
	Dance	GPT-4o-mini	16.48 ± 1.80	8.99 ± 1.07	0.458 ± 0.008
	Dance	GPT-4o	$\textbf{44.18} \pm 1.30$	$\textbf{28.36} \pm 1.07$	$\textbf{0.251} \pm 0.006$
	Dance	DeepSeek-V3	37.47 ± 1.51	23.06 ± 1.14	0.315 ± 0.008
	Dance	Qwen3-4B	14.91 ± 2.14	8.07 ± 1.25	1.192 ± 0.045
	Dance	Qwen3-14B	24.34 ± 1.24	13.86 ± 0.80	1.124 ± 0.018
	Dance	Qwen2.5-72B	32.76 ± 2.17	19.61 ± 1.54	0.327 ± 0.002
	MA	GPT-4o	56.71 ± 0.82	39.58 ± 0.80	0.105 ± 0.005
	MA	DeepSeek-V3	$\textbf{59.07} \pm 1.92$	$\textbf{41.94} \pm 1.92$	$\textbf{0.088} \pm 0.008$
Lap	zs-CoT	GPT-4o	12.80 ± 0.66	6.84 ± 0.38	0.447 ± 0.044
	fs-CoT	GPT-4o	$\textbf{16.61} \pm 0.71$	$\textbf{9.06} \pm 0.42$	$\textbf{0.465} \pm 0.034$
	Dance	GPT-3.5-turbo	1.32 ± 0.39	0.66 ± 0.20	0.734 ± 0.106
	Dance	GPT-4o-mini	6.91 ± 0.27	3.58 ± 0.15	0.451 ± 0.059
	Dance	GPT-4o	$\textbf{29.97} \pm 0.78$	$\textbf{17.63} \pm 0.54$	$\textbf{0.215} \pm 0.008$
	Dance	DeepSeek-V3	26.68 ± 1.10	15.40 ± 0.73	0.295 ± 0.006
	Dance	Qwen3-4B	5.05 ± 0.33	2.59 ± 0.17	0.998 ± 0.099
	Dance	Qwen3-14B	11.49 ± 0.96	6.10 ± 0.54	0.842 ± 0.093
	Dance	Qwen2.5-72B	16.85 ± 0.80	9.20 ± 0.47	0.356 ± 0.030
	MA	GPT-4o	45.21 ± 0.75	29.21 ± 0.62	$\textbf{0.103} \pm 0.008$
	MA	DeepSeek-V3	$\textbf{47.40} \pm 1.21$	$\textbf{31.07} \pm 1.04$	0.105 ± 0.006

Table 2: Performance comparison of Dance, MA, zero-shot (zs) and 5-shot (fs) CoT with various base models on Inf-ABSIA. F1 and Acc represent the evaluation metrics for the ACOS tuple extraction, MAE for SIS. $\pm$ indicates std score. We use boldface to indicate the best performance within each method.

Evaluation Methods

Overall Performance Following prior studies (Gou, Guo, and Yang 2023; Bai et al. 2024; Kim et al. 2024), we adopt F1-score and Accuracy (Acc) as the primary evaluation metrics for the ACOS tuple extraction task. We also use Mean Absolute Error (MAE) for sentiment intensity score (SIS), which captures how well each baseline estimates SIS, especially in the presence of informal expressions. These 3 metrics (F1, Acc, MAE) together offer a comprehensive assessment of model performance on the ACOSI tuple extraction task.

Expert Agents and Sub-tasks Performance of Dance We use accuracy (Acc) as the evaluation metric for each expert agent and its associated sub-task. For convenience, we denote accuracy as P. Given the outputs produced by a Dance team and the ground-truth labels provided by Human-AI

collaboration, we directly evaluate the performance of the Divider agent F_D using $P(a, g)$. Additionally, we assess the sub-tasks of aspect extraction and thought grouping individually using $P(a)$ and $P(g)$, respectively. To estimate the performance of downstream agents we use the probability chain rule (Bertsekas and Tsitsiklis 2008). Specifically, we assess the performance of F_C as:

$$P(c \mid a, g) \approx P(c \mid a) = P(a, c)/P(a) \qquad (1)$$

We make this approximation because a document can be segmented into thought groups in multiple ways. However, as long as the aspect extraction and the downstream predictions for opinion terms and sentiment are correct, the aspect-based thought grouping method can be considered valid. This is reasonable because ACOS evaluation metrics, such as F1-score, do not consider differences in thought group segmentation. We apply the same approximation to F_O. F_S can be further decomposed into its two sub-tasks, sentiment polarity and sentiment intensity analysis, which are evaluated as $P(s \mid a, g, o) \approx P(a, o, s)/P(a, o)$ and $P(i \mid a, g, o) \approx P(a, o, i)/P(a, o)$, respectively.

Experiments
Experimental Setting and Details
Given the lack of comparable work on document-level AB-SIA, we design several baselines to evaluate the performance of our DanceHA framework: 1) LLMs with zero-shot and few-shot (5-shot) Chain-of-Thought (CoT) Reasoning (Wei et al. 2022): We adapt ACOS tuple extraction prompts from prior studies (Bai et al. 2024; Šmíd, Priban, and Kral 2024) and extend them with CoT prompting for document-level ACOSI tuple extraction. This baseline serves to isolate and assess the effectiveness of the Dance component. 2) To examine the influence of the backbones to Dance, we implement the component with 7 LLMs: 3 close-source, GPT-4o, GPT-4o-mini, and GPT-3.5-turbo, and 4 open-source, DeepSeek-V3 (Liu et al. 2024), Qwen2.5-72B-Instruct (Yang et al. 2024), Qwen3-4B (Yang et al. 2025) and Qwen3-14B. They allow us to analyze how different LLM backbones impact the overall ACOSI tuple extraction performance. 3) For MA, we use GPT-4o and DeepSeek-V3 as the backbone model due to their strong overall performance across tasks. We select 3 Dance teams as the inputs to MA with backbones as GPT-4o, DeepSeek-V3 and Qwen2.5-72B as their superior capability among the 7 LLMs used in this study. 4) We use Qwen3-4B and Qwen3-14B models for efficient reasoning fine-tuning with LoRA. The dataset is split into 70% for training and 30% for testing. We train for 3 epochs with a learning rate of 2e-4, a LoRA rank of 32, and a context length of 4096 tokens. We set the maximum token limit to 4096 for all LLMs and the temperature to 1.0 for all models, except for DeepSeek-V3, which is set to 0.6 following the official usage recommendations (DeepSeek-AI 2025).

Results
The effectiveness of Dance and MA. Table 2 summarizes the performance of zero-shot and few-shot CoT method,

Expert Agents	Divider(F_D)			Category(F_C)	Opinion(F_O)	Sentiment Polarity & Intensity(F_S)		
Sub-tasks	Thought Group	Aspect	All	All	All	Polarity	SIS	All
Rest								
GPT-3.5-turbo	17.71 ± 0.16	8.54 ± 1.22	2.40 ± 0.26	75.33 ± 1.88	36.87 ± 3.42	**96.69** ± 0.80	40.06 ± 2.76	38.73 ± 2.68
GPT-4o-mini	20.32 ± 0.19	29.66 ± 1.48	10.29 ± 0.28	78.37 ± 1.29	46.77 ± 2.86	95.20 ± 1.98	59.52 ± 0.48	56.66 ± 1.22
GPT-4o	**54.11** ± 1.49	**54.53** ± 1.73	**36.01** ± 1.21	**82.58** ± 0.69	**68.39** ± 1.07	95.96 ± 0.49	**73.43** ± 1.25	**70.47** ± 1.54
DeepSeek-V3	51.81 ± 1.09	46.47 ± 0.82	28.34 ± 0.35	81.79 ± 0.71	61.06 ± 0.93	95.14 ± 1.64	70.52 ± 2.18	67.09 ± 2.43
Qwen3-4B	24.40 ± 0.17	26.90 ± 2.65	9.53 ± 0.76	72.77 ± 0.33	35.43 ± 2.34	96.59 ± 1.97	39.01 ± 2.88	37.69 ± 3.10
Qwen3-14B	42.19 ± 0.75	46.36 ± 2.09	21.79 ± 1.31	68.95 ± 0.24	43.70 ± 1.33	95.13 ± 1.17	28.56 ± 1.19	27.18 ± 1.38
Qwen2.5-72B	42.22 ± 0.43	48.73 ± 0.40	24.59 ± 0.60	79.20 ± 0.96	55.05 ± 2.22	94.19 ± 1.04	61.05 ± 1.14	57.50 ± 1.17
Hotel								
GPT-3.5-turbo	17.26 ± 0.65	5.07 ± 0.50	1.96 ± 0.08	75.96 ± 2.30	39.30 ± 1.54	93.42 ± 1.80	41.29 ± 8.63	38.53 ± 7.99
GPT-4o-mini	21.40 ± 1.18	24.31 ± 1.83	9.46 ± 0.63	72.29 ± 0.99	50.37 ± 2.47	93.72 ± 0.60	57.47 ± 1.19	53.85 ± 0.76
GPT-4o	**54.17** ± 0.80	**51.91** ± 1.57	**33.28** ± 1.29	77.68 ± 0.84	**68.19** ± 2.13	95.44 ± 0.54	**69.69** ± 0.68	**65.55** ± 0.53
DeepSeek-V3	52.01 ± 1.18	48.28 ± 0.84	29.31 ± 1.02	78.27 ± 1.72	62.65 ± 0.49	95.02 ± 0.98	65.34 ± 0.79	62.08 ± 0.52
Qwen3-4B	29.02 ± 0.53	29.52 ± 2.49	12.34 ± 0.85	66.25 ± 0.50	39.91 ± 2.76	94.20 ± 0.86	27.88 ± 1.10	26.27 ± 1.25
Qwen3-14B	41.38 ± 1.03	41.86 ± 1.81	19.68 ± 1.39	68.89 ± 0.84	47.51 ± 1.74	**95.76** ± 0.54	22.59 ± 1.99	21.63 ± 1.88
Qwen2.5-72B	40.98 ± 0.83	47.38 ± 1.20	23.94 ± 0.70	74.26 ± 1.10	56.73 ± 0.54	94.54 ± 1.04	64.08 ± 1.02	60.58 ± 1.04
Lap								
GPT-3.5-turbo	14.08 ± 1.34	7.16 ± 0.60	1.84 ± 0.42	32.04 ± 3.40	34.12 ± 2.72	94.27 ± 4.10	37.18 ± 4.47	35.21 ± 5.74
GPT-4o-mini	15.01 ± 1.97	19.12 ± 1.01	5.06 ± 0.57	40.76 ± 0.54	46.67 ± 0.69	93.72 ± 0.84	54.52 ± 4.58	51.08 ± 4.10
GPT-4o	**51.03** ± 2.01	**43.98** ± 0.41	**27.22** ± 1.11	64.02 ± 0.48	**61.29** ± 0.85	**94.86** ± 0.49	**73.79** ± 0.31	**70.00** ± 0.33
DeepSeek-V3	50.68 ± 1.31	42.23 ± 1.72	24.51 ± 0.64	**65.79** ± 2.74	55.98 ± 2.61	93.47 ± 0.46	68.68 ± 1.44	64.20 ± 1.66
Qwen3-4B	22.81 ± 0.98	23.19 ± 0.88	7.09 ± 0.37	34.50 ± 0.91	31.07 ± 1.38	93.09 ± 1.48	38.13 ± 4.08	35.54 ± 4.23
Qwen3-14B	40.13 ± 1.70	34.82 ± 1.70	14.26 ± 0.19	44.71 ± 2.11	37.31 ± 1.44	94.72 ± 1.34	35.69 ± 5.45	33.88 ± 5.65
Qwen2.5-72B	38.98 ± 1.52	36.33 ± 1.21	17.86 ± 1.27	51.93 ± 1.86	50.28 ± 0.76	91.93 ± 0.34	62.70 ± 0.97	57.64 ± 1.09

Table 3: Evaluation of Expert Agents in Dance and their sub-tasks. The evaluation metric is accuracy (%). We use boldface to indicate the best performance and underline to indicate the second-best.

Dance with the 7 LLMs, and MA with 2 advanced backbones. The main findings are as follows: 1) Dance equipped with GPT-4o reaches SOTA results among 7 backbones across 3 domains. This finding echoes prior work (Zheng et al. 2023) and reinforces the importance of choosing a strong base model for multi-agent frameworks. 2) Few-shot (5-shot) CoT with GPT-4o significantly outperforms zero-shot CoT, aligning with the findings of (Zhang et al. 2023b). However, it still lags behind our Dance equipped with GPT-4o and DeepSeek-V3. 3) Building on top of Dance teams and advanced LLMs, MA achieves a significant performance boost across 3 domains. For instance, in the restaurant domain, MA with DeepSeek-V3 reaches an F1 score of 63.18%, substantially outperforming the best Dance with GPT-4o, which achieves 47.80%. This finding highlights the importance of designing multi-agent collaboration topology for self-improvement (Subramaniam et al. 2025).
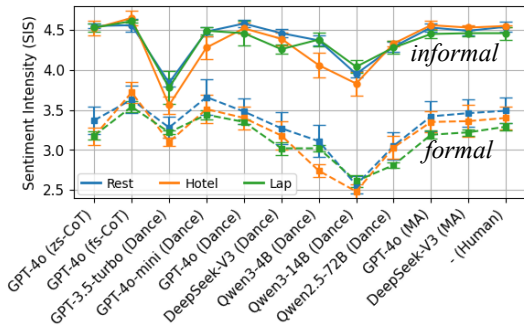

Figure 3: SIS scores for ACOSI tuples with informal and formal styles among different baselines. Error bars indicate standard deviation score. Results demonstrate the influence of informal styles on sentiment intensity and the capability of advanced models to interpret informal expressions.

Importance of Informal Styles. We analyze the differences in SIS between ACOSI tuples with and without informal styles (Figure 3). Across both CoT-based and agent-based methods, and consistently across all three domains, opinions expressed using informal styles (e.g., lengthening, extra punctuation) exhibit higher sentiment intensity score than those using more formal expressions. Notably, advanced models demonstrate a stronger sensitivity to opinions with informal cues, capturing more nuanced intensity variations in sentiment.

	Method	F1(%)	Acc(%)	MAE
Rest	Dance	35.38 ± 1.23	21.50 ± 0.91	0.364 ± 0.007
	w/o D&C	27.19 ± 0.63 (8.19 ↓)	15.73 ± 0.42 (5.76 ↓)	0.473 ± 0.033 (0.109 ↑)
	w/o Teamwork	32.48 ± 1.21 (2.90 ↓)	19.40 ± 0.87 (2.10 ↓)	0.456 ± 0.019 (0.093 ↑)
Hotel	Dance	32.76 ± 2.17	19.61 ± 1.54	0.327 ± 0.002
	w/o D&C	22.28 ± 1.40 (10.48 ↓)	12.54 ± 0.89 (7.07 ↓)	0.490 ± 0.022 (0.163 ↑)
	w/o Teamwork	29.81 ± 2.19 (2.95 ↓)	17.53 ± 1.51 (2.08 ↓)	0.479 ± 0.033 (0.151 ↑)
Lap	Dance	16.85 ± 0.80	9.20 ± 0.47	0.356 ± 0.030
	w/o D&C	10.42 ± 0.13 (6.43 ↓)	5.50 ± 0.07 (3.70 ↓)	0.473 ± 0.048 (0.118 ↑)
	w/o Teamwork	13.97 ± 0.55 (2.88 ↓)	7.51 ± 0.32 (1.69 ↓)	0.476 ± 0.070 (0.120 ↑)

Table 4: Ablation results for the Dance component. w/o D&C removes the Divide-and-Conquer Strategy; w/o Teamwork uses a single agent for the Conquer phase. Red indicates performance drops.

Performance of Expert Agents. We present the accuracy of expert agents and their sub-tasks in Table 3. 1) The Divider agent faces greater challenges, reflected by its relatively lower average accuracy, compared to other expert agents. Moreover, the two sub-tasks handled by the Divider, aspect-based thought grouping and aspect terms extraction, are more difficult. This finding shows Divider agent has potential for further improvement. 2) While the sentiment polarity classification achieves consistently high accuracy (average 94.62%), SIS prediction remains challenging with an average Acc of 60.92%. This indicates that LLMs struggle to interpret the nuanced sentiment intensity conveyed with

	Method	Model	Mode	F1 (%)	Acc (%)	MAE
Rest	zs-CoT	Qwen-4B	thinking	20.22 ± 2.34	11.27 ± 1.44	1.238 ± 0.068
	zs-CoT	Qwen-14B	thinking	19.06 ± 2.80	10.56 ± 1.73	0.740 ± 0.100
	Dance	Qwen3-4B	non-thinking	13.68 ± 1.13	7.34 ± 0.65	0.909 ± 0.051
	Dance	Qwen3-14B	non-thinking	25.87 ± 1.85	14.87 ± 1.23	1.016 ± 0.010
	SFT	Qwen3-4B	thinking	<u>30.82</u> ± 1.14	<u>18.22</u> ± 0.80	<u>0.408</u> ± 0.026
	SFT	Qwen3-14B	thinking	**34.52** ± 1.23	**20.87** ± 0.91	**0.395** ± 0.031
Hotel	zs-CoT	Qwen-4B	thinking	20.38 ± 1.27	11.35 ± 0.79	1.165 ± 0.176
	zs-CoT	Qwen-14B	thinking	24.56 ± 1.67	14.01 ± 1.09	0.940 ± 0.058
	Dance	Qwen3-4B	non-thinking	14.91 ± 2.14	8.07 ± 1.25	1.192 ± 0.045
	Dance	Qwen3-14B	non-thinking	24.34 ± 1.24	13.86 ± 0.80	1.124 ± 0.018
	SFT	Qwen3-4B	thinking	<u>30.38</u> ± 3.53	<u>17.96</u> ± 2.42	<u>0.462</u> ± 0.038
	SFT	Qwen3-14B	thinking	**32.95** ± 1.86	**19.74** ± 1.33	**0.418** ± 0.042
Lap	zs-CoT	Qwen-4B	thinking	9.16 ± 1.12	4.80 ± 0.62	1.209 ± 0.250
	zs-CoT	Qwen-14B	thinking	11.95 ± 1.27	6.36 ± 0.72	0.820 ± 0.153
	Dance	Qwen3-4B	non-thinking	5.05 ± 0.33	2.59 ± 0.17	0.998 ± 0.099
	Dance	Qwen3-14B	non-thinking	11.49 ± 0.96	6.10 ± 0.54	0.842 ± 0.093
	SFT	Qwen3-4B	thinking	<u>13.92</u> ± 0.87	<u>7.48</u> ± 0.50	<u>0.472</u> ± 0.007
	SFT	Qwen3-14B	thinking	**17.19** ± 0.79	**9.40** ± 0.47	**0.452** ± 0.050

Table 5: Results for distillation with reasoning chains. We use boldface to indicate the best performance and underline to indicate the second-best.

	Rest	Hotel	Laptop	Avg
Rater	4.62	4.55	4.25	4.47
IAA	0.65	0.61	0.58	0.61

Table 6: Human evaluation results for data quality.

informal expressions. 3) The Laptop domain has the lowest performance for Category Assignment agent due to its broad and diverse set of categories, which increases task complexity. Specifically, the number of categories is 121 in Laptop compared to 13 in Restaurant and 16 in Hotel.

Distillation with Reasoning Chain. We present the reasoning distillation results with Qwen3-4B and Qwen3-14B in Table 5. 1) Efficient supervised fine-tuning on reasoning chains can effectively distill the knowledge of our DanceHA framework into student models. For instance, SFT Qwen-14B outperforms few-shot CoT GPT-4o across all domains and achieves performance comparable to Dance with Qwen2.5-72B. 2) We configure Dance with Qwen3 models as non-thinking to explore whether multi-agent collaboration in Dance can substitute for inference with explicit reasoning. As noted in the Qwen3 technical report (Yang et al. 2025), enabling thinking mode equips the model with multistep reasoning capabilities for complex tasks. Surprisingly, the results depend on model size. For the Qwen-4B, zeroshot CoT with reasoning outperforms Dance. However, for the larger Qwen-14B, Dance achieves comparable or better performance than zero-shot CoT with reasoning.

Ablation Study

We conduct ablation experiments to evaluate the effectiveness of key modules of Dance. Specifically, we design two experimental conditions: 1) Without Divide-and-Conquer Strategy (w/o D&C)–we create a single agent baseline equipped with a super-prompt, formed by merging the prompts of all expert agents, excluding the aspect-based thought grouping step. Unlike the zero-shot CoT baseline in Table 2, this prompt includes a detailed definition of aspect term. 2) Without Teamwork in the Conquer Phase (w/o Teamwork): we combine prompts from the 3 expert agents in the Conquer phase into a single agent configuration, removing the collaborative setup. GPT-4o is the underlying LLM for all experiments. Table 4 summarizes the results. We observe that the D&C Strategy is critical to Dance's success, removing it results in a significance performance decline: with the average F1-score dropping by 8.37%, accom-

panied by similar decreases in ACOS Accuracy and SIS prediction MAE. As expected, the w/o Teamwork variant outperforms the w/o D&C group, but still experiences a performance drop, with the average F1-score decreasing by 2.91%. This underscores the necessity of decomposing tasks in the Divider phase, enabling expert agents to handle manageable sub-tasks and mitigating information overload.

Human Evaluation for Data Quality

We recruited two annotators to evaluate the quality of the labels generated by DanceHA. Annotators were instructed to rate score on a 1–5 scale (1 = completely disagree, 5 = completely agree) for 90 randomly sampled documents (30 per domain) with a total of 820 ACOSI tuples. Inter-annotator agreement (IAA) is measured using Cohen's Kappa Score (Artstein and Poesio 2008). We achieved an average rating score of 4.47 out of 5 and an IAA score of 0.61. These results indicate DaceHA's effectiveness in balancing fine-grained labeling with high annotation quality.

Conclusion & Future Work

This work presented DanceHA, a multi-agent framework for open-ended, document-level ABSIA with informal styles. It comprises two components: Dance, which follows a divide-and-conquer approach to decompose the complex ABSIA task into manageable sub-tasks handled by expert agents; and HA, Human-AI collaboration to integrate and revise outputs from multiple Dance teams to produce fine-grained, high-accuracy labels. Extensive experiments across three domains and seven LLMs demonstrate that Dance effectively handles complex ABSIA tasks and significantly outperforms few-shot CoT. Moreover, MA achieves even better performance by building upon Dance teams with different backbone models. Our findings reveal the importance of informal styles in ABSIA, which often amplify sentiment intensity for specific aspects and opinions. Our experiments of distillation with reasoning chain show that we can effectively transfer the multi-agents knowledge in DanceHA into student models. Future work may explore reasoning distillation with reinforcement learning paradigms and address the challenges introduced by the complex category options in the laptop domain.

Limitations

We acknowledge several limitations for future exploration. First, the influence of team size on annotation quality remains an open question. Moreover, some sub-tasks might be replaced with small language models such as T5 and RoBERTa to improve efficiency.

Acknowledgments

References

Adamu, M.; Zhang, Q.; Pan, H.; Latecki, L. J.; and Dragut, E. C. 2025. Querying Climate Knowledge: Semantic Retrieval for Scientific Discovery. *arXiv preprint arXiv:2509.10087*.

Aljebreen, A.; Meng, W.; and Dragut, E. 2021. Segmentation of tweets with urls and its applications to sentiment analysis. In *AAAI*, volume 35, 12480–12488.

Artstein, R.; and Poesio, M. 2008. Survey Article: Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4): 555–596.

Bai, Y.; Han, Z.; Zhao, Y.; Gao, H.; Zhang, Z.; Wang, X.; and Hu, M. 2024. Is Compound Aspect-Based Sentiment Analysis Addressed by LLMs? In *Findings of EMNLP*, 7836–7861.

Bertsekas, D.; and Tsitsiklis, J. N. 2008. *Introduction to probability*, volume 1. Athena Scientific.

Bhattacharjee, K.; and Gangadharaiah, R. 2023. Document-Level Supervision for Multi-Aspect Sentiment Analysis Without Fine-grained Labels. arXiv:2310.06940.

Brody, S.; and Diakopoulos, N. 2011. Coooooooooooooooolllllllllllll!!!!!!!!!!!!! Using Word Lengthening to Detect Sentiment in Microblogs. In *EMNLP*, 562–570.

Cai, H.; Xia, R.; and Yu, J. 2021. Aspect-Category-Opinion-Sentiment Quadruple Extraction with Implicit Aspects and Opinions. In *ACL-IJCNLP*, 340–350.

Chen, Z.; Meng, W.; and Dragut, E. 2025. ComCrawler: General Crawling Solution for Article Comments.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948.

Fang, H.; Zhu, X.; and Gurevych, I. 2024. DARA: Decomposition-Alignment-Reasoning Autonomous Language Agent for Question Answering over Knowledge Graphs. arXiv:2406.07080.

Feng, Y.; Htut, P. M.; Qi, Z.; Xiao, W.; Mager, M.; Pappas, N.; Halder, K.; Li, Y.; Benajiba, Y.; and Roth, D. 2025. Rethinking LLM Uncertainty: A Multi-Agent Approach to Estimating Black-Box Model Uncertainty. arXiv:2412.09572.

Gao, T.; Fang, J.; Liu, H.; Liu, Z.; Liu, C.; Liu, P.; Bao, Y.; and Yan, W. 2022. LEGO-ABSA: A Prompt-based Task Assemblable Unified Generative Framework for Multi-task Aspect-based Sentiment Analysis. In *COLING*, 7002–7012.

Gou, Z.; Guo, Q.; and Yang, Y. 2023. MvP: Multi-view Prompting Improves Aspect Sentiment Tuple Prediction. arXiv:2305.12627.

He, L.; Han, C.; Mukherjee, A.; Obradovic, Z.; and Dragut, E. 2020. On the dynamics of user engagement in news comment media. *WIREs Data Mining Knowl Discov*, 10(1): e1342.

Hoang, M.; Bihorac, O. A.; and Rouces, J. 2019. Aspect-Based Sentiment Analysis using BERT. In Hartmann, M.; and Plank, B., eds., *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, 187–196. Turku, Finland: Linköping University Electronic Press.

Hosseinia, M.; Dragut, E.; and Mukherjee, A. 2019. Pro/con: Neural detection of stance in argumentative opinions. In *SBP-BRiMS*, 21–30.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685.

Jin, Y.; Zhao, Q.; Wang, Y.; Chen, H.; Zhu, K.; Xiao, Y.; and Wang, J. 2024. AgentReview: Exploring Peer Review Dynamics with LLM Agents. arXiv:2406.12708.

Jones, H. T.; Maus, N.; Ludan, J. M.; Huan, M. Z.; Liang, J.; Torres, M. D. T.; Liang, J.; Ives, Z.; Barash, Y.; de la Fuente-Nunez, C.; et al. 2025. A Dataset for Distilling Knowledge Priors from Literature for Therapeutic Design. *arXiv preprint arXiv:2508.10899*.

Kim, J.; Heo, R.; Seo, Y.; Kang, S.; Yeo, J.; and Lee, D. 2024. Self-Consistent Reasoning-based Aspect-Sentiment Quad Prediction with Extract-Then-Assign Strategy. In *Findings of ACL*, 7295–7303.

Lan, F.; Aljebreen, A.; and Dragut, E. 2025. UniT: One Document, Many Revisions, Too Many Edit Intention Taxonomies. In *Findings of ACL*, 23005–23024.

Lee, L.-H.; Yu, L.-C.; Wang, S.; and Liao, J. 2024. Overview of the SIGHAN 2024 shared task for Chinese dimensional aspect-based sentiment analysis. In *SIGHAN Workshop*, 165–174.

Li, B.; Luo, F.; Haider, S.; Agashe, A.; Li, T.; Liu, R.; Miao, M.; Ramakrishnan, S.; Yuan, Y.; and Callison-Burch, C. 2025. Multilingual Retrieval Augmented Generation for Culturally-Sensitive Tasks: A Benchmark for Cross-lingual Robustness. arXiv:2410.01171.

Li, J. 2020. Hotel Review Dataset. Access date: 2023-10-11.

Liang, S.; Zhang, B.; Zhao, J.; and Liu, K. 2024a. ABSEval: An Agent-based Framework for Script Evaluation. In *EMNLP*, 12418–12434.

Liang, T.; He, Z.; Jiao, W.; Wang, X.; Wang, Y.; Wang, R.; Yang, Y.; Shi, S.; and Tu, Z. 2024b. Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate. arXiv:2305.19118.

Lin, J.; Diesendruck, M.; Du, L.; and Abraham, R. 2024. BatchPrompt: Accomplish more with less. In *ICLR*.

Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Luo, Y.; Cai, H.; Yang, L.; Qin, Y.; Xia, R.; and Zhang, Y. 2022. Challenges for Open-domain Targeted Sentiment Analysis. arXiv:2204.06893.

Mamta; and Ekbal, A. 2023. Service Is Good, Very Good or Excellent? Towards Aspect Based Sentiment Intensity Analysis. In *ECIR*, 685–700.

Ni, J.; Li, J.; and McAuley, J. 2019. Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects. In *EMNLP-IJCNLP*, 188–197.

Pontiki, M.; Galanis, D.; Papageorgiou, H.; Androutsopoulos, I.; et al. 2016. SemEval-2016 Task 5: Aspect Based Sentiment Analysis. In *SemEval-2016*, 19–30.

Pontiki, M.; Galanis, D.; Pavlopoulos, J.; Papageorgiou, H.; Androutsopoulos, I.; and Manandhar, S. 2014. SemEval-2014 Task 4: Aspect Based Sentiment Analysis. In *SemEval-2014*, 27–35.

Schneider, A.; and Dragut, E. 2015. Towards debugging sentiment lexicons. In *ACL-IJCNLP*, 1024–1034.

Schneider, A.; Male, J.; Bhogadhi, S.; and Dragut, E. 2018. DebugSL: An interactive tool for debugging sentiment lexicons. In *NAACL*, 36–40.

Seo, Y.; Song, S.; Heo, R.; Kim, J.; and Lee, D. 2024. Make Compound Sentences Simple to Analyze: Learning to Split Sentences for Aspect-based Sentiment Analysis. arXiv:2410.02297.

Singh, A.; Blanco, E.; and Jin, W. 2019. Incorporating Emoji Descriptions Improves Tweet Classification. In *NAACL*, 2096–2101.

Šmíd, J.; Priban, P.; and Kral, P. 2024. LLaMA-Based Models for Aspect-Based Sentiment Analysis. In *WASSA*, 63–70.

Song, N.; Cai, H.; Xia, R.; Yu, J.; Wu, Z.; and Dai, X. 2023. A Sequence-to-Structure Approach to Document-level Targeted Sentiment Analysis. In *Findings of EMNLP*, 7687–7698.

Su, G.; Wu, M.; Huang, Z.; Zhang, Y.; Wang, T.; Hu, Y.; and Sha, Y. 2024a. Refine, Align, and Aggregate: Multiview Linguistic Features Enhancement for Aspect Sentiment Triplet Extraction. In *Findings ACL*, 3212–3228.

Su, Y.; Yang, D.; Yao, S.; and Yu, T. 2024b. Language Agents: Foundations, Prospects, and Risks. In *EMNLP: Tutorial Abstracts*, 17–24.

Subramaniam, V.; Du, Y.; Tenenbaum, J. B.; Torralba, A.; Li, S.; and Mordatch, I. 2025. Multiagent Finetuning: Self Improvement with Diverse Reasoning Chains. arXiv:2501.05707.

Triantafyllopoulos, I.; Qu, R.; Giorgi, S.; Curtis, B.; Ungar, L. H.; and Sedoc, J. 2025. Simple Methods Defend RAG Systems Well Against Real-World Attacks. *arXiv preprint arXiv:2508.02296*.

Wang, L.; and Dragut, E. 2024. The Overlooked Repetitive Lengthening Form in Sentiment Analysis. In *Findings of EMNLP*, 16225–16238.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Chi, E. H.; Le, Q.; and Zhou, D. 2022. Chain of Thought Prompting Elicits Reasoning in Large Language Models. *CoRR*, abs/2201.11903.

Xu, H.; Liu, B.; Shu, L.; and Yu, P. S. 2019. BERT Post-Training for Review Reading Comprehension and Aspect-based Sentiment Analysis. arXiv:1904.02232.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; and Chenxu Lv, e. a. 2025. Qwen3 Technical Report. arXiv:2505.09388.

Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Yang, F.; Mukherjee, A.; and Dragut, E. 2017. Satirical news detection and analysis using attention mechanism and linguistic features. *arXiv preprint arXiv:1709.01189*.

Yelp. 2021. Yelp Dataset. Yelp. Retrieved Feb 16, 2021.

Yifei, L. S.; Chang, A.; Malaviya, C.; and Yatskar, M. 2025. Researchqa: Evaluating scholarly question answering at scale across 75 fields with survey-mined questions and rubrics. *arXiv preprint arXiv:2509.00496*.

You, W.; Xue, A.; Havaldar, S.; Rao, D.; Jin, H.; Callison-Burch, C.; and Wong, E. 2025. Probabilistic soundness guarantees in llm reasoning chains. In *EMNLP*, 7517–7536.

Zhang, Q.; Pan, H.; Chen, Z.; Latecki, L. J.; Caragea, C.; and Dragut, E. 2025. DynClean: Training Dynamics-based Label Cleaning for Distantly-Supervised Named Entity Recognition. In *Findings of NAACL*.

Zhang, S.; He, L.; Dragut, E.; and Vucetic, S. 2019. How to invest my time: Lessons from human-in-the-loop entity extraction. In *ACM SIGKDD*, 2305–2313.

Zhang, S.; He, L.; Vucetic, S.; and Dragut, E. 2018. Regular expression guided entity mention mining from noisy web data. In *EMNLP*, 1991–2000.

Zhang, T.; Tham, I.; Hou, Z.; Ren, J.; Zhou, L.; Xu, H.; Zhang, L.; Martin, L. J.; Dror, R.; Li, S.; et al. 2023a. Human-in-the-loop schema induction. In *ACL: Demos*.

Zhang, W.; Deng, Y.; Li, X.; Yuan, Y.; Bing, L.; and Lam, W. 2021. Aspect Sentiment Quad Prediction as Paraphrase Generation. arXiv:2110.00796.

Zhang, W.; Deng, Y.; Liu, B.; Pan, S. J.; and Bing, L. 2023b. Sentiment Analysis in the Era of Large Language Models: A Reality Check. arXiv:2305.15005.

Zhang, X.; Du, C.; Pang, T.; Liu, Q.; Gao, W.; and Lin, M. 2024a. Chain of Preference Optimization: Improving Chain-of-Thought Reasoning in LLMs. arXiv:2406.09136.

Zhang, Y.; Sun, R.; Chen, Y.; Pfister, T.; Zhang, R.; and Arik, S. O. 2024b. Chain of Agents: Large Language Models Collaborating on Long-Context Tasks. In *NeurIPS*.

Zhao, Y.; Zhang, K.; Hu, T.; Wu, S.; Bras, R. L.; Anderson, T.; Bragg, J.; Chang, J. C.; Dodge, J.; Latzke, M.; et al. 2025. SciArena: An Open Evaluation Platform for Foundation Models in Scientific Literature Tasks. *arXiv preprint arXiv:2507.01001*.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; Zhang, H.; Gonzalez, J. E.; and Stoica, I. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *NeurIPS*, volume 36, 46595–46623.

Zhu, A.; and Callison-Burch, C. 2025. Overhearing LLM Agents: A Survey, Taxonomy, and Roadmap. *arXiv preprint arXiv:2509.16325*.

Zhu, A.; Martin, L.; Head, A.; and Callison-Burch, C. 2023. CALYPSO: LLMs as Dungeon Master's Assistants. In *AAAI*, volume 19, 380–390.