Do Natural Language Explanations Represent Valid Logical Arguments? Verifying Entailment in Explainable NLI Gold Standards

Marco Valentino[†], **Ian Pratt-Hartman**[†], **André Freitas**[†‡]
Department of Computer Science, University of Manchester, United Kingdom[†]
Idiap Research Institute, Switzerland[‡]
{marco.valentino,ian.pratt,andre.freitas}
@manchester.ac.uk

Abstract

An emerging line of research in Explainable NLP is the creation of datasets enriched with human-annotated explanations and rationales, used to build and evaluate models with step-wise inference and explanation generation capabilities. While human-annotated explanations are used as ground-truth for the inference, there is a lack of systematic assessment of their consistency and rigour. In an attempt to provide a critical quality assessment of Explanation Gold Standards (XGSs) for NLI, we propose a systematic annotation methodology, named *Explanation Entailment Verification* (EEV), to quantify the logical validity of human-annotated explanations.

The application of EEV on three mainstream datasets reveals the surprising conclusion that a majority of the explanations, while appearing coherent on the surface, represent logically invalid arguments, ranging from being incomplete to containing clearly identifiable logical errors. This conclusion confirms that the inferential properties of explanations are still poorly formalised and understood, and that additional work on this line of research is necessary to improve the way Explanation Gold Standards are constructed.

Worldtree

Question: Which of the following characteristics would best help a tree survive the heat of a forest fire?
[A] large leaves [B] shallow roots
[*C] thick bark [D] thin trunks

Explanation:
Protecting something means preventing harm.
Fire causes harm to trees, forests, and other living things.
Thickness is a measure of how thick an object is.
A tree is a kind of living thing.

QASC

Question: Differential heating of air can be harnessed for what?
[*A] electricity production [B] erosion prevention
[C] transfer of electrons [D] reduce acidity of food

Explanation:
Differential heating of air produces wind.
Wind is used for producing electricity.

e-SNLI

Premise: A man in an orange vest leans over a pickup truck.
Hypothesis: A man is touching a truck.
Label: entailment

Explanation:
Man leans over a pickup truck implies that he is touching it.

Figure 1: Does the answer logically follow from the explanation? While step-wise explanations are used as ground-truth for the inference, there is a lack of assessment of their consistency and rigour. We propose EEV, a methodology to quantify the logical validity of human-annotated explanations.

1 Introduction

Explanation Gold Standards (XGSs) are emerging as a fundamental enabling tool for step-wise and explainable Natural Language Inference (NLI). Resources such as WorldTree (Xie et al., 2020; Jansen et al., 2018), QASC (Khot et al., 2020), among others (Wiegreffe and Marasović, 2021; Thayaparan et al., 2020b; Bhagavatula et al., 2020; Camburu et al., 2018) provide a corpus of linguistic evidence on how humans construct explanations that are perceived as plausible, coherent and complete.

Designed for tasks such as Textual Entailment (TE) and Question Answering (QA), these reference datasets are used to build and evaluate models with step-wise inference and explanation generation capabilities (Valentino et al., 2021; Cartuyvels et al., 2020; Kumar and Talukdar, 2020; Rajani et al., 2019). While these explanations are used as ground-truth for the inference, there is a lack of systematic assessment of their consistency and rigour, introducing inconsistency biases within the models.

This paper aims to provide a critical quality assessment of Eplanation Gold Standards for NLI in terms of their logical inference properties. By

systematically translating natural language explanations into corresponding logical forms, we induce a set of recurring logical violations which can then be used as testing conditions for quantifying quality and logical consistency in the annotated explanations. More fundamentally, the paper reveals the surprising conclusion that a majority of the explanations present in explanation gold standards contain one or more major logical fallacies, while appearing to be coherent on the surface. This study reveals that the inferential properties of explanations are still poorly formalised and understood.

The main contributions of this paper can be summarised as:

1. Proposal of a systematic methodology, named *Explanation Entailment Verification* (EEV), for analysing the logical consistency of NLI explanation gold-standards.

2. Validation of the quality assessment methodology for three contemporary and mainstream reference XGSs.

3. The conclusion that most of the annotated human-explanations in the analysed samples represent logically invalid arguments, ranging from being incomplete to containing clearly identifiable logical errors.

2 Related Work

An emerging line of research in Explainable NLP is focused on the creation of datasets enriched with human-annotated explanations and rationales (Wiegreffe and Marasović, 2021). These resources are often adopted as Explanation Gold Standards (XGSs), providing additional supervision for training and evaluating explainable models capable of generating natural language explanations in support of their predictions (Valentino et al., 2021, 2020; Kumar and Talukdar, 2020; Cartuyvels et al., 2020; Thayaparan et al., 2020a; Rajani et al., 2019).

XGSs are designed to support Natural Language Inference, asking human-annotators to transcribe the reasoning required for deriving the correct prediction (Thayaparan et al., 2020b). Despite the popularity of these datasets, and their application for measuring explainability on tasks such as Textual Entailment (Camburu et al., 2018), Multiple-choice Question Answering (Xie et al., 2020; Jhamtani and Clark, 2020; Khot et al., 2020; Jansen et al., 2018), and other inference tasks (Wang et al., 2020; Ferreira and Freitas, 2020b,a; Bhagavatula et al., 2020), little has been done to provide a clear understanding on the nature and the quality of the reasoning encoded in the explanations.

Previous work on explainability evaluation has mainly focused on methods for assessing the quality and faithfulness of explanations generated by deep learning models (Camburu et al., 2020; Subramanian et al., 2020; Kumar and Talukdar, 2020; Jain and Wallace, 2019; Wiegreffe and Pinter, 2019). Our work is related to this research, but focuses instead on the resources on which explainable models are trained. In that sense, this paper is more aligned to gold standard evaluation methods, which aim to design systematic approaches to qualify the content and the inference capabilities involved in mainstream NLP benchmarks (Lewis et al., 2021; Bowman and Dahl, 2021; Schlegel et al., 2020; Ribeiro et al., 2020; Pavlick and Kwiatkowski, 2019; Min et al., 2019). However, to the best of our knowledge, none of these methods have been adopted to provide a critical assessment of human-annotated explanations present in XGSs.

3 Explanation Gold Standards

Given a generic classification task T, an Explanation Gold Standard (XGS) is a collection of distinct instances of T, $XGS(T) = \{I_1, I_2, \ldots, I_n\}$, where each element of the set, $I_i = \{X_i, s_i, E_i\}$, includes a problem formulation X_i, the expected solution s_i for X_i, and a human-annotated explanation E_i.

In general, the nature of the elements in a XGS can vary greatly according to the task T under consideration. In this work, we restrict our investigation to Natural Language Inference (NLI) tasks, such as Textual Entailment and Question Answering, where problem formulation, expected solution, and explanations are entirely expressed in natural language.

For this class of problems, the explanation is typically a composition of sentences, whose role is to describe the reasoning required to arrive at the final solution. As shown in the examples depicted in Figure 1, the explanations are constructed by human annotators transcribing the commonsense and world knowledge necessary for the correct answer to hold. Given the nature of XGSs for NLI, we hypothesise that a human-annotated explanation represents a valid set of premises from which the expected solution logically follows.

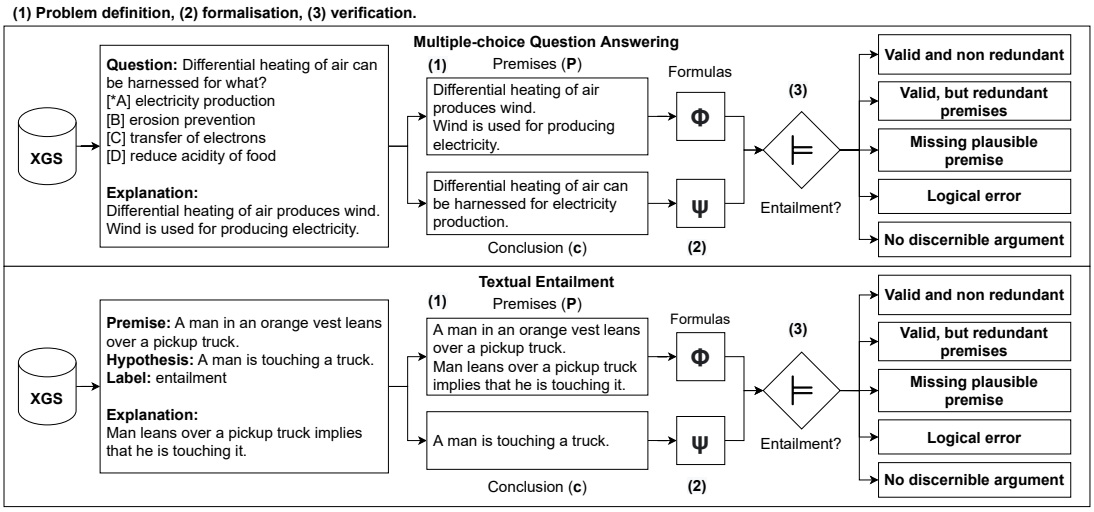

Figure 2: Overview of the Explanation Entailment Verification (EEV) applied to different NLI problems. EEV takes the form of a multi-label classification problem where, for a given NLI problem, a human annotator has to qualify the validity of the inference process described in the explanation through a pre-defined set of classes.

In order to validate or reject this hypothesis, we design a methodology aimed at evaluating XGSs in terms of logical entailment, quantifying the extent to which human-annotated explanations actually entail the final answer.

4 Explanation Entailment Verification

We present a methodology, named Explanation Entailment Verification (EEV), aimed at quantifying and assessing the quality of human-annotated explanations in XGS for NLI tasks, in terms of their logical inference properties.

To this end, we design an annotation framework that takes the form of a multi-label classification problem defined on a XGS. Specifically, the goal of EEV is to label each element in a XGS, $I_i = \{X_i, s_i, E_i\}$, using one of a predefined set of classes qualifying the validity of the inference process described in the explanation E_i.

Figure 2 shows a schematic representation of the annotation pipeline. One of the challenges involved in the design of a standardised methodology for EEV is the formalisation of an annotation task that is applicable to NLI problems with different shapes, such as Textual Entailment (TE) and Multiple-choice Question Answering (MCQA). To minimise the ambiguity in the annotation and make it independent of the specific NLI task, we define a methodology composed of three major steps: (1) *problem definition*; (2) *formalisation*; and (3) *verification*.

In the problem definition step, each example I_i in the XGS is translated into an entailment form ($P \models c$), identifying a set of sentences P representing the premises for the entailment, and a single sentence c representing its conclusion. As illustrated in Figure 2, this step defines an entailment problem with a single surface form that allows abstracting from the NLI task under investigation.

In the formalisation step, the sentences in P and c are translated into a logical form ($\Phi \models \psi$). Specifically, the formalisation is performed using event-based semantics, in which verbs correspond to event-types, and their objects to semantic roles (additional details on the formalism are provided in section 4.3). This step aims to minimise the ambiguity in the interpretation of the meaning of the sentences, supporting the annotators in the identification of logical errors and gaps in the explanations, and maximise the inter-annotator agreement in the downstream verification task.

The final step corresponds to the actual multi-label classification problem. Specifically, the annotators are asked to verify whether the formalised set of premises Φ entails the conclusion ψ ($\Phi \models \psi$) and to classify the explanation in the corresponding example $I_i = \{X_i, s_i, E_i\}$ selecting one of the following classes: (1) *Valid and non redundant*; (2) *Valid, but redundant premises*; (3) *Missing plausible premise*; (4) *Logical error*; (5) *No discernible argument*. The classes are mutually exclusive: each example can be assigned to one and only one label.

After EEV is performed for each instance in the dataset, the frequencies of the classification labels can be adopted to estimate and evaluate the

overall entailment properties of the explanations in the XGS under consideration.

4.1 Problem definition

The problem definition step consists in the identification of the sentences in $I_i = \{X_i, s_i, E_i\}$ that will compose the set of premises P and the conclusion c for the entailment problem $P \models c$.

Here, we describe the procedure adopted for translating a specific NLI task into the entailment problem of interest given its original surface form. In particular, we employ two different translation procedures for Textual Entailment (TE) and Multiple-choice Question Answering (MCQA) problems.

Textual Entailment (TE). For a TE task, the problem formulation X_i is generally composed of two sentences, p and h, representing a premise and a hypothesis (see e-SNLI in figure 1). Each example in a TE task can be classified using one of the following labels: *entailment*, *neutral*, and *contradiction* (Bowman et al., 2015). In this work, we focus on examples where the expected solution s_i is *entailment*, implying that the hypothesis h is a consequence of the premise p. Therefore, to define the entailment verification problem, we simply include the premise p in P and consider the hypothesis h as a the conclusion c. For this class of problems, the explanation E_i describes additional factual knowledge necessary for the entailment $p \models h$ to hold (Camburu et al., 2018). Specifically, the sentences in E_i can be interpreted as a further set of premises for the entailment verification problem and are included in P.

Multiple-choice Question Answering (MCQA). In the case of MCQA, X_i is typically composed of a question $Q_i = \{c_1, \ldots, c_n, q\}$, and a set of mutually exclusive candidate answers $A_i = \{a_1, \ldots, a_m\}$ (see QASC and Worldtree in figure 1). In this case, the expected label s_i corresponds to one of the candidate answers in A_i (Jansen et al., 2018; Khot et al., 2020). Q_i can include a set of introductory sentences $c_1, \ldots, c_n$ acting as a context for the question q. We consider each sentence c_i in the context as a premise for q and include it in P. Similarly to TE, we interpret the explanation E_i for a MCQA example as a set of premises that entails the correct answer s_i. Therefore, the sentences in E_i are included in P. The question q takes the form of an elliptical assertion, and the candidate answers are possible substitutions for the ellipsis. Therefore, to derive the conclusion c, we adopt the correct answer s_i as a substitution for the ellipsis in q. Details on the formalisation adopted for MCQA problems are described in section 4.3.

4.2 Verification

In the verification step, the annotators adopt the formalised set of premises Φ and conclusion ψ to classify the entailment problem in one of the following categories:

1. **Valid and non-redundant:** The argument is formally valid, and all premises are required for the derivation.

2. **Valid, but redundant premises:** The argument is formally valid, but some premises are not required for the derivation. This includes the cases where more than one premise is present, and the conclusion simply repeats one of the premises.

3. **Missing plausible premise:** The argument is formally invalid, but would become valid on addition of a reasonable premise, such as, for example, *"If x affects y, then a change to x affects y"*, or *"If x is the same height as y and y is not as tall as z then x is not as tall as z"*.

4. **Logical error:** The argument is formally invalid, apparently as a result of confusing *"and"* and *"or"* or *"some"* and *"all"*, or of illicitly changing the direction of an implication.

5. **No discernible argument:** The argument is invalid, no obvious rescue exists in the form of a missing premise, and no simple logical error can be identified.

4.3 Formalisation

In this section, we describe an example of formalisation for a MCQA problem. A typical multiple-choice problem is a triple consisting of a *question* Q together with a set of *candidate answers* $A_1, \ldots, A_m$. It is understood that Q takes the form of a elliptical assertion, and the candidate answers are possible substitutions for the ellipsis. The task is to determine which of the candidate answers would result in an assertion entailed by some putative knowledge-base. The corpora investigated feature a list of multiple-choice textual entailment problems together, in each case, with a specification of a correct answer and an *explanation* in the form of a set of assertions Φ from the knowledge

base providing a justification for the answer. For example, the following problem together with its resolution is taken from the Worldtree corpus (Jansen et al., 2018).

Question: A group of students are studying bean plants. All of the following traits are affected by changes in the environment except ...

Candidate answers: [A] leaf color. [B] seed type. [C] bean production. [D] plant height.

Correct answer: B

Explanation: (i) The type of seed of a plant is an inherited characteristic; (ii) Inherited characteristics are the opposite of learned characteristics; acquired characteristics; (iii) An organism's environment affects that organism's acquired characteristics; (iv) A plant is a kind of organism; (v) A bean plant is a kind of plant; (vi) Trait is synonymous with characteristic.

In formalising such problems, we represent the question as a sentence of first-order logic featuring a schematic formula variable P (corresponding to the ellipsis), and the candidate answers as first-order formulas. In the above example, we assume that the essential force of the question to find a characteristic of plants *not* affected by those plants' environments. That is, we are asked for a P making the schematic formula

$$\begin{aligned}\forall xyzwe(&\text{bnPlnt}(x) \wedge \text{env}(y,x) \wedge \\ &\text{changeIn}(z,y) \wedge \text{trait}(w,x) \wedge \text{affct}(e) \wedge \\ &\text{agnt}(e,z) \wedge P \rightarrow \neg\text{ptnt}(e,w)). \quad (1)\end{aligned}$$

into a true statement. We formalise the correct answer (B) by the atomic formula $\text{sdTp}(w,x)$ "w is the seed type of x", with the other candidate answers formalised similarly. In choosing predicates for formalisation, we typically render common noun-phrases using predicates, taking these to be relational if the context demands (e.g. "environment/seed type *of* a plant x"). In addition, we typically render verbs as predicates whose arguments range over eventualities (events, processes, etc.), related to their participants via a standard list of binary "semantic role" predicates (agent, patient, theme) etc. Thus, to say that "x affects y" is to report the existence of an eventuality e of type "affecting", such that x is the agent of e and y its patient. This approach, although somewhat strained in many general contexts, aids standardization and, more importantly, also makes it easier to deal with adverbial phrases. Of course, many choices in formalisation strategy inevitably remain.

The knowledge-base excerpt Φ is formalised straightforwardly as a finite set of first-order formulas, following the same general rendering policies. In the case of the above example, sentences (i), (ii) and (iv)–(vi) in Φ might be formalised as:

$$\begin{aligned}&\forall xy(\text{plnt}(x) \wedge \text{sdTp}(y,x) \rightarrow \text{char}(y,x) \wedge \text{inhtd}(y)) \\ &\forall xy(\text{char}(x,y) \wedge \text{inhtd}(x) \rightarrow \neg\text{acqrd}(x)) \\ &\forall x(\text{plnt}(x) \rightarrow \text{orgnsm}(x)) \\ &\forall x(\text{bnPlnt}(x) \rightarrow \text{plnt}(x)) \\ &\forall xy(\text{trait}(x,y) \leftrightarrow \text{char}(x,y)),\end{aligned}$$

with the more complicated sentence (iii) formalised as

$$\begin{aligned}\forall xyw(&\text{orgnsm}(x) \wedge \text{env}(y,x) \wedge \\ &\text{char}(w,x) \wedge \text{acqrd}(w) \rightarrow \quad (2) \\ &\exists e(\text{affct}(e) \wedge \text{agnt}(e,y) \wedge \text{ptnt}(e,w)))\end{aligned}$$

Denoting by ψ the result of substituting $\text{sdTp}(w,x)$ for P in (1), we ask ourselves: Does Φ entail ψ? A moment's thought shows that it does not. At the very least, statement (iii) in the explanation, whose *prima facie* formalisation is (2), must instead be read as asserting that an organism's environment affects *only* that organism's acquired characteristics, that is to say:

$$\begin{aligned}\forall xyw(&\text{orgnsm}(x) \wedge \text{env}(y,x) \wedge \text{char}(w,x) \wedge \\ &\exists e(\text{affct}(e) \wedge \text{agnt}(e,y) \wedge \text{ptnt}(e,w)) \rightarrow \quad (3) \\ &\text{acqrd}(w)).\end{aligned}$$

This is not unreasonable, of course. Generalizations in natural language are notoriously vague as to the direction of implication; let Φ' be the result of substituting (3) for (2) in Φ. Does Φ' entail ψ? Again, no. The problem this time is that, model-theoretically speaking, just because something is affected by a *change in* its environment, that does not mean to say it is affected by its environment. An assertion to the effect that it is would have to be postulated:

$$\begin{aligned}\forall xyzw(&\text{env}(y,x) \wedge \text{changeIn}(z,y) \wedge \\ &\exists e(\text{affct}(e) \wedge \text{agnt}(e,z) \wedge \text{ptnt}(e,w)) \rightarrow \\ &\exists e(\text{affct}(e) \wedge \text{agnt}(e,y) \wedge \text{ptnt}(e,w))).\end{aligned}$$

Let Φ'' be the result of augmenting Φ' in this way. Then Φ'' does indeed entail ψ.

Feature	Worldtree	QASC	e-SNLI
Task	MCQA	MCQA	TE
Multi-hop	yes	yes	no
Crowd-sourced	no	yes	yes
Explanation type	generated + composed	composed	generated
Avg. number of sentences	6	2	1

Table 1: Features of the datasets selected for the Explanation Entailment Verification (EEV).

Applying a general principle of charity, it is reasonable to take the interpretation of the explanation to be given by Φ'. However, the additional premise required to obtain Φ'' seems to have been forgotten. Although not a logical truth, it has the status of a plausible general principle of the kind that is frequently explicitly articulated in the Worldtree database. Therefore, we classify this example as a *missing plausible premise*.

5 Corpus Analysis

We employ EEV to analyse a set of contemporary XGSs designed for Textual Entailment and Multiple-choice Question Answering.

In the following sections, we describe the methodology adopted for extracting a representative sample from the selected XGSs, and for implementing the annotation pipeline efficiently. Finally, we present the results of the annotation, reporting the frequency of each entailment verification class and presenting a list of qualitative examples to provide additional insights on the logical properties of the analysed explanations.

5.1 Selected Datasets

We select three contemporary XGSs with different and complementary characteristics. In particular, we apply our methodology to two MCQA datasets (Worldtree (Jansen et al., 2018), QASC(Khot et al., 2020)) and one TE benchmark (e-SNLI (Camburu et al., 2018)).

The main features of the selected XGSs are reported in Table 1. *Multi-hop* indicates whether the problem requires step-wise reasoning, combining more than one sentence to compose the final explanation. *Crowd-sourced* indicates whether the resource is curated using standard crowd-sourcing platforms. *Explanation type* represents the methodology adopted to construct the explanations. *Generated* means that the sentences in the explanations are entirely created by human annotators. On the other hand, *composed* means that the sentences are retrieved from an external knowledge resource. Finally, the last row reports the *average number of sentences* composing the explanations.

5.2 Annotation Task

The bottleneck of the annotation framework lies in the formalisation phase, which is generally time consuming and requires trained experts in the field. In order to make the application of EEV efficient in practice, we extract a sub-set of $n = 100$ examples from each XGS (Worldtree, QASC, and e-SNLI). To maximise the representativeness of the explanations in the subset, given a fixed size n, we combine a set of sampling methodologies with effect size analysis. The details of the sampling methodology are described in section 5.3 while the results are presented in section 5.4. Code and data adopted for the experiments are available online [1].

The extracted examples are randomly assigned to 2 annotators with an overlap of 20 instances to compute the inter-annotator agreement. All the annotators are active researchers in the field of Natural Language Processing and Computational Semantics. Table 2 reports the inter-annotator agreement achieved on each dataset separately. Overall, we observe an average of 72% accuracy in the multi-label classification task, computed considering the percentage of overlaps between the final entailment verification classes chosen by the annotators.

5.3 Sampling Methodology

To maximise the representativeness of the explanations for the subsequent annotation task, while analysing a fixed number n of examples for each dataset, we combine a set of sampling methodologies with effect size analysis. In this section, we describe the sampling techniques adopted for each dataset.

A stratified sampling methodology has been adopted for the Worldtree corpus (Xie et al., 2020; Jansen et al., 2018). The stratified sampling con-

[1] https://github.com/ai-systems/explanation-entailment-verification/

sists in partitioning the dataset using a set of classes and performing random sampling from each class independently. This strategy guarantees that the same amount of examples is extracted from each class. The stratified technique requires the classes to be collectively exhaustive and mutually exclusive – i.e, each example has to belong to one and only one class. To apply stratified sampling on Worldtree, we consider the high-level topics introduced in (Xu et al., 2020), which are used to classify each question in the dataset according to one of the following categories: Life, Earth, Forces, Materials, Energy, Scientific Inference, Celestial Objects, Safety, Other. The same technique cannot be applied to e-SNLI (Camburu et al., 2018) and QASC (Khot et al., 2020) since the examples in these datasets are not partitioned using any abstract set of classes. In this case, therefore, we use random sampling on the whole dataset to extract a fixed number n of examples.

Once a fixed number of examples n is extracted from each dataset, we consider the annotated explanation sentences of each example to verify whether the extracted set of explanations is representative of the whole dataset. To perform this analysis, we assume the predicates in the explanation sentences to be the expression of the type of knowledge of the whole explanation. Therefore, we consider the extracted sample of explanations representative if the distribution of predicates in the sample is correlated with the same distribution in the whole dataset. To this end, we compute the frequencies of the verbs appearing in the explanation sentences from the extracted sub-set and original dataset separately. Subsequently, we compare the frequencies in the sub-sample with the frequencies in the whole dataset computing a Pearson correlation coefficient. In this case, a coefficient greater than .7 indicates a strong correlation between the types of explanations in the sample and the types of explanations in the original dataset. After running the sampling for t times independently, we select the subset of explanations for each dataset with the highest Pearson correlation coefficient. Table 3 reports the Pearson correlation for the subsets adopted in our analysis with fixed sample size $n = 100$.

5.4 Results

The quantitative analysis presented in this section aims to empirically assess the hypothesis that human-annotated explanations in XGSs constitute

Dataset	Agreement Accuracy
Worldtree	.70
QASC	.70
e-SNLI	.75

Table 2: Inter-annotator agreement computed in terms of accuracy in the multi-label classification task considering the first annotator as a gold standard.

Dataset	Correlation Coefficient
Worldtree	.964
QASC	.958
e-SNLI	.987

Table 3: Effect size analysis of the samples extracted from each XGS for the downstream EEV annotation.

valid and non-redundant logical arguments for the expected answers. We report the quantitative results of the explanation entailment verification in Table 4. Specifically, the table reports the percentage of the frequency of each verification class in the analysed samples. The column *AVG* reports the average for each class.

Overall, we observe that the results of the annotation task tend to reject our research hypothesis, with an average of only 20.42% of analysed explanations being classified as *valid and non redundant* arguments. When considering also *valid, but redundant* explanations (21.91%), the average percentage of valid arguments reaches a total of 42.33%. Therefore, we can conclude that the majority of the explanations represent invalid arguments (57.66%).

We observed that the majority of invalid arguments are classified as *missing plausible premise*. This finding implies that a significant percentage of annotated explanations are incomplete arguments (26.00%), that can be made valid on addition of a reasonable premise. We attribute this result to the tendency of human explainers to take for granted part of the world knowledge required in the explanation (Walton, 2004).

A lower but significant percentage of explanations contain identifiable logical errors (11.19%), which result from confusing the set of quantifiers and logical operators, or from illicitly changing the direction of an implication. Similarly, 20.47% of the explanations where labeled as *no discernible arguments*, where no obvious premise can be added to make the argument valid and no simple logical error can be detected. This result can be attributed partly to natural errors occurring in a gold standard

Entailment Verification Class	Worldtree	QASC	e-SNLI	AVG
Valid and non-redundant	12.24	17.65	**31.37**	20.42
Valid, but redundant premises	26.53	7.84	**31.37**	21.91
Missing plausible premise	**38.78**	21.57	17.65	**26.00**
Logical error	6.12	17.65	9.80	11.19
No discernible argument	16.33	**35.29**	9.80	20.47
Valid argument	38.77	25.49	**62.74**	42.33
Invalid argument	61.23	**74.51**	37.25	**57.66**

Table 4: Results of the application of EEV for each entailment verification category.

creation process, partly to the effort required for human-annotators to identify logical fallacies in their explanations. In the remaining of this section, we analyse the results obtained on each XGS.

Worldtree. The analysed sample contains the highest percentage of incomplete arguments, with a total of 38.78% explanations classified as *missing plausible premise*. This result can be explained by the fact that the questions in Worldtree require complex forms of reasoning, facilitating the construction of arguments containing implicit world knowledge and missing premises. At the same time, the dataset contains the smallest percentage of logical errors (6.12%). We attribute this outcome to the fact that Worldtree is not crowd-sourced, implying that the quality of the annotated explanations is more easily controllable using internal verification methods.

QASC. This XGS contains the highest rate of invalid arguments (62.74%), with 35.29% of the explanations classified as *no discernible argument*. One of the factors contributing to these results might be related to the length of the constructed explanations, which is limited to 2 facts extracted from a predefined corpus of sentences. The high rate of no discernible arguments and missing premises (35.29% and 21.57% respectively) suggests that the majority of the questions require additional world knowledge and more detailed explanations. This conclusion is also supported by the percentage of *valid, but redundant* arguments, which is the lowest among the analysed samples (7.84%). Finally, the highest rate of logical errors (17.65%) might be due to a combination of factors, including the complexity of the question answering task and the adopted crowd-sourcing mechanism, which prevent a thorough quality assessment.

e-SNLI. The sample includes the highest percentage of valid arguments with a total of 31.37%. However, we noticed that the complexity of the reasoning involved in e-SNLI is generally lower than Worldtree and QASC, with most of the textual entailment problems being an example of *monotonicity reasoning*. This observation is supported by the highest percentage of *valid, but redundant* cases (31.37%), where the explanation simply repeats the content of the conclusion. This occurrs quite often for examples of lexical entailment, where the words in the conclusion are a subset of the words in the premise. The lexical entailment instances, in fact, do not require any additional world knowledge, making any attempt of constructing an explanation redundant. Despite these characteristics, our evaluation suggests that a significant percentage of arguments are invalid (37.25%). Again, this percentage might be the results of different factors, including the errors produced by the crowd-sourcing process.

Table 5 reports a set of representative cases extracted from the evaluated samples. For each entailment verification class, we report an example extracted from the XGS with the highest percentage of instances in that class.

5.5 Contrastive Explanations

Previous studies highlight the fact that explanations are *contrastive* in nature, that is, they describe why an event *P* happened instead of some counterfactual event *Q* (Miller, 2019; Lipton, 1990). Following this definition, we perform an additional analysis to verify whether the explanations contained in MCQA datasets are *contrastive* with respect to the wrong candidate answers – i.e., the explanation supports the validity of the correct answer while excluding the set of alternative choices. In order to quantify this aspect, we asked the annotators to label the questions with more than one plausible answer, whose explanations do not mention any discriminative commonsense or world knowledge that explains why the gold answer is correct instead of the alternative choices.

Problem Formulation	Explanation	XGS
Valid and non-redundant (20.42%)		
Premise: A smiling woman is playing the violin in front of a turquoise background. **Hypothesis:** A woman is playing an instrument.	A violin is an instrument.	e-SNLI
Valid, but redundant premises (21.91%)		
Premise: Four people are bandaging a head wound. **Hypothesis:** People are bandaging an injured head.	People are bandaging an injured head wound.	e-SNLI
Missing plausible premise (26.00%)		
Question: A group of students are studying bean plants. All of the following traits are affected by changes in the environment except [A] Leaf color [*B] Seed type [C] Bean production [D] Plant height	The type of seed of a plant is an inherited characteristic. Inherited characteristics are the opposite of learned characteristics; acquired characteristics. An organism's environment affects that organism's acquired characteristics. A plant is a kind of organism. Trait is synonymous with characteristic.	Worldtree
Logical error (11.19%)		
Question: What group of animals do chordates belong to? [A] graptolites [B] more abundant [C] warm-blooded [D] four limbs [E] epidermal [*F] Vertebrates [G] animals [H] insects	Chordates have a complete digestive system and a closed circulatory system. Vertebrates have a closed circulatory system.	QASC
No discernible argument (20.47%)		
Question: What do plants require for reproduction? [A] energy [B] nutrients [C] bloom time [*D] animals [E] sunlight [F] Energy. [G] food [H] hormones	Plants require seed dispersal for reproduction. Seeds are probably dispersed by animals.	QASC

Table 5: Examples of explanations classified with different entailment verification categories.

Dataset	Non contrastive explanations
Worldtree	26.53
QASC	**49.02**

Table 6: Percentage of explanations in the MCQA sample labeled as non contrastive.

The results of this experiment are reported in Table 6. Overall, we found that a significant percentage of explanations are labeled as non contrastive. This outcome is particularly evident for QASC. We attribute these results to the presence of multi-adversary answer choices in QASC, which are generated automatically to make the dataset more challenging for language models. However, we found that this mechanism can produce questions with more than one plausible correct answer, which can cause the explanation to loose its contrastive function (see QASC examples in Table 5).

6 Conclusion and Future Work

This paper proposed a systematic annotation methodology to quantify the logical validity of human-annotated explanations in Explanation Gold Standards (XGSs). The application of the framework on three mainstream datasets led us to the conclusion that a majority of the explanations represent logically invalid arguments, ranging from being incomplete to containing clearly identifiable logical errors.

The main limitation of the framework lies in the scalability of its current implementation, which is generally time consuming and requires trained semanticists. One way to improve its efficiency is to explore the adoption of supporting tools for the formalisation, such as semantic parsers and/or automatic theorem provers.

Despite the current limitations, this study offers some important pointers for future work. On the one hand, the results suggest that logical errors can be reduced by a careful design of the gold standard, such as authoring explanations with internal verification strategies or reducing the complexity of the reasoning task. On the other hand, the finding that a large percentage of curated explanations still represent incomplete arguments has a deeper implication on the nature of explanations and on what annotators perceive as a valid and complete logical argument. Therefore, we argue that future progress on the design of XGSs will depend, among other things, on a better formalisation and understanding of the inferential properties of explanations.

References

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *International Conference on Learning Representations*.

Samuel R. Bowman, 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. Association for Computational Linguistics.

Samuel R. Bowman and George E. Dahl. 2021. What will it take to fix benchmarking in natural language understanding?

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 9539–9549. Curran Associates, Inc.

Oana-Maria Camburu, Brendan Shillingford, Pasquale Minervini, Thomas Lukasiewicz, and Phil Blunsom. 2020. Make up your mind! adversarial generation of inconsistent natural language explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4157–4165, Online. Association for Computational Linguistics.

Ruben Cartuyvels, Graham Spinks, and Marie-Francine Moens. 2020. Autoregressive reasoning over chains of facts with transformers. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6916–6930, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Deborah Ferreira and André Freitas. 2020a. Natural language premise selection: Finding supporting statements for mathematical text. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2175–2182, Marseille, France. European Language Resources Association.

Deborah Ferreira and André Freitas. 2020b. Premise selection in natural language mathematical texts. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7365–7374, Online. Association for Computational Linguistics.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Peter Jansen, Elizabeth Wainwright, Steven Marmorstein, and Clayton Morrison. 2018. WorldTree: A corpus of explanation graphs for elementary science questions supporting multi-hop inference. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Harsh Jhamtani and Peter Clark. 2020. Learning to explain: Datasets and models for identifying valid reasoning chains in multihop question-answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 137–150, Online. Association for Computational Linguistics.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. Qasc: A dataset for question answering via sentence composition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8082–8090.

Sawan Kumar and Partha Talukdar. 2020. NILE : Natural language inference with faithful natural language explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8730–8742, Online. Association for Computational Linguistics.

Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. 2021. Question and answer test-train overlap in open-domain question answering datasets. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1000–1008, Online. Association for Computational Linguistics.

Peter Lipton. 1990. Contrastive explanation. *Royal Institute of Philosophy Supplement*, 27:247–266.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

Sewon Min, Eric Wallace, Sameer Singh, Matt Gardner, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2019. Compositional questions do not necessitate multi-hop reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4249–4257, Florence, Italy. Association for Computational Linguistics.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent Disagreements in Human Textual Inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself!

leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Viktor Schlegel, Marco Valentino, Andre Freitas, Goran Nenadic, and Riza Batista-Navarro. 2020. A framework for evaluation of machine reading comprehension gold standards. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5359–5369, Marseille, France. European Language Resources Association.

Sanjay Subramanian, Ben Bogin, Nitish Gupta, Tomer Wolfson, Sameer Singh, Jonathan Berant, and Matt Gardner. 2020. Obtaining faithful interpretations from compositional neural networks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5594–5608, Online. Association for Computational Linguistics.

Mokanarangan Thayaparan, Marco Valentino, and André Freitas. 2020a. Explanationlp: Abductive reasoning for explainable science question answering.

Mokanarangan Thayaparan, Marco Valentino, and André Freitas. 2020b. A survey on explainability in machine reading comprehension.

Marco Valentino, Mokanarangan Thayaparan, and André Freitas. 2021. Unification-based reconstruction of multi-hop explanations for science questions. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 200–211, Online. Association for Computational Linguistics.

Marco Valentino, Mokanarangan Thayaparan, and André Freitas. 2020. Explainable natural language reasoning via conceptual unification.

Douglas Walton. 2004. A new dialectical theory of explanation. *Philosophical Explorations*, 7(1):71–89.

Cunxiang Wang, Shuailong Liang, Yili Jin, Yilong Wang, Xiaodan Zhu, and Yue Zhang. 2020. SemEval-2020 task 4: Commonsense validation and explanation. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 307–321, Barcelona (online). International Committee for Computational Linguistics.

Sarah Wiegreffe and Ana Marasović. 2021. Teach me to explain: A review of datasets for explainable nlp.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

Zhengnan Xie, Sebastian Thiem, Jaycie Martin, Elizabeth Wainwright, Steven Marmorstein, and Peter Jansen. 2020. WorldTree v2: A corpus of science-domain structured explanations and inference patterns supporting multi-hop inference. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5456–5473, Marseille, France. European Language Resources Association.

Dongfang Xu, Peter Jansen, Jaycie Martin, Zhengnan Xie, Vikas Yadav, Harish Tayyar Madabushi, Oyvind Tafjord, and Peter Clark. 2020. Multi-class hierarchical question classification for multiple choice science exams. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5370–5382, Marseille, France. European Language Resources Association.