# PuDy at FadeIT: Enhancing Fine-Grained Fallacy Detection with Hierarchy-Aware Training and Contextual Enrichment

Duy Dang Phu[1,*], Son Bui Hong[1] and Thin Van Dang[1]

[1]University of Information Technology - VNUHCM, Ho Chi Minh City, Vietnam

**Abstract**

Fine-grained fallacy detection-identifying specific spans of argumentative errors within text is a critical yet challenging task in argument mining, characterized by data scarcity and high reasoning demands. Addressing these challenges within the **FadeIT** shared task requires not only robust extraction capabilities but also adaptation to the specific hierarchical taxonomy defined by the competition. In this paper, we adopt a span-based model architecture as the foundation for precise fallacy localization. To significantly enhance performance, we introduce two key contributions: (1) a data augmentation strategy based on *contextual enrichment* to mitigate data scarcity while preserving semantics; and (2) a *hierarchy-aware training mechanism* that utilizes label propagation to align the optimization process with the Soft F1 metric. Experimental results on the FadeIT benchmark demonstrate the efficacy of our composite approach, achieving a Soft F1 score of 50.92 and securing the Top 1 ranking in Subtask B. These findings highlight the importance of integrating data augmentation with metric-aligned training objectives in competitive argument analysis tasks.

**Keywords**

Span-based fallacy detection, Hierarchy-Aware Training, Data Augmentation

## 1. Introduction

Fallacious reasoning poses a significant challenge due to its potential to spread misinformation and distort public debates on social media. The informal, diverse, and often implicit nature of user-generated language further complicates the identification of subtle argumentative flaws. As a result, accurately detecting fallacious reasoning at a fine-grained level remains a difficult problem in computational argumentation.

Shared tasks in fine-grained semantic analysis are essential for advancing research on structured linguistic interpretation. In this context, Subtask B of the FadeIT shared task [1], organized as part of the Evalita 2026 evaluation campaign [2], focuses on span-level fallacy detection. This task is particularly challenging due to a combination of factors, including limited annotated data and a hierarchical evaluation metric.

Existing approaches often struggle to address these constraints simultaneously, particularly in terms of generalization under limited data and alignment with hierarchical evaluation metrics. To address these challenges, our approach is guided by two key design principles concerning data augmentation and training objective alignment.

First, to improve robustness and generalization under limited annotated data, we introduce a data augmentation strategy based on *Contextual Enrichment*. Unlike standard paraphrasing methods that may alter the semantics of fallacious spans, our approach strictly preserves fallacious segments while diversifying the surrounding context, encouraging the model to reduce reliance on spurious contextual cues.

Second, to better align training objectives with hierarchical evaluation metrics such as Soft F1, we incorporate a *hierarchy-aware training strategy*. By leveraging label propagation, this method enforces

*EVALITA 2026: 9th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Feb 26 – 27, Bari, IT*

*Corresponding author.

24520010@gm.uit.edu.vn (D. D. Phu)

0009-0006-3795-7932 (D. D. Phu); 0009-0006-7420-9212 (S. B. Hong); 0000-0001-8340-1405 (T. V. Dang)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

consistency between fine-grained fallacy labels and their superordinate categories during optimization.

The effectiveness of these design choices is demonstrated by our experimental results on the FadeIT benchmark. Our composite approach achieves strong performance and secures the **Top 1 ranking** in Subtask B, highlighting the importance of aligning data augmentation and training objectives with the semantic structure of the task and the hierarchical evaluation setting.

## 2. Related work

Previous work on fallacy detection has mainly focused on sentence-level classification and benchmark datasets such as Logic and LogicClimate [3] and MAFALDA [4]. More recently, fine-grained datasets with span-level annotations like FAINA have been proposed to capture nuanced fallacy expressions in social media posts [5]. Other works explore transformer-based classification in political debates [6]. Techniques to enhance reasoning, such as logical-structure-based prompts, also show performance gains [7]. However, accurately identifying fallacious spans in informal social media text remains challenging, motivating span-based models that incorporate richer contextual augmentation.

**Span-based modeling.** Independent of fallacy detection, span-based modeling has been widely studied across various NLP tasks. SpanBERT [8] introduces a span-aware pre-training objective by masking and predicting contiguous spans, yielding stronger representations for span selection tasks such as question answering and coreference resolution. In named entity recognition, SpanNER [9] reformulates sequence labeling as a span prediction problem, demonstrating the benefits of directly classifying candidate spans rather than assigning token-level tags. More generally, DyGIE++ [10] proposes a unified information extraction framework that enumerates, refines, and scores candidate spans for entities, relations, and events, showing that contextualized span representations can effectively capture boundary-sensitive and overlapping phenomena. These span-based paradigms provide useful inductive biases for fine-grained prediction and motivate our span-based approach to fallacy detection.

**Data augmentation.** Data augmentation for fine-grained sequence labeling is a non-trivial challenge, as transformations must introduce sufficient distributional diversity while strictly preserving the semantic integrity and boundary precision of annotated spans. Prior work has explored contextual augmentation through word-level substitutions based on paradigmatic relations [11], as well as LLM-based rewriting strategies for text classification tasks [12].

However, for argumentation-centric tasks such as fallacy detection, naive lexical perturbations may disrupt the underlying argument structure, leading to label noise. Recent studies highlight that models often rely on spurious correlations between labels and contextual surface cues rather than causal reasoning patterns [13], motivating augmentation strategies that explicitly disentangle invariant semantic signals from environmental noise.

## 3. Task Description

Subtask B of the FadeIT shared task addresses the challenge of fine-grained fallacy detection in Italian social media texts. The dataset is provided in CoNLL format with token-level annotations following a BIO tagging scheme. A key characteristic of this task is the presence of overlapping fallacies, where a single token may be associated with multiple labels corresponding to distinct fallacy types.

Formally, given an input sequence of tokens

$$\mathbf{x} = (x_1, x_2, \ldots, x_n),$$

the objective is to identify all contiguous spans that exhibit fallacious reasoning and assign them the corresponding fallacy labels. Although annotations are provided at the token level, the evaluation is conducted at the span level. We therefore formulate this task as a multi-label span classification problem, which naturally supports variable-length and overlapping fallacy expressions.

To mitigate the impact of confusions among closely related fallacy categories, the shared task adopts a soft evaluation strategy based on the hierarchical taxonomy proposed by [5]. Under this scheme, predictions that correctly identify the *direct parent category* of the gold fallacy label are awarded partial credit. Consequently, our system is primarily optimized to maximize the **Soft Micro F1 score**, which explicitly accounts for hierarchical consistency between predicted and gold labels.

# 4. Methodology

## 4.1. Span-Based Modeling

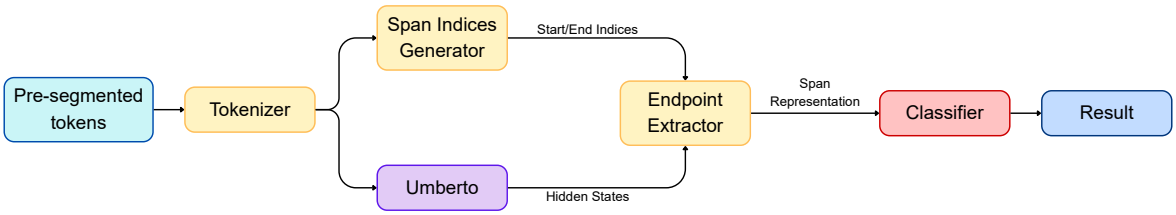


**Figure 1:** System Overview

We implement a span-based modeling approach leveraging UmBERTo [14] as the backbone encoder. Given the CoNLL-formatted input with pre-segmented tokens and BIO annotations, we re-tokenize the input using the UmBERTo tokenizer to obtain subword-level units, which are then fed into the encoder to produce contextualized hidden representations $\mathbf{H} = \{h_1, \ldots, h_n\}$.

Following standard span-based modeling approaches [9], we enumerate all candidate spans defined by start ($i$) and end ($j$) indices. To ensure computational efficiency, we impose a maximum span length constraint such that $j - i < 20$, which covers most of gold spans in the training set while significantly reducing the computational cost.

For each valid span, we construct a span representation $\mathbf{s}i, j$ by concatenating the corresponding boundary embeddings and a learned width embedding. Specifically, the boundary embeddings $h_i$ and $h_j$ correspond to the contextualized token representations at positions $i$ and $j$, obtained from the final hidden layer of the encoder. These embeddings capture both the internal semantics of the span and its surrounding contextual information. The learned width embedding $\mathbf{w}j - i + 1$ encodes the span length and provides the model with explicit structural information. The final span representation is computed as:

$$\mathbf{s}i, j = [h_i; h_j; \mathbf{w}j - i + 1] \tag{1}$$

where $[;]$ denotes the concatenation operation and $\mathbf{w}_k$ represents the embedding for a span of width $k$.

The resulting span representations are passed through a linear classification layer followed by a Sigmoid activation to obtain per-label probabilities for each fallacy type.

**Loss Function.** Due to the severe class imbalance among fallacy types and the predominance of non-fallacious spans, we employ *Focal Loss* [15] to reduce the contribution of frequent negative examples and to focus learning on rare fallacious spans, which is particularly effective in our multi-label span classification setting.

## 4.2. Hierarchy-Aware Training Strategy

To better align the optimization process with the Soft F1 metric, we employ a **label propagation strategy**. Since the evaluation metric awards partial credit for superordinate categories, we modify the training targets to explicitly encode this hierarchy.

Specifically, whenever a fine-grained fallacy label is active, we enforce a soft constraint on its immediate parent category $y_{\text{parent}}$ using the update rule:

$$y_{\text{parent}} \leftarrow \max(y_{\text{parent}}, \lambda) \tag{2}$$

where $\lambda \in [0, 1]$ is a hyperparameter acting as a minimum confidence floor. This ensures the model receives supervision for broad categories even when specific sub-types are distinguished, thereby improving robustness against hierarchical inconsistencies.

### 4.3. Contextual Enrichment

Motivated by the observation that span-based models may overfit to contextual lexical cues instead of argumentative structure, we propose a *Contextual Enrichment* strategy that selectively perturbs non-fallacious context while strictly preserving fallacious spans. This constraint is critical because fallacious reasoning often hinges on precise phrasing; altering the span content risks inadvertently invalidating the ground-truth label.

Formally, we decompose an input sequence $x$ into a fallacious span $s$ (tokens with labels $B$ and $I$) and its surrounding context $c$ (tokens with label $O$), such that $x = c \oplus s$. Our objective is to define an augmentation function $f_{\text{aug}}$ that transforms only the contextual component:

$$x' = f(x) = c' \oplus s \tag{3}$$

where $c'$ is a semantically equivalent but lexically diversified paraphrase of $c$, and the span $s$ remains bit-wise immutable.

To instantiate $f$, we leverage the generative capabilities of the Gemini API through a controlled prompt engineering strategy. The LLM is instructed to act as a constrained re-writer, applying syntactic restructuring, synonym replacement, and mild noise injection exclusively to non-fallacious tokens, while explicitly forbidding any modification to the annotated fallacious span.

This targeted augmentation generates diverse training instances that encourage the model to focus on invariant reasoning patterns within $s$ rather than spurious correlations in the surrounding context $c$, thereby improving robustness and generalization to unseen data.

### 4.4. Span-Relaxed Augmentation

In contrast to the strict preservation in Contextual Enrichment, this strategy permits modifying the entire sequence to introduce lexical diversity. Here, the LLM is instructed to lightly paraphrase the fallacious span $s$ while strictly adhering to a class-consistent semantic constraint. Specifically, the transformation must preserve the exact reasoning flaw associated with the ground-truth label, thereby preventing any shift in the fallacy category.

Formally, this relaxes the token-invariance constraint, defining the transformation as:

$$x' = g(x) = c' \oplus s' \tag{4}$$

where $c'$ denotes the rewritten context and $s'$ represents the linguistically altered, yet semantically equivalent, version of the original fallacy.

While this approach significantly enhances the lexical diversity of the training data-preventing the model from overfitting to specific fallacy phrasings-it introduces a distinct challenge. The modification of the span boundaries necessitates distinct alignment mechanisms to ensure $s'$ is correctly re-annotated, as the relaxation of strict constraints risks introducing label noise or subtle semantic drift.

## 5. Experiments

### 5.1. Analysis on the Development Set

In this section, all analyses are conducted on the original train-development split provided by the organizers.

### 5.1.1. Hierarchy-Aware Training Strategy

We investigate the impact of the proposed hierarchy-aware label propagation strategy by varying the confidence floor parameter $\lambda$. Table 1 reports the corresponding precision, recall, and Soft Micro F1 scores on the development set.

As $\lambda$ increases, the model exhibits a clear trade-off between precision and recall. Specifically, higher values of $\lambda$ enforce stronger supervision on parent categories, which substantially improves recall at the expense of precision. This behavior is consistent with the design of the Soft F1 evaluation metric, which assigns partial credit to correct superordinate predictions.

Notably, setting $\lambda = 1.0$-corresponding to full activation of the parent label whenever a child label is present-yields the highest Soft Micro F1 score. This indicates that, under soft evaluation and for this dataset, aggressively propagating hierarchical supervision most effectively aligns the training objective with the evaluation criterion.

**Table 1**
Effect of Hierarchy-Aware Label Propagation with Different Confidence Floors ($\lambda$) on the development set.

| Confidence Floor ($\lambda$) | Precision (%) | Recall (%) | Soft Micro F1 (%) |
|---|---|---|---|
| 0.00 | 50.35 | 35.80 | 41.85 |
| 0.50 | 50.20 ↓0.15 | 37.99 ↑2.19 | 43.25 ↑1.40 |
| 0.75 | 43.02 ↓7.33 | 51.76 ↑15.96 | 46.99 ↑5.14 |
| 1.00 | 41.72 ↓8.63 | 57.48 ↑21.68 | 48.35 ↑6.50 |

### 5.1.2. Contextual Enrichment vs. Span-Relaxed Augmentation

To isolate the effect of data augmentation, all experiments are conducted with the confidence floor fixed at $\lambda = 1.0$, which has been previously identified as optimal for Soft Micro F1. This controlled setting ensures that any observed performance differences can be attributed solely to the augmentation strategies. Table 2 summarizes the results under varying augmentation intensities.

Across both augmentation strategies, introducing additional augmented samples consistently improves performance over the non-augmented baseline. In particular, Recall exhibits substantial gains, suggesting that data augmentation enhances the model's coverage of fallacious instances. These improvements indicate that moderate augmentation acts as an effective regularizer by exposing the model to a broader range of contextual realizations.

Despite these initial gains, the benefits of augmentation do not scale indefinitely. For **Contextual Enrichment**, Soft Micro F1 reaches its peak at 800 augmented samples (49.96%) before declining at 1400 samples. Notably, while Recall continues to increase, Precision decreases, leading to an overall drop in F1. This divergence suggests that excessive augmentation introduces distributional artifacts and weakens the alignment between augmented samples and the original supervision signal, thereby limiting further gains in generalization.

This effect is considerably more pronounced for **Span-Relaxed Augmentation**. Although this strategy achieves large Recall improvements—reaching 68.62% with 200 augmented samples—these gains come at the cost of a substantial Precision drop, falling below 40%. Moreover, performance degradation emerges earlier and progresses more rapidly than in Contextual Enrichment.

By directly reformulating the fallacious span itself ($s \rightarrow s'$), Span-Relaxed Augmentation increases the likelihood of semantic drift. Since fallacy identification demands high-level reasoning to discern subtle flaws, even minor linguistic perturbations in the span can unintentionally rectify the argument or shift the specific fallacy category. Consequently, the augmented instances become less faithfully aligned with their original labels.

In contrast, **Contextual Enrichment** preserves the exact wording of the fallacious span while varying only the surrounding context. This conservative design better maintains the semantic integrity of the fallacious error, resulting in more stable Precision and a higher peak Soft Micro F1 score. Given

that fallacy detection relies on fine-grained and fragile cues, this strategy proves more suitable for the task.

**Table 2**
Impact of Data Augmentation on Model Performance on the development set

| Strategy | Samples | Precision (%) | Recall (%) | Soft Micro F1 (%). |
|---|---|---|---|---|
| No Augmentation | - | 41.72 | 57.48 | 48.35 |
| Contextual Enrichment | 200 | 42.06 ↑0.34 | 60.21 ↑2.73 | 49.52 ↑1.17 |
| Contextual Enrichment | 800 | 42.08 ↑0.36 | 61.48 ↑4.00 | 49.96 ↑1.61 |
| Contextual Enrichment | 1400 | 40.19 ↓1.53 | 61.91 ↑4.43 | 48.73 ↑0.38 |
| Span-Relaxed Augmentation | 200 | 38.25 ↓3.47 | 68.62 ↑11.14 | 49.12 ↑0.77 |
| Span-Relaxed Augmentation | 800 | 39.34 ↓2.38 | 63.47 ↑5.99 | 48.57 ↑0.22 |

### 5.1.3. Prompting Strategy Analysis

Given that **Contextual Enrichment** consistently outperformed **Span-Relaxed Augmentation** in previous experiments, we adopt this strategy as the primary data augmentation framework and further investigate the impact of different prompting schemes under this setting. Our initial configuration employs a **strict-instruction + few-shot** prompt, where the model is guided by a clearly defined and constrained instruction accompanied by two illustrative examples, randomly sampled from the training set, following prior findings that random few-shot selection remains a strong and reliable default for LLM-based data augmentation [16].

Building upon this baseline, we explore two alternative prompting strategies. First, we examine a **chain-of-thought (CoT)** prompting setup, in which the model is encouraged to explicitly reason through the task under a strict instruction, but without providing few-shot demonstrations. Second, we consider a **soft-instruction + few-shot** variant that remains compliant with the Contextual Enrichment constraints, but uses a less detailed and less prescriptive instruction to allow greater generative flexibility, while retaining the same two-shot examples as the baseline.

The results, summarized in Table 3, indicate that the original **strict-instruction + few-shot** configuration remains the most effective overall. Although the **soft-instruction + few-shot** strategy yields higher Recall, this improvement comes at the cost of a substantial decrease in Precision, suggesting that looser constraints introduce additional semantic noise that is detrimental to fine-grained fallacy detection. In contrast, the **CoT** prompting strategy fails to deliver consistent gains, which we attribute to the absence of few-shot exemplars that could otherwise anchor the model's reasoning process to task-specific patterns.

Based on these findings, we retain **strict-instruction + few-shot** as the default prompting strategy for Contextual Enrichment in all subsequent experiments.

**Table 3**
Impact of Prompting Strategy on Model Performance on the development set.

| Strategy | Precision (%) | Recall (%) | Soft Micro F1 (%) |
|---|---|---|---|
| Context Enrichment strict-instruction + few-shot | 42.08 | 61.48 | 49.96 |
| Context Enrichment strict-instruction + CoT | 40.96 ↓1.12 | 61.20 ↓0.28 | 49.07 ↓0.89 |
| Context Enrichment soft-instruction + few-shot | 39.80 ↓2.28 | 64.94 ↑3.46 | 49.35 ↓0.61 |

## 5.2. Official Results

We submitted three system runs to the shared task. All runs employ Contextual Enrichment with a strict-instruction + few-shot prompting scheme, which consistently outperformed alternative configurations on the development set.

The first run uses 800 augmented samples, selected as a trade-off between maintaining a stable training loss and achieving strong $F_1$ performance on the development set. The second run adopts

the same augmentation configuration but corresponds to the checkpoint that attained the highest validation $F_1$ score. The third run is initialized from the first configuration and further fine-tuned for one additional epoch on the validation set.

Table 4 reports the performance of our submissions under both strict and soft evaluation settings. Among the three runs, Run 1 achieves the best Soft $F_1$ score (50.92), while Run 3 demonstrates a more balanced precision–recall trade-off under the strict evaluation setting.

**Table 4**
Performance Comparison under Strict and Soft Evaluation Modes on the official test set.

| Team | Run | Strict Mode | | | Soft Mode | | |
|---|---|---|---|---|---|---|---|
| | | P | R | $F_1$ | P | R | $F_1$ |
| PuDy | 1 | 27.68 | 37.83 | 31.97 | 43.76 | 60.88 | **50.92** |
| PuDy | 2 | 25.96 | 40.80 | 31.73 | 40.30 | 64.79 | 49.69 |
| PuDy | 3 | 29.90 | 32.96 | 31.36 | 45.75 | 52.36 | 48.83 |
| TiGRO | 3 | 47.82 | 37.67 | **42.13** | **51.01** | 40.25 | 44.98 |
| TiGRO | 1 | 38.33 | 40.35 | 39.30 | 42.50 | 45.20 | 43.80 |
| TiGRO | 2 | 38.23 | 40.05 | 39.11 | 42.68 | 44.87 | 43.74 |
| RBG-AI | 3 | 19.47 | 25.25 | 21.99 | 24.18 | 32.44 | 27.71 |
| RBG-AI | 2 | 19.21 | 18.24 | 18.71 | 24.67 | 23.96 | 24.31 |
| RBG-AI | 1 | 17.13 | 11.01 | 13.41 | 20.70 | 13.27 | 16.17 |
| **Baseline** | – | 60.94 | 3.05 | 5.80 | 65.97 | 3.28 | 6.25 |

## 6. Error Analysis

From Table 5, several fallacy types consistently obtain an F1 score of 0.00 in the official results, including *Causal-oversimplification*, *Cherry-picking*, *Circular-reasoning*, *False-analogy*, *False-dilemma*, and *Strawman*. Additionally, categories such as *Slippery-slope* and *Appeal-to-authority* exhibit unstable performance, dropping to an F1 of 0.00 in certain runs. A common pattern among these categories is the scarcity of fallacious spans in the development set whose length is less than or equal to 20 tokens. As our model follows a span-based formulation with a maximum span length of 20 tokens, these classes provide insufficient supervision signals for the model to effectively learn discriminative representations given the semantic complexity of these fallacies.

The imposed span length constraint is a deliberate design choice to balance computational efficiency and overall performance. However, fallacy types that are typically expressed through longer textual spans are consequently underrepresented during training, which leads to systematic prediction failures in the official evaluation for these categories.

Furthermore, this limitation lies beyond the scope of *Contextual Enrichment*. An essential requirement of the contextual enrichment strategy is the strict preservation of the original fallacious spans. As a result, it cannot introduce additional supervision or compensate for the scarcity of valid short-span instances without violating span boundaries.

One potential remedy would be to employ separate models or specialized components for fallacy types characterized by short versus long average span lengths. However, such a design would substantially increase training and inference costs, as well as model complexity, making it impractical under typical computational constraints. This observation highlights an inherent trade-off in span-based fallacy detection between computational feasibility and coverage of long-span fallacy types.

## 7. Conclusion

In this paper, we introduce a **Hierarchy-Aware Training Strategy** that aligns model optimization with soft evaluation metrics for fine-grained fallacy detection. By explicitly incorporating hierarchical label relationships into the training objective, the proposed strategy consistently improves performance

**Table 5**
Per-class statistics of fallacy spans on the development set (number of spans, average span length, and count of spans with length $\leq 20$), together with per-class F1 scores from three independent runs on the official test set.

| Fallacy Type | #Spans | Avg. Len. | #Len $\leq 20$ | F1 Run1 | F1 Run2 | F1 Run3 |
|---|---|---|---|---|---|---|
| Ad-hominem | 247 | 14.57 | 176 | 16.86 | 14.14 | 15.21 |
| Appeal-to-authority | 170 | 5.15 | 169 | 8.03 | 3.33 | 0.00 |
| Appeal-to-emotion | 1618 | 4.08 | 1598 | 42.47 | 42.89 | 40.85 |
| Causal-oversimplification | 113 | 18.51 | 77 | 0.00 | 0.00 | 0.00 |
| Cherry-picking | 76 | 28.18 | 24 | 0.00 | 0.00 | 0.00 |
| Circular-reasoning | 16 | 27.69 | 5 | 0.00 | 0.00 | 0.00 |
| Doubt | 397 | 15.30 | 304 | 37.04 | 31.72 | 37.07 |
| Evading-the-burden-of-proof | 318 | 15.47 | 239 | 12.60 | 9.61 | 5.31 |
| False-analogy | 187 | 20.36 | 107 | 0.00 | 0.00 | 0.00 |
| False-dilemma | 70 | 15.31 | 55 | 0.00 | 0.00 | 0.00 |
| Flag-waving | 316 | 3.27 | 311 | 33.15 | 35.46 | 42.19 |
| Hasty-generalization | 372 | 9.50 | 344 | 21.43 | 21.81 | 17.60 |
| Loaded-language | 2016 | 1.54 | 2010 | 46.90 | 43.48 | 43.55 |
| Name-calling-or-labelling | 880 | 1.61 | 880 | 57.42 | 56.28 | 57.06 |
| Red-herring | 199 | 11.51 | 174 | 5.14 | 4.92 | 4.29 |
| Slippery-slope | 140 | 9.48 | 131 | 31.58 | 0.00 | 0.00 |
| Slogan | 311 | 2.45 | 310 | 64.07 | 57.19 | 59.88 |
| Strawman | 87 | 36.07 | 17 | 0.00 | 0.00 | 0.00 |
| Thought-terminating-cliches | 221 | 4.26 | 220 | 28.86 | 41.24 | 35.16 |
| Vagueness | 1059 | 8.24 | 966 | 16.56 | 16.85 | 5.77 |

under soft evaluation, demonstrating its effectiveness in modeling graded relationships among fallacy types as defined by the label hierarchy.

In addition, we systematically investigate two data augmentation strategies tailored for fine-grained fallacy detection. Through extensive experiments, we show that while both approaches initially improve model generalization, their effectiveness differs substantially as augmentation intensity increases. In particular, **Contextual Enrichment**, which preserves the original fallacious span while varying the surrounding context, yields more stable Precision and achieves the highest overall Soft Micro F1 score. In contrast, more aggressive span-level augmentation is prone to semantic drift, leading to weaker alignment between augmented instances and their original labels.

Taken together, our findings highlight the importance of respecting label hierarchy and semantic fidelity when training models for challenging reasoning tasks such as fallacy detection. We hope this work provides practical guidance for future research on hierarchy-aware learning and data augmentation in fine-grained reasoning tasks. Notably, with these results, our system secured the 1st place in the official ranking for Subtask B of the FadeIT shared task.

## 8. Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT and Gemini for grammar and spelling check and paraphrase/re-wording. After using these tools, the author(s) reviewed and edited all generated content as needed and take(s) full responsibility for the publication's content.

## References

[1] A. Ramponi, S. Tonelli, FadeIT at EVALITA 2026: Overview of the fallacy detection in italian social media texts task, in: Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026), CEUR.org, Bari, Italy, 2026.

[2] F. Cutugno, A. Miaschi, A. P. Aprosio, G. Rambelli, L. Siciliani, M. A. Stranisci, Evalita 2026: Overview of the 9th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026), CEUR.org, Bari, Italy, 2026.

[3] Z. Jin, A. Lalwani, T. Vaidhya, X. Shen, Y. Ding, Z. Lyu, M. Sachan, R. Mihalcea, B. Schölkopf, Logical fallacy detection, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2022, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 7180–7198. URL: https://aclanthology.org/2022.findings-emnlp.532/. doi:10.18653/v1/2022.findings-emnlp.532.

[4] C. Helwe, T. Calamai, P.-H. Paris, C. Clavel, F. Suchanek, MAFALDA: A benchmark and comprehensive study of fallacy detection and classification, in: K. Duh, H. Gomez, S. Bethard (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 4810–4845. URL: https://aclanthology.org/2024.naacl-long.270/. doi:10.18653/v1/2024.naacl-long.270.

[5] A. Ramponi, A. Daffara, S. Tonelli, Fine-grained fallacy detection with human label variation, in: L. Chiruzzo, A. Ritter, L. Wang (Eds.), Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Albuquerque, New Mexico, 2025, pp. 762–784. URL: https://aclanthology.org/2025.naacl-long.34/. doi:10.18653/v1/2025.naacl-long.34.

[6] E. Cantín Larumbe, A. Chust Vendrell, Argumentative fallacy detection in political debates, in: E. Chistova, P. Cimiano, S. Haddadan, G. Lapesa, R. Ruiz-Dolz (Eds.), Proceedings of the 12th Argument mining Workshop, Association for Computational Linguistics, Vienna, Austria, 2025, pp. 369–373. URL: https://aclanthology.org/2025.argmining-1.36/. doi:10.18653/v1/2025.argmining-1.36.

[7] Y. Lei, R. Huang, Boosting logical fallacy reasoning in LLMs via logical structure tree, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 13157–13173. URL: https://aclanthology.org/2024.emnlp-main.730/. doi:10.18653/v1/2024.emnlp-main.730.

[8] M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, O. Levy, SpanBERT: Improving pre-training by representing and predicting spans, Transactions of the Association for Computational Linguistics 8 (2020) 64–77. URL: https://aclanthology.org/2020.tacl-1.5/. doi:10.1162/tacl_a_00300.

[9] J. Fu, X. Huang, P. Liu, SpanNER: Named entity re-/recognition as span prediction, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), 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), Association for Computational Linguistics, Online, 2021, pp. 7183–7195. URL: https://aclanthology.org/2021.acl-long.558/. doi:10.18653/v1/2021.acl-long.558.

[10] D. Wadden, U. Wennberg, Y. Luan, H. Hajishirzi, Entity, relation, and event extraction with contextualized span representations, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 5784–5789. URL: https://aclanthology.org/D19-1585/. doi:10.18653/v1/D19-1585.

[11] S. Kobayashi, Contextual augmentation: Data augmentation by words with paradigmatic relations, in: M. Walker, H. Ji, A. Stent (Eds.), Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 452–457. URL: https://aclanthology.org/N18-2072/. doi:10.18653/v1/N18-2072.

[12] H. Dai, Z. Liu, W. Liao, X. Huang, Y. Cao, Z. Wu, L. Zhao, S. Xu, F. Zeng, W. Liu, N. Liu, S. Li, D. Zhu, H. Cai, L. Sun, Q. Li, D. Shen, T. Liu, X. Li, AugGPT: Leveraging ChatGPT for Text Data Augmen-

tation , IEEE Transactions on Big Data 11 (2025) 907–918. URL: https://doi.ieeecomputersociety.org/10.1109/TBDATA.2025.3536934. doi:10.1109/TBDATA.2025.3536934.
[13] D. Kaushik, E. Hovy, Z. C. Lipton, Learning the difference that makes a difference with counterfactually augmented data, International Conference on Learning Representations (ICLR) (2020).
[14] L. Parisi, S. Francia, P. Magnani, Umberto: an italian language model trained with whole word masking, https://github.com/musixmatchresearch/umberto, 2020.
[15] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2999–3007. doi:10.1109/ICCV.2017.324.
[16] J. Cegin, B. Pecher, J. Simko, I. Srba, M. Bielikova, P. Brusilovsky, Use random selection for now: Investigation of few-shot selection strategies in LLM-based text augmentation, in: C. Christodoulopoulos, T. Chakraborty, C. Rose, V. Peng (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2025, Association for Computational Linguistics, Suzhou, China, 2025, pp. 5533–5550. URL: https://aclanthology.org/2025.findings-emnlp.296/. doi:10.18653/v1/2025.findings-emnlp.296.

# A. Prompt Templates

To ensure reproducibility, we report the full prompt templates used for data augmentation.

## A.1. Contextual Enrichment (strict-instruction + few-shot)

```
You are an expert text augmenter. Your task is to increase the contextual diversity of a
     given Italian sentence
while keeping the meaning and the labels of non-'O' tokens unchanged.

Instructions:
1. The sentence is provided as a list of tokens with their labels.
2. Tokens labeled 'O' are context tokens and can be modified.
3. Tokens with other labels must remain unchanged.
4. Keep the order of tokens logical and coherent.
5. Output the new sentence as a list of tokens with the same labels.

Example Input:
<A RANDOMLY CHOSEN EXAMPLE>

Example Output:
<A RANDOMLY CHOSEN EXAMPLE>

Now apply this process to the following Italian input:
<INPUT>
```

## A.2. Span-Relaxed Augmentation (strict-instruction + few-shot)

```
You are an expert in fallacies and Italian text augmentation.
Your task is to rewrite a given sentence to increase diversity while preserving the
    specific fallacies present.

Input Format:
A list of tokens with BIO labels (e.g., B-Ad-hominem, I-Hasty-generalization, O).

Goal:
Paraphrase the sentence naturally in Italian.
```

```
Instructions:
1. Context ("O" labels):
   - Significantly modify these tokens.
   - You may rephrase, change sentence structure, or use synonyms to create a diverse
       context.

2. Fallacy Spans (B-X, I-X labels):
   - You SHOULD lightly paraphrase or modify the wording of the fallacy segments to
       introduce variety.
   - You MUST preserve the specific meaning of the fallacy type.
     For example:
       * If the label is "Ad-hominem", the new phrase must still be a personal attack.
       * If the label is "Hasty-generalization", it must still be a broad claim based on
           insufficient evidence.
   - Do NOT change the category of the fallacy.

3. BIO Tagging Integrity:
   - If you rephrase a fallacy span (e.g., changing 1 word to 2 words), ensure you apply
       the labels correctly
     (B-Label for the first token, I-Label for subsequent tokens).
   - Ensure the output is a valid list of JSON objects.

Example Input:
<A RANDOMLY CHOSEN EXAMPLE>


Example Output:
<A RANDOMLY CHOSEN EXAMPLE>

Now apply this process to the following Italian input:
<INPUT>
```

## A.3. Context Enrichment (Strict Instruction + Chain-of-Thought)

```
You are an expert Italian Data Augmenter specialized in fallacies.
Your goal is to paraphrase the context ("O" labels) of a sentence while strictly
    preserving all fallacy segments.

Input Format:
A list of JSON objects with fields:
- "token"
- "label"

Your Task Process:
1. Analyze: Identify which tokens form fallacy spans (non-"O") and which form contextual
     parts ("O").
2. Plan: Decide how to rewrite the context tokens to change tone or style (e.g., more
    formal, colloquial, or emphatic),
   while keeping grammatical correctness in Italian.
3. Draft: Mentally construct the rewritten sentence, ensuring that all fallacy tokens
    remain exactly unchanged.
4. Generate: Output the final JSON list.

Constraints:
- You MUST NOT modify any token labeled with "B-" or "I-".
- You MUST return ONLY the JSON list as the final output.
```

```
Now apply this reasoning process step by step, and output the final JSON list:
<INPUT>
```

## A.4. Context Enrichment (Soft Instruction + Few-shot)

```
Task: Paraphrase the contextual part of an Italian sentence without altering any fallacy
     triggers or labeled spans.

Rules:
- Tokens with label "O": rewrite or paraphrase.
- Tokens with labels starting with "B-" or "I-": keep exactly unchanged.

Example Input:
<A RANDOMLY CHOSEN EXAMPLE>

Example Output:
<A RANDOMLY CHOSEN EXAMPLE>

Now apply the same transformation to the following Italian input:
<INPUT>
```