Explainability and Subjectivity in Textual Entailment: the *e-RTE-3-it* Dataset

Andrea Zaninello*
Libera Università di Bozen-Bolzano
Fondazione Bruno Kessler

Sofia Brenna**
Libera Università di Bozen-Bolzano
Fondazione Bruno Kessler

Bernardo Magnini†
Fondazione Bruno Kessler

We introduce the 'e-RTE-3-it' dataset, an enriched version of the Italian RTE-3 dataset, where each text-hypothesis pair, in addition to the 'entailment', 'contradiction', or 'neutrality' label, has been combined with an explanation for the relation. Moreover, the dataset includes the level of confidence with which the annotators wrote the explanation as well as an optional alternative label, along with its explanation, which the annotators could express when they did not agree with the original label. This offers the opportunity to analyse cases of uncertainty in annotation and take into account different perspectives on natural language understanding and generation.

1. Introduction

Recently, Large Language Models (LLMs) like T5 (Raffel et al. 2020), GPT-3.5/4 (OpenAI 2023), LLama-2 (Touvron et al. 2023), or the Italian It5 (Sarti and Nissim 2022), Camoscio (Santilli 2023), Minerva[1], etc. have demonstrated remarkable performances across various natural language processing (NLP) tasks. Despite their success, these LLMs also face limitations and risks, such as lack of factuality (Honovich et al. 2022), hallucinations (Ji et al. 2022), and poor transparency (Guidotti et al. 2019).

As a result, there is a growing demand for "inherent explainability" (Longo et al. 2024), which refers to the ability of models to provide human-like, natural language explanations for their predictions. This interest is demonstrated by the high number of studies focusing on natural language explanations, such as the ones presented at the *1st Workshop on Natural Language Reasoning and Structured Explanations* at ACL 2023[2] and by the numerous datasets collected and created for this task (see Wiegreffe and Marasovic (2021) for comprehensive review). Many of these resources focus on English and are based on extensive crowd-sourcing, thus leaving a gap for non-English languages, including Italian, and for carefully curated data.

* Fondazione Bruno Kessler, NLP Research Unit, Via Sommarive 18, Povo, Trento (TN), 38123, Italy. E-mail: azaninello@fbk.eu
** Fondazione Bruno Kessler, NLP Research Unit, Via Sommarive 18, Povo, Trento (TN), 38123, Italy. E-mail: sbrenna@fbk.eu
† Fondazione Bruno Kessler, NLP Research Unit, Via Sommarive 18, Povo, Trento (TN), 38123, Italy. E-mail: magnini@fbk.eu
1 https://nlp.uniroma1.it/minerva/
2 https://nl-reasoning-workshop.github.io/

To fill this void, this paper introduces the *e-RTE-3-it* dataset, the first Italian dataset for textual entailment[3] enriched with free-form, human-written explanations for the relationship between each *text* and *hypothesis*. Additionally, the dataset includes confidence scores for each explanation and a subset of pairs with alternative relation labels from annotators, whenever they even partially disagreed with the provided label. This aspect of the annotation scheme makes the *e-RTE-3-it* dataset also a potential resource for exploring subjectivity and variability in language and reasoning[4].

2. Background and Related Work

2.1 Defining Entailment: the RTE Task

The Recognizing Textual Entailment (RTE) task emerged in 2005 (Dagan, Glickman, and Magnini 2006) as the problem of determining if two sentences stand in an *entailment* or *not-entailment* relationship. A common definition of "semantic entailment" (also referred to as *presupposition* in some studies) is that "A sentence S presupposes a proposition p if p must be true in order for S to have a truth-value (to be true or false)." (Chierchia and McConnell-Ginet 2000). A text t is said to entail another text (*hypothesis*, h) if h is true in every circumstance (possible world) in which t is true. RTE, however, suggests a more empirical definition, allowing for cases in which the truth of the hypothesis is *highly plausible, for most practical purposes*, rather than certain. According to Dagan et al. (2010), this "shallow" definition better accounts for the types of uncertain inferences that are typically expected from text-based applications.

Recognising Textual Entailment was formalised through a series of successful challenges and workshops that began in 2005 (Dagan, Glickman, and Magnini 2006) and lasted until 2012. Starting from the RTE-3 edition, the task was extended from two labels to a three-label classification, splitting the not-entailment label into two classes, *contradiction* and *neutrality*. Given the interest on the task, an Italian version of the RTE-3 dataset was developed to explore language comprehension and textual entailment (Magnini, Lavelli, and Magnolini 2020). The English RTE-3 dataset was translated into Italian by professional translators, with the goal of maintaining as much as possible of the semantic labels of the English RTE-3.

2.2 Natural Language Explanations

The collection of textual explanations has increasingly expanded in the last few years, contributing to the offspring of *Explainable Natural Language Processing* (ExNLP). In ExNLP, explanations are collected for three main purposes (Wiegreffe and Marasovic 2021):

- *data augmentation*, i.e., creation of synthetic additional data to enhance performance on predictive tasks;
- *explanation generation*, where collected explanations serve as supervision for training models to generate justifications for their predictions;

3 Also referred to as *natural language inference*.
4 We make the e-RTE-3-it dataset available at the following link: `https://nlplab.fbk.eu/tools-and-resources/lexical-resources-and-corpora/e-rte-3-ita`

- *evaluation* of generated explanations, providing ground-truth examples used to assess the quality of model-generated explanations.

Wiegreffe and Marasovic (2021) analyse about a hundred datasets containing textual explanations and identify three main types of explanations: highlights (a selection of the input string that justifies the prediction), structured explanations (like tables or tree graphs), and free-text explanations (usually self-standing utterances of different kind and informativeness).

Free-text explanations, like the one we collected for our *e-RTE-3-it* dataset, play a central role in enhancing understanding and interpretability of NLP models, as they are also the most common kinds of explanations found in human communication (Lombrozo 2007). Free-text explanations are free-form textual justifications of a statement, label, prediction, etc., generally consisting of one or few sentences per instance. These explanations are not constrained by elements of the input instance: as a result, they are best suited for explaining reasoning problems that require information beyond the given input (Zaninello and Magnini 2023).

For example, the CODAH dataset (Chen et al. 2019) comprises a wide range of commonsense reasoning problems. It includes explanations that were adversarially constructed by humans, who had access to feedback from a pre-trained model. These explanations were intentionally designed to create challenging commonsense questions.

The COPA-SSE (Brassard et al. 2022) presents crowd-sourced explanations for the Choice of Plausible Alternatives without superficial cues (COPA) benchmark (Kavumba et al. 2019), with explanations given as a set of triple-like common sense statements with ConceptNet relations and free-text concepts. The COS-E dataset (Rajani et al. 2019) couples commonsense reasoning problems with explanations, thereby providing valuable insights into how humans approach commonsense reasoning tasks. Brahman et al. (2021) explore various methods for automatically generating rationales using pre-trained language models, and demonstrate their approach on the defeasible inference task, a form of non-monotonic reasoning in which an inference can be either reinforced or diminished when new information is introduced.

The most similar resource to our *e-RTE-3-it* is the e-SNLI dataset (Camburu et al. 2018) (Table 1), an enriched version of the Stanford Natural Language Inference (SNLI) corpus (Bowman et al. 2015). The e-SNLI dataset comprises human-produced English explanations, manually written by crowd-workers and post-edited both automatically and manually, giving three explanations for each of the 570k *text-hypothesis* pairs, explaining their *entailment*, *contradiction*, and *neutral* relationship labels. While this resource is very valuable for tasks requiring large amounts of training data, it focuses on the English language and explanations in the dataset tend to be short and lack the variety of ecological data. As can be seen from the examples reported in Table 1, they sometimes tend to repeat patterns that rephrase the input without adding new information, so that models trained on them tend to present recurrent linguistic structures instead of elaborating more complex reasoning paths (Becker, Liang, and Frank 2021; Zaninello and Magnini 2023).

Another common feature of explanation datasets is that annotators apply an explanation to an existent label, which is assumed to be valid or likely. However, while this assumption may not hold for every task, ambiguity and subjectivity have also emerged as salient aspects of human communication. For example, Creanga and Dinu (2024) argue that the stagnation in Natural Language Inference (NLI) progress is due to neglecting the subjective nature of meaning, tied to individual worldviews, and propose creating datasets that capture annotator demographics and values to model diverse

perspectives. Their initial experiments with the SBIC dataset indicate that including annotator metadata can enhance model performances.

Table 1
Example explanations from the *e-SNLI* dataset for *entailment* (YES) and *contradiction* (NO) labels. Three explanations are provided for each pair.

Text	Hypothesis	Label	Explanations
A man in a blue shirt standing in front of a garage-like structure painted with geometric designs.	A man is wearing a blue shirt.	YES	**Expl. 1**: "in a blue shirt" is inferred as "wearing a blue shirt". **Expl. 2**: In a blue shirt is a paraphrasing of wearing a blue shirt. **Expl. 3**: "In a blue shirt" is a rephrasing of "is wearing a blue shirt".
A man in a blue shirt standing in front of a garage-like structure painted with geometric designs.	A man is wearing a black shirt.	NO	**Expl. 1**: Blue is a different color than black. **Expl. 2**: A person cannot be wearing a black shirt while in a blue shirt. **Expl. 3**: A man cannot be wearing a blue shirt and a black shirt at the same time.

3. Original Dataset Annotation Layers

The RTE-3-it dataset (Magnini, Lavelli, and Magnolini 2020), our starting point, is an XML-coded dataset comprising a development and a test split of 800 pairs of *text* (element *t*) and *hypothesis* (element *h*) each, totalling 1600 pairs (Figure 1).

Each text-hypothesis *pair* element is complemented with the following attributes: an *id*, the original *task* from which the pair was taken, the *length* of the text (long or short), and the *label* for the entailment relation, which can take 3 values:

- "YES", when the hypothesis is *entailed* by the text, i.e., given the facts stated in the text, it follows that the hypothesis must be true.
- "NO", when the hypothesis is *contradicted* by the text, i.e., given the facts stated in the text, it follows that the hypothesis must be false.
- "UNKNOWN", when the hypothesis is *neutral* to the text (neither *entailed* not *contradicted* by the text), i.e., given the facts stated in the text, the hypothesis could either be true or false.

4. Data Collection and Additional Annotation Layers

To collect the explanations, we deployed a two-fold writing-and-editing strategy. We recruited 40 annotators among students at the University of Bologna (from undergraduate to PhD level), native Italian speakers, and fluent in at least one other language; each annotator took at least one linguistics university course, ensuring their meta-linguistic proficiency as well as broader cultural understanding. They were also instructed about the *Recognizing Textual Entailment* task and about XML mark-up, so that they could work directly on the original data through an IDE (Integrated Development Environment).

```
<pair id="224" entailment="UNKNOWN" task="IR" length=
     "short">
<t>Basandosi su uno studio mondiale [...] gli
     epidemiologi [...] dimostrano che il fumo e' la
      causa principale degli incendi e delle morti
     per incendi nel mondo.</t>
<h>Gli incendi domestici sono una causa importante
     delle morti da incendio.</h>
</pair>
```

Figure 1
Example of datapoint in the original RTE-3-it dataset.

Annotators had access to the original labels for the text-hypothesis pair, and were asked to provide

- an *explanation* (element *e*) for the entailment label in the original dataset;
- their *confidence* in providing that explanation, expressed by an attribute on a 5-point Likert scale, with 1 = *not confident* and 5 = *very confident*.

We also encouraged diversity in perspectives by allowing annotators to disagree with the original label. In such cases, they were still asked to provide an explanation and their confidence for the original label, but optionally they could provide

- an *alternative label* (attribute *new_label*);
- a corresponding *alternative explanation* (element *a*) for the new label;
- the level of *confidence* for the explanation of the alternative label as an attribute of *a*.

4.1 Guidelines

Each annotator was provided with 50 text-hypothesis pairs from either the development or the test set, each labelled with an entailment relationship. They were asked to write *one free-form, natural language explanation in Italian* clarifying why the two sentences stood in that particular entailment, contradiction, or neutrality relationship.

To ensure language variety as well as uniformity across annotators, the following guidelines were given:

1. please write an explanation in the form of one or two self-contained sentences for each `<pair, label>`;
2. you can refer back to, quote, or paraphrase chunks of both the text and the hypothesis;
3. use case marking and punctuation consistently with the original sentences;

4. you can use metalanguage to refer back to the original sentences with phrases such as "in the text, it is stated that...", "the hypothesis does not mention...", etc.;
5. please provide your level of confidence (i.e., how sure you are about the reasons provided in your explanation) on a scale from 1 to 5;
6. if you (even partially) disagree with the given label, provide a new label for the pair, an explanation for the new label, and your level of confidence in the new explanation.

4.2 Post-editing and Data Curation

Finally, two linguistics experts manually curated the data via a post-editing phase, where they validated the explanations addressing any logical or grammatical fallacies and ensuring uniformity in spelling and punctuation.

Cases when the experts edited explanations include explanations containing grammatical, semantic or logical errors (Table 2, Example 1)[5], uninformative or vague content, like explanations that only paraphrased the hypothesis without making the text-hypothesis entailment relationship explicit (Example 2)[6], or explanations that did not consider or misinterpreted the entailment relationship details, the text and/or the hypothesis (Example 3)[7]. In the latter case, we asked another annotator to write a new explanation from scratch.

4.3 Original Dataset Correction

While editing the explanations, the experts also detected and corrected some errors in the original dataset. In few cases, these included missing information that made it impossible to infer the right label for *t* and *h*. For example, consider the following text-hypothesis pair from the test set (id = 52):

> **Text**: *Oscar Chisini (nato il 4 marzo 1889 a Bergamo, morto il 10 aprile *****1967****** a Milano) fu un matematico italiano. Lui introdusse la media Chisini nel 1929.*
> **Hypothesis**: *Oscar Chisini morì nel 1967.*
> **Entailment**: *"YES".*[8]

5 **Example 1 Text:** Police in Rio de Janeiro arrested five men and recovered millions of dollars worth of art stolen earlier this month, including works by Salvador Dali and Henri Matisse. Police recovered all the stolen art except two Chinese ceramic sculptures from the 7th Century and a collection of silverware. **Hypothesis**: Millions of dollars of art were recovered, including works by Dali. **Original Explanation:** There is no evidence that among the stolen art were the works of Dali. **Edited Explanation:** If police in Rio de Janeiro have recovered millions of dollars worth of stolen art, including works by Salvador Dali, it means millions of dollars worth of art has been recovered, including works by Dali.

6 **Example 2 Text:** Between March and June, scientific observers say, up to 300,000 seals are killed. In Canada, seal-hunting means jobs, but opponents say it is vicious and endangers the species, also threatened by global warming. **Hypothesis**: Hunting endangers seal species. **Original Explanation:** If seals were not hunted, their species would not be endangered. **Edited Explanation:** If it is true that seal hunting is cruel and endangers the species, then it is true that hunting endangers the seal species.

7 **Example 3 Text:** Based on a worldwide study of smoking-related fire and disaster data, UC Davis epidemiologists show smoking is a leading cause of fires and death from fires globally. **Hypothesis**: Domestic fires are the major cause of fire death. **Original Explanation:** Smoking is the major cause of domestic fires. **Edited Explanation:** Smoking is the major cause of fires and fire deaths, but it is not specified whether another major cause of fire deaths is actually domestic fires.

8 **Text:** Oscar Chisini (born March 4, 1889 in Bergamo, died April 10, 1967 in Milan) was an Italian mathematician. He introduced the Chisini mean in 1929. **Hypothesis**: Oscar Chisini died in 1967.

Table 2
Examples of post-editing of the collected explanations.

Example 1	**Text**: La polizia di Rio de Janeiro ha arrestato cinque uomini e ha recuperato arte rubata per milioni di dollari all'inizio di questo mese, tra cui lavori di Salvador Dali e Henri Matisse. La polizia ha recuperato tutta l'arte rubata eccetto due sculture di ceramica cinese del 7° secolo e una collezione di argenteria. **Hypothesis**: Milioni di dollari d'arte sono stati recuperati, tra cui lavori di Dali. **Entailment**: YES **Original Explanation**: Non c'è evidenza che tra arte rubata c'erano i lavori di Dali. **Edited Explanation**: Se la polizia di Rio de Janeiro ha recuperato arte rubata per milioni di dollari, tra cui lavori di Salvador Dali, significa che milioni di dollari d'arte sono stati recuperati, tra cui lavori di Dali.
Example 2	**Text**: Tra marzo e giugno, gli osservatori scientifici dicono che vengono uccise più di 300.000 foche. In Canada la caccia alle foche significa lavoro, ma gli oppositori affermano che è crudele e che mette a rischio d'estinzione la specie, minacciata anche dal riscaldamento globale. **Hypothesis**: La caccia mette a rischio d'estinzione le specie delle foche. **Entailment**: YES **Original Explanation**: Se le foche non venissero cacciate, la loro specie non sarebbe a rischio. **Edited Explanation**: Se è vero che la caccia alle foche è crudele e mette a rischio d'estinzione la specie, allora è vero che la caccia metta a rischio d'estinzione la specie delle foche.
Example 3	**Text:** Basandosi su uno studio mondiale [...] gli epidemiologi [...] dimostrano che il fumo è la causa principale degli incendi e delle morti per incendi nel mondo. **Hypothesis**: Gli incendi domestici sono una causa importante delle morti da incendio. **Entailment**: UNKNOWN **Original Explanation**: Il fumo è la causa principale degli incendi domestici. **Edited Explanation**: Il fumo è la causa principale degli incendi e delle morti per incendio, ma non è specificato se un'altra causa importante di morti da incendio siano proprio gli incendi domestici.

The string within stars ** (the year of death) was missing from the Italian dataset but was present in the original English RTE-3 dataset. This information was essential to infer the entailment relationship, and was re-introduced by checking the original English version and restoring the original label.

Moreover, the Italian RTE-3 dataset, as reported in the description[9], changed the original label (from "YES": entailment, to "NO": contradiction) in 15 pairs, creating a mismatch with the English dataset. To ensure comparability, we decided to restore the original label provided by the English dataset, as our annotators were still able to express an alternative label in case they did not agree with it.

9 The original RTE 3 Italian dataset description can be found at `https://nlplab.fbk.eu/tools-and-resources/lexical-resources-and-corpora/rte-3-ita`

Within the 15 restored labels, annotators expressed an alternative to the original label only in the development set, where they provided an alternative label "NO" in pairs 51, 490, 549, and a label "UNKNOWN" in pair 604. In all other cases, they agreed with the original, restored label. For these reasons, the e-RTE-3-it dataset can also be regarded as an emended, manually curated version of the original RTE-3-it dataset.

5. Dataset Description and Analysis

The final dataset comprises 1600 text-hypothesis pairs, maintaining two dev/test splits of 800 pairs each. Each pair inherits the original dataset's attributes (pair's ID, entailment relation, original task and length) and is complemented with additional annotations layers, specifically one explanation for the original label and a confidence score for each pair, and alternative labels with their respective explanations and confidence scores for about 9% of the cases. In Figure 2, we provide a snippet from the test set (and its translation on the right side), exemplifying alternative labels and explanations.

```
<?xml version="1.0" encoding="UTF-8"?>
<pair id="201" entailment="YES" task="IR" length="
     short">
<t>Berlino ha un nuovo punto di riferimento. Sopra le
      gru che ancora dominano l'orizzonte della
     nuova capitale dell'Europa adesso c'e' una
     cancelleria, dove vivra' il capo del governo
     Gerhard Schroeder e il governo tedesco terra' i
      suoi incontri regolari.</t>
<h>Nuovi edifici sono stati eretti a Berlino.</h>
<e confidence="4">La frase "sopra le gru... adesso c'
     e' una cancelleria" e' da intendersi in modo
     figurato, e indica che e' stato costruito un
     nuovo edificio dove ha sede la cancelleria.</e>
<a confidence="5" new_label="UNKNOWN">Il fatto che
     ora sopra le gru c'e' una cancelleria, non
     implica che nuovi edifici sono stati eretti a
     Berlino.</a>
</pair>
```

```
<?xml version="1.0" encoding="UTF-8"?>
<pair id="201" entailment="YES" task="IR" length="
     short">
<t>Berlin has a new landmark. Among the cranes which
     still dominate the skyline of Europe's newest
     capital now stands a chancellery, where the
     head of government Gerhard Schroeder will live
     and the German cabinet will hold its regular
     meetings.</t>
<h>New buildings have been erected in Berlin.</h>
<e confidence="4">The statement "among the cranes...
     now stands a chancellery" is meant figuratively
     , and indicates that a new building has been
     built where the chancellery is located.</e>
<a confidence="5" new_label="UNKNOWN">The fact that
     there is now a chancellery among the cranes
     does not imply that new buildings have been
     erected in Berlin.</a>
</pair>
```

Figure 2
Example for the e-RTE-3-it dataset, with explanations, confidence scores and alternative label.

Table 3
Statistics for the enriched e-RTE-3-it dataset. Number of labels are in absolute values, confidence values are averaged over all pairs on a scale from 1 to 5. "New labels from" indicate times when the original label (row) was changed, and to which label (column).

	Total/Average	Entailment	Contradiction	Neutrality
Original label	1600	800	150	650
New labels in ⟨a⟩	147	48	62	37
New labels from entailment	41	–	10	31
New labels from contradiction	6	0	–	6
New labels from neutrality	100	48	52	–
Confidence (mean) in ⟨e⟩	4.00	4.17	4.03	3.81
Confidence (mean) in ⟨e⟩w/o ⟨a⟩	4.01	4.24	4.05	3.91
Confidence (mean) in ⟨e⟩with ⟨a⟩	3.14	2.78	3.33	3.28
Confidence (mean) in ⟨a⟩	3.48	3.35	3.47	3.65

5.1 Subjectivity in RTE: Alternative Labels

Table 3 reports a detailed description of the dataset's annotation. The original labels in the dataset exhibit a distribution of 50% 'entailment', 10% 'contradiction', and 40% 'neutrality'. If we consider disagreements from the original label (147 pairs) as per our annotation, we observe an increase of the contradiction relationship to 13% and a decrease to 37% of the neutrality label. As an example, consider the following:

> ***Text***: *Finora non ci sono segnalazioni di qualche parente che abbia reclamato i corpi dei quattro uomini delle forze armate che sono presumibilmente morti quando l'aereo si schiantò.*
> ***Hypothesis***: *Quattro uomini delle forze armate morirono in uno schianto aereo.*[10]

The original label in the Italian RTE-3 dataset was 'YES'. However, an annotator disagreed and assigned the alternative label 'NO', explaining: *Affermando che quattro uomini delle forze armate sono presumibilmente morti quando l'aereo si schiantò, si manifesta una mancata certezza totale dell'episodio*.[11] The annotator rated their confidence in this explanation as 4.

We observed that among the cases where annotators disagreed with the original label, the most frequent changes were from the "neutrality" label 'UNKNOWN' to either "contradiction" ('NO', 52 changes) or to "entailment" ('YES', 48). Upon examining the explanations provided for these revised labels, a common theme emerged: they often stated that the interpretation of the hypothesis needed to assign the original neutrality label was too narrow, and did not match with commonsense reasoning and inferences commonly made in discourse.

For example, in a case when an annotator changed the label from neutrality to entailment, 't' and 'h' stated that:

> ***Text***: *[...] Michael Howard non riuscì a scalzare il Governo Laburista, sebbene i Conservatori avessero guadagnato 33 seggi.*
> ***Hypothesis***: *i Conservatori ottennero 33 seggi".*[12]

Here, the usual interpretation would be that they obtained *at least*, and not *exactly* 33 seats, explaining that *Guadagnare in questo caso è sinonimo di ottenere*[13].

In another case when the annotator changed the label from "UNKNOWN" to "NO" with confidence 4, 't' and 'h' stated:

> ***Text***: *I proprietari di Phinda, l'Ente per la Conservazione con base in Sud Africa, non avrebbero potuto pagare per una pubblicità migliore per la loro filosofia di tutela della natura: un approccio alla tutela basato sulle persone, che sta lentamente guadagnando terreno in Africa poiché le riserve di caccia sono sempre piü minacciate dalle popolazioni locali affamate, povere e arrabbiate.*
> ***Hypothesis***: *L'Ente per la Conservazione con base in Sud Africa minaccia la popolazione locale.*
> ***Explanation for new label "NO"***: *L'Ente per la Conservazione con base in Sud Africa basa il*

10 **Text:** There are no reports so far as to whether any relatives have claimed the bodies of the four military men who were reportedly killed when the plane crashed. **Hypothesis**: Four military men died in a plane crash.

11 Claiming that four military men allegedly died when the plane crashed manifests a lack of certainty about the incident.

12 **Text:** [...] Michael Howard failed to unseat the Labour Government, although the Conservatives did gain 33 seats. **Hypothesis:** In the May 2005 general election Conservatives got 33 seats.

13 In this case *gain* is a synonym for *obtain*.

suo rapporto di tutuela sulle persone, sulle popolazioni povere e affamate, quindi aiutandole non minacciandole.[14]

Cases like these underline the subtleties involved in the inference process, and how tightly it connects to the interpretation of words in context, which may also be influenced by some level of subjectivity, an observation that paves the way for further investigation.

5.2 Uncertainty in RTE: Confidence Judgements

As can be seen in Table 3, the confidence scores assigned by annotators to explanations were generally high, with a mean score of <e> of 4 on a 5-point Likert scale and the highest score being given to the entailment label. However, when annotators disagreed with the original label (and an alternative label was given) the mean confidence score for <e> decreased to 3.14 and the entailment label became the label with the lowest score (2.78). When an alternative label is expressed, the mean confidence in <a> (3.48) is higher than confidence in <e> (3.14). However, if we consider overall confidence, scores for <a> are lower than those of <e>, indicating that while annotators felt confident in their judgements when they agreed with the label, cases involving label revision posed more challenges and involved a higher degree of uncertainty.

5.3 Properties of Explanations: Lexical Variety

We were also interested in the lexical variety of both the original sentences and the collected explanations. As displayed in Table 4, we notice that while the overall mean type/token ratio for each sentence is very high, indicating that few words are repeated in the same sentence, if we look at the lexical overlap between the various components, we notice a low overlap of the alternative-label explanation and the fields, which seems to indicate that the alternative explanations introduce new information compared to the input. The high overlap from 'e' to 'h' seems to indicate that original-label explanations rely on information that is in the hypothesis more than the explanations for the alternative label. This preliminary analysis encourages us to further look into the properties of explanations, in order to understand how different explanations, carrying different information, interact with the input and with explanation label-predictive power (see Section 6).

6. Explanations in Practice: Experiments and Baselines

One of the straightforward applications of our dataset is *explanation generation*, where human-written explanations usually represent the "reference".

One way to indirectly evaluate explanation quality is *simulatability* (Hase et al. 2020), i.e., the delta between the performance of a model without the explanation and the

14 **Text:** The owners of Phinda, the South African-based Conservation Corporation, could not have paid for a better advertisement for their philosophy of wildlife conservation: a people-based approach to conservation, which is slowly gaining ground in Africa as game reserves come under ever greater threat from hungry, poor and angry local populations. **Hypothesis:** The South African-based Conservation Corporation threatens the local population. **Explanation for new label "NO":** The South African-based Conservation Corporation bases its conservation approach on people, on poor and hungry people, therefore helping them, and not threatening them.

Table 4
Lexical variety in the dataset. The lexical overlapping is calculated by taking the word types of the element in the row also present in the element in the column, divided by the number of types of the element in the row.

mean length	**t**	**h**	**e**	**a**
length (**tokens**)	34	9	22	23
length (**types**)	30	9	19	20
types/tokens **ratio**	0.9	0.99	0.88	0.89
lexical overlapping	**t**	**h**	**e**	**a**
t	1.000	0.726	0.665	0.764
h	0.220	1.000	0.421	0.530
e	0.389	0.813	1.000	0.769
a	0.082	0.188	0.141	1.000

performance of the same model using the explanation. This section reports baseline experiments conducted on our dataset to study the effects of explanations on model predictions.

For our experiments we use two multilingual SOTA LLMs: GPT-3.5 (OpenAI 2023) and Mistral 8x7b (Jiang et al. 2024). Firstly, we prompt the language models to solve the RTE task on the test set without any additional information, and evaluate their accuracy (`no-exp` setting). Secondly, we provide the model with either the explanation for the original label or that for the alternative label (`human`) by inserting it into the prompt as a "hint" for the model to predict the label, to see if they help the models increase their prediction accuracy. To avoid that explanation effectiveness be biased by explanations that simply suggest the answer without really explaining, we substitute any direct reference to the labels with a placeholder (`XXX`). Moreover, to check that the results are not simply due to a larger size of the input text, we also define another baseline (*dummy*), where instead of an actual explanation we inject a copy of the *hypothesis* as hint, which does not add any new information.

We report results in Table 6 considering three subsets: 1. the instances where no alternative label was expressed; 2. the subset with alternative labels where the original explanation was injected, and results are evaluated on the original label; 3. the subset with alternative labels where the alternative explanation was injected, and results are evaluated on the original label.

Results suggest that, while non informative explanations (dummy) can hurt performance, using quality explanations (human) positively affects predictions, with Mistral benefiting most of explanation injection. This is confirmed by the experiments on the subset with alternative labels, showing even stronger influence for Mistral of explanation injection compared to baselines, where figures are lower than the original-label subset, confirming that the subset featuring alternative explanations is indeed a particularly challenging one. GPT-3.5, on the other hand, seems to be less sensitive to explanations, having positive but weaker gains after explanation injection.

7. Conclusions and Future Work

In this paper, we presented the *e-RTE-3-it* dataset, an enriched and emended version of the Italian RTE-3 dataset featuring human-written explanations for each label, the level

Table 5
Effects of explanation injection on accuracy on test sets of the two models.

	MODEL	no-exp	dummy	human
No alternative explanation	GPT-3.5	63	58	65
	Mistral 8x7b	73	64	86
Original explanation - evaluated on original label	GPT-3.5	44	44	54
	Mistral 8x7b	54	52	86
Alternative explanation - evaluated on alternative label	GPT-3.5	40	42	56
	Mistral 8x7b	36	36	82

of confidence in writing the explanation, and an optional layer of alternative labels, explanations and confidence scores. The insights derived from the 'e-RTE-3-it' dataset pave the way for multifaceted research directions. The provided explanations can serve as a gold standard for training models to generate human-like explanations. Further, the alternative labels and explanations open avenues for investigating the subjectivity in language understanding. The rich layers of the dataset also allow for the study of correlation between the original and alternative labels, the confidence score, and the degree of disagreement among annotators. Future work includes utilising the data to develop models capable of providing explanations for their entailment decisions and conducting a deeper analysis into the dynamics of subjectivity in the entailment task.

Acknowledgements

This work has been partially supported by the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU and by the ANTIDOTE project (CHIST-ERA grant of the Call XAI 2019 of the ANR with the grant number Project-ANR-21-CHR4-0002).

References

Becker, Maria, Siting Liang, and Anette Frank. 2021. Reconstructing implicit knowledge with language models. In Eneko Agirre, Marianna Apidianaki, and Ivan Vulić, editors, *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 11–24, Online, June. Association for Computational Linguistics.

Bowman, Samuel R., Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, September. Association for Computational Linguistics.

Brahman, Faeze, Vered Shwartz, Rachel Rudinger, and Yejin Choi. 2021. Learning to rationalize for nonmonotonic reasoning with distant supervision. In *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI-21) Conference on Artificial Intelligence*, volume 35(14), pages 12592–12601, Online, Febuary.

Brassard, Ana, Benjamin Heinzerling, Pride Kavumba, and Kentaro Inui. 2022. COPA-SSE: semi-structured explanations for commonsense reasoning. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, pages 3994–4000. European Language Resources Association.

Camburu, Oana-Maria, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31.

Chen, Michael, Mike D'Arcy, Alisa Liu, Jared Fernandez, and Doug Downey. 2019. CODAH: An adversarially-authored question answering dataset for common sense. In Anna Rogers, Aleksandr Drozd, Anna Rumshisky, and Yoav Goldberg, editors, *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 63–69, Minneapolis, USA, June. Association for Computational Linguistics.

Chierchia, Gennaro and Sally McConnell-Ginet. 2000. *Meaning and grammar: An introduction to semantics*. MIT press Cambridge, MA.

Creanga, Claudiu and Liviu P. Dinu. 2024. Designing NLP systems that adapt to diverse worldviews. In Gavin Abercrombie, Valerio Basile, Davide Bernardi, Shiran Dudy, Simona Frenda, Lucy Havens, and Sara Tonelli, editors, *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 95–99, Torino, Italia, May. ELRA and ICCL.

Dagan, Ido, Bill Dolan, Bernardo Magnini, and Dan Roth. 2010. Recognizing textual entailment: Rational, evaluation and approaches–erratum. *Natural Language Engineering*, 16(1):105–105.

Dagan, Ido, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In Joaquin Quiñonero-Candela, Ido Dagan, Bernardo Magnini, and Florence d'Alché Buc, editors, *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190, Berlin, Heidelberg. Springer Berlin Heidelberg.

Guidotti, Riccardo, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2019. A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5):93:1–93:42.

Hase, Peter, Shiyue Zhang, Harry Xie, and Mohit Bansal. 2020. Leakage-adjusted simulatability: Can models generate non-trivial explanations of their behavior in natural language? In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: Empirical Methods in Natural Language Processing 2020*, pages 4351–4367, Online, November. Association for Computational Linguistics.

Honovich, Or, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. True: Re-evaluating factual consistency evaluation. *arXiv preprint arXiv:2204.04991*.

Ji, Ziwei, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2022. Survey of hallucination in natural language generation. *ACM Computing Surveys*.

Jiang, Albert Q., Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of Experts, January. arXiv:2401.04088 [cs].

Kavumba, Pride, Naoya Inoue, Benjamin Heinzerling, Keshav Singh, Paul Reisert, and Kentaro Inui. 2019. When choosing plausible alternatives, clever hans can be clever. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, pages 33–42, Hong Kong, China, November. Association for Computational Linguistics.

Lombrozo, Tania. 2007. Simplicity and probability in causal explanation. *Cognitive Psychology*, 55(3):232–257.

Longo, Luca, Mario Brcic, Federico Cabitza, Jaesik Choi, Roberto Confalonieri, Javier Del Ser, Riccardo Guidotti, Yoichi Hayashi, Francisco Herrera, Andreas Holzinger, Richard Jiang, Hassan Khosravi, Freddy Lecue, Gianclaudio Malgieri, Andrés Páez, Wojciech Samek, Johannes Schneider, Timo Speith, and Simone Stumpf. 2024. Explainable artificial intelligence (xai) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion*, 106:102301.

Magnini, Bernardo, Alberto Lavelli, and Simone Magnolini. 2020. Comparing machine learning and deep learning approaches on NLP tasks for the Italian language. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2110–2119, Marseille, France, May. European Language Resources Association.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Rajani, Nazneen Fatema, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy, July. Association for Computational Linguistics.

Santilli, Andrea. 2023. Camoscio: An italian instruction-tuned llama. `https://github.com/teelinsan/camoscio`.

Sarti, Gabriele and Malvina Nissim. 2022. It5: Large-scale text-to-text pretraining for italian language understanding and generation. *ArXiv preprint 2203.03759*, mar.

Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Wiegreffe, Sarah and Ana Marasovic. 2021. Teach me to explain: A review of datasets for explainable natural language processing. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, Online, December.

Zaninello, Andrea and Bernardo Magnini. 2023. A smashed glass cannot be full: Generation of commonsense explanations through prompt-based few-shot learning. In Bhavana Dalvi Mishra, Greg Durrett, Peter Jansen, Danilo Neves Ribeiro, and Jason Wei, editors, *Proceedings of the 1st Workshop on Natural Language Reasoning and Structured Explanations (NLRSE)*, pages 18–29, Toronto, Canada, June. Association for Computational Linguistics.