Journal of Computing and Security

Research Article, 2026, Volume 13, Number 1 (pp. 83--97)

http://www.jcomsec.org

Aspect-Based Sentiment Analysis in Persian Using a Fine-Tuned LLaMA3 Model

Niloofar Ranjbar [a,*] Hamed Baghbani [a]

[a] *Department of Computer Engineering, Jam Faculty of Engineering, Persian Gulf University, Bushehr, Iran.*

ARTICLE INFO.

Article history:
Received: 12 July 2025
Revised: 28 June 2026
Accepted: 22 July 2026
Published Online: 28 July 2026

Keywords:
Aspect-Based Sentiment Analysis, Persian Language, Sentiment Analysis, LLaMA3 Model, ChatGPT, Natural Language Processing.

ABSTRACT

Aspect-Based Sentiment Analysis (ABSA) is less developed for Persian than for English, particularly for extracting complete aspect–category–opinion–sentiment structures rather than classifying sentiment for predefined aspects. This paper presents a generative Persian ABSA framework based on fine-tuned LLaMA3-family instruction models. Given a Persian review, the model generates a JSON array of quadruples containing aspect terms, categories, opinion spans, and sentiment labels. The training data are drawn from a GPT-3.5-assisted Persian ABSA corpus whose annotations were reviewed and filtered by human annotators; we therefore treat it as GPT-assisted/human-validated rather than purely human-authored. To reduce circular evaluation, we also construct an independent human-only test set from raw Persian reviews. Three annotators manually annotated 240 reviews from scratch, and a two-of-three consensus procedure produced 555 gold aspect-level labels from 239 reviews. We evaluate zero-shot LLaMA3, fine-tuned LLaMA3 models, and Dorna-LLaMA3-8B-Instruct as a Persian-oriented generative baseline under the same schema, parser, and metrics. On the human-only test set, zero-shot LLaMA3 obtains 0.1426 aspect F1 and 0.1185 aspect–sentiment F1. Fine-tuned LLaMA3-CG improves the five-seed mean scores to 0.5477±0.0229 and 0.4394±0.0212, while Dorna-CG reaches 0.5809±0.0123 and 0.4897±0.0194. Adding ParsiNLU substantially improves in-domain ParsiNLU performance but reduces human-only performance, revealing a domain and annotation-style trade-off.

1 Introduction

Social networks, e-commerce platforms, and online review websites have become major sources of user feedback. Consumers frequently express opinions about products, services, applications, movies, and public issues. For decision makers, the most useful information is often not the overall polarity of a review, but the sentiment attached to specific aspects such as price, delivery, design, quality, taste, story, or service.

The volume of this user-generated content makes manual analysis impractical. Sentiment analysis provides computational methods for identifying positive, negative, or neutral attitudes in text. Although many studies have addressed document-level or sentence-

* Corresponding author.
Email addresses: nranjbar@pgu.ac.ir (N. Ranjbar), baghbani.hamed@gmail.com (H. Baghbani)
https://dx.doi.org/10.22108/jcs.2026.145865.1175
ISSN: 2322-4460

level sentiment classification in English [1–3] and Persian [4–6], aspect-level analysis remains more challenging and less developed for Persian.

Aspect-Based Sentiment Analysis (ABSA) aims to identify opinion targets and determine the sentiment expressed toward each target [**?**]. For example, in a product review, the same sentence may contain a positive opinion about screen quality and a negative opinion about battery life. Existing Persian ABSA studies often formulate the task as classification over predefined aspect categories or known aspect terms [7–9]. This formulation is useful when the aspect inventory is fixed, but less suitable for open-domain reviews, where users mention new, implicit, or domain-specific aspects.

This paper addresses this limitation by formulating Persian ABSA as a generative review-to-quadruple task. Given a Persian review, the model outputs a JSON array in which each item contains four fields: `aspect`, `category`, `opinion`, and `sentiment`. This design allows the model to extract aspect terms directly from text, associate them with opinion spans, assign categories when available, and predict sentiment labels without requiring a predefined aspect list at inference time.

A central challenge is data construction. We use a Persian ABSA corpus that was initially pre-annotated by GPT-3.5 and subsequently reviewed by three Persian annotators. To avoid overstating the annotation source, we refer to this resource throughout the paper as a *GPT-assisted/human-validated* corpus, not as a purely human-authored dataset. Since evaluating primarily on data that follows the same GPT-assisted annotation style could inflate performance, we also introduce an independent human-only test set. This test set was annotated manually from scratch by three Persian annotators without model-generated pre-annotations and is used only for evaluation.

We explicitly position this work as an applied and resource-oriented contribution rather than as a new transformer architecture. The innovation lies in combining a generative Persian ABSA formulation, richer aspect–category–opinion–sentiment labels, human validation of GPT-assisted annotations, independent human-only evaluation, leakage-aware test construction, and comparison between general and Persian-oriented LLaMA3-family instruction models. In addition to a general LLaMA3-family 8B instruction model [10], we evaluate Dorna-LLaMA3-8B-Instruct [11], a Persian-oriented LLaMA3-based instruction model. Recent Persian LLM evaluation work has also included Dorna-family models among evaluated Persian-oriented models [12], making Dorna a relevant generative baseline for Persian ABSA.

The primary contributions of this paper are as follows:

(1) We introduce a generative Persian ABSA framework that jointly produces aspect terms, aspect categories, opinion spans, and sentiment labels as structured quadruples.
(2) We construct and evaluate a GPT-assisted/human-validated Persian ABSA corpus with richer labels than existing Persian classification-style ABSA resources.
(3) We add an independent human-only Persian ABSA test set consisting of 240 manually annotated raw reviews, yielding 555 consensus aspect-level labels from 239 reviews.
(4) We provide leakage-aware evaluation across GPT-assisted/human-validated, Pars-ABSA, ParsiNLU, and human-only test sets.
(5) We compare zero-shot and fine-tuned general LLaMA3 models with Dorna-LLaMA3-8B-Instruct as a Persian-oriented generative baseline under the same ABSA output schema, parser, and evaluation metrics.
(6) We analyze domain shift, annotation-style effects, and recurrent error types such as over-generation, generic categorization, ambiguous opinions, and sarcasm.

Throughout this paper, we use the following abbreviations: Aspect-Based Sentiment Analysis (ABSA), large language model (LLM), parameter-efficient fine-tuning (PEFT), Low-Rank Adaptation (LoRA), Quantized LoRA (QLoRA), JavaScript Object Notation (JSON), and graphics processing unit (GPU).

2 Related Works

To provide a comprehensive background for our study on ABSA, we review related work in English and Persian ABSA, as well as research on fine-tuning large language models for downstream natural language processing tasks.

2.1 Aspect-Based Sentiment Analysis

Aspect-based sentiment analysis has advanced considerably in high-resource languages. An et al. [13] proposed a heterogeneous graph neural network for ABSA by constructing graphs in which nodes represent aspects, sentiments, and contextual words. Chen et al. [14] introduced adversarial training for a multi-channel graph convolutional network for aspect-sentiment triplet extraction, thereby improving robustness under noisy conditions.

More recently, Mohammadkhani et al. [15] proposed E2TP, a two-step prompting method that predicts individual ABSA elements and combines them into tuples using a T5 model. This line of work shows the potential of text-to-text and generative formulations for ABSA. In parallel, vsmid et al. [16] investigated LLaMA-based models for English ABSA benchmarks and showed that fine-tuned LLaMA-style systems can perform competitively on ABSA tasks, while zero-shot and few-shot settings remain weaker than task-specific fine-tuning.

For Persian ABSA, Jafarian et al. [7] used BERT-style contextual representations, while Vazan and Razmara [17] modeled Persian movie-review aspect category detection and polarity as a multi-label classification problem over predefined aspect categories. Shangipour ataei et al. [8] introduced Pars-ABSA, a benchmark dataset for Farsi product reviews. Ariai et al. [9] used ParsBERT [18] for Persian ABSA. These studies are important for Persian sentiment analysis, but they generally assume a fixed aspect inventory or classify known aspect categories rather than generating aspect terms and opinion spans directly from raw review text.

In contrast, the present work uses LLaMA3-family instruction models for a generative Persian ABSA task. The model is trained to produce aspect terms, categories, opinion expressions, and sentiment labels directly from the input review. To the best of our knowledge, this combination of generative quadruple extraction, GPT-assisted/human-validated Persian ABSA data, human-only evaluation, and Persian-oriented LLaMA3-family comparison has not previously been reported for Persian.

2.2 Fine-Tuning LLMs for Downstream Tasks

Large language models such as LLaMA2 and LLaMA3 have shown strong adaptability across downstream tasks. Efficient fine-tuning frameworks such as LlamaFactory [19] combine quantization, adapter-based training, and PEFT techniques such as LoRA and QLoRA. These methods make it possible to adapt large models to limited hardware. MeteoRA [20] further explores mixture-of-experts-style adapter selection.

Fine-tuning LLaMA-family models has also been explored for sentiment and ABSA tasks. For example, LLaMA2 fine-tuning projects for sentiment analysis [21, 22] show the practical value of adapting instruction models for classification and extraction tasks. SetFitABSA [23] demonstrates that careful few-shot and prompt-based methods can remain competitive even with smaller models. More recent Persian-oriented LLM resources and evaluation efforts, including Dorna-LLaMA3-8B-Instruct [11] and MELAC [12], motivate evaluating whether Persian-oriented instruction models provide additional gains over general LLaMA3-family models [10]. However, most existing LLM-based ABSA studies focus on English or other higher-resource languages. The present paper extends this direction to Persian and evaluates the resulting systems under an explicit human-only test protocol.

3 Data Collection and Annotation

For constructing the GPT-assisted/human-validated corpus, we began with reviews derived from the Pars-ABSA data source. To reduce annotation noise caused by very long inputs, we filtered out reviews longer than 30 words. This reduced the source pool from 5,602 unique reviews to 2,050 candidate reviews.

The candidate reviews were pre-annotated using GPT-3.5 through the API. The prompt asked the model to identify all ABSA components for each review: aspect term, aspect category, short opinion phrase, and sentiment label (positive, neutral, or negative), and to return only a JSON-formatted output. Table 1 presents examples of the resulting ABSA components.

Because these annotations were generated by an LLM, they were not treated as gold labels without verification. Three native Persian annotators independently reviewed the generated components. Each component was scored using a 0–2 scale:

- **0**: the component is incorrect;
- **1**: the component is partially correct;
- **2**: the component is correct.

The scoring scheme was applied to aspect terms, categories, opinion spans, and sentiment labels. For training, we retained aspect-level items that received score 2 from at least two annotators. For the GPT-assisted/human-validated test split, we used the stricter validated subset that satisfied the agreement filters. After filtering, this corpus contained 1,513 unique training reviews and 321 unique GPT-assisted/human-validated test reviews before leakage filtering.

Annotation Guidelines. Annotators received written guidelines specifying how to identify aspect terms, map aspects to categories, select minimal opinion spans, and assign positive, negative, or neutral sentiment labels. The guidelines included examples of multi-aspect reviews, ambiguous cases, implicit opinions, and comments about product photos or website information. These cases were included because Per-

Table 1. Sample outputs of aspect-based sentiment analysis components generated during GPT-assisted pre-annotation and translated to English.

Text	Aspect	Category	Opinion	Sentiment
بهترین انتخاب، کیفیت بسیار بالا و مقرون‌به‌صرفه	کیفیت	کیفیت	بسیار بالا	positive
	قیمت	قیمت	مقرون‌به‌صرفه	positive
Best choice, very high quality, and affordable	*Quality*	*Quality*	*Very high*	
	Price	*Price*	*Affordable*	
کیفیت خوبی داره؛ تو شگفت‌انگیز ارزش داره خریدش	کیفیت	کیفیت	خوب	positive
It has good quality; it is worth buying during the Amazing Offer.	*Quality*	*Quality*	*Good*	
کپی بسیار ناقص از جنس مرغوب؛ متریال ضعیف؛ به هیچ وجه نخرید	مدل	کیفیت	ضعیف	negative
A very poor copy of good material; weak, do not buy at all	*Model*	*Quality*	*Weak*	

sian product reviews often mix product-level opinions with contextual or platform-level comments.

Inter-annotator Agreement. The original GPT-assisted corpus was filtered using annotator agreement rather than accepted as fully reliable automatic annotation. In addition, to address the risk of circular evaluation, we created a separate independent human-only test set. This set consists of 240 raw Persian reviews that were not used for training. The reviews were manually annotated from scratch by three Persian annotators without ChatGPT, machine translation, automatic labelers, or model-generated pre-annotation. We then grouped aspect mentions using normalized Persian text and fuzzy aspect matching. An aspect was retained when at least two of the three annotators supported it. This produced 555 consensus aspect-level rows from 239 reviews with at least one consensus aspect; one review had no retained consensus aspect. The human-only set is used only for evaluation and is never mixed into the training data.

Important terminology. Throughout the rest of this paper, we call the main corpus the *GPT-assisted/human-validated* corpus. This terminology is important: GPT-3.5 generated the initial pre-annotations, and humans validated or filtered them. Therefore, this corpus is not claimed to be a purely human-authored annotation set.

Figure 1 illustrates an example review and compares the aspect-based sentiment analysis performed by Pars-ABSA and our richer representation. Pars-ABSA provides a more limited aspect-sentiment structure, whereas our annotation format captures aspect terms, categories, opinions, and sentiments.

Dataset and Code Availability. The preprocessing, training, structured-output parsing, and evaluation code are available in the project repository at https://github.com/nranjbar/persian_absa. Third-party datasets and individual annotation records are not redistributed in the repository and remain subject to their original licenses, consent conditions, and data-governance requirements.

4 Methodology

4.1 Preparing Data for Fine-Tuning LLaMA3

We formulate Persian ABSA as a conditional generation task. Given a Persian review x, the model generates a JSON array y in which each item has exactly four keys: `aspect`, `category`, `opinion`, and `sentiment`. The sentiment value is restricted to `positive`, `negative`, or `neutral`. If a category or opinion is unavailable in a source dataset such as ParsiNLU, the corresponding field is represented as `null`. This unified output format allows the same model and parser to be applied across datasets with different annotation completeness.

The GPT-assisted/human-validated training set contains 1,513 unique reviews. We also use the ParsiNLU ABSA data [24] as an optional additional training source. ParsiNLU provides aspect and sentiment labels but does not consistently provide opinion spans or the same category inventory as our corpus. When ParsiNLU is used, missing fields are therefore serialized as `null`. Combining GPT-assisted/human-validated training data with ParsiNLU yields 3,006 training reviews.

We evaluate the following fine-tuned settings:

- **LLaMA3-CG**: a general LLaMA3-family instruction model fine-tuned only on the 1,513-review GPT-assisted/human-validated training set.
- **LLaMA3-CG+ParsiNLU**: the same model fine-tuned on the GPT-assisted/human-validated training set combined with ParsiNLU.

Review Text:

از لحاظ تنظیم زاویه نور بسیار عملکرد خوبی داره ولی پایه چراغ در پایه‌اش ثابت نیست و قفل نمیشه و در می‌آید

Pars-ABSA Dataset:
Aspect Term: پایه چراغ; ***Sentiment:*** Negative

Our Dataset:
Aspect Term: تنظیم زاویه نور; ***Category:*** عملکرد; ***Opinion:*** خوبی داره; ***Sentiment:*** Positive.

Aspect Term: پایه چراغ; ***Category:*** کیفیت ; ***Opinion:*** نامطلوب; ***Sentiment:*** Negative

Figure 1. Detailed ABSA analysis of a Persian review using Pars-ABSA and our richer aspect–category–opinion–sentiment representation.

Instruction: *Analyze the following review to identify ALL aspects mentioned. For each aspect, determine the category it belongs to, the specific opinion expressed, and the sentiment. Provide the results in a structured format listing each aspect, its category, opinion, and sentiment. Find ALL aspects.*

Input: مودم خوبیه ولی حجم رایگان نداره

Response:
Aspect: مودم, ***Category:*** عمومی, ***Opinion:*** خوبیه, ***Sentiment:*** positive;

Aspect: حجم رایگان, ***Category:*** هزینه, ***Opinion:*** نداره, ***Sentiment:*** negative; <***end of text***>

Figure 2. Example of an instruction-style training entry. The input contains the task instruction and Persian review; the target is a JSON array of aspect–category–opinion–sentiment quadruples.

- **Dorna-CG**: Dorna-LLaMA3-8B-Instruct [11], a Persian-oriented LLaMA3-family instruction model, fine-tuned on the GPT-assisted/human-validated training set.
- **Dorna-CG+ParsiNLU**: Dorna-LLaMA3-8B-Instruct fine-tuned on the combined training set.

We also evaluate a zero-shot LLaMA3 baseline using the same instruction prompt but no task-specific fine-tuning.

Before evaluation, exact train–test text overlaps are removed for the fine-tuned models. In the final rerun, overlap filtering removed 4 reviews from the GPT-assisted/human-validated test set and 406 reviews from Pars-ABSA. No overlap was found for ParsiNLU food, ParsiNLU movie, or the independent human-only test set.

4.2 Fine-Tuning LLaMA3

One of the main advantages of the proposed approach is the ability of instruction-tuned LLaMA3-family models to generate aspects directly from the input review. Unlike traditional Persian encoder-based systems such as ParsBERT, which generally rely on predefined aspect categories or known aspect terms, the generative model can produce aspect terms, categories, opinion spans, and sentiments in a single structured output.

Figure 3 provides a high-level overview of the proposed architecture. A Persian review is converted into an instruction-style prompt and tokenized. During training, the model is optimized with teacher forcing and next-token prediction to generate the target JSON array. During inference, the model autoregressively generates the ABSA quadruples directly from the review.

As presented in Table 2, we use 4-bit QLoRA to reduce memory requirements. The base 8B instruction model is loaded in 4-bit quantized form, and only LoRA adapter parameters are updated while the remaining base-model parameters remain frozen. The LoRA rank is $r = 16$, the scaling factor is $\alpha = 32$, and the dropout rate is 0.05. Adapters are applied to the attention projection matrices (`q_proj`, `k_proj`, `v_proj`, `o_proj`) and the MLP projection matrices (`gate_proj`, `up_proj`, `down_proj`) in every transformer block. This updates approximately 41.9 million adapter parameters.

We train for one epoch with a per-device batch size of 2 and gradient accumulation over 4 steps, corresponding to an effective batch size of 8. With 1,513 training reviews, this gives approximately 190 optimization steps; with 3,006 training reviews, this gives approximately 376 optimization steps. We use AdamW with 8-bit optimizer states, a learning rate of 1×10^{-3}, weight decay of 0.01, and linear warmup over the first 10% of training steps followed by linear decay. The maximum prompt length is 256 tokens, the maximum training sequence length is 512 tokens, and at inference time the model generates up to 128

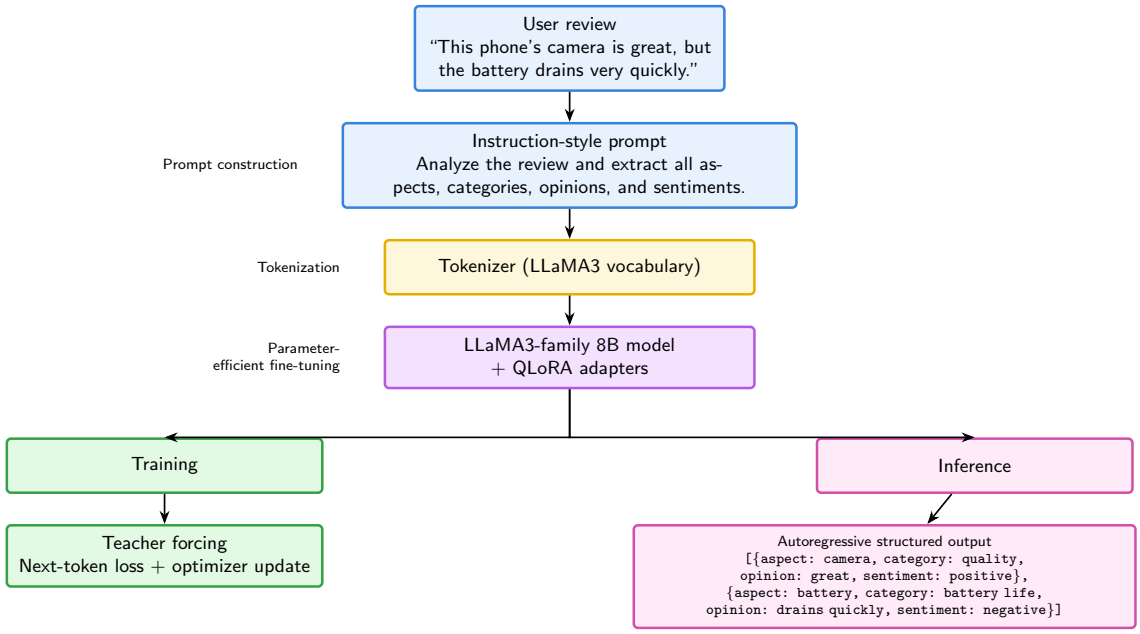

Figure 3. Overview of the proposed LLaMA3-based Persian ABSA architecture. A user review is converted into an instruction-style prompt, tokenized, and fed to a fine-tuned LLaMA3-family 8B instruction model with QLoRA adapters. During training, the model is optimized with teacher forcing and next-token loss; at inference time, it autoregressively generates aspect–category–opinion–sentiment quadruples.

new tokens per review using greedy decoding.

The final reviewer-response rerun was conducted on a single NVIDIA RTX 3090 GPU. Generation is the main computational bottleneck because each output is produced autoregressively.

4.3 Aspect-Based Sentiment Analysis (ABSA) Approach

Aspect-Based Sentiment Analysis identifies specific features or targets in a text and determines the sentiment associated with each one. Unlike traditional sentiment analysis, which classifies an entire review as positive, negative, or neutral, ABSA provides a more detailed representation by linking individual aspects to opinions and sentiments.

In this work, each predicted item contains four fields. The `aspect` field identifies the target expression in the review. The `category` field assigns a broader semantic category when available. The `opinion` field records a short opinion span. The `sentiment` field is one of `positive`, `negative`, or `neutral`. The model is instructed to return only a JSON array so that predictions can be parsed and evaluated automatically.

4.3.1 Aspect Extraction, Categorization, and Challenges in Persian Text

Persian ABSA is challenging because of orthographic variation, clitics, compound verbs, informal spelling, and frequent omission of explicit opinion targets. Before annotation aggregation, leakage checking, training, and fuzzy evaluation, Persian text is normalized to reduce superficial variation. Arabic characters are mapped to Persian equivalents, Arabic diacritics and elongation marks are removed, punctuation is standardized or removed for matching, and repeated whitespace is collapsed. Reviews are otherwise preserved; we do not stem, lemmatize, translate, or remove content words, because these operations can remove valid aspect terms or opinion expressions.

Traditional models such as ParsBERT generally assume fixed aspect labels, whereas the proposed generative formulation can identify aspects directly from the review. This flexibility is important for Persian reviews, where users may mention domain-specific or colloquial aspects do not present in a fixed inventory. However, the model can still struggle with implicit aspects, sarcasm, and ambiguous opinion expressions. These limitations are analyzed in Section 5.5.

To evaluate the proposed model, we use five distinct test sets that vary in domain, average review length, and sentiment distribution. Table 3 summarizes the key characteristics of these datasets, highlighting the differences between the newly introduced

Table 2. Main fine-tuning and inference hyperparameters used in the final rerun.

Component	**Setting**
Base models	General LLaMA3-family 8B instruction model; Dorna-LLaMA3-8B-Instruct
Quantization	4-bit QLoRA
Trainable modules	LoRA adapters
LoRA rank / alpha / dropout	16 / 32 / 0.05
Trainable parameters	approximately 41.9M adapter parameters
Frozen parameters	base-model parameters frozen
Optimizer	AdamW with 8-bit optimizer states
Learning rate / weight decay	1×10^{-3} / 0.01
Warmup	10% linear warmup followed by linear decay
Epochs	1
Batch size	2 per device, gradient accumulation 4, effective batch size 8
Maximum prompt / sequence / generation length	256 / 512 / 128 tokens
Decoding	greedy decoding, no sampling
Hardware	single NVIDIA RTX 3090 GPU

GPT-assisted/human-validated and human-only sets versus the existing Pars-ABSA and ParsiNLU benchmarks.

5 Evaluation and Results

5.1 Evaluation Metrics

We report two main metrics. The first is aspect extraction F1, computed after matching predicted aspects to gold aspects. Because Persian aspect spans may differ slightly in spelling, spacing, or inflection, we normalize Persian text and use fuzzy matching based on string similarity. A predicted aspect is counted as matching a gold aspect when its normalized similarity score is at least 50.

The second main metric is aspect–sentiment F1 (AS-F1), which requires both a matched aspect and the correct sentiment label. We also report matched sentiment accuracy, defined as sentiment accuracy over matched aspect pairs. Category and opinion fields are generated by the model, but they are not equally available across all datasets. For example, Pars-ABSA and ParsiNLU do not provide the same category and opinion-span annotations as our GPT-assisted/human-validated and human-only data. Therefore, category and opinion scores are used only as diagnostic analyses and are not treated as the main cross-dataset metrics.

For LLaMA3-CG and LLaMA3-CG+ParsiNLU, we report mean±standard deviation over five independent seeds. For Dorna-CG and Dorna-CG+ParsiNLU, we report mean±standard deviation over three independent seeds. Zero-shot LLaMA3 is reported as a single run. All scores are rounded to four decimal places to avoid the inconsistent reporting criticized in the previous version.

5.2 Datasets

The training data used for fine-tuning is summarized in Table 4. This includes the GPT-assisted/human-validated training split and the ParsiNLU training split. For evaluation, we use five test sets. The GPT-assisted/human-validated test set measures performance on the same annotation style as the validated training data, but after exact leakage removal. Pars-ABSA and ParsiNLU provide external benchmark comparisons. The independent human-only test set is the most important evaluation set for assessing whether the fine-tuned models generalize beyond GPT-assisted labels.

5.3 Baseline Models

The main controlled baseline is zero-shot LLaMA3, which receives the same instruction prompt as the fine-tuned models but is not trained on Persian ABSA data. This directly tests whether the observed performance comes from task-specific fine-tuning rather than from generic instruction-following ability.

We also compare with Dorna-LLaMA3-8B-Instruct [11], a Persian-oriented LLaMA3-family instruction model, under the same generative output schema. Dorna is included because the reviewer re-

Table 3. Summary of ABSA test datasets. Pars-ABSA and ParsiNLU are existing benchmarks; the GPT-assisted/human-validated test set and the human-only test set are used to evaluate the proposed generative schema.

Dataset	Unique reviews	Avg aspects/review	Avg words/review	Positive	Negative	Neutral
GPT-assisted/human-validated test	321	1.88	18.00	59.60%	24.50%	15.89%
Pars-ABSA [8]	1,719	1.16	65.59	51.08%	30.62%	18.31%
ParsiNLU food [24]	186	2.01	18.80	50.00%	47.86%	2.14%
ParsiNLU movie [24]	96	2.02	31.33	52.06%	45.36%	2.58%
Human-only test	239	2.32	26.81	66.13%	29.37%	4.50%

Table 4. Training datasets used in the fine-tuning experiments.

Training set	Unique reviews	Avg aspects/review	Avg words/review	Positive	Negative	Neutral
GPT-assisted/human-validated	1,513	2.13	18.01	59.63%	23.95%	16.41%
ParsiNLU train [24]	1,493	2.51	18.73	52.71%	45.85%	1.44%

Table 5. Feature comparison between classification-based Persian ABSA models and the proposed generative LLaMA3-family models.

Feature	ParsBERT-style classifier	Fine-tuned LLaMA3-family model
Requires predefined aspect terms	Yes	No
Generates new aspect terms	No	Yes
Generates opinion spans	No	Yes
Generates categories	Usually fixed-label only	Yes
End-to-end review-to-quadruple output	No	Yes

quested comparison with generative models beyond the originally fine-tuned LLaMA3, and because Persian LLM evaluation work has treated Dorna-family models as relevant Persian-oriented baselines [12]. A multilingual mT5-base run was also attempted under the same strict JSON parser, but in the final rerun it produced zero parseable predicted aspects on the evaluated splits. Since this does not constitute a meaningful generative ABSA comparison, mT5 is not included in the main result tables. Instead, we focus on Dorna as the additional generative baseline.

ParsBERT, TD-LSTM, and other classification-based models are discussed as related Persian ABSA baselines. However, these systems generally assume predefined aspects or perform classification rather than open-ended generation of aspect–category–opinion–sentiment quadruples. Therefore, direct raw-score comparison with them is not fully fair unless gold or candidate aspects are supplied. Table 5 summarizes this functional difference.

5.4 Results

5.4.1 Aspect Finding

Table 6 reports aspect extraction precision, recall, and F1. The strongest evidence against circular evaluation comes from the independent human-only test set. Zero-shot LLaMA3 obtains only 0.1426 aspect F1 on this set. Fine-tuning the general LLaMA3-family model on the GPT-assisted/human-validated corpus improves the five-seed mean aspect F1 to 0.5477±0.0229. The Persian-oriented Dorna-CG model obtains a stronger three-seed mean of 0.5809±0.0123 on the same human-only benchmark. This indicates that the model is not merely learning to reproduce GPT-assisted test labels; task-specific fine-tuning transfers to a manually annotated human-only evaluation set.

Adding ParsiNLU produces a clear domain-specific effect. It strongly improves ParsiNLU food and movie performance, especially for Dorna-CG+ParsiNLU on food reviews, but it reduces performance on the human-only test set compared with training only on the GPT-assisted/human-validated corpus. This shows that additional ABSA supervision is beneficial

when its annotation scheme and domain match the target test set, but it can shift the model away from the human-only annotation policy.

5.4.2 Sentiment Analysis of Aspects

Table 7 reports aspect–sentiment F1 and matched sentiment accuracy. The zero-shot model performs poorly on the human-only test set, with aspect–sentiment F1 of 0.1185. LLaMA3-CG improves this to 0.4394±0.0212, and Dorna-CG further improves it to 0.4897±0.0194. This result supports the use of a Persian-oriented LLaMA3-family model as a meaningful generative baseline and shows that the proposed training procedure generalizes beyond GPT-assisted labels.

The ParsiNLU results again show a domain-specific pattern. LLaMA3-CG+ParsiNLU obtains 0.8075±0.0147 aspect–sentiment F1 on ParsiNLU food, while Dorna-CG+ParsiNLU obtains 0.8695±0.0067. On ParsiNLU movie, LLaMA3-CG+ParsiNLU remains stronger than Dorna-CG+ParsiNLU in aspect extraction and aspect–sentiment F1, suggesting that the advantage of a Persian-oriented instruction model is not uniform across all domains. Overall, adding ParsiNLU improves in-domain ParsiNLU performance but reduces human-only performance relative to GPT-assisted-only training.

5.4.3 Performance of the GPT-Assisted/Human-Validated Dataset

The GPT-assisted/human-validated test set remains useful because it measures how well the model follows the richer annotation schema used during training. However, it is not the only evidence for generalization. The revised evaluation therefore gives special importance to the independent human-only test set. On the GPT-assisted test set, LLaMA3-CG obtains 0.6875±0.0078 aspect F1, while Dorna-CG obtains 0.7051±0.0087. On the human-only test set, the corresponding values are lower, but still much higher than zero-shot performance. This gap shows both the value and the limitation of GPT-assisted/human-validated training: it teaches the model the task format and improves extraction, but generalization is still affected by annotation style and domain shift.

The Pars-ABSA scores are lower than the GPT-assisted and ParsiNLU scores. This is expected because Pars-ABSA reviews are longer on average, exact overlap removal substantially reduces the fine-tuned evaluation subset, and the aspect definitions differ from the generative annotation scheme used in our training data. We therefore interpret the lower Pars-ABSA results as a genuine cross-dataset generalization challenge rather than as a simple failure of sentiment classification.

5.5 Error Analysis

Manual inspection reveals several recurrent error patterns. The first is over-generation: the model sometimes extracts near-duplicate aspects such as repeated “quality” or “appearance” entries with slightly different opinions. The second is generic categorization: the model may assign a broad category such as “quality” when a more specific category such as design, brand, size, or delivery would be preferable. The third is context confusion: comments about product photos, the website interface, or seller behavior can be incorrectly interpreted as product aspects. The fourth is implicit sentiment: when the opinion is indirect, sarcastic, or figurative, the model may assign the literal sentiment instead of the intended sentiment.

These errors are more frequent in the human-only and movie-review settings than in the ParsiNLU food setting. Food reviews tend to contain concrete aspects such as taste, packaging, portion size, or delivery, and their sentiment cues are often explicit. Movie reviews more often involve abstract aspects such as story, acting, emotional impact, or atmosphere, which makes aspect boundaries and sentiment anchoring harder.

The independent human-only test set also exposes the limitation of relying on GPT-assisted training data. Although fine-tuning improves performance substantially compared with zero-shot prompting, the model still misses some human-annotated aspects and sometimes predicts aspects that are plausible but not included in the consensus gold set. This confirms that GPT-assisted/human-validated data are useful for training, but they do not eliminate the need for independent human-only evaluation.

Effect of pretraining and language coverage. Several observed errors are consistent with the limited representation of Persian in many general-purpose LLM pretraining corpora. General LLaMA3 models can misinterpret colloquial Persian expressions, compound verbs, or sarcasm. Dorna-LLaMA3-8B-Instruct partly addresses this limitation by using a Persian-oriented instruction model, and its three-seed human-only results are stronger than the corresponding general LLaMA3-CG results. However, the gains are not uniform across all domains: the general LLaMA3-CG+ParsiNLU model remains stronger on ParsiNLU movie, while Dorna-CG+ParsiNLU is strongest on ParsiNLU food.

Why is performance on Pars-ABSA lower? The model performs worst on Pars-ABSA. Several factors

Table 6. Aspect extraction results. LLaMA3-CG and LLaMA3-CG+ParsiNLU are reported as mean±standard deviation over five seeds. Dorna-CG and Dorna-CG+ParsiNLU are reported as mean±standard deviation over three seeds. Zero-shot LLaMA3 is a single-run baseline.

Model	Test set	Reviews	Precision	Recall	Aspect F1
Zero-shot LLaMA3	GPT-assisted test	321	0.2624	0.2533	0.2578
Zero-shot LLaMA3	Pars-ABSA	1,719	0.0698	0.1416	0.0935
Zero-shot LLaMA3	ParsiNLU food	186	0.1158	0.0882	0.1002
Zero-shot LLaMA3	ParsiNLU movie	96	0.0347	0.0361	0.0354
Zero-shot LLaMA3	Human-only	239	0.1610	0.1279	0.1426
LLaMA3-CG	GPT-assisted test	317	0.6575 ± 0.0140	0.7208 ± 0.0173	0.6875 ± 0.0078
LLaMA3-CG	Pars-ABSA	1,313	0.2393 ± 0.0075	0.5321 ± 0.0254	**0.3297 ± 0.0020**
LLaMA3-CG	ParsiNLU food	186	0.2832 ± 0.0318	0.2508 ± 0.0230	0.2656 ± 0.0242
LLaMA3-CG	ParsiNLU movie	96	0.2882 ± 0.0219	0.3196 ± 0.0382	0.3024 ± 0.0265
LLaMA3-CG	Human-only	239	0.5993 ± 0.0091	0.5049 ± 0.0340	0.5477 ± 0.0229
LLaMA3-CG+ParsiNLU	GPT-assisted test	317	0.6228 ± 0.0118	0.6930 ± 0.0311	0.6556 ± 0.0112
LLaMA3-CG+ParsiNLU	Pars-ABSA	1,313	0.2346 ± 0.0128	0.5027 ± 0.0252	0.3194 ± 0.0097
LLaMA3-CG+ParsiNLU	ParsiNLU food	186	0.8821 ± 0.0192	0.8503 ± 0.0145	0.8658 ± 0.0138
LLaMA3-CG+ParsiNLU	ParsiNLU movie	96	**0.6945 ± 0.0847**	**0.5845 ± 0.0284**	**0.6319 ± 0.0297**
LLaMA3-CG+ParsiNLU	Human-only	239	0.5353 ± 0.0268	0.4436 ± 0.0325	0.4845 ± 0.0231
Dorna-CG	GPT-assisted test	317	**0.6685 ± 0.0191**	**0.7472 ± 0.0341**	**0.7051 ± 0.0087**
Dorna-CG	Pars-ABSA	1,313	0.2326 ± 0.0015	0.5411 ± 0.0268	0.3252 ± 0.0051
Dorna-CG	ParsiNLU food	186	0.2956 ± 0.0234	0.2594 ± 0.0117	0.2758 ± 0.0110
Dorna-CG	ParsiNLU movie	96	0.2769 ± 0.0227	0.3196 ± 0.0089	0.2965 ± 0.0155
Dorna-CG	Human-only	239	**0.6138 ± 0.0103**	**0.5520 ± 0.0263**	**0.5809 ± 0.0123**
Dorna-CG+ParsiNLU	GPT-assisted test	317	0.6461 ± 0.0099	0.7371 ± 0.0301	0.6883 ± 0.0092
Dorna-CG+ParsiNLU	Pars-ABSA	1,313	0.2424 ± 0.0056	0.4854 ± 0.0436	0.3227 ± 0.0069
Dorna-CG+ParsiNLU	ParsiNLU food	186	**0.8897 ± 0.0106**	**0.8975 ± 0.0111**	**0.8935 ± 0.0009**
Dorna-CG+ParsiNLU	ParsiNLU movie	96	0.5980 ± 0.0478	0.5515 ± 0.0052	0.5733 ± 0.0246
Dorna-CG+ParsiNLU	Human-only	239	0.5778 ± 0.0106	0.5231 ± 0.0436	0.5486 ± 0.0271

(a) Mobile Phone Aspects.

(b) Mobile Phone Categories.

Figure 4. Word clouds of mobile-phone aspects and categories extracted by the generative model.

Table 7. Aspect–sentiment results. AS-F1 requires both a matched aspect and the correct sentiment label. Sent. Acc. denotes sentiment accuracy over matched aspect pairs. LLaMA3 rows are mean±standard deviation over five seeds; Dorna rows are mean±standard deviation over three seeds.

Model	Test set	Reviews	AS-F1	Sent. Acc.
Zero-shot LLaMA3	GPT-assisted test	321	0.2123	0.8235
Zero-shot LLaMA3	Pars-ABSA	1,719	0.0780	0.8339
Zero-shot LLaMA3	ParsiNLU food	186	0.0910	0.9091
Zero-shot LLaMA3	ParsiNLU movie	96	0.0303	0.8571
Zero-shot LLaMA3	Human-only	239	0.1185	0.8310
LLaMA3-CG	GPT-assisted test	317	0.6372 ± 0.0109	0.9269 ± 0.0091
LLaMA3-CG	Pars-ABSA	1,313	**0.2879 ± 0.0058**	0.8730 ± 0.0149
LLaMA3-CG	ParsiNLU food	186	0.2315 ± 0.0215	0.8716 ± 0.0176
LLaMA3-CG	ParsiNLU movie	96	0.2671 ± 0.0293	0.8819 ± 0.0356
LLaMA3-CG	Human-only	239	0.4394 ± 0.0212	0.8021 ± 0.0123
LLaMA3-CG+ParsiNLU	GPT-assisted test	317	0.6039 ± 0.0144	0.9211 ± 0.0109
LLaMA3-CG+ParsiNLU	Pars-ABSA	1,313	0.2769 ± 0.0088	0.8667 ± 0.0039
LLaMA3-CG+ParsiNLU	ParsiNLU food	186	0.8075 ± 0.0147	0.9328 ± 0.0139
LLaMA3-CG+ParsiNLU	ParsiNLU movie	96	**0.5466 ± 0.0227**	0.8655 ± 0.0256
LLaMA3-CG+ParsiNLU	Human-only	239	0.3943 ± 0.0232	0.8138 ± 0.0264
Dorna-CG	GPT-assisted test	317	**0.6591 ± 0.0099**	0.9348 ± 0.0031
Dorna-CG	Pars-ABSA	1,313	0.2852 ± 0.0057	0.8768 ± 0.0040
Dorna-CG	ParsiNLU food	186	0.2493 ± 0.0073	0.9041 ± 0.0118
Dorna-CG	ParsiNLU movie	96	0.2679 ± 0.0194	0.9029 ± 0.0187
Dorna-CG	Human-only	239	**0.4897 ± 0.0194**	0.8428 ± 0.0163
Dorna-CG+ParsiNLU	GPT-assisted test	317	0.6438 ± 0.0153	0.9353 ± 0.0102
Dorna-CG+ParsiNLU	Pars-ABSA	1,313	0.2871 ± 0.0074	0.8895 ± 0.0040
Dorna-CG+ParsiNLU	ParsiNLU food	186	**0.8695 ± 0.0067**	**0.9732 ± 0.0080**
Dorna-CG+ParsiNLU	ParsiNLU movie	96	0.5269 ± 0.0248	0.9190 ± 0.0055
Dorna-CG+ParsiNLU	Human-only	239	0.4769 ± 0.0194	0.8697 ± 0.0115

Table 8. Examples of model behavior on Persian texts with highlighted aspects. Positive aspects are highlighted in green, neutral in yellow, and negative in red.

Review (Persian)	Translation (English)	Aspects Identified (Category)	Model Output Analysis
این لپ‌تاپ از نظر سرعت پردازنده عالیه، ولی باتری‌اش خیلی زود خالی می‌شه	This laptop has excellent processor speed, but its battery drains very quickly.	Processor speed (Performance), Battery (Performance)	Correct identification of a positive performance aspect and a negative battery-related aspect.
دوربین موبایل فوق‌العاده است، اما طراحی ظاهری اصلاً جالب نیست	The mobile camera is fantastic, but its physical design is not appealing at all.	Mobile camera (Performance), Physical design (Appearance)	Correct handling of contrasting sentiments in one review.
رنگش قشنگه، اما نمی‌دونم به دکور خونه‌ام میاد یا نه	Its color is nice, but I do not know whether it suits my home decor.	Color (Appearance), Home decor (Neutral)	Ambiguous opinion about suitability can lead to unstable aspect or sentiment assignment.
برنامه کاربردی و سریع بود، اما گاهی دچار اشکالات فنی می‌شد	The application was practical and fast, but it occasionally had technical issues.	Application (Performance), Technical issues (Performance)	The model may merge distinct aspects into a broad performance category.
این گوشی شاهکاره، فقط اگر کار کنه	This phone is a masterpiece, if it works!	Phone (Functionality)	Sarcasm is difficult; literal positive words can mask negative intent.

(a) Battery. (b) Camera.

Figure 5. Word clouds of people's opinions on the battery and camera of a mobile phone.

contribute to this gap. First, Pars-ABSA reviews are much longer than the GPT-assisted and ParsiNLU reviews, and they often contain multiple clauses and densely packed aspect mentions. Second, the aspect definitions and categories in Pars-ABSA differ from our generative annotation scheme. Third, exact overlap removal reduces the evaluation subset for fine-tuned models. These differences explain why the model generalizes less well to Pars-ABSA than to the GPT-assisted, human-only, and ParsiNLU settings.

Domain-specific challenges. The model performs better on the ParsiNLU food domain than on the movie domain after ParsiNLU training. Food reviews typically focus on concrete aspects such as taste, packaging, portion size, or delivery, and their sentiment cues are often explicit. Movie reviews often involve abstract aspects such as story, acting, emotional impact, or atmosphere, which makes aspect boundaries and sentiment anchoring harder.

5.6 Case Study: Generative Capabilities of Our Model

In this section, we illustrate the generative capability of the fine-tuned model on product reviews. The model autonomously extracts aspects, opinions, categories, and sentiments without relying on a predefined aspect list. These qualitative examples are intended to complement the quantitative results and show how the generated quadruples can support exploratory analysis of user feedback.

5.6.1 Mobile Phone

For mobile-phone reviews, the model frequently extracts aspects such as battery, camera, display quality, speed, memory, price, and design. The word clouds in Figures 4 and 5 illustrate how the model groups repeated user concerns. Battery-related opinions often include negative expressions about fast discharge or charging, whereas camera-related opinions often include positive expressions about quality, clarity, or resolution. These patterns demonstrate the practical value of open-ended aspect extraction for e-commerce analysis.

5.6.2 Book Reader

In book-reader or digital-reading-device reviews, extracted aspects tend to include screen, size, battery, light, price, and usability. The model can identify both product-level aspects and usage-related opinions, such as whether the screen is comfortable for reading or whether the device is easy to carry. At the same time, some errors occur when users discuss reading habits or delivery conditions rather than the product itself.

5.6.3 Clothing

For clothing reviews, the model commonly extracts aspects such as material, color, size, stitching, design, and price. These aspects are often domain-specific and may not appear in predefined aspect inventories. The generative formulation is therefore useful because it can adapt to product categories where the relevant aspects differ substantially from electronics or food.

5.7 Comparative Analysis and Practical Implications

Compared with classification-based Persian ABSA systems, the proposed generative formulation provides richer outputs. It not only classifies sentiment for known aspects; it also extracts new aspect terms, opinion spans, and categories. This is useful for e-commerce, market analysis, and social-media mon-

itoring, where users may mention unexpected or emerging concerns.

However, the approach is computationally more expensive than encoder-based classifiers because inference requires autoregressive generation. It is therefore more suitable when rich structured output is needed than when a lightweight polarity classifier is sufficient. The revised experiments also show that training data composition matters: adding ParsiNLU strongly improves ParsiNLU evaluation but can reduce performance on the human-only set. Practitioners should therefore match the training data to the target domain and annotation policy.

6 Conclusions

This paper presented a generative framework for Persian Aspect-Based Sentiment Analysis using fine-tuned LLaMA3-family instruction models. The model generates structured JSON quadruples containing aspect terms, categories, opinion spans, and sentiment labels. The work is positioned as an applied and resource-oriented contribution rather than a new transformer architecture.

To address the risk of circularity in GPT-assisted annotation, we introduced an independent human-only test set annotated from scratch by three Persian annotators. The results show that task-specific fine-tuning substantially improves over zero-shot prompting on this human-only benchmark. LLaMA3-CG improves from 0.1426 to 0.5477±0.0229 aspect F1 and from 0.1185 to 0.4394±0.0212 aspect–sentiment F1. The Persian-oriented Dorna-CG model further obtains 0.5809±0.0123 aspect F1 and 0.4897±0.0194 aspect–sentiment F1, suggesting that Persian-oriented LLaMA3-family models are promising for this task.

The experiments also reveal important limitations. Performance remains sensitive to domain shift, annotation-scheme mismatch, implicit opinions, sarcasm, and over-generation. Adding ParsiNLU improves in-domain ParsiNLU performance but does not universally improve the independent human-only test set. Future work should expand the human-only dataset, improve structured-output constraints, evaluate additional Persian-oriented LLMs, and explore continued Persian pretraining or retrieval-augmented annotation tools.

References

[1] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135, 2008. ISSN 1554-0669. doi:10.1561/1500000011. URL http://dx.doi.org/10.1561/1500000011.

[2] Zhengjie Gao, Ao Feng, Xinyu Song, and Xi Wu. Target-dependent sentiment classification with bert. *IEEE Access*, 7:154290–154299, 2019. doi:10.1109/ACCESS.2019.2946594.

[3] Margarita Rodríguez-Ibánez, Antonio Casánez-Ventura, Félix Castejón-Mateos, and Pedro-Manuel Cuenca-Jiménez. A review on sentiment analysis from social media platforms. *Expert Systems with Applications*, 223:119862, 2023. ISSN 0957-4174. doi:10.1016/j.eswa.2023.119862. URL https://www.sciencedirect.com/science/article/pii/S0957417423003639.

[4] Saif M. Mohammad and Xiaodan Zhu. Sentiment analysis of social media texts. In Lucia Specia and Xavier Carreras, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, Doha, Qatar, October 2014. Association for Computational Linguistics. URL https://aclanthology.org/D14-2001.

[5] Kia Dashtipour, Mandar Gogate, Ahsan Adeel, Cosimo Ieracitano, Hadi Larijani, and Amir Hussain. Exploiting deep learning for persian sentiment analysis. In Jinchang Ren, Amir Hussain, Jiangbin Zheng, Cheng-Lin Liu, Bin Luo, Huimin Zhao, and Xinbo Zhao, editors, *Advances in Brain Inspired Cognitive Systems*, pages 597–604, Cham, 2018. Springer International Publishing. ISBN 978-3-030-00563-4. doi:10.1007/978-3-030-00563-4_58.

[6] Yazdan Zandiye Vakili, Avisa Fallah, and Soodabeh Zakeri. Enhancing sentiment analysis of persian tweets: A transformer-based approach. In *2024 10th International Conference on Web Research (ICWR)*, pages 226–230, 2024. doi:10.1109/ICWR61162.2024.10533353.

[7] Hamoon Jafarian, Amir Hossein Taghavi, Alireza Javaheri, and Reza Rawassizadeh. Exploiting bert to improve aspect-based sentiment analysis performance on persian language. In *2021 7th International Conference on Web Research (ICWR)*, pages 5–8, 2021. doi:10.1109/ICWR51868.2021.9443131.

[8] Taha Shangipour ataei, Kamyar Darvishi, Soroush Javdan, Behrouz Minaei-Bidgoli, and Sauleh Eetemadi. Pars-ABSA: a manually annotated aspect-based sentiment analysis benchmark on Farsi product reviews. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7056–7060, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.763.

[9] Farid Ariai, Maryam Tayefeh Mahmoudi, and Ali Moeini. Enhancing aspect-based sentiment analysis with ParsBERT in persian language. *Journal of AI and Data Mining*, 12(1):1–14, 2024. doi:10.22044/jadm.2023.13666.2482.

[10] AI at Meta Llama Team. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

[11] Part AI. Dorna-llama3-8b-instruct. Hugging Face model card, 2024. URL https://huggingface.co/PartAI/Dorna-Llama3-8B-Instruct. Accessed: 2026-06-27.

[12] Farhan Farsi, Farnaz Aghababaloo, Shahriar Shariati Motlagh, Parsa Ghofrani, MohammadAli SadraeiJavaheri, and etc. Melac: Massive evaluation of large language models with alignment of culture in persian language, 2025. URL https://arxiv.org/abs/2508.00673.

[13] Wenbin An, Feng Tian, Ping Chen, and Qinghua Zheng. Aspect-based sentiment analysis with heterogeneous graph neural network. *IEEE Transactions on Computational Social Systems*, 10(1):403–412, 2023. doi:10.1109/TCSS.2022.3148866.

[14] Yinuo Chen, Jing Zhao, and Haoran Gao. Adversarial training enhanced multi-channel graph convolutional network for aspect sentiment triplet extraction. In *Proceedings of the 2023 6th International Conference on Signal Processing and Machine Learning*, SPML '23, page 301–307, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400707575. doi:10.1145/3614008.3614054. URL https://doi.org/10.1145/3614008.3614054.

[15] Ghiasvand Mohammadkhani, Niloofar Ranjbar, and Saeedeh Momtazi. E2tp: Element to tuple prompting improves aspect sentiment tuple prediction. *Neural Networks*, 191:107847, 2025. ISSN 0893-6080. doi:10.1016/j.neunet.2025.107847. URL https://www.sciencedirect.com/science/article/pii/S0893608025007270.

[16] Jakub vsmid, Pavel Priban, and Pavel Kral. Llama-based models for aspect-based sentiment analysis. pages 63–70. Association for Computational Linguistics, 2024. doi:10.18653/v1/2024.wassa-1.6. URL https://aclanthology.org/2024.wassa-1.6/.

[17] Milad Vazan and Jafar Razmara. Jointly modeling aspect and polarity for aspect-based sentiment analysis in persian reviews. *arXiv preprint*, arXiv:2109.07680, 2021. doi:10.48550/arXiv.2109.07680. URL https://arxiv.org/abs/2109.07680.

[18] Mehrdad Farahani, Mohammad Gharachorloo, Marzieh Farahani, and Mohammad Manthouri. Parsbert: Transformer-based model for persian language understanding. *Neural Processing Letters*, 53(6):3831–3847, Dec 2021. ISSN 1573-773X. doi:10.1007/s11063-021-10528-4. URL https://doi.org/10.1007/s11063-021-10528-4.

[19] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. Llama factory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 400–410, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi:10.18653/v1/2024.acl-demos.38. URL https://aclanthology.org/2024.acl-demos.38/.

[20] Jingwei Xu, Junyu Lai, and Yunpeng Huang. Meteora: Multiple-tasks embedded lora for large language models, 2024. URL https://arxiv.org/abs/2405.13053.

[21] Luca Massaron. Fine-tune llama 2 for sentiment analysis, kaggle. URL https://www.kaggle.com/code/lucamassaron/fine-tune-llama-2-for-sentiment-analysis.

[22] Furquan. furquan/llama2-sentiment-prompt-tuned. URL https://huggingface.co/furquan/llama2-sentiment-prompt-tuned.

[23] Hugging Face. Setfitabsa: Few-shot aspect based sentiment analysis using setfit. URL https://huggingface.co/models?pipeline_tag=absa.

[24] Daniel Khashabi, Arman Cohan, Siamak Shakeri, Pedram Hosseini, Pouya Pezeshkpour, Malihe Alikhani, and etc. Parsinlu: A suite of language understanding challenges for persian. *Transactions of the Association for Computational Linguistics*, 9:1147–1162, 2021. doi:10.1162/tacl_a_00419.

Niloofar Ranjbar is a faculty member in the Department of Computer Engineering, Jam Faculty of Engineering, Persian Gulf University. Her research interests include natural language processing, explainable and interpretable artificial intelligence, sentiment analysis, large language models, and responsible AI. She has several years of university teaching experience and has contributed to Persian and multilingual NLP research, including semantic disambiguation, emotion classification, and aspect-based sentiment analysis.

Hamed Baghbani is a data scientist and researcher with experience in machine learning, natural language processing, and large language models. His work includes Persian language modeling, multilingual NLP, sentiment and emotion analysis, and applied data science. He has contributed to research and open-source projects on Persian language resources and generative models and has worked on production-oriented analytics and machine-learning systems. His current interests include efficient model adaptation and responsible evaluation of language technologies.