Fine-Grained Perspectives: Modeling Explanations with Annotator-Specific Rationales

Olufunke O. Sarumi[1], **Charles Welch**[2], **Daniel Braun**[1]
[1]Marburg University, Marburg, Germany
[2]McMaster University, Hamilton, Ontario, Canada
sarumio,daniel.braun@uni-marburg.de, cwelch@mcmaster.ca

Abstract

Beyond exploring disaggregated labels for modeling perspectives, annotator rationales provide fine-grained signals of individual perspectives. In this work, we propose a framework for jointly modeling annotator-specific label prediction and corresponding explanations, fine-tuned on the annotators' provided rationales. Using a dataset with disaggregated natural language inference (NLI) annotations and annotator-provided explanations, we condition predictions on both annotator identity and demographic metadata through a representation-level User Passport mechanism. We further introduce two explainer architectures: a post-hoc prompt-based explainer and a prefixed bridge explainer that transfers annotator-conditioned classifier representations directly into a generative model. This design enables explanation generation aligned with individual annotator perspectives. Our results show that incorporating explanation modeling substantially improves predictive performance over a baseline annotator-aware classifier, with the prefixed bridge approach achieving more stable label alignment and higher semantic consistency, while the post-hoc approach yields stronger lexical similarity. These findings indicate that modeling explanations as expressions of fine-grained perspective provides a richer and more faithful representation of disagreement. The proposed approaches advance perspectivist modeling by integrating annotator-specific rationales into both predictive and generative components.

Keywords: Explanation, Perspectives, Annotator

1. Introduction

Perspectivist NLP argues that annotations should reflect the specific judgments of individual annotators rather than converge on a single consensus label (Pavlick and Kwiatkowski, 2019). In tasks such as natural language inference (NLI), stance detection, and hate speech classification (Xu et al., 2024), it is legitimate for annotators to disagree due to differences in background, interpretation, or socio-demographic perspective. Modeling such disagreement has become an important focus of recent shared tasks (Leonardelli et al., 2023; Uma et al., 2021) and research initiatives, shifting the emphasis away from majority voting toward preserving variation.

Most perspectivist approaches implicitly model perspectives using linguistic and contextual signals such as sociodemographic information, user IDs, and group affiliations (Davani et al., 2022; Plepi et al., 2022). These signals are used to infer the potential sources of variation and diversity in annotations. However, beyond disaggregated labels, annotators' perspectives often remain abstracted and only indirectly represented in the model. Although some datasets require annotators to provide rationales for selecting a particular label (Weber-Genzel et al., 2024), these explanations are rarely integrated explicitly into perspectivist modeling. Incorporating annotator rationales enables more nuanced and fine-grained representations of perspective.

Explainability in perspectivist approaches has gradually emerged as a critical component of trustworthy NLP systems, as it supports model-level interpretability through the analysis of attention patterns or internal structures used to justify predictions (Mastromattei et al., 2022a). In recommendation systems, for example, natural language generation (NLG) methods have been proposed to generate flexible, free-text explanations based on user-generated content (Li et al., 2021b). While such approaches demonstrate the potential of generative models to produce fluent and varied explanations, they also expose limitations: generated content may be off-topic, insufficiently grounded in the input, repetitive Li et al. (2021a), or insufficiently personalized. These challenges highlight the need for controllable and faithful explanation generation, particularly when explanations are expected to reflect specific user or annotator viewpoints.

Within perspectivist NLP, explainability has been approached in different ways. Some studies treat it as post-hoc model interpretation Mastromattei et al. (2022a), identifying linguistic features or structural patterns that influence perspective-aware predictions (Muscato et al., 2025). Others rely on prompting strategies to simulate user perspectives in large language models (Hayati et al., 2024). However, relatively little work has explicitly modeled annotator-

Proceedings of NLPerspectives @ LREC 2026, pages 66–75
12 May 2026.

specific explanations alongside disaggregated labels, partly because few datasets contain both disagreement and distinct rationales.

In this study, we integrate perspectivist modeling with perspectivist explanation by explicitly conditioning explanation generation on annotator-specific representations. Using a dataset with disaggregated NLI labels and annotator-provided explanations, we model perspective in both label prediction and rationale generation. We explore a prompt-based post-hoc explainer and a representation-prefix bridge that transfers classifier representations enriched with annotator information into a generative model. In doing so, we treat explanations as expressions of perspective rather than merely post-hoc justifications of a model's decisions.

2. Related works

Perspectivist NLP aims to preserve the nuanced information hidden within disagreement by modeling annotator-specific labels rather than aggregating them into a single label (Cabitza et al., 2023). However, explainability within this paradigm remains relatively understudied and fragmented (Frenda et al., 2025). Existing research primarily approaches explainability either through model interpretability as in Mastromattei et al. (2022a) or by explicitly prompting Large Language Models (LLMs) for explanations (Orlikowski et al., 2025). However, most work has yet to explore annotator-specific rationales grounded in internal representations as a primary approach for perspectivist explainability.

2.1. Current Approaches to Perspective-Aware Explanations

One line of research addresses explainability in perspectivist models by identifying the linguistic components in Hate speech tasks with the use of recognizers that incorporates syntactic dependency trees to provide post-hoc justifications for classifications (Mastromattei et al., 2022a). In these instances, explainability focuses on revealing the mechanics of the model's prediction rather than capturing the annotator's subjective reasoning. Similarly, Mastromattei et al. (2022b) explored explainable syntax-based models within hate speech detection to identify trigger words that influence target classification. In a different vein, Nirmal et al. (2024) implicitly extracted user rationales from input text using LLMs to guide classifier outcomes, aiming for a more interpretable architectural framework.

2.2. Personalized Generation and Recommendation

A shift toward personalized explanation is evident in the work of Li et al. (2021b), who designed a specialized Transformer for explainable recommendation. This model utilizes user IDs and items alongside linguistic cues to generate recommendations and justifications that reflect individual user interests. Similarly, Li et al. (2020) utilized a neural template approach to address user ratings within recommender systems. More recently, Plepi et al. (2024) introduced twin-encoder architectures that separately encode auxiliary user information to facilitate perspective-taking in conflict situations. This allows the model to conceptualize user viewpoints through self-disclosure statements. While this approach structurally integrates user context, it does not explicitly disentangle annotator-specific explanatory reasoning in disaggregated datasets, where annotators might agree on a label but diverge significantly in their underlying logic. In this study, we address explainability through the lens of annotator rationales, seeking to understand the *why* behind a label from the human's perspective. Our approach models annotator perspectives at both the classification and explanation levels. Furthermore, we introduce a representation-level bridge that conditions explanation generation directly on annotator-specific internal representations. By doing so, we treat explanation not merely as a post-hoc interpretability tool, but as an explicit expression of annotators perspectives tied directly to disaggregated labels they represent.

3. Methods and data

We study perspectivism in generative explainability using the VariErrNLI dataset (Weber-Genzel et al., 2024), which contains disaggregated annotator labels and annotator-specific rationales. Unlike most existing disaggregated datasets, VariErrNLI preserves both label disagreement and explanation diversity, making it suitable for modeling fine-grained perspectives.

Our framework consists of two components: (i) an annotator-aware classifier that predicts label sets for each annotator, and (ii) an annotator-conditioned explainer that generates corresponding rationales. We explicitly model annotator identity using learned embeddings and metadata features, which are fused with the contextual representation of the input (context and statement) to produce annotator-specific predictions.

We compare two explanation approaches. The first is a post-hoc, prompt-based explainer that generates explanations from textual inputs. The second is a prefixed bridge explainer that conditions

generation on the classifier's internal annotator-specific representations. This allows the model to incorporate both predicted labels and underlying annotator-specific reasoning signals.

3.1. VariErrNLI Dataset

We use VariErrNLI (Variation vs. Error), a perspectivist NLI dataset designed to disentangle human label variation from annotation error. VariErrNLI contains approximately 500 NLI items sampled from ChaosNLI (MNLI subset) and annotated in two rounds by four independent annotators.

In Round 1, annotators assigned one or more NLI labels, Entailment (E), Neutral (N), or Contradiction (C) to each item and provided a one-sentence explanation for each label assigned, preserving fine-grained reasoning diversity. This round of annotation produced 1,933 label-explanation pairs.

In Round 2, annotators independently evaluated the validity of each label–explanation pair (including their own) by judging whether the explanation plausibly supports the assigned label. This second stage enables distinguishing plausible human label variation from annotation errors. The dataset, therefore, provides not only disaggregated labels and rationales but also meta-judgments about their validity.

Although VariErrNLI was originally designed to study annotation error versus variation, we use it for a different purpose. Specifically, we leverage its disaggregated labels and annotator-specific explanations to model and generate annotator-conditioned reasoning. For this study, we use the version released for the Learning with Disagreement (LeWiDi) 2025 Shared Task (Leonardelli et al., 2026), which provides predefined training, development, and test splits. The dataset statistics are presented in Table 1

3.2. Problem Formulation

We formalize annotator-specific prediction and explanation as a joint task. Each instance in the VariErrNLI dataset consists of a context c, a statement s, and annotations from annotators $a \in \mathcal{A}$. Each annotator provides a judgment (in some instances, multi-label) over the label set

$$\mathcal{L} = \{C, E, N\}, \tag{1}$$

corresponding to contradiction (C), entailment (E), and neutral (N); and an explanation that justifies their labeling decision. For each annotator a, there is an annotator-specific label

$$y_a \subseteq \mathcal{L}, \tag{2}$$

and an associated explanation r_a, where r_a is a short sentence describing the reasoning for y_a. Because two annotators can assign the same label for different reasons, we treat explanation generation as an explicitly perspectivist problem. Our model therefore has two goals: (i) predict the annotator-specific label set for each annotator a, and (ii) generate the corresponding annotator explanation r_a, which we define as the annotator's expressed perspective. For each instance (c, s) and annotator a, we learn an annotator-aware classifier and an annotator-conditioned explainer trained on the provided human rationales.

3.3. Annotator-Aware Classification

We implement the *User Passport* method to explicitly model annotator-specific perspectives within our classification framework (Sarumi et al., 2025), using DeBERTa-v3-base as the backbone encoder. This approach incorporates annotator identity and metadata directly at the representation level rather than through input text modification or token-based methods (Welch et al., 2022). The resulting classifier serves as the underlying prediction component for both the post-hoc and prefixed bridge explanation models.

Formally, we consider an annotated dataset defined by $\mathcal{D} = (X, A, Y)$, where X is the set of text instances $\{x_1, x_2, \ldots, x_n\}$. Each instance $x_i \in \mathcal{X}$ is a pair (c_i, s_i) representing the context and statement. The set $A = \{a_1, a_2, \ldots, a_k\}$ represents unique annotators, and the annotation matrix is defined as:

$$Y : X \times A \to \{0, 1\}^3 \tag{3}$$

To handle varying annotator coverage, a masking mechanism is applied during training and evaluation. The annotator-level loss is computed only for instances where a label exists, using a binary mask to ensure missing annotations do not contribute to the training objective.

The encoder extracts a pooled representation $h \in \mathbb{R}^H$ capturing the relationship between c_i and s_i. To incorporate individual variation, we define a learnable embedding space where each annotator a_j is mapped to a unique, d-dimensional vector $u_j \in \mathbb{R}^E$:

$$u_j = \text{Embedding}(a_j) \tag{4}$$

Simultaneously, each annotator's structured demographic metadata is transformed into a fixed-size vector m_j and projected into the latent space of the text encoder. We then perform a *representation-level fusion* by concatenating the instance representation, the annotator embedding, and the metadata projection. The resulting fused representation z_{ij} is passed to the classification head:

$$z_{ij} = [\, h \,;\, u_j \,;\, m_j] \tag{5}$$

This allows the model to explicitly account for both the annotator's identity and their demographic context by learning systematic patterns between these

Statistic	Train	Dev	Test	Total
Split-level statistics				
Instances	388	50	50	488
Annotators	4	4	4	4
Annotations	1,505	187	199	1,891
Avg. annotations / instance	3.88	3.74	3.98	3.88
Explanations	1,505	187	199	1,891
Avg. explanation length (words)	13.90	13.12	14.28	13.86
Label distribution (count, %)				
Entailment	446 (29.6%)	34 (18.2%)	61 (30.7%)	541 (28.6%)
Neutral	767 (51.0%)	96 (51.3%)	93 (46.7%)	956 (50.6%)
Contradiction	292 (19.4%)	57 (30.5%)	45 (22.6%)	394 (20.8%)
Annotations per annotator (count) and demographics				
Ann1 (F,22,CN,MSc)	367	45	47	459
Ann2 (M,33,DE,Postdoc)	376	45	47	468
Ann3 (F,25,CN,MSc)	379	46	54	479
Ann4 (M,25,CN,MSc)	383	51	51	485

Table 1: VariErrNLI dataset statistics by split. Demographics are abbreviated as Gender, Age, Nationality, Education (CN=Chinese, DE=German; MSc=Master student).

features and labeling behavior through latent feature fusion.

3.4. Annotator Explanation Modeling

To generate annotator-specific rationales, we implement two explanation approaches that produce an explanation r_a but differ in how they incorporate classifier information.

3.4.1. Post-hoc Explainer

Our first approach trains a standard encoder-decoder model $Flan-T5$ (Chung et al., 2024) to generate an annotator explanation using a text-only prompt. For each training record, we construct an input prompt that contains: the context and statement, the annotator's gold labels, annotators persona: derived from the annotator metadata information, and an annotator control token. The annotator control token is linked to the annotator ID in the dataset and prepended to the prompt by extending the tokenizer vocabulary with a unique, learnable special token (Sarumi et al., 2024; Plepi et al., 2022). At inference time, we insert the classifier's predicted probabilities (p_C, p_E, p_N) into the prompt. The explainer then generates a short explanation. In this setup, there is no differentiable connection between the classifier and the explainer.

3.4.2. Prefixed Bridge Explainer

Our second approach introduces a stronger coupling between classification and explanation using the classifier's continuous internal representation, rather than text-only features. We first run the annotator-aware classifier on (c, s) to obtain the fused representation z_{ij}. We then learn a small neural Prefixed Bridge (a 2-layer MLP) that projects this vector into a sequence of prefix embeddings with the same dimensionality as the T5 encoder embedding space. These prefix embeddings are prepended to the T5 encoder input embeddings before encoding. We then train by freezing the classifier parameters and optimizing the bridge and the T5 parameters to minimize explanation generation loss. At inference time, explanation generation is performed using the prefix produced by the bridge, which is concatenated with the prompt token embeddings before encoding and generation (see Figure 1).

4. Experiments

In our experiments, we used two base models that follow the encoder-decoder architecture. We also implemented the User Passport method for incorporating annotator-meta information.

4.1. Experimental set-up

We train the annotator-aware classifier for 50 epochs using the AdamW optimizer with a learning rate of 2×10^{-5} and weight decay 0.01. A linear scheduler with warmup (ratio 0.06), gradient clipping (max norm 1.0), and early stopping on development macro-F1 (patience 3) is applied. The backbone model is DeBERTa-v3-base (He et al., 2023), with a maximum input length of 256 and batch size 32. To model annotator-specific predictions, we incorporate annotator information through a learnable annotator embedding (dimension 64) and a projected metadata representation, fused

Explainer	F1 (Macro)	Exact Match	ROUGE-L	Semantic Similarity
User Passport (Sarumi et al., 2025)	70.5	—	—	—
Post-hoc Explainer	92.3	92.2	**24.5**	51.0
Prefixed Bridge Explainer	**93.9**	**92.4**	24.0	**53.4**

Table 2: Aggregated evaluation scores across all annotators. We report the results of the User Passport model from previous work, without explanation, as the baseline. Bold values indicate the best scores. All scores are reported as the mean of three runs.

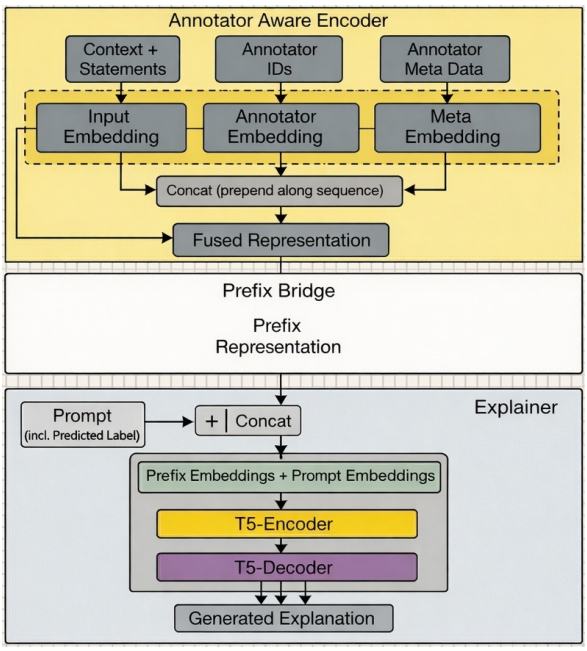

Figure 1: The Prefixed Bridged Explainer

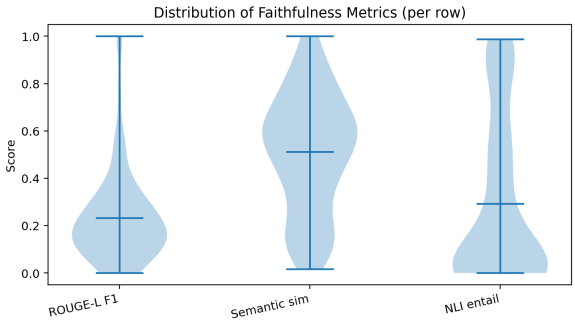

Figure 2: Prefixed Bridged Faithfulness Evaluation

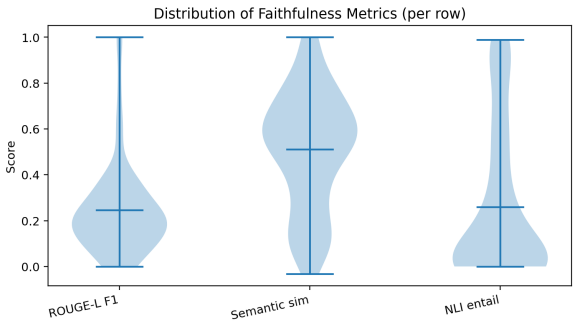

Figure 3: Post-hoc Faithfulness Evaluation

with the instance representation at the feature level. Training uses masked binary cross-entropy with an auxiliary soft-label alignment objective ($\lambda_{\text{soft}} = 1.0$). Class imbalance is handled using masked focal BCE with class-specific positive weighting.

For explanation generation, both explainer variants are trained using Flan-T5-base with a maximum input length 512 and target length 128. Models are trained for up to 50 epochs with early stopping on validation loss, using AdamW with learning rate 8×10^{-5} and weight decay 0.01. Label thresholds are tuned on the development set with grid search over $[0.1, 0.9]$, selecting the configuration that maximizes mean Jaccard similarity with gold annotator label-sets.

All experiments are conducted on a single NVIDIA A100 80GB PCIe GPU (CUDA 13.1). Average end-to-end runtime (training and evaluation) is approximately 15-20 minutes per model. All reported results are averaged over three runs.

5. Result and Discussion

Aggregated Evaluation of Explainers

Table 2 presents the aggregated performance comparison between the baseline annotator-aware classifier (User Passport) from previous work, the Post-hoc Explainer, and the Prefixed Bridge Explainer. The baseline achieves a Macro-F1 score of 70.5, indicating that incorporating explanation modeling substantially improves classification performance.

Both explanation-based approaches outperform the baseline, with the Prefixed Bridge Explainer achieving the highest Macro-F1 (93.9) and Exact Match (92.4), indicating stronger agreement with the gold labels. The Post-hoc Explainer also performs well with Macro-F1(92.3), but remains slightly below the bridge model.

In terms of explanation quality, ROUGE-L is marginally higher for the Post-hoc Explainer, suggesting better lexical overlap with reference explanations, which is consistent with its text-based ap-

Prefixed Bridge Explainer								
Annotator	Gender	Age	Nationality	Education	Macro F1	Exact Match	ROUGE-L	Semantic Sim
Ann1	Female	22	Chinese	MSc.	**94.3**	**94.2**	23.8	55.2
Ann2	Male	33	German	Postdoc	92.5	92.0	**34.1**	**59.6**
Ann3	Female	25	Chinese	MSc	92.0	88.7	21.2	53.5
Ann4	Male	25	Chinese	MSc	**95.5**	**94.7**	17.9	45.8
Post-hoc Explainer								
Annotator	Gender	Age	Nationality	Education	Macro F1	Exact Match	ROUGE-L	Semantic Sim
Ann1	Female	22	Chinese	MSc.	87.9	90.6	24.5	49.9
Ann2	Male	33	German	Postdoc	**96.7**	**97.1**	**31.3**	**55.9**
Ann3	Female	25	Chinese	MSc	90.4	88.7	22.9	50.0
Ann4	Male	25	Chinese	MSc	92.4	92.7	19.6	48.7

Table 3: Comparison of Explainers per annotator: A descriptive Analysis. Bold values highlight key patterns discussed in section 5: improved predictive performance with the bridge model (Ann1, Ann4), stronger lexical overlap and semantic strength with both prefixed and post-hoc model (Ann2), and overall Macro-F1 score, Exact Match (notably Ann2).

proach. In contrast, the Prefixed Bridge Explainer achieves higher semantic similarity, indicating that its generated explanations are better aligned in meaning. This improvement can be attributed to its use of the classifier's internal representations, which provide richer contextual features for generation.

These results show that explanation modeling significantly improves performance over the baseline, while tighter integration between prediction and generation further enhances classification consistency and semantic alignment.

Faithfulness Distribution and Qualitative Analysis

The faithfulness distributions in Figures 2 and 3 show that, for both models, semantic similarity scores cluster around moderate values (median $\sim$ 0.5), while ROUGE-L remains relatively low (median $\sim$ 0.24), indicating lexical divergence despite semantic alignment. However, the Prefixed Bridge Explainer exhibits a more balanced NLI entailment distribution, with a larger proportion of high-entailment cases compared to the Post-hoc model, suggesting stronger inferential alignment between predictions and explanations. To further examine these differences, we present qualitative examples in Figures 4 and 5, focusing on cases where the predicted label is consistent but the generated explanations differ in structure and depth. In both examples, the two models correctly identify that the context supports investment in information technology rather than the financial sector. However, the nature of the generated explanations differs. The Prefixed Bridge Explainer produces explanations that are more concise and directly grounded in the key contrast between the context and the statement, closely mirroring the underlying reasoning required for the prediction. In contrast, the Post-hoc Explainer tends to generate more verbose explanations, introducing additional statements that are not explicitly stated in the context. While both explanations are semantically aligned with the gold rationale, the prefixed bridged explanation more precisely captures the core inference without introducing random words. This qualitative difference supports the distributional findings. The Prefixed Bridge Explainer demonstrates stronger alignment between prediction and explanation, not only quantitatively through higher scores, but also qualitatively in the clarity and focus of the generated explanation.

Comparison of Explainers per annotator: A descriptive Analysis.

Table 3 presents a detailed examination across the four annotators. Differences are observed in Macro-F1, Exact Match, ROUGE-L, and Semantic similarity, suggesting that both models interact differently with individual annotator patterns.

The Prefixed Bridge Explainer generally produces more stable performance across annotators in terms of Macro-F1 and Exact Match. In particular, Ann1 and Ann4 show improvements in both metrics compared to the Post-hoc Explainer, indicating that incorporating classifier-level representations contributes to more reliable alignment between predictions and annotator-specific labels. This suggests that the shared representation between the two encoders better captures variability in annotator decision patterns, especially when explanations differ in structure or clarity.

Ann2 achieves the highest overall performance, particularly under the Post-hoc Explainer (Macro-F1: 96.7, Exact Match: 97.1), outperforming the Prefixed Bridge model. Ann3 and Ann4 exhibit comparatively lower or more variable performance across certain metrics, particularly in ROUGE-L. For Ann4, while Macro-F1 and Exact Match improve under the Prefixed Bridge Explainer, ROUGE-

```
"statement": "Investment in the financial sector explains most of the acceleration
in labor productivity.",
    "context": "Indeed, recent economic research suggests that investment in
information technology explains most of the acceleration in labor productivity
growth-a major component of overall economic growth-since 1995.",
    "gen_expl": "The context suggests that investment in information technology
explains most of the acceleration in labor productivity growth, which suggests that
the financial sector is not involved in this explanation. We don't know whether the
investment in the financial sector is involved in this explanation.",
    "gold_expl": "Investment in information technology explains most of the
acceleration in labor productivity, not investment in financial sector.\nWe don't
know whether investment in information technology is a subsector of investment in
financial sector, or reverse."
  },
```


Figure 4: The Prefixed Bridged Explanation Example


```
"statement": "Investment in the financial sector explains most of the
acceleration in labor productivity.",
    "context": "Indeed, recent economic research suggests that investment
in information technology explains most of the acceleration in labor
productivity growth-a major component of overall economic growth-since
1995.",
    "gen_expl": "The context suggests that investment in information
technology explains most of the acceleration in labor productivity growth,
not the financial sector. The model probabilities are (0.25, 0.22, 0.55)",
    "gold_expl": "The reason of the acceleration in labor productivity is
the investment in information technology, not in the financial sector."
  },
```


Figure 5: The Post-hoc Explanation Example

L and Semantic similarity remain relatively low across both models.

A closer examination of the VariErrNLI dataset (Weber-Genzel et al., 2024) provides important context for interpreting these results. The dataset explicitly distinguishes between variation and annotation error through a second round of self- and peer-validation, where explanations are assessed for whether they plausibly support the assigned labels. As shown in the original study, agreement increases substantially after validation, indicating that a portion of annotator disagreement is attributable not to genuine perspectives and differences, but to inconsistencies and errors.

This distinction is reflected in our findings. Annotators whose explanations are more consistently grounded in the input text and validated by peers are more reliably modeled by both approaches. In

particular, Ann2 achieves the highest predictive performance across metrics, especially under the Post-hoc Explainer, aligning with the dataset's validation framework where more coherent and text-aligned reasoning leads to more stable label–explanation pairs. Notably, Ann2 is also the annotator with the highest age (33) and level of education (Postdoc) in the dataset. While this may be associated with clearer or more structured explanations, stronger task understanding or domain expertise, we do not draw definitive conclusions from this observation due to the limited number of annotators. Instead, this serves as an indicative pattern that can be further investigated in larger and more controlled settings.

In contrast, annotators exhibiting more variability in explanation quality are more challenging to model. For example, Ann4 shows comparatively lower or less consistent performance across certain metrics, particularly in lexical overlap (ROUGE-L), despite improvements in predictive performance under the Prefixed Bridge Explainer. This pattern is consistent with the dataset observations, where some explanations may be less well-aligned with the assigned labels or expressed in ways that deviate from reference formulations. As a result, the model relies more heavily on underlying representations rather than surface-level cues.

A consistent pattern across annotators is the divergence between lexical and semantic metrics. The Post-hoc Explainer tends to produce higher ROUGE-L scores, indicating closer surface-level similarity to reference explanations. In contrast, the Prefixed Bridge Explainer achieves higher or comparable semantic similarity across most annotators, suggesting better alignment in meaning. This reflects the underlying modeling difference: the Post-hoc approach relies primarily on textual prompts, whereas the bridge model leverages classifier-derived internal representations, enabling richer contextual grounding of explanations.

These per-annotator differences highlight that identical labels do not imply identical reasoning processes. The variation observed across metrics suggests that annotators may express similar decisions through different explanatory structures, levels of detail, or linguistic forms. By incorporating explanation generation, both models move beyond label prediction and provide additional insight into how annotator perspectives are represented. The Prefixed Bridge Explainer, in particular, better preserves the relationship between predictions and underlying reasoning, especially when explanations are less consistent in form.

It is important to note that these observations are based on a small number of annotators, with limited demographic diversity and a relatively small test set. As such, we do not perform statistical significance testing and instead rely on descriptive analysis. The patterns observed should therefore be interpreted as indicative trends rather than generalizable findings.

Overall, the per-annotator analysis suggests that incorporating explanation modeling improves the representation of annotator perspectives, and that tighter integration between prediction and explanation, as in the Prefixed Bridge Explainer, provides more consistent and semantically aligned outputs across diverse annotator behaviors.

6. Conclusion

This work demonstrates the importance of Modeling fine-grained annotator perspectives jointly with explanation generation in natural language inference. Rather than treating explanations as post-hoc rationalizations, we show that integrating annotator-expressed rationales into the predictive architecture enables more robust modeling of human diversity. By leveraging explanation-level supervision tied to individual annotations, the model captures not only label outcomes but also the reasoning patterns underlying them, allowing for more faithful representation of disagreement and interpretative nuance.

Methodologically, we implement an encoder-to-encoder bridge architecture that explicitly connects prediction and explanation modules. This structural coupling enables the model to condition its explanatory representations on the same signals that drive classification decisions, thereby improving macro-level stability and inferential alignment across annotators. Our results show that Modeling perspectives through annotator rationales strengthens semantic consistency and predictive robustness, particularly in semantically complex categories. Overall, this work highlights the value of integrating explanation modeling into annotator-aware architectures for developing more transparent and perspective-sensitive NLP systems.

7. Limitation

A primary limitation of this work is the dataset used, which was originally constructed to investigate annotation errors in human label variation. Although the inclusion of annotator-specific rationales represents a substantial step toward preserving individual reasoning patterns, the explanations were not designed to systematically capture controlled variations in demographic, linguistic, or cultural background, but were instead targeted toward annotation error detection. As a result, the scope remains limited, which may constrain the generalizability of the proposed encoder-to-encoder bridge framework.

We initially proposed extending the ChaOSNLI instances used in VariErrNLI with explanations written by native English speakers to systematically examine how bilingual versus native annotator rationales affect modeling outcomes. This extension would allow a more controlled investigation of linguistic background effects on explanation faithfulness and predictive performance. Future work will focus on expanding the dataset in this direction to strengthen the empirical foundation of perspective modeling.

Additionally, an ensemble of the Post-hoc and Prefixed Bridge approaches presents an interesting direction for future work, as it could leverage the strengths of the individual models to produce a more well-rounded output.

All code and resources developed for this study are publicly available[1] to facilitate reproducibility and further research.

Bibliographical References

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868.

Aida Mostafazadeh Davani, Markos Markatou, Tommaso Fornaciari, Silviu Paun, Dirk Hovy, Joel Tetreault, and Cecilia Ovesdotter Alm. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide Bernardi. 2025. Perspectivist approaches to natural language processing: a survey. *Language Resources and Evaluation*, 59:1719–1746.

Shirley Anugrah Hayati, Minhwa Lee, Dheeraj Rajagopal, and Dongyeop Kang. 2024. How far can we extract diverse perspectives from large language models? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5336–5366, Miami, Florida, USA. Association for Computational Linguistics.

[1] https://github.com/Responsible-NLP/LRECNLPerspectives2026-Fine-Grained-Perspective

Elisa Leonardelli, Gavin Abercrombie, Dina Almanea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Verena Rieser, Alexandra Uma, and Massimo Poesio. 2023. SemEval-2023 task 11: Learning with disagreements (LeWiDi). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2304–2318, Toronto, Canada. Association for Computational Linguistics.

Belinda Z. Li, Maxwell Nye, and Jacob Andreas. 2021a. Implicit representations of meaning in neural language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1813–1827, Online. Association for Computational Linguistics.

Lei Li, Yongfeng Zhang, and Li Chen. 2020. Generate neural template explanations for recommendation. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20)*, pages 755–764, New York, NY, USA. ACM.

Lei Li, Yongfeng Zhang, and Li Chen. 2021b. Personalized transformer for explainable recommendation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4947–4957, Online. Association for Computational Linguistics.

Matteo Mastromattei, Leonardo Ranaldi, Francesca Fallucchi, and Fabio Massimo Zanzotto. 2022a. Syntax and prejudice: ethically-charged biases of a syntax-based hate speech recognizer unveiled. *PeerJ Computer Science*, 8:e859.

Michele Mastromattei, Valerio Basile, and Fabio Massimo Zanzotto. 2022b. Change my mind: How syntax-based hate speech recognizer can uncover hidden motivations based on different viewpoints. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 117–125, Marseille, France. European Language Resources Association.

Benedetta Muscato, Lucia Passaro, Gizem Gezici, and Fosca Giannotti. 2025. Perspectives in play: A multi-perspective approach for more inclusive nlp systems. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, IJCAI-2025, page 9827–9835. International Joint Conferences on Artificial Intelligence Organization.

Ayushi Nirmal, Amrita Bhattacharjee, Paras Sheth, and Huan Liu. 2024. Towards interpretable hate speech detection using large language model-extracted rationales. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 223–233, Mexico City, Mexico. Association for Computational Linguistics.

Matthias Orlikowski, Jiaxin Pei, Paul Röttger, Philipp Cimiano, David Jurgens, and Dirk Hovy. 2025. Beyond demographics: Fine-tuning large language models to predict individuals' subjective text perceptions.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Joan Plepi, Béla Neuendorf, Lucie Flek, and Charles Welch. 2022. Unifying data perspectivism and personalization: An application to social norms. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7391–7402, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Joan Plepi, Charles Welch, and Lucie Flek. 2024. Perspective taking through generating responses to conflict situations. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6482–6497, Bangkok, Thailand. Association for Computational Linguistics.

Olufunke O. Sarumi, Béla Neuendorf, Joan Plepi, Lucie Flek, Jörg Schlötterer, and Charles Welch. 2024. Corpus considerations for annotator modeling and scaling. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1029–1040, Mexico City, Mexico. Association for Computational Linguistics.

Olufunke O. Sarumi, Charles Welch, and Daniel Braun. 2025. NLP-ResTeam at LeWiDi-2025:performance shifts in perspective aware models based on evaluation metrics. In *Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP*, pages 219–227, Suzhou, China. Association for Computational Linguistics.

Alexandra Uma, Tommaso Fornaciari, Anca Dumitrache, Tristan Miller, Jon Chamberlain, Barbara Plank, Edwin Simpson, and Massimo Poesio. 2021. SemEval-2021 task 12: Learning with disagreements. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 338–347, Online. Association for Computational Linguistics.

Charles Welch, Chenxi Gu, Jonathan K. Kummerfeld, Veronica Perez-Rosas, and Rada Mihalcea. 2022. Leveraging similar users for personalized language modeling with limited data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1742–1752, Dublin, Ireland. Association for Computational Linguistics.

Jin Xu, Mariët Theune, and Daniel Braun. 2024. Leveraging annotator disagreement for text classification. In *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*, pages 1–10, Trento. Association for Computational Linguistics.

Language Resource References

Chung, Hyung Won and Hou, Le and Longpre, Shayne and Zoph, Barret and Tai, Yi and Fedus, William and Li, Yunxuan and Wang, Xuezhi and Dehghani, Mostafa and Brahma, Siddhartha and Webson, Albert and Gu, Shixiang Shane and Dai, Zhuyun and Suzgun, Mirac and Chen, Xinyun and Chowdhery, Aakanksha and Castro-Ros, Alex and Pellat, Marie and Robinson, Kevin and Valter, Dasha and Narang, Sharan and Mishra, Gaurav and Yu, Adams and Zhao, Vincent and Huang, Yanping and Dai, Andrew and Yu, Hongkun and Petrov, Slav and Chi, Ed H. and Dean, Jeff and Devlin, Jacob and Roberts, Adam and Zhou, Denny and Le, Quoc V. and Wei, Jason. 2024. *Scaling instruction-finetuned language models*. JMLR.org.

Pengcheng He and Jianfeng Gao and Weizhu Chen. 2023. *DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing*.

Elisa Leonardelli and Silvia Casola and Siyao Peng and Giulia Rizzi and Valerio Basile and Elisabetta Fersini and Diego Frassinelli and Hyewon Jang and Maja Pavlovic and Barbara Plank and Massimo Poesio. 2026. *LeWiDi-2025 at NLPerspectives: Third Edition of the Learning with Disagreements Shared Task*.

Weber-Genzel, Leon and Peng, Siyao and De Marneffe, Marie-Catherine and Plank, Barbara. 2024. *VariErr NLI: Separating Annotation Error from Human Label Variation*. Association for Computational Linguistics.