Proceedings of the Twentieth International AAAI Conference on Web and Social Media (ICWSM 2026)

Towards Transparent Stance Detection: A Zero-Shot Approach Using Implicit and Explicit Interpretability

Apoorva Upadhyaya, Wolfgang Nejdl, Marco Fisichella

L3S Research Center, Leibniz Universität Hannover, Hannover, Germany
{upadhyaya, nejdl, mfisichella}@l3s.de

Abstract

Zero-Shot Stance Detection (ZSSD) identifies the attitude of the post toward unseen targets. Existing research using contrastive, meta-learning, or data augmentation suffers from generalizability issues or lack of coherence between text and target. Recent works leveraging large language models (LLMs) for ZSSD focus either on improving unseen target-specific knowledge or generating explanations for stance analysis. However, most of these works are limited by their over-reliance on explicit reasoning, provide coarse explanations that lack nuance, and do not explicitly model the reasoning process, making it difficult to interpret the model's predictions. To address these issues, in our study, we develop a novel interpretable ZSSD framework, IRIS. We provide an interpretable understanding of the attitude of the input towards the target implicitly based on sequences within the text (implicit rationales) and explicitly based on linguistic measures (explicit rationales). IRIS considers stance detection as an information retrieval ranking task, understanding the relevance of implicit rationales for different stances to guide the model towards correct predictions without requiring the ground-truth of rationales, thus providing inherent interpretability. In addition, explicit rationales based on communicative features help decode the emotional and cognitive dimensions of stance, offering an interpretable understanding of the author's attitude towards the given target. Extensive experiments on the benchmark datasets of VAST, EZ-STANCE, P-Stance, and RFD using 50%, 30%, and even 10% training data prove the generalizability of our model, benefiting from the proposed architecture and interpretable design.

1 Introduction

Zero-shot stance detection (ZSSD) is about recognizing the author's stance towards the unknown targets. Current advances in ZSSD have mainly focused on adversarial, meta-learning, or data augmentation approaches (Zhang et al. 2024b; Wang, Zhang, and Wang 2024). However, these techniques suffer from generalization problems due to the uneven distribution of targets or the lack of high-quality annotated training data. In addition, large language models (LLMs) have been used for stance detection (SD) (Ding et al. 2024b; Zhang et al. 2024c). Most of the existing studies using LLMs aim to improve target-specific knowledge for the unseen targets with the input text (Guo, Jiang, and Liao 2024; Zhang et al. 2024c). Some of the more recent work has focused on interpretable models, such as (Zhang et al. 2024a) developed a knowledge-augmented interpretable network using LLM to provide perspectives towards targets, (Saha, Lakshmanan, and Ng 2024) constructs a stance tree to provide explanations for claim-based stance, and (Ding et al. 2024a) uses instruction-based chain-of-thought with LLMs to generate explanations for stance analysis.

While these works provide a high-level perspective, they often lack fine-grained insight into the word-level influences that may affect the overall decision-making process for identifying stance, thus reducing interpretability at a detailed level. Moreover, most of these studies rely on explicit inference or acquisition of target-specific knowledge, which can overlook nuanced or implicit cues in the input text, especially subtle attitudes. To address these issues, in contrast to existing works that are limited by over-reliance on coarse explanations (Saha, Lakshmanan, and Ng 2024; Ding et al. 2024a), we develop an interpretable ZSSD system that incorporates both implicit rationales based on sequences within the text and explicit rationales based on linguistic measures. This provides a more fine-tuned, interpretable method that enables a more comprehensive understanding of the author's stance by capturing both subtle and overt cues, ultimately improving both transparency and granularity.

To achieve this, we first focus on deciphering the short snippets in the input post (DeYoung et al. 2020), which provide enough evidence to support the classification results and are referred to as *implicit rationales*. Furthermore, it is possible that a single post contains different rationales that correspond to different attitudes toward the given target. For example: *Target:* "nuclear mission"; *Text:* "..I was against Nuclear power..but now it seems that nuclear should be in the mix. Fission technology is better..to be explored.."; *Favor:* "nuclear should be in the mix", "Fission technology is better"; *Against:* "I was against Nuclear power". Hence, we focus on extracting all possible implicit rationales from the given input. Our preliminary investigations with LLMs motivated us to leverage LLMs to extract rationales rather than to identify stances and best-possible rationale for predictions (Section 5). Moreover, due to cost and time constraints, it is not possible to achieve ground-truth stance labels for all rationales for all inputs. Therefore, we develop different stages

of our model, namely the "relevance ranking", which automatically interprets the relevance of each rationale to favor, against, and neutral stance using ranking algorithms, while the "grouping and selection" stage selects the top-k most diverse relevant and irrelevant rationales, thus guiding the model to correctly predict stance with the reasoning for such prediction, thus ensuring inherent interpretability.

In addition, the linguistic measures of communication dynamics, such as empathy, absolutism, action, abstract, concrete, communion language, etc., provided by the LIWC framework (Boyd et al. 2022) could offer deeper insights into how and why a user adopts a certain attitude. For example, empathic or communicative language may indicate a supportive stance, while absolutist or action-oriented language could indicate strong opposition. For instance: *Target:* "prices"; *Stance:* "Against", *Text:* "...Increased..electricity..gas prices again..incompetent, stupid, useless..Pwehsident..! Leche!"; *Linguistic assessment (LLM):* "The post exhibits low empathy as it contains insulting language..confrontational and critical, indicating absolutist thinking..approach language reflects anger and frustration to the issues of rising price..no signs of allure or persuasion, instead, the language is direct and attacking.". Hence, these reasonings as *explicit rationales* help to uncover hidden biases, motivations, or attitudes in a post to provide a more interpretable understanding of the user's attitude towards the target. We utilize the reasoning capabilities of LLMs to assess the input based on several linguistic features (refer Section 3.1). The final concatenated relevant implicit and explicit rationales help to classify the final attitude of the given input.

The main contributions of our work are as follows:
(i.) To the best of our knowledge, this is the first study that considers stance detection as an information retrieval ranking task while providing inherent interpretability by focusing on both implicit and explicit reasonings to uncover the hidden stance-related information in the post. ***(ii.)*** We propose a novel interpretable ZSSD framework, consisting of 3 main stages, *relevance ranking, grouping and selection, and classification* that leverages implicit rationales as a subsequence of words and explicit rationales based on linguistic measures to classify the stance. We refer to our model as **I**nterpretable **R**ationales for **S**tance Detection as **IRIS**. ***(iii.)*** In our proposed relevance ranking stage, stance detection is considered as a ranking task to automatically assign each implicit rationale towards stances, while the grouping and selection phase selects the k most diverse rationales. In this way, the disadvantage of the lack of human-annotated ground truth for rationale stances is overcome, leading the model towards relevant rationales and supporting interpretability by design. ***(iv.)*** The classification phase of IRIS uses the encoded representations of the implicit and explicit arguments to predict the stance of each selected rationale and classifies the final stance of the inputs using majority voting. ***(v.)*** Extensive experiments are conducted on VAST and EZ-STANCE datasets for ZSSD and P-Stance and RFD datasets for generalizability analysis using *50%*, *30%*, and even *10%* training data to demonstrate the efficiency of our IRIS. Our code and datasets are present here[1].

2 Related Work

2.1 Stance Detection

A wealth of work has been explored in the field of in-target and cross-target SD, where training and test targets are the same or closely related (Upadhyaya, Fisichella, and Nejdl 2023c,b,a). However, recognizing the stance of an unknown target appears to be of greater importance as training data is not available for all targets. Therefore, recent literature aims to address the ZSSD. Recent works on ZSSD have either focused on target-specific and invariant features (Zhang et al. 2024b) or used contrastive learning to improve stance detection in zero-shot scenarios (Yao, Yang, and Wei 2024). Building on advancements in prompt learning, Yao, Yang, and Wei (2024) integrated prompt learning with contrastive learning to improve zero-shot stance detection. However, Wang, Zhang, and Wang (2024) introduced a meta-learning algorithm combined with data augmentation to address generalizability challenges associated with stance detection. Additionally, Zhao et al. (2024) proposed a collaborative feature learning framework with multiple experts for zero-shot stance detection. Despite these innovations, these methods continue to face limitations in performance on unseen targets, often struggling to capture nuanced stance-related information and lacking transparency or contextual understanding between targets and text. This motivated us to design a framework capable of uncovering the underlying reasoning and motivations behind the author's attitude toward a given target. By extracting both implicit and explicit reasons, our proposed approach provides a clear explanation for the stance expressed in the post.

2.2 LLMs for Stance Detection

Recently, LLMs have proven to be beneficial for various number of classification tasks (Upadhyaya, Nejdl, and Fisichella 2024, 2025a). Recent trends are also moving towards the use of LLMs for ZSSD (Zhang et al. 2024a; Ma et al. 2024; Upadhyaya, Nejdl, and Fisichella 2025b). (Zhang et al. 2024c) utilized LLM to explicitly extract the relationship between paired texts and targets, while (Hu et al. 2024) extracted target information from the web to improve the attitude task. In addition, (Saha, Lakshmanan, and Ng 2024) focused on generating explanations for predicted stances by capturing the pivotal argumentative structure embedded in a document. However, our approach formulates stance detection as a ranking task that aims to inherently capture the relevance of implicit rationales for different stances while using communicative features to decode the emotional and cognitive dimensions of the attitude. This eliminates the need for ground truth for rationales while providing a more comprehensive understanding of the author's attitude. Even recent work has highlighted the challenges of interpretability in the era of LLMs (Singh et al. 2024). Hence, our work offers advantages over existing works in

[1]https://osf.io/ja97s/?view_only=bb6402c6ddce4bc482deeb879a0096d5

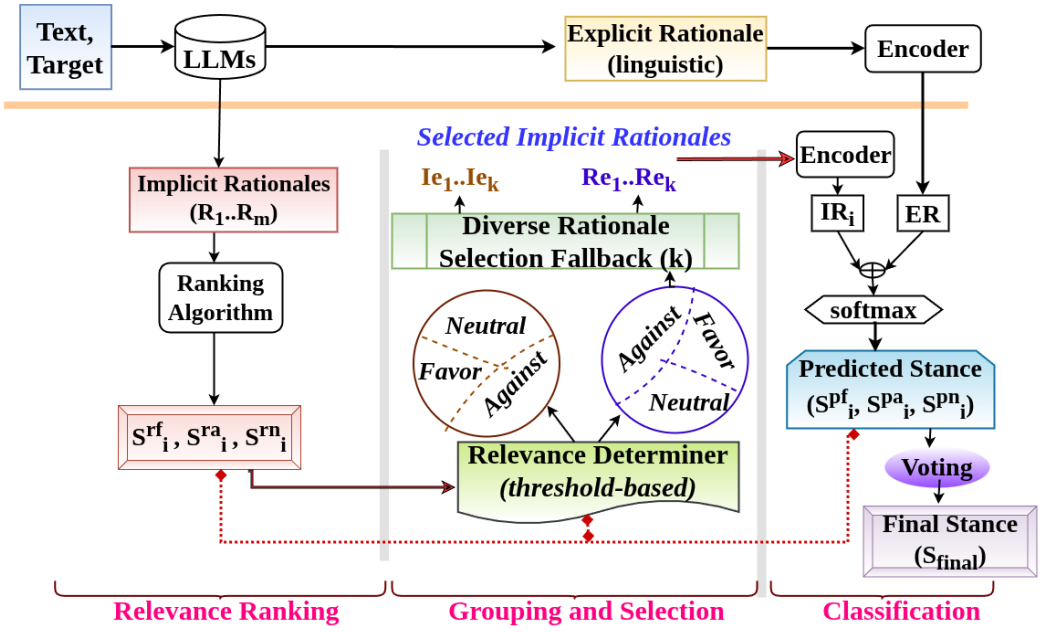

Figure 1: Stage-wise detailed architecture design. *[Left]: Relevance Ranking; [Center]: Grouping and Selection; [Right]: Classification*. Notations: R_i: LLM-generated i^{th} implicit rationale; $\{S_i^{rf}, S_i^{ra}, S_i^{rn}\}$: Relevance scores of i^{th} implicit rationale towards favor, against, and neutral stances; Re_k, Ie_k: k relevant and irrelevant implicit rationales; IR_i: Encoded i^{th} relevant implicit rationale; ER: Encoded explicit linguistic rationale; $\{S_i^{pf}, S_i^{pa}, S_i^{pn}\}$: stance scores for each encoded relevant i^{th} implicit rationale and explicit rationale; S_{final}: Final predicted stance. The details are clearly mentioned in Section 3.

terms of interpretability, rationale analysis, and considering specific text subsequences in relation to stance and linguistic cues.

3 Methodology

Problem Definition: Given a training dataset containing P samples with text, target, and the corresponding stance towards the target $((x_i, t_i, s_i)_{i=1}^{P})$ and a set of Q samples as a test set with text and targets that cannot be seen during training $((x_i, t_i)_{i=1}^{Q})$. The aim is to design an interpretable ZSSD to determine the stance of the unseen test set (favor/against/neutral), achieved by inherently focusing on relevant implicit rationales together with the explicit reasonings without the human annotations (ground truth) for rationales.

Figures 1 and 4 represent the flow of our proposed approach, IRIS, which consists of 3 main stages of *Relevance Ranking, Grouping and Selection, and Classification*. The input post is first fed into the LLM to extract the explicit rationales based on linguistic features and the implicit rationales, i.e. the subsequence of words from the input text that lead to the different stance categories towards the given target. The implicit rationales are then fed to the *relevance ranking* stage which determines the ranking of the implicit rationales for a favorable, against, or neutral attitude. Each rationale is then categorized into relevant or irrelevant groups and the k most informative and diverse rationales are then selected in the *grouping and selection* phase and encoded using sentence transformer. In parallel, the explicit rationales are also encoded separately to capture meaningful context. The selected implicit rationales together with the explicit reasoning then pass through the softmax layer to determine the stance for all rationales. To classify the final stance of the input, the attitudes predicted by the relevant rationales are guided by the consensus and result in the final stance label. Next, we describe the different stages in detail.

3.1 Rationale Generation

This stage ensures that both explicit and implicit reasoning behind stance decisions is captured for interpretability. Here, the input text with the target is fed into the LLM to extract rationales. As the same input text can have different stances depending on the target, we apprehend the target with the input text. Using the prompt given in Figure 5, we ask the LLM to assess the given post based on the linguistic measures and provide a more interpretable understanding of the user's attitude towards the target, by viewing them as explicit rationales. To ensure inherent interpretability in stance task, we use the prompt (Figure 6) to extract the subsequence of words associated with different attitudes as a set of implicit rationales. In the following, we explain the components that help in processing the implicit and explicit rationales.

3.2 Relevance Ranking

The main responsibility of this stage is to provide transparency by designing a ranking algorithm that inherently interprets the relevance of each rationale towards different stance labels as we do not rely on ground truth for rationales (Figure 1[Left]). This helps in minimizing human efforts and

Instruction for Ranking Algorithms

Given a query consisting of a sequence of words directed towards given target, evaluate how relevant the query is to the document based on how document's sentences address alignment with their corresponding targets.

Figure 2: Instruction for Ranking Algorithm

automatically captures the context of each rationale towards different stances and targets, thus directly contributing to our model's decision-making process. Here, we use the LLM extracted implicit rationales retrieved from the previous *rationale generation* stage (Section 3.1) as a set of m rationales $\{R_1, R_2, ...R_m\}$ ("m" may vary for each input depending on the LLM responses). We consider each *rationale* with the *target* as a *query* and prepare a set of 3 documents containing various posts and their targets corresponding to "favor", "against", and "neutral" stances without explicitly mentioning the stances in the documents.

Document Preparation The documents are articulated using the publicly available benchmark zero-shot stance datasets. To ensure novelty and minimize overlap, we only include statements in the documents that have a cosine similarity of less than 0.05 with training and test data. This filtering process prevents the use of information that is similar to the samples used for model evaluation, thus reducing the risk of data leakage and maintaining the integrity of the zero-shot framework. We intentionally exclude any stance labels from the documents, ensuring that there is no bias introduced into the training process. The documents may contain statements that align with favor, against, or neutral perspectives, but these are kept implicit and without stance labels to avoid any task-specific bias. Hence, these documents act as a source of external knowledge for the ranker to compute relevance scores that indicate how closely a given implicit rationale as a query aligns with the content of each document. By excluding stance labels and ensuring novelty through cosine similarity filtering, the documents preserve the zero-shot nature of the stance detection task.

Instruction In addition to the query and set of documents, we first construct an instruction for the ranker (refer Figure 2). This helps the rankers understand contextual relationships between document-target connections and allows the ranker to infer relevance without revealing explicit stance labels. The query-document pair together with the instruction is then fed to the ranker.

Ranking Algorithm We experiment with different pretrained rankers and LLM, as described in Section 5. We then select the best ranking algorithm for our study as FlagReranker (Xiao et al. 2023). Inspired by (Chen et al. 2024), we also employ a publicly available bgereranker[2] to obtain the relevance scores of each rationale towards favor, against and neutral documents. The ranker provides a list of 3 relevance scores which are often raw and unbounded, therefore passed to the softmax function to obtain a probabilistic distribution, as probabilities provide a clearer understanding of the relative confidence for each stance, compared to raw values, resulting in $\{S_i^{rf}, S_i^{ra}, S_i^{rn}\}$ for all R_i. We utilize the well-established pre-trained FlagReranker to ensure that scores are grounded in the underlying semantic relationships. By identifying and ranking rationales, the system inherently highlights key supporting or opposing evidence, making the stance decision more explainable. Thus, the use of ranker adds a layer of inherent interpretability to the stance detection pipeline.

[2]https://huggingface.co/BAAI/bge-reranker-large

3.3 Grouping and Selection

This stage identifies relevant and irrelevant rationales towards stances (Figure 1[Center]). This results in improving interpretability by clearly indicating both the positive (relevant) and negative (irrelevant) influences on the stance, making the model's final predictions more explainable.

Relevance Determiner We use the threshold-based difference to determine whether a rationale is relevant to a particular stance and irrelevant to others. Given a relevance score of the rationale (R_i) $S_i^{rf}, S_i^{ra}, S_i^{rn}$ obtained from the previous ranking stage, we determine if $(S_i^{rf} - max(S_i^{ra}, S_i^{rn})) > threshold$, then consider the rationale as relevant for the favor stance and irrelevant for against and neutral stance; a similar process is repeated for all stance labels of all rationales. We also consider other approaches for grouping that use clustering methods, however, we already have relevance scores from the previous phase. Moreover, our component provides an unsupervised approach with transparency to identify relevant rationale based on score distributions without requiring ground truth labeling, which led us to utilize this method in determining the relevance of all generated rationales. The output of "Relevance Determiner" results in 2 groups of relevant and irrelevant rationales consisting of favor, against, and neutral stances as sub-groups.

Diverse Rationale Selection Fallback (k) aims to select diverse rationales from two main groups (relevant and irrelevant) while balancing the representation of subgroups within each group (favor, against, and neutral). In a real environment, it is often impossible to have the same number of rationales for each subgroup (favor, against, neutral). To achieve this, we use KL-divergence to minimize the deviation from a dynamic target desired distribution based on available rationales. The detailed steps are explained in Algorithm 1. The algorithm starts by computing a target distribution V based on the proportion of rationales available in each subgroup (steps 1-3 of Algorithm 1). Using KL-divergence, the algorithm iteratively selects rationales to minimize the divergence between the current distribution of selected rationales and the target distribution. Initially, no rationales are selected, so the current distribution is initialized with small non-zero values to avoid computation errors (step 4 of Algorithm 1). At each step, a rationale is temporarily added, the new distribution is calculated, and the rationale

that best aligns the updated distribution with the target is chosen (steps 8-13 of Algorithm 1). The process continues until the desired number (k) of rationales is selected. A fallback mechanism ensures progress even if one subgroup is exhausted by selecting from the remaining groups (step 15 of Algorithm 1). This process is applied separately for relevant and irrelevant implicit rationales to maintain balanced coverage.

This stage finally leads to k relevant ($Re_1..Re_k$) and irrelevant ($Ie_1..Ie_k$) selected implicit rationales.

3.4 Classification

Here, the main aim is to classify input's stance (Figure 1[Right]). Loss functions improve the IRIS performance and emphasize clarity in the decision-making of predictions.

Rationale Encodings As the previous grouping stage results in implicit rationales k relevant ($Re_1..Re_k$) and irrelevant ($Ie_1..Ie_k$) in a given context, these implicit rationales, together with their input targets, are then passed through the sentence encoder (Lee et al. 2024) of the embedding dimension (d_e) followed by a dense layer (d_d), resulting in $IR_i \in \mathbb{R}^{1\times d_d}$, helping to understand the meaningful context within the rationales. These embeddings have been effective in various retrieval and classification tasks (Yang 2024; Lee et al. 2024) (different embeddings have been examined in Section 5). In parallel, the explicit rationales extracted by LLM are separately encoded using sentence embeddings (d_e) (Lee et al. 2024) and passed through a dense layer (d_d). This captures the essence of the context between the input and the linguistic characteristics, leading to $ER \in \mathbb{R}^{1\times d_d}$.

Stance Prediction For each selected implicit rationale, the implicit and explicit encoded rationales are concatenated ($R^{1\times 2(d_d)}$) and passed through softmax, resulting in the stance prediction of each rationale towards favor, against, or neutral stance $\{S_i^{pf}, S_i^{pa}, S_i^{pn}\}$. We select the stance predictions of concatenated explicit and relevant implicit rationales and classify the final stance (S_{final}) of the input based on the majority vote. If no decision is made, we classify the output as neutral. Various loss functions are as follows.

Loss functions ***Stance Loss (L_{ce}^s):*** The categorical cross-entropy (ce) loss is calculated for the final predicted stance of input S_{final} and the true stance label for ZSSD.
Rationale Usefulness Reward Punish Loss (L_{rp}): This custom loss calculates how much the relevance ranking phase should reward/penalize depending on whether or not the rationales selected resulted in the correct stance in the final stage. It is based on the idea that the selection of relevant rationales should lead to correct stances, while irrelevant rationales should result in incorrect predictions. We focus on the loss in the relevance phase (L_{ce}^{rel}), where $S_i^{rf}, S_i^{ra}, S_i^{rn}$ are prediction scores versus ground truth stance. We also examine the relevance labels of the rationales (relevant/irrelevant) based on the grouping stage. $L_{rp} = L_{ce}^{rel} \times (1-R)$, where R is a reward/punishment term defined using the stance predictions of the classification stage as follows: $R = \beta(reward)$: if relevant rationale leads to correct stance or irrelevant rationale leads to incorrect stance predictions; while $R = -\beta(punish)$: if relevant rationale results in incorrect stance or irrelevant rationale leads to correct predictions [here the predictions are $\{S_i^{pf}, S_i^{pa}, S_i^{pn}\}$ of classification stage compared to ground truth stance] (refer Figure 1). This adjustment encourages the relevance ranking stage to improve its rationale selection strategy based on stance prediction outcomes in the classification stage.
Total loss: of our proposed approach is, $L = L_{ce}^s + qL_{rp}$, where q is ratio of rationale usefulness loss to total loss.

	Train	Dev	Test
# Examples	13477	2062	3006
# Unique Comments	1845	682	786
# Zero-shot Topics	4003	383	600
# Few-shot Topics	638	114	159

Table 1: Dataset statistics of VAST

	Train	Val	Test
# Samples of noun-phrase targets	13756	2354	2663
# Samples of claim targets	18879	4349	5135

Table 2: Dataset statistics of EZ-STANCE

4 Experimental Setup

4.1 Datasets

To evaluate our IRIS model, we conduct experiments on two **zero-shot** stance detection (ZSSD) benchmarks: VAST and the recently curated EZ-STANCE. VAST comprises comments from The New York Times' Room for Debate section, while EZ-STANCE contains tweets collected from social media discourse. These datasets together offer a rich variety of targets, including both noun phrases and claim-based target-text pairs, thus enabling a broad and comprehensive evaluation of model performance across informal and structured contexts. ***VAST (Allaway and McKeown 2020):*** ZSSD dataset with *pro, con, neutral* labels for both zero-shot and few-shot targets in the train and test set. The dataset statistics are present in Table 1. ***EZ-STANCE (EZ) (Zhao and Caragea 2024):*** recently curated tweets including noun-phrase (N), claim-based (C), and mixed (M) targets with *Favor, Against, Neutral* stance labels. The statistics are available in Table 2.

Generalizability Analysis: To further assess IRIS's robustness across domains, stance types, and linguistic styles, we extend our evaluation to two additional datasets: P-Stance and the RFD News Articles dataset. P-Stance focuses on political figures, enabling domain-specific generalization analysis within political discourse. Importantly, we evaluate IRIS on both **in-target** and **zero-shot** stance settings for P-Stance, thereby testing its adaptability to both seen and unseen targets. RFD, on the other hand, provides long-form opinion articles from The New York Times' Room for Debate section—similar in source to VAST but distinct in content length and writing style. While VAST samples average around 100 words, RFD articles average 416 words, the highest among all datasets (EZ-STANCE: 40; P-Stance:

30). This extended length introduces greater discourse complexity and nuanced stance expression, making RFD particularly valuable for evaluating IRIS's generalizability under **in-target** settings in formal, context-rich environments. ***P-Stance (Li et al. 2021):*** is a political stance dataset containing $7,953$ annotated tweets for "Donald Trump", $7,296$ for "Joe Biden", and 6,325 for "Bernie Sanders" as target domains with favor and against stances. ***RFD (Saha, Lakshmanan, and Ng 2024):*** consists of a total of 764 claim-based article pairs with pro (44%), con (47%), and balanced (9%) stance labels. Despite its small size, it has the highest article/text length (416 average words) when compared with all other datasets, as mentioned above (Average number of words- VAST: 100; EZ-STANCE: 40; P-Stance: 30).

All datasets consist of English-language posts and do not include any personally identifiable information; however, they may contain offensive content due to their explicit stances on topics such as religion, politics, immigrants, etc. We strictly adhere to the requirements of the respective licenses for all datasets used in our study. ***Please note that VAST and EZ-STANCE serve as the primary datasets for evaluating ZSSD in our study, while P-Stance and RFD are utilized to assess the generalizability of our approach across domains and settings.***

4.2 Rationale Generation and Annotation for LLM-Based ZSSD

LLM Prompts for Rationale Generation A pilot study was conducted on a randomly selected small sample of 100 posts from the VAST and EZ datasets to assess the efficacy of our prompts. This exploratory analysis enabled us to evaluate prompt responsiveness in a controlled setting, informing subsequent refinements that optimized the prompts utilized in our experimental design. Figures 5 and 6 illustrate the final prompts we used to ask Llama 3.1 to generate explicit and implicit rationales in zero-shot settings respectively. As mentioned in Section 5, Llama 3.1 performs better than Mistral on the ZSSD task. We take advantage of the text generation capabilities to use Llama for rationales generation. For the case of *explicit rationales* (Figure 5), we derive definitions for various linguistic measures from LIWC (Boyd et al. 2022) and Receptiviti's LIWC framework using the 'liwc_extension' object[3]. This allows us to extract scores for the following attributes: empathy, allure, absolutist, action, concrete, agency_language, communion_language, and approach. For full definitions, please refer to the documentation[3] and Figure 5 (Steps 5-13). We omit the avoidance measure as it is simply the inverse of approach emotions and does not provide additional linguistic value. These linguistic measures focused on understanding the dynamics of communication and the determinants of interpersonal support, thus helping us to decipher the reason for the stance of the text. We then prompted Llama 3.1 to assess the following linguistic characteristics based on the given textual post and target (Steps 1-4 of Figure 5) and generate a summarized response in 3-5 lines (Step 14 of Figure 5).

To understand the relevance of each rationale for the different viewpoints, we use the prompt mentioned in Figure 6 to extract the all possible implicit rationales based on given input post and target (Steps 1-4 of Prompt 6) and retrieve their relevance scores towards favor, against, and neutral stances (Step 5 of Prompt 6). Please note that we use these probabilistic scores directly in the relevance ranking phase of our IRIS as part of the ablation study to investigate LLM scores being used to determine the relevance of rationales as described in Section 5.3. We manually analyzed the LLM responses for extracting implicit and explicit rationales for 100 randomly selected posts. A team of two PhD students and one postdoctoral researcher with a background in linguistics and computer science were assigned to examine the LLM responses and were given clear instructions for the task. The annotators observed satisfactory and convincing results, therefore, we used LLMs in the rationale generation stage of our IRIS.

Rationale Annotation for LLM's Correct and Incorrect Predictions for ZSSD As mentioned in Section 5, to understand the interpretability provided by LLMs for zero-shot stance task, we investigate the random 100 correct and incorrect predictions of Llama 3.1 and prompt the LLM to extract the best possible subsequence of words from the input text leading to the predicted stance by LLM. In order to evaluate the LLM's performance on implicit rationale extraction, we manually annotate those 100 samples, considering them as ground truth. The annotation team consists of three annotators who are proficient in English and are doctoral and postdoctoral researchers in the field of linguistics and computer science. They have a strong background and in-depth knowledge of natural language processing, computational social sciences, and information retrieval tasks, which predestines them for this task. We consider rationale identification as a binary classification task. The aim is to give each input token a binary label, suggesting whether the token is present in the rationale or not. These clear instructions were given to the annotators to manually annotate the ground truth for selected 100 samples. We then compared the ground truth versus the Llama responses for the rationale and observed a significantly lower F1 score of 0.617, which motivated us to find other better ranking algorithms that can be beneficial in providing relevance scores for each rationale towards different stances (Section 5.1). To further assess the quality of implicit and explicit rationales by IRIS for ZSSD, we conduct both automatic and human evaluations in Section 5.4.

4.3 Implementation Details

We use the publicly available training and test splits for all datasets. For EZ, we focus on Subtask A, which involves identifying target-based ZSSD across noun-phrase, claim-based, and mixed targets. Following previous work (Zhao and Caragea 2024), we also train IRIS under 3 settings on EZ: (1) using train set with mixed targets, (2) using only noun-phrase targets, and (3) using only claim targets. IRIS is then evaluated on corresponding test set: (1) mixed-targets, (2) noun-phrases only, and (3) claims only (refer Table 2 for this setup). As described in Section 4.1, VAST and EZ are used for evaluating ZSSD, while P-Stance

[3]https://docs.receptiviti.com/frameworks/liwc-extension

and RFD are included to demonstrate the generalizability of IRIS model across different domains and task types. Specifically, P-Stance covers both in-target and zero-shot settings within the political domain, while RFD focuses on in-target stance detection in long-form news articles. To assess the efficacy of IRIS under limited supervision, we train using **10%**, **30%**, and **50%** of the available training data for each dataset. VAST and EZ datasets use 3 stance labels, whereas P-Stance and RFD use the binary stance setup (favor/against and pro/con) for experiments, consistent with their original papers (Li et al. 2021; Saha, Lakshmanan, and Ng 2024).

Hyperparameters: Embedding dimension (d_e): 4096; dense layer dimension (d_d) [with ReLu activation]: 128; threshold *(Relevance Determiner)*: 0.3; k *(Diverse Rationale Selection)*: 3; R (β) [reward/punish metric]: 0.1; output neurons [softmax activation]: 3 [VAST and EZ]/ 2 [P-Stance and RFD]; optimizer: Adam (0.0001 learning rate), batch_size: 32. The best parameter values are selected using TPE in Hyperopt (Bergstra, Yamins, and Cox 2013), which minimizes loss functions. We fine-tune the loss weight using Grid Search from Scikit-learn (q=0.5 (L_{rp})).

Evaluation Metrics: We perform 5 independent runs of our IRIS to account for variability and report average metric scores and standard deviation. Macro-F1 scores are used for VAST, EZ, and RFD datasets as proposed in their work (Allaway and McKeown 2020; Zhao and Caragea 2024; Saha, Lakshmanan, and Ng 2024). Following previous work (Li et al. 2021; Upadhyaya, Fisichella, and Nejdl 2023c), we calculate F_{avg}, which represents the average of F1 scores for *Favor* and *Against* for the P-Stance dataset.

Implementation of LLMs for ZSSD: Since Unsloth's quantized models lead to lower GPU VRAM consumption and have been used in recent research (Kumar 2024), we also work with Unsloth's pre-quantized 4bit Llama-3.1-8B-Instruct[4] and 4bit mistral-7b-instruct-v0.3[5] models for zero-shot, few-shot, and fine-tuning of the LLMs for the stance task. For fine-tuning the LLMs, the specified hyperparameters resulted in the best fine-tuning performance using LoRA: rank of 8, alpha of 16, and dropout rate of 0.1. We employ the better-performing Llama 3.1 model on the stance task (refer Table 3) to extract explicit and implicit rationales by using prompts given in Section 4.2.

Environment Details: GPU Model: NVIDIA A100 GPU servers with carbon efficiency of 0.42 kgCO_2eq/kWh, Library Version: tensorflow 2.12.0, torch 2.4.0+cu121, transformers 4.43.2. The training times for various IRIS variants and baseline models are provided in Appendix A.

4.4 Baselines

We adopt recent state-of-the-art methods as baselines, focusing on those that report results across the datasets used in our study. To ensure a statistically fair comparison with our IRIS model, we re-run each reproducible baseline over five independent runs using our GPU setup (described in Section 4.3) and report the average performance metrics. However, for the VAST and RFD datasets, some baselines could not be reproduced due to the lack of publicly available code or incomplete details regarding hyperparameters and implementation. In such cases, we report their results as stated in the respective original publications without re-implementation.

VAST: We re-run the following baselines for five rounds: COLA (Lan et al. 2024), KAI (Zhang et al. 2024a), EDDA with GPT (EDDA-GPT) and LLaMa (EDDA-LLaMa) (Ding et al. 2024b), Infuse (Yan, Joey, and Ivor 2024), and LKI-BART (Zhang et al. 2024c). The results of the following are taken directly from their original papers: LOT (Hu et al. 2024), RoBERTa-base-PV-P1 (Motyka and Piasecki 2024), CNet-Ad (Zhang et al. 2024b), MCLDA (Wang, Zhang, and Wang 2024), and S-ESD (Ding et al. 2024a).

EZ-STANCE: We implement the top three performing models from the original EZ dataset paper (Zhao and Caragea 2024) as baselines: BART-MNLI-e_p, BART-MNLI-e, and BART-MNLI.

P-Stance: We include the methods that report results on the P-Stance dataset for in-target [Infuse (Yan, Joey, and Ivor 2024) and LOT (Hu et al. 2024)] and zero-shot [COLA (Lan et al. 2024) and KAI (Zhang et al. 2024a)] stance task.

RFD: Due to the lack of available code, we report the baseline results as provided in the original RFD dataset paper (Saha, Lakshmanan, and Ng 2024), including: MoLE (Hardalov et al. 2021) + XSD_p, MTDNN (Schiller, Daxenberger, and Gurevych 2021) + XSD_p, and HSD (Sepúlveda-Torres et al. 2021) + XSD_p. However, we implement Infuse (Yan, Joey, and Ivor 2024), the top-performing method on the VAST dataset, as a baseline on RFD to enable a fair comparison with IRIS across five runs.

Models	Technique	VAST	EZ
Mistral	**Zero-shot(ZS)**	67.09	42.58
	Few-shot(FS)	69.76	44.06
	Fine-tune	71.77	53.4
Llama 3.1	**Zero-shot(ZS)**	63.03	55.13
	Few-shot(FS)	67.55	58.54
	Fine-tune	**72.81**	**63.28**

Table 3: Results of LLMs on VAST (zero-shot targets) and EZ (noun-phrase targets) for ZSSD.

5 Results and Analysis

We begin by analyzing the performance of LLMs, followed by a comparison between baselines and our IRIS on VAST and EZ datasets for ZSSD. We then conduct ablation studies using our primary datasets, VAST and EZ. To further examine the interpretability of IRIS, we evaluate both implicit and explicit rationales using human judgments and automated metrics. Next, we assess IRIS's generalizability on P-Stance and RFD datasets. Finally, we provide case studies to offer deeper insight into IRIS's predictive behavior.

5.1 Performance of LLMs

Table 3 presents the results of LLMs on VAST (zero-shot targets) and EZ (noun-phrase targets) datasets. It can be seen from Table 3 that LLMs perform better when fine-tuned with the training datasets than zero- and few-shot settings.

[4]unsloth/Meta-Llama-3.1-8B-Instruct-bnb-4bit

[5]unsloth/mistral-7b-instruct-v0.3-bnb-4bit

Models	VAST (Zero-shot)			
	Pro	**Con**	**Neu**	**All**
COLA†	71.85	70.77	80.38	74.33
EDDA-GPT†	67.5	67.3	89.6	74.8
EDDA-LLaMA†	67.4	71.2	90.1	76.23
GPT-TiDA	-	-	-	76.1
KAI†	63.58	71.29	87.66	74.17
LKI-BART†	74.6	72.8	90.2	79.2
Cnet-Ad	63.3	65.3	91.1	73.2
MCLDA	67.0	68.2	90.1	75.1
S-ESD	62.0	68.1	88.6	72.9
LOT	-	-	-	78.6
RoBERTa-PV-P1	-	-	-	77.6
Infuse†	74.7	75.1	**94.5**	81.43
Our IRIS (% of training data)				
IRIS (10%)†	73.48	75.63	86.92	78.68
IRIS (30%)†	77.19*	79.28*	90.19	82.22
IRIS (50%)†	**81.15***	**82.06***	93.47	**85.56***

Table 4: Results of IRIS and baselines on VAST (zero-shot targets) for ZSSD. † denotes baselines re-run over 5 rounds; results for the remaining are taken directly from their original papers (Section 4.4). Standard deviations are in Table 10. $*$ indicates IRIS outperforms baselines† at $p < 0.05$ (paired t-test). [**bold**: best, underline: second best.]

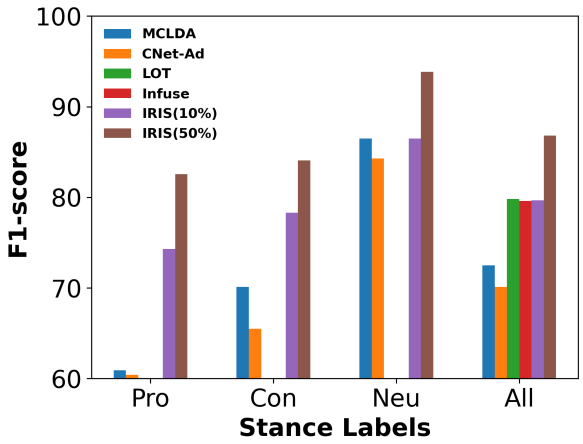

Figure 3: Results of IRIS with few-shot targets VAST

Llama 3.1 outperforms Mistral, showcasing its better alignment with the instructions compared to Mistral for the SD task. However, compared to other state-of-the-art methods, LLMs are still lagging (Tables 3, 4, and 5). Furthermore, we manually examined the 100 randomly selected correct and incorrect predictions of Llama from both datasets and prompted it to provide the best subsequence of words leading to the predicted stance (best implicit rationale), which resulted in a significantly lower F1 score of 0.617 compared to the ground-truth (refer Section 4.2 for analysis of 100 samples). However, we observed satisfactory results when we prompted Llama to generate all possible implicit rationales and linguistic-based explicit rationales (detailed in Section 4.2). This motivated us to take advantage of the Llama model for generating implicit rationales as a subsequence of words and linguistic arguments rather than using LLMs for the task of recognizing attitudes directly.

5.2 Comparison with Baselines (ZSSD)

VAST Table 4 indicates that our IRIS approach trained on 50% data outperforms other baselines with an overall averaged F1-score of 85.56 over 5 rounds of execution, resulting in an average improvement of 12.62% on the VAST for ZSSD. The standard deviation of IRIS and re-run baselines are reported in Table 10. We also report the results for different variants of our IRIS with 10% and 30% training data. Our IRIS (10%) also performs better than most baselines on VAST data with the exception of LKI-BART and Infuse, which either used LLM to analyze the author's implied emotions or infused target-related knowledge using the web. However, in contrast to these works, our IRIS prioritizes interpretability by automatically identifying relevant word subsequences and leveraging external documents to understand the contextual relationships. This helps in guiding the model toward accurate stances since we do not rely on ground truth for rationales. Thus, it surpasses LKI-BART and Infuse even with 30% and 50% training data. ***Few-Shot Stance Detection*** Figure 3 shows that our IRIS (50%) significantly outperforms the other baseline methods with an overall F1 result of 86.81 when tested with few-shot targets from VAST due to the efficient architecture guided by relevant rationale learning. It can be further seen that our IRIS trained on 10% data with F1 score of 79.68 outperforms the meta-contrastive learning-based MCDLA (72.5 F1) and commonsense based adversarial learning CNet-Ad (70.1 F1) and also results in comparable performance with Infuse (79.60 F1) and LOT (79.80 F1), thus highlighting the importance of our proposed ranking and selection stages that not only provide inherent interpretability, but lead the model to correct predictions, resulting in better performance even when trained with a smaller amount of data.

EZ Table 5 presents the average F1 scores and standard deviations of IRIS and all baselines over 5 runs. IRIS (50%) consistently outperforms all baselines across nearly all train→test target combinations and ranks second-best in one case, demonstrating strong and robust performance (Table 5). Training on mixed targets and evaluating on specific ones (M→N and M→C) results in better performance with 71.15 and 90.58 F1 scores, compared to cross-target settings such as N→C and C→N with 67.29 and 60.09 F1, thus reflecting similar trends observed in baselines. This indicates that IRIS also benefits from mixed-target training data that includes both noun and claim targets. Further, results for M→N and M→C are also comparable to those for N→N and C→C (Table 5), demonstrating IRIS's ability to generalize well under limited data. One exception is BART-MNLI, which outperforms IRIS on N→C, suggesting a need for more diverse training targets. This may be due to IRIS's reduced ability to transfer from noun-based training data to claim-based test inputs, possibly limited by incorrect rationales in relevance ranking stage. Nevertheless, IRIS achieves the best results on C→C (89.46 F1) and M→C (90.58 F1) when trained on claim and mixed targets

Train/Val	Mixed targets (M)			Noun-phrase targets (N)			Claim targets (C)		
Test	**M**	**N**	**C**	**M**	**N**	**C**	**M**	**N**	**C**
BART-MNLI	65.51/1.07	31.14/0.42	79.51/1.10	64.29/1.19	32.24/2.01	**81.45/0.16**	67.75/1.08	31.05/2.14	80.44/1.02
BART-MNLI-e	79.29/1.15	66.24/0.75	87.43/0.69	45.4/0.35	67.06/0.25	30.05/2.31	71.9/0.13	35.46/1.35	88.72/0.50
BART-MNLI-e$_p$	81.01/0.54	65.48/0.71	88.37/0.65	44.81/0.15	67.49/1.21	32.52/0.26	-	-	-
Our IRIS (% of training data)									
IRIS (10%)	75.37/1.49	60.68/2.16	82.03/2.24	65.05/1.31	67.53/1.11	58.13/0.72	66.28/0.56	50.22/2.45	79.42/2.25
IRIS (30%)	80.01/1.26	65.43/2.02	86.39/1.57	69.11/2.15*	71.34/1.15*	62.28/1.5	69.91/1.67	55.47/2.31*	85.38/1.19
IRIS (50%)	**82.18/0.75**	**71.15/1.11***	**90.58/2.09**	**72.38/2.03***	**74.32/1.02***	67.29/1.11	**75.50/1.05***	**60.09/0.35***	**89.46/1.63**

Table 5: Results (Average/Std. Dev) of IRIS and baselines on EZ for ZSSD. All methods are run over 5 rounds. $*$ denotes IRIS outperforms baselines at $p < 0.05$ with paired t-test. [**bold**: best, underline: second best.]

Model	VAST	EZ
Rational Gen. (RG) [Mistral]+ Relevance Ranking (RR)	73.02/2.49	55.69/2.13
RG [Llama] + RR	79.43/1.86	68.01/2.05
RG+RR+Grouping &Selection (GS)	83.06/1.47	71.52/1.65
RG+RR+GS +Classification(C) [Majority(Rel. Imp., Irr. Imp., Exp.)]	84.26/0.76	72.29/1.31
RG+RR+GS+C [Majority(Rel. Imp., Exp.)] (IRIS(50%))	85.56/0.33	74.32/1.02

Table 6: Results (Avg./Std. Dev) for different IRIS components on VAST (zero-shot) and EZ (noun-phrase) for ZSSD.

Model	VAST	EZ
Relevance Ranking (RR) (FlagReranker)	79.43/1.86	68.01/2.13
RR $\sim$ instruction	77.06/1.01	66.35/1.27
RR $\sim$ target	76.25/0.93	65.72/1.58
RR $\sim$ (Flag)+LLM scores	75.79/2.55	66.2/2.36
RR $\sim$ (Flag)+LLM as ranker	74.91/2.37	65.59/2.19
RR$\sim$ (Flag)+flashrank	75.42/0.84	73.50/1.22
RR$\sim$ (Flag)+rank-bm25	68.15/1.04	59.72/1.46
RR$\sim$ (Flag)+cosine sim	65.30/0.57	57.24/0.32

Table 7: Results (Avg./Std. Dev) of various ranking algorithms in ranking stage on VAST and EZ for ZSSD.

and evaluated on claim targets. Furthermore, even with only 10% or 30% of training data, IRIS outperforms all baselines in combinations such as N→M, N→N, and C→N, and exceeds the performance of BART-MNLI on various combinations of M→M, M→N, M→C, N→M, N→N, and C→N, highlighting the effectiveness of combining relevance ranking and linguistic features for stance detection when trained with a smaller amount of data.

5.3 Ablation Experiments

We conduct ablation studies to analyze the contribution of individual IRIS components, explore alternative ranking strategies, assess sensitivity to various parameters, and evaluate the impact of different embedding choices. For ease of comparison, we focus on the primary datasets of VAST with zero-shot targets and EZ with noun phrase targets as part of our ablation experiments. We report average F1 metrics and standard deviations over 5 runs.

Significance of Different IRIS Components Table 6 presents the ablation study that justifies the importance of the different components of our IRIS approach. It can be seen that the rationales generated by Llama lead to better performance than those generated by Mistral LLM. From Table 6, the addition of the "Grouping and Selection" component with the "Relevance Ranking" leads to an improvement of the overall F1 result by 4.57% and 5.16% on VAST and EZ datasets respectively because of segregating the relevant and irrelevant rationales and extracting the diverse relevant rationale. It is also shown that concatenating the linguistic evaluation of posts together with implicit relevant and irrelevant rationales improves IRIS performance with 84.26 (VAST) and 72.29 (EZ) F1 scores. Moreover, stance classification based on the majority vote of explicit and implicit relevant rationales further guides IRIS to classify attitudes effectively (refer Table 6).

Different Ranking Algorithms Table 7 represents the results when the relevance ranking (RR) stage is used exclusively for ZSSD without grouping and selection and explicit rationales. It is noted that our RR stage alone performs better than fine-tuned LLMs, proving the significance of our approach and justifying better performance than relying only on LLMs for ZSSD (Table 3). We find out that removing instruction and target from the query and document reduces task performance by 1.81% and 3.68% on average for both datasets. This suggests that the instruction helps the rankers understand the hidden requirement of aligning the query with the given target based on how the sentences in the documents address their targets. Moreover, the same sentences in datasets belong to different attitudes towards different targets, indicating the need to capture the target information in the query. We investigate LLM as a ranker providing relevance score of query w.r.t our 3 articulated documents and also LLM directly providing probabilistic scores of given query towards favor, against, and neutral stance (Prompt 6). From Table 7, FlagReranker outperforms LLM by leveraging its powerful architecture based on well-established language models and document relevance along with enhanced retrieval capability, thus benefitting the stance detection pipeline. We also explore the use of flashrank[6] cross-encoder model fine-tuned on Amazon shopping query dataset (model_name='ce-esci-MiniLM-L12-v2'), rank-bm25 (Brown 2020) and cosine similarity (torch.nn.functional.cosine_similarity) instead of FlagReranker that we used as a ranking algorithm, and find that the FlagReranker was more effective than others

[6]https://github.com/PrithivirajDamodaran/FlashRank

Rationales	Suff.	Comp.	Faith.	Plaus.
Implicit	4.23	3.55	4.36	3.49
Explicit	4.35	3.89	4.25	4.37

Table 8: Human evaluation of rationales extracted by IRIS

because our datasets contain similar statements for different targets leading to different attitudes, and these methods failed to capture the implicit and underlying relationships of the statements with the targets.

Sensitivity Analysis *Figure 7* plots the overall F1 score for different values of k, i.e. for the selection of k relevant and k irrelevant rationales extracted using the "diverse rationale selection fallback" component. It can be seen that the selection of k=3 for VAST and EZ achieves superior performance. Beyond this point, there is a noticeable decline in performance, as the average number of rationales retrieved per input is 6.25 for VAST and 4.15 for EZ, possibly leading to more irrelevant or empty subgroups during the grouping and selection stage of IRIS. *Figure 8* plots the overall F1 score for different values of β used as reward/punish metric for Rationale Usefulness Reward Punish loss function (L_{rp}). The parameter β controls the trade-off between rewarding useful rationales and punishing less useful or irrelevant ones. According to various experimental analyses, it is found that $\beta = 0.1$ performs best for both VAST and EZ and was therefore chosen as the hyperparameter for our model. This could be due to how β balances the reward and punishment dynamics in training. A value of β=0.1 encourages IRIS to explore and learn useful patterns without being overly constrained by penalties. If β is larger (e.g. $\beta = 0.2$), the model is rewarded more, but at the same time the penalty for selecting less useful rationales becomes stronger, which could discourage the model from exploring fewer specific areas of the data and potentially missing informative but noisy patterns. The observed result when $\beta = 0.1$ strikes a balance in scenarios with competing objectives (reward vs. punishment), ensuring robust learning of useful rationales.

Different Embeddings Appendix B and Table 13.

5.4 Evaluation of Rationales

To evaluate the interpretability of IRIS, we conduct both automatic and human evaluations of the implicit and explicit rationales extracted by IRIS (50%) for decoding stance.

Setup: We selected 100 random test samples from VAST (zero-shot) and EZ (noun-phrase). These are the same samples previously used to assess rationale quality from LLM, which initially motivated our use of more effective ranking algorithms within IRIS (Section 4.2). For *implicit rationales*, ground truth annotations were already created by annotators, framing this as a binary classification task: each token is labeled as 1 (part of rationale) or 0 (not part of rationale) (Section 4.2). We use the relevant implicit rationales extracted by IRIS after the grouping and selection stage for evaluation. For *explicit linguistic rationales*, the same team of 3 annotators performed the linguistic annotations. They were provided the same linguistic rationale-generation prompt as used for LLMs (Figure 5). Two doctoral researchers annotated 50 samples each, while a third postdoctoral researcher reviewed all annotations, requesting revisions where necessary to ensure high-quality labels.

Models	Trump	Biden	Sanders
In-target Stance Detection			
LOT	86.1	83.7	80.5
Infuse†	86.2/0.39	84.5/0.95	80.53/0.86
IRIS (50%)	**90.49/2.51***	**91.19/2.06***	**89.61/2.19***
Zero-Shot Stance Detection (ZSSD)			
COLA†	87.1/0.57	84.6/0.35	80.4/0.61
KAI†	72.45/2.05	85.16/1.51	78.69/1.73
IRIS (50%)	**87.98/1.15**	**89.05/1.08***	**85.77/0.82***

Table 9: Results (Average/Std. Dev) of IRIS (50%) and baselines on P-Stance for in-target and zero-shot stance. † indicates baselines are re-run over 5 rounds; results for the remaining are taken directly from their original papers (Section 4.4). $*$ denotes IRIS outperforms baselines† at $p < 0.05$.

Automatic Evaluation: *Implicit Rationales:* We evaluate IRIS extracted relevant implicit rationales against the ground truth using F1 as rationale identification is considered a binary classification task. IRIS achieves an F1 of 0.859, significantly outperforming Llama's score of 0.617 (Section 4.2), demonstrating the efficacy of IRIS in identifying the most relevant subsequences for interpretability. *Explicit linguistic Rationales:* We use BLEU (Papineni et al. 2002) and BERTScore (Zhang et al. 2020) to evaluate the quality of explicit rationales in IRIS. We observe BLEU 1, BLEU 2, BLEU 3, and BLEU 4 scores of 21.1, 18.54, 11.08, and 5.71, respectively, reflecting n-gram precision from unigrams to four-grams, resulting in an average BLEU score of 14.10, while a BERTScore of 84.01 indicates strong semantic alignment between IRIS-generated and ground-truth rationales. Please note that we simply use LLM-generated linguistic measures as explicit rationales in IRIS, reviewed by annotators (Section 4.2), showing that while LLMs may struggle with classification, they are effective in generating high-quality textual explanations.

Human Evaluation A separate team of annotators consisting of 3 master students with a computational social science background rated IRIS-generated implicit and explicit rationales on a 5-point Likert scale across four criteria: *Sufficiency* (how well it justifies the prediction), *Comprehensiveness* (coverage of relevant aspects), *Faithfulness* (alignment with the model's true reasoning), and *Plausibility* (how convincing it is to a human), where 1 indicates poor and 5 indicates strong alignment. As shown in Table 8, both rationale types scored similarly on sufficiency and comprehensiveness. Explicit (linguistic) rationales were more plausible due to their detailed explanations, while implicit rationales were rated higher in faithfulness, better reflecting the model's reasoning (Table 8). This highlights the complementary interpretability of both rationale types in IRIS.

5.5 Generalizability Analysis

Tables 9 and 12 demonstrate that IRIS, when trained on 50% of the data, outperforms all baselines on both P-Stance (in-target and zero-shot) and RFD (in-target), respectively. On

P-Stance, IRIS improves F1 scores by 11.29% and 7.83% in in-target and zero-shot stance tasks. Notably, the zero-shot version of our IRIS surpasses even in-target baselines across all domains of P-Stance, highlighting the strength of our relevance ranking and rationale grouping stages, which enable effective rationale selection even without access to ground-truth labels. Similarly, on RFD (Table 12), IRIS outperforms all baselines, demonstrating its ability to manage longer contextual articles by leveraging linguistic cues and identifying the most relevant subsequences to determine the correct stance.

5.6 Qualitative Analysis

To better understand IRIS's predictive capabilities for stance detection, we examine case studies in Table 14. Examples 1–3 demonstrate correct predictions across different datasets, supported by relevant implicit rationales (subsequences of text) and explicit rationales (communicative text features). Example 3 from Table 14 shows that IRIS correctly identifies the implicit rationale, in contrast to the XSD_p approach proposed in (Saha, Lakshmanan, and Ng 2024), which misidentifies the relevant rationale for the same instance. Due to space constraints, we could not include Figure 7 that illustrates this error in the XSD_p method (refer (Saha, Lakshmanan, and Ng 2024)). This example further highlights the strength of IRIS in handling longer and more complex texts, such as those in RFD, where multiple statements may imply different stances. IRIS effectively extracts the top k most relevant implicit rationales during the grouping and selection stage to guide accurate classification—where k is 7 for RFD due to its longer text length, compared to 3 for other datasets. Examples 4 and 5 illustrate error scenarios. In Example 4, IRIS focuses on positive sentiment toward "Smolensk" while referencing negative actions by others, resulting in a "favor" prediction. In Example 5, although the predicted stance is correct, the implicit rationales include redundancy and miss the most relevant justification when compared with human annotations. However, explicit rationales still provide sufficient reasoning, reinforcing the value of using both types of rationales in IRIS for accurate and interpretable stance prediction. To further understand the benefits of our prompts and IRIS approach, we looked into step-by-step processing for one of the sample predictions from VAST (Example 1 of Table 14) by LLMs only and our IRIS model, detailed in Appendix C.

6 Conclusion

In this work, we present a novel zero-shot stance detection system that offers an interpretable understanding of the attitude of the input by focusing on implicit sequence-based and explicit linguistic-based rationales. Our approach understands the context of the rationales toward the different stances while ensuring that the relevant rationales are captured to guide the model in correcting stance predictions. Extensive experiments on various benchmark datasets using 50%, 30%, and even 10% training data validate the significance of the different components of our approach. By making our framework implicitly and explicitly interpretable, we ensure that the model is usable in complex environments like online content moderation, and social media monitoring, and help identify biased rationales in sensitive areas such as politics or climate change to ensure the ethical use of AI. In the future, we plan to explore these advances and focus on the interpretability of low-resource stance detection.

Acknowledgments

This work is partially supported by the research project RECITALS, funded by the European Commission with grant agreement number 101168490.

References

Allaway, E.; and McKeown, K. R. 2020. Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP*. Association for Computational Linguistics.

Bergstra, J.; Yamins, D.; and Cox, D. 2013. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *International conference on machine learning*, 115–123. PMLR.

Boyd, R. L.; Ashokkumar, A.; Seraj, S.; and Pennebaker, J. W. 2022. The development and psychometric properties of LIWC-22. *Austin, TX: University of Texas at Austin*, 10.

Brown, D. 2020. Rank-BM25: A Collection of BM25 Algorithms in Python. *https://pypi.org/project/rank-bm25/*.

Chen, J.; Zhou, P.; Hua, Y.; Loh, Y.; Chen, K.; Li, Z.; Zhu, B.; and Liang, J. 2024. FinTextQA: A Dataset for Long-form Financial Question Answering. *arXiv preprint arXiv:2405.09980*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the North American chapter of the association for computational linguistics: human language technologies*.

DeYoung, J.; Jain, S.; Rajani, N. F.; Lehman, E.; Xiong, C.; Socher, R.; and Wallace, B. C. 2020. ERASER: A Benchmark to Evaluate Rationalized NLP Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL*.

Ding, D.; Dai, G.; Peng, C.; Peng, X.; Zhang, B.; and Huang, H. 2024a. Distantly Supervised Explainable Stance Detection via Chain-of-Thought Supervision. *Mathematics*, 12(7).

Ding, D.; Dong, L.; Huang, Z.; Xu, G.; Huang, X.; Liu, B.; Jing, L.; and Zhang, B. 2024b. EDDA: A Encoder-Decoder Data Augmentation Framework for Zero-Shot Stance Detection. *arXiv preprint arXiv:2403.15715*.

Guo, M.; Jiang, X.; and Liao, Y. 2024. Improving Zero-Shot Stance Detection by Infusing Knowledge from Large Language Models. In *International Conference on Intelligent Computing*, 121–132. Springer.

Hardalov, M.; Arora, A.; Nakov, P.; and Augenstein, I. 2021. Cross-Domain Label-Adaptive Stance Detection. In Moens, M.; Huang, X.; Specia, L.; and Yih, S. W., eds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP*. ACL.

Hu, K.; Yan, M.; Chong, W. H.; Yap, Y. K.; Guan, C.; Zhou, J. T.; and Tsang, I. W. 2024. Ladder-of-thought: Using knowledge as steps to elevate stance detection. In *2024 International Joint Conference on Neural Networks (IJCNN)*.

Kumar, S. 2024. Overriding Safety protections of Open-source Models. *arXiv preprint arXiv:2409.19476*.

Lan, X.; Gao, C.; Jin, D.; and Li, Y. 2024. Stance detection with collaborative role-infused llm-based agents. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, 891–903.

Lee, C.; Roy, R.; Xu, M.; Raiman, J.; Shoeybi, M.; Catanzaro, B.; and Ping, W. 2024. NV-Embed: Improved Techniques for Training LLMs as Generalist Embedding Models. *arXiv preprint arXiv:2405.17428*.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Li, Y.; Sosea, T.; Sawant, A.; Nair, A. J.; Inkpen, D.; and Caragea, C. 2021. P-stance: A large dataset for stance detection in political domain. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*.

Ma, J.; Wang, C.; Xing, H.; Zhao, D.; and Zhang, Y. 2024. Chain of Stance: Stance Detection with Large Language Models. In *Natural Language Processing and Chinese Computing - 13th Conference, NLPCC*. Springer.

Motyka, D.; and Piasecki, M. 2024. Target-Phrase Zero-Shot Stance Detection: Where Do We Stand? In *International Conference on Computational Science*. Springer.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*.

Reimers, N.; and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Saha, R. R.; Lakshmanan, L. V.; and Ng, R. T. 2024. Stance Detection with Explanations. *Computational Linguistics*.

Schiller, B.; Daxenberger, J.; and Gurevych, I. 2021. Stance detection benchmark: How robust is your stance detection? *KI-Künstliche Intelligenz*, 1–13.

Sepúlveda-Torres, R.; Vicente, M.; Saquete, E.; Lloret, E.; and Palomar, M. 2021. Exploring summarization to enhance headline stance detection. In *International Conference on Applications of Natural Language to Information Systems*, 243–254. Springer.

Singh, C.; Inala, J. P.; Galley, M.; Caruana, R.; and Gao, J. 2024. Rethinking interpretability in the era of large language models. *arXiv preprint arXiv:2402.01761*.

Upadhyaya, A.; Fisichella, M.; and Nejdl, W. 2023a. Intensity-Valued Emotions Help Stance Detection of Climate Change Twitter Data. In *IJCAI*, 6246–6254.

Upadhyaya, A.; Fisichella, M.; and Nejdl, W. 2023b. A multi-task model for emotion and offensive aided stance detection of climate change tweets. In *Proceedings of the ACM Web Conference 2023*, 3948–3958.

Upadhyaya, A.; Fisichella, M.; and Nejdl, W. 2023c. Toxicity, morality, and speech act guided stance detection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 4464–4478.

Upadhyaya, A.; Nejdl, W.; and Fisichella, M. 2024. Harnessing empathy and ethics for relevance detection and information categorization in climate and covid-19 tweets. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 4091–4095.

Upadhyaya, A.; Nejdl, W.; and Fisichella, M. 2025a. Enhancing Online Climate Discourse: A Two-Stage Framework for Climate Content Categorization and Moderation. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, 9890–9898.

Upadhyaya, A.; Nejdl, W.; and Fisichella, M. 2025b. Interpretable zero-shot stance detection with proactive content intervention. *Information Processing & Management*, 62(6): 104223.

Wang, C.; Zhang, Y.; and Wang, S. 2024. A meta-contrastive learning with data augmentation framework for zero-shot stance detection. *Expert Systems with Applications*, 123956.

Xiao, S.; Liu, Z.; Zhang, P.; Muennighoff, N.; Lian, D.; and Nie, J.-Y. 2023. C-pack: Packaged resources to advance general chinese embedding. *arXiv preprint arXiv:2309.07597*.

Yan, M.; Joey, T. Z.; and Ivor, W. T. 2024. Collaborative knowledge infusion for low-resource stance detection. *Big Data Mining and Analytics*, 7(3): 682–698.

Yang, X. 2024. Diagnosing Hate Speech Classification: Where Do Humans and Machines Disagree, and Why? *arXiv preprint arXiv:2410.10153*.

Yao, Z.; Yang, W.; and Wei, F. 2024. Enhancing Zero-Shot Stance Detection with Contrastive and Prompt Learning. *Entropy*, 26(4): 325.

Zhang, B.; Ding, D.; Huang, Z.; Li, A.; Li, Y.; Zhang, B.; and Huang, H. 2024a. Knowledge-Augmented Interpretable Network for Zero-Shot Stance Detection on Social Media. *IEEE Transactions on Computational Social Systems*.

Zhang, H.; Li, Y.; Zhu, T.; and Li, C. 2024b. Commonsense-based adversarial learning framework for zero-shot stance detection. *Neurocomputing*, 563: 126943.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2020. BERTScore: Evaluating Text Generation with BERT. In *8th International Conference on Learning Representations, ICLR*.

Zhang, Z.; Li, Y.; Zhang, J.; and Xu, H. 2024c. LLM-Driven Knowledge Injection Advances Zero-Shot and Cross-Target Stance Detection. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, 371–378.

Zhao, C.; and Caragea, C. 2024. EZ-STANCE: A Large Dataset for English Zero-Shot Stance Detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Zhao, X.; Ma, G.; Pang, S.; Guo, Y.; Zhao, J.; and Miao, J. 2024. Zero-shot stance detection based on multi-expert collaboration. *Scientific Reports*, 14(1): 18092.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, we focus on developing a system that provides an interpretable understanding of the attitudes of the post. We use public datasets that do not violate any rights and do not imply disrespect for other cultures.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes, we clearly highlight the work's objective and contribution in the abstract and introduction.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, we described the methodology in Section 3 and further state the implementation details in Section 4.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? NA, as we used publicly available datasets for zero-shot stance task and did not require population-specific distributions.
   (e) Did you describe the limitations of your work? Yes, in Section Limitations.
   (f) Did you discuss any potential negative societal impacts of your work? Yes, we clearly mention the limitations of our work in Sections Limitations and Ethics Statement.
   (g) Did you discuss any potential misuse of your work? Yes, we briefly addressed the potential misuse of our work in Sections Limitations and the Ethics Statement, particularly in cases where inaccurate attitude predictions could unduly influence public opinion or decision-making processes in critical scenarios. However, these issues could be addressed by further improving the performance of our relevance ranking stage to steer the model in the right direction and ensure better interpretability.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, we share the required code publicly and utilize public datasets only without any personally identifiable information.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes, we include URL for the code and mention all implementation details in Section 4.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes, we include all details in Section 4.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes, we ran 5 independent runs of our model to account for variability and report average metric scores.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes, we included all these details in Section 4.
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes, we have justified all details in Section 5 and demonstrated the significance of our approach on both the datasets used in our work together with various other experiments.
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes
   (b) Did you mention the license of the assets? Yes
   (c) Did you include any new assets in the supplemental material or as a URL? We provide the URL for the code.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? No, because we only use open-sourced datasets.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, we mention this information in Section 4.

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? NA

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? NA

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? We have used annotators for a very small subset of 100 samples. Details are found in Section 4.2.

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? We have used annotators for a very small subset of 100 samples from public datasets to investigate the performance of LLMs qualitatively. The annotators were doctoral and postdoctoral researchers of our team and did the job within their normal working hours. These details are specified in Section 4.2.

(d) Did you discuss how data is stored, shared, and deidentified? NA

Limitations

Despite the considerable success we have achieved with our study, we must acknowledge the limitations of our work for future refinement. We have used datasets that are available in English. We may need to explore other embeddings and pre-trained rankers if used for other languages. However, the basis of our approach, namely using relevance ranking and filtering out relevant and irrelevant diverse reasoning might still be suitable for multilingual or low-resource stance detection after we have optimized some embeddings. As we focus on target-based stance, we can expand our scope to target-independent and domain-based stance detection. Further enhancement of the quality of our external knowledge base with higher quality assertions of stance can improve the task. It might be possible that incorrect predictions can influence public opinion or decision-making processes in critical social issues, for instance, our approach can be used by political parties during their election campaigns to regularly monitor how public opinion on social issues such as COVID or climate change has shifted either for or against them in order to gain an advantage for their own interests. Therefore, in our work, we aim to achieve interpretability. However, further improving the performance of the relevance ranking stage could steer the model in the right direction. Consequently, future work should consider these shortcomings and work towards the refinement process.

Ethics Statement

All the datasets that we utilize for this research are open-access datasets. The VAST and EZ-Stance datasets provide full-text data directly.

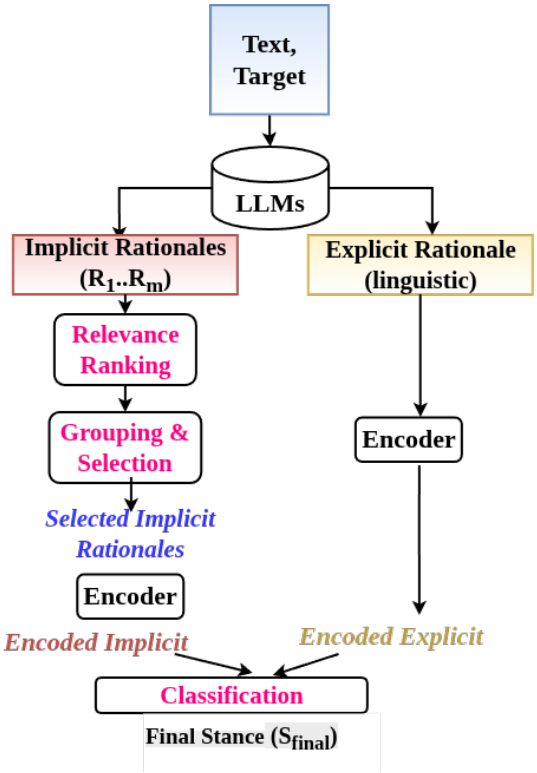

Figure 4: Flow diagram of IRIS.

We acknowledge the potential risks associated with the misuse of our technology, as is the case with many innovations. Specifically, there is a possibility that it could be leveraged for unethical purposes, such as suppressing critics or identifying and targeting dissenting voices on social media by certain entities. We strongly encourage users of our technology to adhere to principles of ethical and responsible use.

A Computation Time

Table 11 presents the training times of different stages of IRIS (50%) compared to baseline methods. Since all baselines leverage external resources—such as Wikipedia

Models	VAST (Zero-shot)			
	Pro	Con	Neu	All
COLA	0.42	0.78	1.01	0.45
EDDA-GPT	0.29	0.11	0.64	0.24
EDDA-LLaMA	0.52	0.33	0.29	0.39
KAI	1.44	1.61	2.13	1.01
LKI-BART	1.17	0.61	0.37	0.79
Infuse	0.42	1.35	0.54	0.54
Our IRIS (% of training data)				
IRIS (10%)	3.10	3.09	2.23	1.68
IRIS (30%)	2.01	1.55	1.69	1.02
IRIS (50%)	0.77	0.36	0.50	0.33

Table 10: Standard deviation of IRIS and baselines run over 5 rounds on VAST with zero-shot targets.

Prompt for Explicit Rationale

1: Consider the following post and the target.
2: ***Post:*** '{text}'.
3: ***Target:*** {target}.

4: Based on the given post, assess the post on the following ***linguistic characteristics*** in the language of the post.

5: ***Empathy*****:** Language that indicates concern for others and sympathy. Reflects shared distress and interest in others' thoughts and feelings.
6: ***Allure:*** A measure derived from advertising that reflects language that is intended to persuade or attract. Includes words that are attention-grabbing and stimulate people's needs and desires.
7: ***Absolutist:*** Language that reflects black-and-white, all-or-nothing thinking. Indicates cognitive rigidity and dislike of ambiguity.
8: ***Action:*** Language related to increasing physical or mental activity; aiming to do something or do more.
9: ***Abstract:*** Language that reflects nonspecific, amorphous, or big-picture ideas. Indicates a higher level of construal.
10: ***Concrete:*** Language that reflects specific or tangible actions, objects, or traits.
11: ***Agency_language:*** Language that suggests a person is exerting willpower to pursue personal goals. Indicates doing things as an individual for personal motivations or desires.
12: ***Communion_language:*** Language that suggests a person is cooperating and connecting with others to improve social relationships. Indicates they are likely doing things with other people to help meet the group's goals.
13: ***Approach:*** Language related to emotions that motivate people to move towards an emotional trigger.

14: Write a summarized response in 3-5 lines.

Figure 5: LLM Prompt for Explicit Rationale

Prompt for Implicit Rationale

1: Consider the following post and the target.
2: ***Post:*** '{text}'.
3: ***Target:*** {target}.

4: Extract ***all possible rationales*** from the post that are subsequences of words leading to favor, against, or neutral stances toward the target '{target}'.
5: For each rationale, provide the ***probabilities of favor, against, and neutral*** stances towards the target, formatted as follows:
6: [[rationale1, favor_prob1, against_prob1, neutral_prob1], ... [rationaleN, favor_probN, against_probN, neutral_probN]].

7: Only return the list in this exact format. Do not include any additional text or explanations.

Figure 6: LLM Prompt for Implicit Rationale

Algorithm 1: Diverse Rationale Selection Fallback

Dynamic Target Distribution:

Let N_f, N_a, N_n be the number of available rationales in the favor, against, and neutral groups, respectively.

The total number of rationales in the group (relevant or irrelevant) is: $N_t = N_f + N_a + N_n$

The target distribution V is computed as $V = (\frac{N_f}{N_t}, \frac{N_a}{N_t}, \frac{N_n}{N_t})$

KL-Divergence Based Selection

Inputs: G: set of relevant/irrelevant rationales; k: the number of relevant and irrelevant rationales to be selected; V: Target distribution for diversity across subgroups.

Outputs: G_k: A set of k selected rationales with diverse coverage over the subgroups.

Set Initial Distribution: Since no rationales are initially selected, the current choice distribution is initialized with $U_{current} = (\epsilon, \epsilon, \epsilon)$, where $\epsilon = 10^{-6}$ to avoid division by zero in the KL-Divergence calculation.

Initialize selected rationales set: $G_k = \{\}$

while $|G_k| < k$ **do**

  for candidate $r \in \mathcal{R}$ **do**

    Add the candidate rationale r to the selected set temporarily: $U_{new} = U_{current}$ + update counts of favor, against, neutral subgroups.

    Calculate KL-divergence between the new distribution and the target distribution: $D_k(U_{new}||Q) = \sum_{j \in \{f,a,n\}} U_{new}(j) log(\frac{U_{new}(j)}{V(j)})$

    Select the rationale r^* that that minimizes the KL-divergence: $r^* = \text{argmin}_r\, D_k(U_{new}||V)$

    Update the selected rationales: $G_k = G_k \cup r^*$

    Update the distribution $U_{current}$ based on the selected rationale's subgroup (favor, against, neutral)

    If a subgroup has no rationales left, skip it, but continue selecting from other subgroups.

  end for

  Fallback: If it is not possible to maintain diversity (e.g., one of the subgroups is empty), continue selecting from the available groups to reach the total count k.

end while

return G_k.

Process is repeated for both relevant and irrelevant sets of rationales.

Here the output G_k will be denoted by Re_k in case of k relevant $Re_1, Re_2 ... Re_k$ set of implicit rationales and Ie_k for the case of k irrelevant $Ie_1, Ie_2 ... Ie_k$ implicit rationales. =0

Model	Training Time
Infuse	32.66 (1960)
KAI	26.38 (1583)
EDDA	13.25 (795)
LKI-BART	23.13 (1388)
Rational Generation (RG)	
+Relevance Ranking(RR)	18.41 (1105)
RG+RR+Grouping & Selection(GS)	26.28 (1577)
RG+RR+GS+Classification	34.73 (2084)

Table 11: Training time [minutes (seconds)] of IRIS (50%) and baselines on VAST with zero-shot targets.

Model	Macro F1
Mole+XSD$_p$	0.69
MTDNN+XSD$_p$	0.52
HSD+XSD$_p$	0.40
Infuse†	0.67/2.02
IRIS (50%)	0.73/1.69*

Table 12: Results (Average/Std. Dev) of IRIS (50%) and baselines on RFD for in-target stance. † indicates baselines are re-run over 5 rounds; results for the remaining are taken directly from their original papers (Section 4.4). $*$ denotes IRIS outperforms baselines† at p $<$ 0.05 (paired t-test).

or LLMs—for information retrieval, similar to our rationale generation stage, we treat these steps as preprocessing and exclude them from the reported training times for fairness. All models were run on our GPU setup, as detailed in Section 4.3). While IRIS generally requires more training time compared to other baselines (with comparable runtime to the Infuse method), it consistently delivers superior performance. This highlights a reasonable trade-off between computational cost and predictive effectiveness. Additionally, the training time of IRIS can be further optimized through parallel processing or more powerful GPU infrastructure—options we intentionally limited due to server availability constraints and our commitment to reducing carbon emissions.

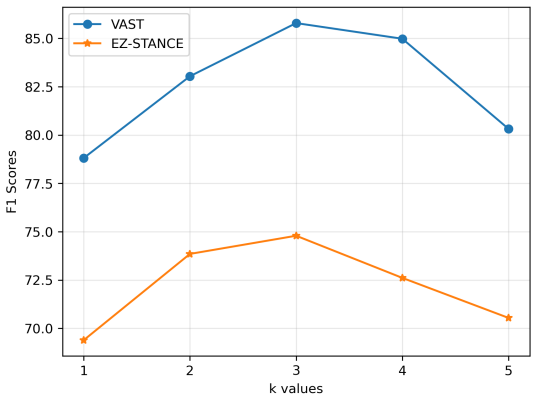

Figure 7: Different 'k' values Performance for ZSSD

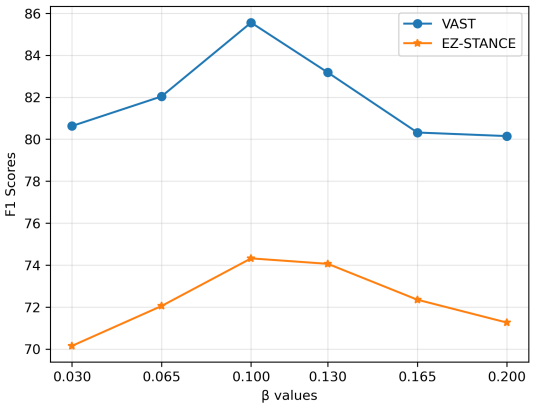

Figure 8: Different 'β' values Performance for ZSSD

Model	VAST	EZ
NV-Embed-v2	85.56/0.33	74.32/1.02
bge-base-en-v1.5	83.21/0.45	71.66/0.29
stella_en_1.5B_v5	83.09/2.05	72.31/2.41
Sentence-BERT	79.63/2.53	68.28/2.39
BART	75.50/3.17	65.45/3.03
BERT	76.36/3.09	66.18/3.20

Table 13: Results (Average/Std. Dev) for different embeddings on VAST (zero-shot) and EZ (noun-phrase) for ZSSD.

B Different Embeddings

We evaluate our approach with different sentence embeddings using HuggingFace MTEB Leaderboard[7], recent works (Lee et al. 2024) and other popular word embeddings. NV-Embed-v2 (https://huggingface.co/nvidia/NV-Embed-v2) leverages a decoder-only Mistral LLM with latent attention pooling and a high embedding dimension 4096. bge-base-en-v1.5 (https://huggingface.co/BAAI/bge-base-en-v1.5) is widely used for retrieval and ranking tasks, offering robust support for longer texts with a 768-dimensional embedding. "stella_en_1.5B_v5" (https://huggingface.co/NovaSearch/stella_en_1.5B_v5) is based on Alibaba-NLP's gte-large-en-v1.5 and gte-Qwen2-1.5B-instruct huggingface models, providing embeddings with a default dimension of 1024. We also evaluate popular baseline embeddings such as Sentence-BERT (Reimers and Gurevych 2019), BART (Lewis et al. 2019), and BERT (Devlin et al. 2019). It can be observed from Table 13 that "NV-Embed-v2" embeddings excel at capturing underlying relationships more effectively than other embeddings. They demonstrate superior performance in identifying implicit relationships, subtle semantic nuances, and contextual cues, which are critical for tasks such as stance detection.

C Qualitative Analysis

To understand the benefits of our prompts and IRIS approach, we looked into step-by-step processing for one of the sample predictions from the VAST dataset by LLMs only and our model: *Target:* "Justice Ginsburg"; *Text:* "If any of the dire prediction of a Trump presidency turning into a fascist dictatorship should start to occur... I do not find Justice Ginsburg's comments any less dignified..justices were to condemn Ms. Clinton's trustworthiness over her lies... The dignity of the court.. is lessened when..political maelstrom."
True Stance: "Against";
LLM stance prediction (Llama 3.1 zero/few/fine-tuned): "Neutral"

However, when asked Llama using prompt 6 to extract implicit rationales and prompt 5 for linguistic measures, we obtain the below implicit and explicit rationales:
Implicit rationales (IR_i): [IR_1: "I do not find Justice Ginsburg's comments any less dignified", IR_2: "less dignified than if one of the more conservative justices were to condemn Ms. Clinton's trustworthiness", IR_3: "The dignity of the court, as well as the appearance of impartiality, is lessened when the justices lower themselves into the political maelstrom.", IR_4: "than if one of the more conservative justices were to condemn", IR_5: "I do not find Justice Ginsburg's comments"]
Explicit rationale (ER): "The post exhibits low empathy, as it criticizes Justice Ginsburg's comments without considering her perspective or feelings. The language is also relatively concrete, focusing on specific events and actions rather than abstract ideas. There is a moderate level of agency language, as the author argues for justices to speak out and take action. However, there is no apparent allure or absolutist language in the post"

After passing through *relevance ranking* stage:
Corresponding ranking scores for implicit rationales $\{S_i^{rf}, S_i^{ra}, S_i^{rn}\}$*:* [["0.85", "0.05", "0.10"], ["0.05", "0.75", "0.20"], ["0.25", "0.55", "0.20"], ["0.3", "0.2", "0.5"], ["0.4", "0.3", "0.3"]]

Grouping and selection stage result in the following set of relevant rationales:
Relevant rationales:IR_2: "less dignified than if one of the more conservative justices were to condemn Ms. Clinton's trustworthiness", IR_3: "The dignity of the court, as well as the appearance of impartiality, is lessened when the justices lower themselves into the political maelstrom"; IR_5: "I do not find Justice Ginsburg's comments any less dignified"
Stance prediction of the concatenation of each relevant implicit and explicit implicit:(IR_1, ER): Against; (IR_2, ER): Against; (IR_3, ER): Favor
Final Stance by IRIS (majority vote): "Against"

This justifies that our IRIS was able to focus on the relevant rationales and the concatenation of explicit and implicit rationales followed by a majority vote leads to a correct prediction of the attitude towards the given target.

[7]https://huggingface.co/spaces/mteb/leaderboard

S.No.	Text+Target	Stance	Implicit Rationale	Explicit Rationale
1.	**VAST:** If any of the dire prediction of a Trump presidency turning into a fascist dictatorship should start to occur... I do not find Justice Ginsburg's comments any less dignified..justices were to condemn Ms. Clinton's trustworthiness over her lies... The dignity of the court.. is lessened when..political maelstrom. **Target:** Justice Ginsburg	**True:** against **Predicted:** against	"less dignified than if one of the more conservative justices were to condemn Ms. Clinton's trustworthiness"; "The dignity of the court, as well as the appearance of impartiality, is lessened when the justices lower themselves into the political maelstrom"; "I do not find Justice Ginsburg's comments any less dignified"	The post exhibits low empathy, as it criticizes Justice Ginsburg's comments without considering her perspective or feelings. The language is also relatively concrete, focusing on specific events and actions rather than abstract ideas. There is a moderate level of agency language, as the author argues for justices to speak out and take action. However, there is no apparent allure or absolutist language in the post.
2.	**EZ:** My girl is a dual Saudi American. In women s rights activists get detained tortured see: LoujainHathloul. We escaped saudi made it to the USA in 2019. My daughters teacher texted me today saying she is leading girls rights protests at school I AM SO PROUD!!!! **Target:** women's rights activists	**True:** favor **Predicted:** favor	"We escaped saudi made it to the USA in 2019"; "she is leading girls rights protests"; "I AM SO PROUD"	The post exhibits high levels of Emotion (specifically, pride and concern), Agency_language (the author is proud of their daughter taking action), and Concrete language (specific events and actions are described). There is also a sense of Approach language, as the author is motivated by positive emotions towards their daughter's actions. The tone is enthusiastic and celebratory.
3.	**RFD:**When I got admitted to Yale...African-Americans rarely end up in elite institutions...Racism is not a zero-sum game, but opportunities for limited resources are...Given how much worse off African-Americans are than whites...President Obama's "yes, we can" fervor seems to be contagious... affirmative action, U.S. Supreme Court Justice Antonin Scalia wrote "there are no debtor or creditor races." But he was wrong. African-Americans still are owed big time. **Traget:** Anti-White Bias Is a Problem.	**True:** against **Predicted:** against	"But, lest anyone worry, white folks, comparatively speaking, are doing just fine"; "Blacks are twice as likely to be unemployed"; "Six African-Americans head Fortune 500 companies. Of those businesses, 480 are run by whites"; "We have one black president, and almost one million black people in prison"; "African-Americans still are owed big time."; "the most surprising thing is that many blacks believe there has been so much racial progress"	The post exhibits low empathy, as it challenges the target statement and presents a opposing view. The language used has a high allure, aiming to persuade readers of the author's perspective.Absolutist language is present, with statements like "Racism is not a zero-sum game." The text promotes action, encouraging readers to acknowledge and address racism. Abstract ideas are discussed, such as the concept of racism and its impact on society. Concrete examples, like statistics on unemployment and prison rates, support the author's claims. The language used is more agentic, with the author presenting their personal perspective and experiences. Overall, the post takes an approach of confronting and challenging the target statement.
4.	**EZ:** On June 13, 1611, Smolensk fell to the Poles after the Russians defended it for 20 months. This occurred during the 1598-1613 Time of Troubles when Poland took advantage of Russia's weakness by engaging in regime change and military campaigns such as the occupation of Moscow. **Target:** Smolensk	**True:** neutral **Predicted:** favor	"On June 13, 1611, Smolensk fell to the Poles"; "Russians defended it for 20 months"; "Poland took advantage of Russia's weakness	The post demonstrates Concrete language through historical specifics like "Smolensk fell," "20 months," and "occupation of Moscow." There's limited Empathy or Communion_language, as the tone is factual and detached. Absolutist and Allure elements are absent, however, a little focus is on invoking motivation or ideology.
5.	**VAST:** People used to have to have children in their teens and 20s because people didn't live very long...Maybe it is just natural part of human evolution as people have become more educated and science.. **Target:** age childbirth	**True:** neutral **Predicted:** neutral	Missing Relevant Rationales: **by IRIS:** "people didn't live very long"; 'now people live a lot longer so there is no hurry"; 'science has solved many of the problems of dieing off" **by Human:** "people live a lot longer so there is no hurry"; "many died in childhood, modern medicine has put an end to the need for that"; "science has solved many of the problems of dieing off before life could be fully lived"	Based on the linguistic characteristics of the post, the post has low Empathy and Absolutist language, moderate Abstract and Concrete language, and high Agency_language and Approach language. The tone is informative and neutral, with no emotional appeal or attempt to persuade. The language suggests a focus on individual perspective and personal goals rather than social relationships or collective well-being.

Table 14: Case studies of IRIS stance prediction including stance labels and rationales extraction.