Does External Knowledge Help Explainable Natural Language Inference? Automatic Evaluation vs. Human Ratings

Hendrik Schuff[1,2] and **Hsiu-Yu Yang**[2] and **Heike Adel**[1] and **Ngoc Thang Vu**[2]
[1]Bosch Center for Artificial Intelligence, Renningen, Germany
[2]Institut für Maschinelle Sprachverarbeitung, University of Stuttgart
{Hendrik.Schuff,Heike.Adel}@de.bosch.com
{Hsiu-Yu.Yang,Thang.Vu}@ims.uni-stuttgart.de

Abstract

Natural language inference (NLI) requires models to learn and apply commonsense knowledge. These reasoning abilities are particularly important for explainable NLI systems that generate a natural language explanation in addition to their label prediction. The integration of external knowledge has been shown to improve NLI systems, here we investigate whether it can also improve their explanation capabilities. For this, we investigate different sources of external knowledge and evaluate the performance of our models on in-domain data as well as on special transfer datasets that are designed to assess fine-grained reasoning capabilities. We find that different sources of knowledge have a different effect on reasoning abilities, for example, implicit knowledge stored in language models can hinder reasoning on numbers and negations. Finally, we conduct the largest and most fine-grained explainable NLI crowdsourcing study to date. It reveals that even large differences in automatic performance scores do neither reflect in human ratings of label, explanation, commonsense nor grammar correctness.

1 Introduction

Natural language inference (NLI) is closely related to real-world applications, such as fact checking. Given two sentences (premise and hypothesis), the task is to decide whether (a) the first sentence entails the second sentence, (b) the two sentences contradict each other or (c) they have a neutral relation. Figure 1 shows an example for an entailment relation. Solving the task requires models to not only reason over the provided information but also to link it with commonsense knowledge.

As for other natural language tasks, state-of-the-art NLI systems rely on deep neural architectures which do not easily expose their inner workings. However, following a model's reasoning process is valuable to machine learning engineers as well as end-users. The former can use the insights to improve models and the latter can base their decision on them whether to trust the system or not. One approach to gain insight into a system is to train it to generate explanations as an additional output (Camburu et al., 2018; Atanasova et al., 2020). Such self-explaining models are particularly interesting for NLI because the explanation can indicate the commonsense knowledge which was utilized during prediction. The integration of external knowledge was shown to improve NLI systems (Jijkoun and de Rijke, 2005; Chen et al., 2018; Li et al., 2019; Faldu et al., 2021). However, the following question remains: **Does the positive effect of external knowledge on the inference ability transfer to the generation of explanations? (R1)** Figure 1 shows an NLI example for which external knowledge potentially helps to infer the correct label and explanation. In the example, the system needs to link *dog* to *animal*, *jumping for a Frisbee* to *playing*, *Frisbee* to *plastic toy* and *snow* to *outside* as well as to *cold weather*. The predicted explanation needs to explicitly state this reasoning chain and thus would be expected to benefit from external knowledge.

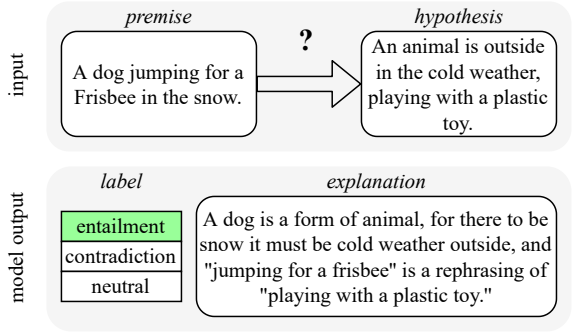


Figure 1: NLI example instance from e-SNLI (Camburu et al., 2018). The system needs to include commonsense knowledge, such as "snow → cold weather".

Especially recently, pre-trained language models, such as BERT (Devlin et al., 2019) or GPT (Rad-

Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, pages 26–41

ford et al., 2019), became popular. It was shown that they are able to learn and store commonsense knowledge implicitly (Petroni et al., 2019). As a result, an open question is: **How effective is the implicit commonsense knowledge of language models compared to symbolic sources of knowledge, such as knowledge base triplets? (RQ2)**

To evaluate NLI models, mainly automatic measures, such as accuracy, are used. However, model weaknesses can stay unnoticed using automatic scores alone. Moreover, Schuff et al. (2020) showed that automatic scores are not necessarily correlated to human-perceived model quality. Thus, human evaluation is a crucial step in the development of user-centered explainable AI systems. Therefore, we ask the question: **How do humans perceive explanation quality of state-of-the-art natural language inference models? (R3)**

In this paper, we investigate the three previously mentioned research questions R1–R3. To answer them, we analyze the impact of external knowledge from multiple sources, such as knowledge graphs, embeddings and language models and propose novel architectures to include and combine them into explainable NLI systems. Further, we conduct an extensive automatic analysis as well as a user study. To the best of our knowledge, our study exceeds previous human evaluations of explainable NLI models regarding the number of participants as well as the variety of rated explanation criteria.

For R1, we find that the positive effect of external knowledge on label accuracy in the standard NLI setting can also be observed in the explainable NLI setting and external knowledge can improve the BLEU scores of the generated explanations. In regard of R2, we observe that pre-trained language models are the most promising source of commonsense knowledge but at the same time identify weaknesses with respect to negations and numerical reasoning abilities which, however, can be mitigated through combination with additional knowledge sources. Despite the improvements in accuracy, BLEU or BLEURT scores, our user study shows for R3 that these do not reflect in human ratings of explanation correctness, commonsense inclusion or grammar and label correctness.

Our results urge caution to solely rely on automatic scores for explainability. Therefore, we expect our paper to motivate the development of dedicated evaluation tasks and scores and further emphasize the importance of the user within explainable AI. To facilitate future work, we release our model's predictions as well as the crowdsourced human ratings.[1]

2 Related Work

2.1 External Knowledge for NLI

External knowledge was shown to help across a wide variety of NLP tasks (Shi et al., 2016; Seyler et al., 2018; Pan et al., 2019; Lin et al., 2019). While early sources for external knowledge are WordNet and NomBank (Jijkoun and de Rijke, 2005; MacCartney et al., 2008), today a large variety of possibilities exist: From COMET (Bosselut et al., 2019) over ConceptNet (Speer et al., 2017) to language models. Chen et al. (2018) show that enriching an NLI system with external lexical-level semantic knowledge increases accuracy scores on SNLI and enhances transfer to MultiNLI. Wang et al. (2019) show the potential of knowledge from ConceptNet for NLI systems. Li et al. (2019) find that external knowledge from pretraining helps NLI and suggest to combine it with external knowledge from human-curated resources. Li and Sethy (2019) propose knowledge-enhanced attention modifications for Transformers and decomposable methods and show that their methods improve model robustness. Faldu et al. (2021) extend BERT by extracting entities from the input text and adding their projected KG embeddings derived from ConceptNet and WordNet as sequential input to a modified BERT layer. Bauer et al. (2021) present ERNIE-NLI, a modified ERNIE (Zhang et al., 2019) model using NLI-specific knowledge embeddings and find that it improves performance over a non-adapted ERNIE model using general-domain TransE embeddings. To compare different possibilities of integrating external knowledge, we propose various models in this paper. Further, we address the question whether external knowledge also improves explanation generation.

2.2 Explainable NLI

The task of explainable NLI consists of (i) predicting the correct entailment label and (ii) providing an explanation that allows the user to assess the model's reasoning. In general, such explanation can take various forms, such as weights and gradients over the input (Simonyan et al., 2014; Ribeiro et al., 2016; Lundberg and Lee, 2017) and

[1] https://github.com/boschresearch/external-knowledge-explainable-nli

text spans or snippets from the input or external text (Zaidan and Eisner, 2008; Lei et al., 2016; Yang et al., 2018). Beyond that, there exists various resources and approaches designed to generate textual explanations. Rajani et al. (2019) present a dataset that contains free-text explanations for multiple-choice commonsense reasoning and Bhagavatula et al. (2020) provide a dataset for abductive multiple choice answering as well as abductive NLG. Camburu et al. (2018) provide the e-SNLI dataset, which adds free-text explanations as an additional layer on the SNLI dataset (Bowman et al., 2015). As numerous models with and without external knowledge have been developed on the SNLI dataset, we use its explainable extension e-SNLI to conduct our analysis and train our models on. Various models have been proposed on e-SNLI including systems based on alignment (Swanson et al., 2020), label-specific explanation generators (Kumar and Talukdar, 2020) and fine-tuned text-to-text models (Narang et al., 2020). In contrast to those, our focus is not on proposing a new architecture or paradigm to develop a high-scoring system. Much more, we seek to conduct a broad comparison across knowledge sources and isolate their effect on automatic scores as well as human perception.

2.3 Evaluation and Human Ratings

Explainable NLI system performance is typically scored using (i) accuracy with respect to annotated gold labels on a reference dataset and (ii) BLEU scores (Papineni et al., 2002) between the generated explanations and the ground truth explanations (Camburu et al., 2018; Kumar and Talukdar, 2020; Narang et al., 2020). BLEU scores can only quantify explanation quality loosely (Narang et al., 2020). Therefore, previous work evaluates explanation quality either by manual annotation (Camburu et al., 2018; Kumar and Talukdar, 2020) or crowdsourcing (Narang et al., 2020). However, previous human evaluations regarding explainable NLI are limited to assess label and/or explanation correctness. In contrast, we additionally evaluate commonsense inclusion as well as grammatical correctness of explanations. As Clinciu et al. (2021) find automatic BLEURT scores to have distinctly stronger correlations to human ratings of generated textual explanations than BLEU, we investigate whether BLEURT is a viable replacement for a user study.

3 Method

In the following, we describe our base model and then present the models we analyze.

3.1 Base Model

We combine a state-of-the-art attention-based inference model with an explainable NLI model that predicts entailment labels and generates explanations. In particular, we use the encoder part of the enhanced sequential inference model (ESIM), which has a cross-attention layer to capture relevant semantics between premise and hypothesis (Chen et al., 2017) and the prediction part of the PRED-EXPL model of Camburu et al. (2018). We represent the input sentences with BERT embeddings (Devlin et al., 2019) which we fine-tune on the SNLI dataset.[2] Throughout the paper, we refer to this model as VANILLA.

3.2 Integration of Knowledge Sources

External knowledge can be found in various formats. We aim at covering a possibly broad variety and focus on state-of-the-art sources and methods. We include the natural language knowledge base COMET (Bosselut et al., 2019), the ConceptNet Numberbatch embeddings (Speer et al., 2017) and the GPT-2 language model (Radford et al., 2019).

3.2.1 Background Knowledge from COMET

As our example in Figure 1 shows, resolving natural language entailment can require reasoning over multiple concepts and relations, such as inferring *cold weather* and *outside* from *snow*. We seek to facilitate this resolvement by providing the model with related words (and phrases) that can be seen as a natural language extension of premise and hypothesis. We use the COMmonsEnse Transformers (COMET) (Bosselut et al., 2019) as a natural language knowledge base to query background knowledge for premise and hypothesis. COMET is based on a transformer language model that is fine-tuned on a knowledge base completion task on ConceptNet. Given an input sentence and a ConceptNet relation, it generates a phrase to complete the object in a knowledge statement expressed in the (subject, relation, object) format. Instead of feeding in the whole premise and hypothesis, we

[2] We pass inputs of the form *[CLS] premise [SEP] hypothesis* to BERT and use a softmax layer on top of the CLS token's embedding to predict the entailment label and fine-tune the model for up to 2 epochs.

find that chunking them into noun and verb sub-phrases based on POS tags patterns yields better object phrase generations.[3] Thus, for each sentence (premise/hypothesis) we generate #chunks × #relations object phrases.[4]

Afterwards, we embed each object phrase (with the respective relation string prependend) with Sentence-BERT (Reimers and Gurevych, 2019) and quantify its similarity to the embedding of the source sentence using cosine similarity. For each relation, we keep the object phrase with the highest similarity score.

Given the relation *HasA* and the chunked sentence *The dog | is walking in the snow*, for example, COMET will generate *bone* and *effect of freeze* for the two sub-phrases, respectively. We only preserve the object phrase *effect of freeze* as it has a higher similarity to the source sentence.

To condense the object phrases into a fixed-length vector representation, we average the respective Sentence-BERT embeddings. This procedure yields one vector representing the background knowledge regarding the premise and one regarding the hypothesis. We combine them with the local inference vector representation (Chen et al., 2017). Following Camburu et al. (2018), this vector is passed to the label prediction module as well as the explanation decoder. We refer to this model as COMET.

3.2.2 Modified Attention with ConceptNet

Following Li and Srikumar (2019), we use knowledge-driven rules to modify the attention weights within the cross-attention layer between premise and hypothesis in the encoder. This enforces the attention mechanism to align word pairs p_i and h_j from premise and hypothesis based on world knowledge. The rules proposed by Li and Srikumar (2019) are shown in Equation 1 and 2. In R_1, the antecedent K_{p_i,h_j} indicates that a word pair p_i and h_j is of a certain relation within ConceptNet. If the condition of the antecedent is true, the consequent A'_{p_i,h_j} that aligns the word pair follows. R_2 is a relatively conservative rule that additionally takes the model's own decision into account. The antecedent $K_{p_i,h_j} \wedge A_{p_i,h_j}$ in R_2 is a conjunctive condition that becomes true if a word pair is both in a relation and aligned by a model's original attention. If such a conjunctive condition is true, the word pair must be aligned which results in a new alignment as the consequent A'_{p_i,h_j} indicates.

$$R_1 : K_{p_i,h_j} \rightarrow A'_{p_i,h_j} \quad (1)$$

$$R_2 : K_{p_i,h_j} \wedge A_{p_i,h_j} \rightarrow A'_{p_i,h_j} \quad (2)$$

Different from the approach of Li and Srikumar (2019) that checks a word pair's relation in a binary fashion, we hypothesize that knowledge-aware embeddings might capture more fine-grained word relationship that exists in multi-hop relational edges. Considering *playground* and *playroom*, for example, the former is usually located outdoors whereas the latter is located indoors. We generalize the binary relational inclusion from Li and Srikumar (2019) to continuous relation scores. For this, we replace the binary rule antecedent with the absolute cosine similarity between the ConceptNet Numberbatch (Speer et al., 2017) vector representations of p_i and h_j. We empirically confirm that our continuous formulation outperforms the binary version regarding label accuracy as well as explanation correctness. In the following, we refer to these modified rules as continuous constraints and use CONT to refer to the respective model.

3.2.3 All-Text Prediction with GPT-2

Similar to Kumar and Talukdar (2020), we fine-tune a pre-trained GPT-2 language model on the e-SNLI dataset. In contrast to Kumar and Talukdar (2020), we use a single GPT-2 model to generate explanations for all three entailment labels instead of training a separate model for each of them. This allows us to directly integrate the label prediction into the language model instead of training an additional model which predicts the label on top of the three explanations. Therefore, we propose two models, which both are GPT-2-large models, but differ regarding their training setting. In the label-first setting (GPT-LF), the model is trained on text following the structure *Premise: <premise> Hypothesis: <hypothesis> [LAB] [label] [EXP] <explanation> EOS*. In the explanation-first setting (GPT-EF) it is trained on text following the structure *Premise: <premise> Hypothesis: <hypothesis> [EXP] <explanation> [LAB] <label> EOS*.

[3] We manually find that feeding in the whole sentence predominantly relates the output to the last tokens of the sentence and fails to include information from tokens earlier in the sentence.

[4] We consider the relations AtLocation, CapableOf, DefinedAs, HasA, HasProperty, HasSubevent, InheritsFrom, InstanceOf, IsA, LocatedNear, MadeOf, PartOf, SymbolOf, UsedFor and LocationOfAction.

3.3 Combined Models

COMET and ConceptNet. We combine COMET with CONT to benefit from both integrated background information from COMET and a knowledge-enhanced attention mechanism based on ConceptNet Numberbatch. We expect this to help the model focus on important relations between premise and hypothesis.

Knowledge-Enhanced Ensembles. We combine the world knowledge of BERT (VANILLA), ConceptNet Numberbatch (CONT), COMET (COMET) and the combined model COMET+CONT with the language model abilities of GPT-2 (GPT-LF and GPT-EF). For this, we propose an ensemble that not merely aggregates label votes, but combines the models with respect to their different strengths.

The label predictions of VANILLA, CONT, COMET, COMET+CONT as well as GPT-LF are passed to a majority voting. In the *basic ensemble*, the GPT-LF model is then conditioned on the voted label and generates the final explanation. We refer to this model as ENSEMBLE.

In the *filtered ensemble*, the majority voting only allows models to vote if their generated explanation lets the GPT-EF model predict the same label prediction as the original model. In other words, we fix the input as well as the generated explanation and only let the GPT-EF model predict the label. This step can be interpreted as a consistency filter which prevents models from voting if their label prediction does not match their explanation prediction. In the following, we refer to this model as FILTERED-ENS. We include a depiction of the model architecture in the appendix.

4 Automatic Evaluation

First, we evaluate the discussed knowledge-enhanced models with respect to commonly used scores on e-SNLI and a stress test evaluation. In addition to our constructed models, we also include PRED-EXPL (Camburu et al., 2018), which is basically our VANILLA baseline without cross-attention but with GloVe embeddings instead of fine-tuned BERT embeddings. Further, we include two recent models proposed for e-SNLI: NILE:post-hoc, which is the highest performing model from Kumar and Talukdar (2020), and WT5-11B from Narang et al. (2020), which holds the current state-of-the-art performance. While NILE:post-hoc is based on GPT-2 as well, WT5-11B is a fine-tuned version of the T5 language model (Raffel et al., 2020). We train all non-LM models with five random seeds and report scores of the median model based on label accuracy. Table 2 shows predicted explanations for the subset of models that we investigate within the human evaluation in Section 5. Further examples are provided in the appendix.

Type	Model	Label Acc.	BLEU	BLEURT
non-LM	PRED-EXPL	84.21	19.77	-0.871
	VANILLA	89.20	19.71	-0.820
	COMET	88.97	18.84	-0.822
	CONT	89.02	20.1	-0.799
	COMET+CONT	89.07	19.66	-0.809
LM-based	GPT-EF	87.89	21.70	-0.624
	GPT-LF	89.70	26.90	-0.577
	ENSEMBLE	90.24	27.10	-0.576
	FILTERED ENS	90.24	27.09	-0.577
	NILE:POST-HOC	91.49	26.26	-0.577
	WT5-11B	**92.3**	**29.01**	**-0.511**

Table 1: Automatic evaluation metrics on the e-SNLI test set. Label accuracy quantifies NLI performance. BLEU and BLEURT score the similarity between predicted and ground truth explanation texts. Higher values are better.

4.1 Performance on e-SNLI

Following prior work on e-SNLI, we report label accuracy as well as BLEU scores (Papineni et al., 2002) for explanations. We additionally evaluate BLEURT scores (Sellam et al., 2020), which is a reference-based learned evaluation metric to model human judgements of text generation. BLEURT is of particular interest for explanation evaluation as Clinciu et al. (2021) compare how various automatic scores such as BLEU, ROUGE and METEOR correlate to human ratings of generated explanations and find that embedding-based methods and particularly BLEURT scores show distinctly higher correlations than, e.g., BLEU.

Table 1 shows the respective scores for all considered models.[5] The upper block lists models that share or extend the PRED-EXPL architecture. Compared to PRED-EXPL, the VANILLA model achieves a notable increase in label accuracy as well as BLEURT scores. Surprisingly, COMET reduces all scores and even decreases the BLEU score be-

[5]For NILE:post-hoc (Kumar and Talukdar, 2020) and WT5-11B (Narang et al., 2020) we report the label accuracy from their paper and calculate BLEU/BLEURT scores based on the explanation predictions provided by the authors. Narang et al. (2020) calculate BLEU scores using SacreBLEU v1.3. (Post, 2018) leading to a higher reported score of 33.7.

Model	Predicted Explanation
GROUND-TRUTH	a man is either playing the accordion or performs a mime act while happy people pass by or angry people glare at him.
VANILLA	a man can not be playing and a mime at the same time
COMET	the man is either playing the accordion or a mime
CONT	people can not be playing and angry at the same time
COMET+ CONT	the man can not be playing the accordion and the mime at the same time
GPT-LF	Happy people are not angry people.
WT5-11B	The man cannot be playing the accordion and performing a mime act at the same time.

Table 2: Explanation predictions of the models used within the human evaluation for the premise "*A man on a sidewalk is playing the accordion while happy people pass by*" and the hypothesis "*A man on the sidewalk performs a mime act while angry people glare at him*". All models correctly predict the class *contradiction* but generate different explanations. The predicted explanation of the FILTERED-ENS model is identical to the explanation of the GPT-LF model as GPT-LF is used to predict the ensemble's explanation.

low the PRED-EXPL score. In contrast, knowledge-enhanced cross attention (CONT) improves BLEU and BLEURT scores and reaches a label accuracy close to VANILLA. Combining CONT with COMET retains the CONT label accuracy but again slightly decreases BLEU and BLEURT scores. The lower block contains models that are or include language models. All language model-based models increase BLEU and BLEURT scores. All except GPT-EF outperform all non-language model models.

To analyze whether the performance differences of models can be really attributed to a better reasoning and commonsense knowledge ability instead of merely different model capacity, we next evaluate our models on the NLI stress test evaluation.

4.2 Stress Test Evaluation

Table 3 shows the results of our models on the NLI stress test evaluation proposed by Naik et al. (2018). The dataset contains multiple subsets of which each subset is used to evaluate the robustness of the system against a specific type of perturbation, e.g., spelling errors, negations, numerical reasoning and more. On average, all models distinctly improve performance compared to the PRED-EXPL baseline. With respect to VANILLA, all models except GPT-EF improve average performance. Further, both COMET and CONT improve average label accuracy, while their combination decreases performance. Surprisingly, GPT-LF outperforms the ensemble methods on average. While COMET+CONT reaches the best performance in terms of e-SNLI label accuracies, it performs worst on the stress tests. The same effect can be observed for the FILTERED-ENS. While it reaches top performance for the spelling error test, its performance drops for numerical reasoning, where it performs worse than any other model. These results show that combining different knowledge sources does not result in a consistent combination of their weaknesses and strengths. Instead, the sources of external knowledge have to be carefully adjusted to the target domain and our results paint a rather pessimistic picture regarding a cure-all solution. Further, a model's reasoning capabilities have to be assessed in detail as evaluation across different reasoning types easily masks model weaknesses.

Finally, we assess whether language models reach their higher performance due to better reasoning: For most of the assessed reasoning types — with exception of numerical reasoning and negation — the best non-ensemble model in fact is GPT-LF. Also, GPT-LF reaches the highest accuracy on average. Therefore one could generally recommend to include external knowledge in form of a pre-trained language model as the foremost option. However, our results also show that language models are not necessarily the best choice for all reasoning needs and can, e.g., severely decrease performance for numerical reasoning and negations, where models based on language models perform worse than all other models.

5 Human Evaluation

While automatic scores, such as BLEU, provide a valuable starting point for evaluating explanations, they fall short in capturing the model's real explanation capabilities. We, therefore, conduct a large-scale crowdsourcing study to complement our automatic evaluations on e-SNLI and the stress tests. Following related work (Narang et al., 2020), we assess explanation quality based on ratings from crowdworkers on Mechanical Turk. While previ-

			Competence Test		Distraction Test			Noise Test
Type	Model	Total	Antonymy	Numerical	Word Overlap	Length Mismatch	Negation	Spelling
non-LM	PRED-EXPL	48.69	36.36	36.55	47.17	53.44	45.31	52.42
	VANILLA	56.94	37.94	32.24	55.46	65.21	52.03	62.90
	COMET	57.05	34.54	35.48	57.31	64.15	52.85	62.33
	CONT	57.09	32.50	**40.28**	52.10	64.35	**53.38**	62.77
	COMET+CONT	56.26	44.43	34.16	51.34	64.39	49.36	63.03
LM-based	GPT-EF	52.74	51.81	31.33	55.91	60.97	38.44	58.20
	GPT-LF	**59.28**	**54.84**	28.80	**64.06**	**68.72**	42.82	67.07
	ENSEMBLE	59.19	37.97	34.03	58.13	67.45	52.51	65.92
	FILTERED-ENS	58.99	52.53	28.54	63.70	68.02	42.18	**67.10**

Table 3: Label accuracies (higher is better) for all categories in the NLI stress test tasks (Naik et al., 2018). The six rightmost columns show (i) the model's reasoning abilities (competence), (ii) how sensitive it is to lexical distractors (distraction) and (iii) how robust it is against noise from different perturbations (noise). Each column corresponds to one dataset. For datasets with matched and mismatched subsets, we report the accuracy over all labels within the group. Similarly, the total accuracy is calculated over all labels.

ous work limited evaluation to rating explanation correctness, we additionally ask participants to provide fine-grained ratings of commonsense inclusion and grammatical correctness. A screenshot of the interface is shown in the appendix. We release the full data of our study.

5.1 Conditions

In order to evaluate effects across the discussed sources of external knowledge, we include seven models in our human evaluation: VANILLA, COMET, CONT, COMET+CONT, GPT-LF, FILTERED-ENS and WT5-11B. Additionally we evaluate the e-SNLI ground truth labels and explanations. Table 2 displays the different explanations the models predict for an exemplary input as well as its ground truth explanation annotation.

5.2 Dependent Variables

We evaluate the models' predicted labels and explanations along four self-reported dimensions.

Label Correctness. Following Kumar and Talukdar (2020) and Narang et al. (2020), we ask participants to rate if the predicted label is correct.

Explanation Correctness. Similar to Camburu et al. (2018), Kumar and Talukdar (2020) and Narang et al. (2020), we collect subjective yes/no explanation correctness ratings.

Grammatical Correctness. We ask participants to rate if the generated explanation is grammatical.

Commonsense Inclusion. We ask participants whether the explanation includes commonsense knowledge that is needed to answer the question. We collect responses on an item with the options *yes*, *no* and *no need*.

5.3 Study Design

In order to evaluate the effect of the level of required external knowledge, we compile, like Kumar and Talukdar (2020) and Narang et al. (2020), a set of 100 premise-hypothesis pairs. In contrast to them, we compose the 100 pairs to contain 50 pairs that require a low level of external knowledge and 50 pairs that require a high level. To gather pairs of both categories, we let two annotators rate 250 premise-hypothesis pairs from the e-SNLI test set. We sample 50 pairs per level of external knowledge from the 179 pairs on which the annotators agree. We provide details on the annotation criteria in the appendix. During the study, we, like Narang et al. (2020), collect 5 crowdsourced ratings for each condition and for each of the 10 input pairs per batch, i.e., 500 ratings per model and a total of 4000 ratings for each variable. We provide ratings of exemplary model predictions in the appendix.

5.4 Analysis

We collect responses from 290 crowdworkers and discard those that were entered in less then 5 minutes (31%) as this might indicate arbitrary answer selection. Note that the repeated measures design of our study possibly introduces inter-dependencies within ratings as, e.g., certain participants can have a tendency to rate explanations as correct more often than others or a certain question might elicit more label correctness ratings. Thus, we use gener-

alized linear mixed models (GLMM) to account for the potentially confounding variables (worker ID, question ID and level of required commonsense knowledge). As our response variables are binary,[6] we use binomial GLMMs. We include fixed effects (model and commonsense level) as well as random intercepts (worker and question IDs). Figure 2 shows effect displays for the collected ratings in relation to the predictor *model type*.

We conduct likelihood ratio tests between the full model and the model without the evaluated predictor to test the effects of *model type* and *commonsense level* on all four rating variables. As *model type* contains more than two factors, we additionally conduct single-step corrected Tukey HSD post-hoc tests for all four variables.

Label Correctness. We do not observe a significant main effect of *model type* ($\chi^2(7) = 13.00$, $p = 0.0723$) but a significant main effect of *commonsense level* ($\beta = 0.28$, $\chi^2(1) = 4.54$, $p < 0.0331$).[7]

Explanation Correctness. We observe a main effect of *model type* ($\chi^2(7)$=24.06, p<0.0012) and *commonsense level* ($\beta = 0.27$, $\chi^2(1) = 7.79$, $p < 0.0053$). For *model type*, a post-hoc Tukey test showed significant differences between FILTERED-ENS and VANILLA ($p < 0.0055$) as well as FILTERED-ENS and COMET+CONT ($p < 0.0029$).

Grammatical Correctness. We observe a main effect of *model type* ($\chi^2(7) = 14.20$, $p < 0.0479$). However, a post-hoc Tukey test did not reveal significant differences between any model type pair. No significant main effect of *commonsense level* was observed ($\beta = 0.02$, $\chi^2(1) = 0.02$, $p = 0.8803$).

Commonsense Correctness. We observe a main effect of *model type* ($\chi^2(7) = 20.63$, $p < 0.0044$). However, a post-hoc Tukey test did not reveal significant differences between any model type pair. No significant main effect of *commonsense level* was observed ($\beta = 0.07$, $\chi^2(1) = 0.25$, $p = 0.6163$).

Overall, these results show surprisingly few significant differences between the different model types and conflict with the large differences within automatic evaluation scores.

[6] We do not consider "no need" commonsense ratings during the respective model estimation.

[7] β refers to the estimate of a *high* commonsense level.

6 Discussion

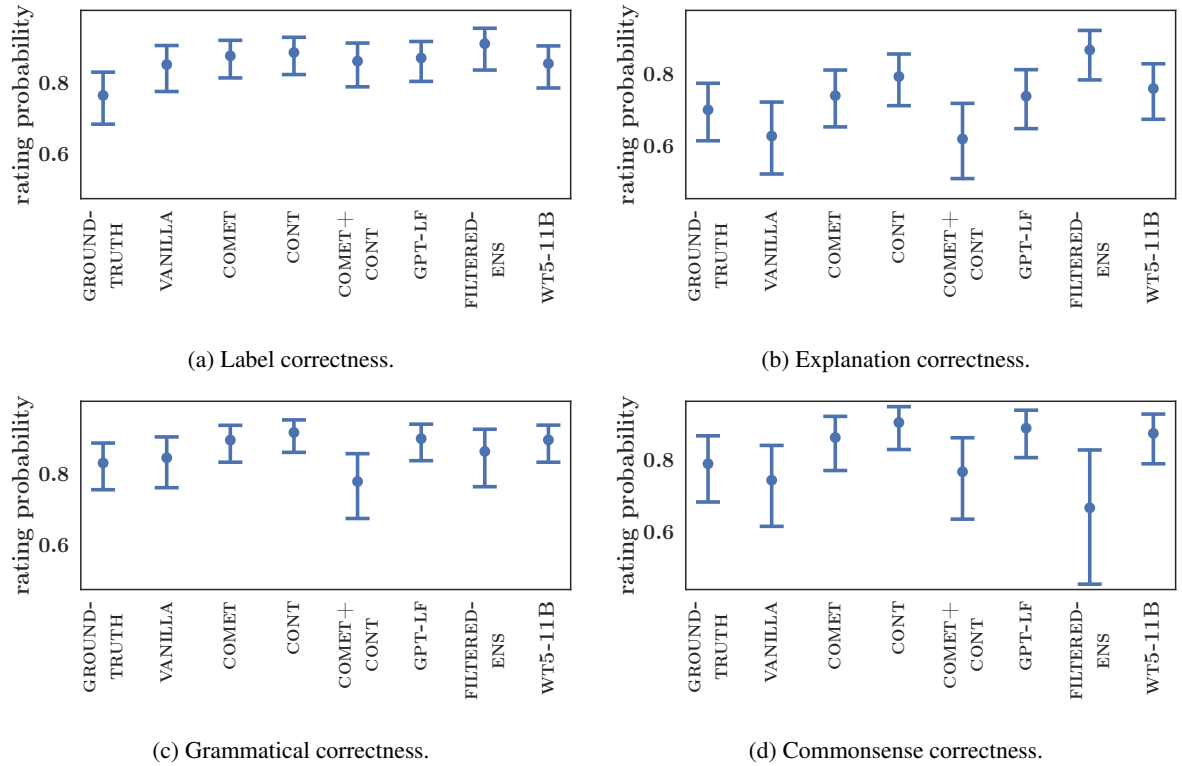
R1: Effect of External Knowledge. We showed that external knowledge can increase label accuracies on e-SNLI as well as on the stress tests. In addition, we found external knowledge to increase BLEU/BLEURT scores and thus help explanation generation in terms of automatic evaluation.

R2: Implicit Knowledge in Language Models. While language models achieve the best scores on general e-SNLI performance, the stress tests showed that they do not succeed in all reasoning types. Thus, for choosing the best way of integrating commonsense knowledge, the final reasoning goal of the model needs to be considered.

R3: Perceived Explanation Quality by Humans. We expected the large differences in e-SNLI label accuracy (up to 3.23%), BLEU (up to 10.17) and BLEURT (0.31) to reflect in human ratings, but none of these maximal differences in scores leads to a significantly different rating for any dependent variable. Regarding the observed significant differences, FILTERED-ENS is not the best model included in the study with respect to e-SNLI (WT5-11B reaches distinctly higher values for all scores) and, similarly, neither VANILLA nor COMET+CONT are the worst models on any score in Table 1. Thus, large accuracy gains do not necessarily imply better models when used in real-world applications with users. In the following, we will further discuss these results.

Superhuman Model or Noisy Ground Truth? It is particularly remarkable that the ground truth ratings do not significantly differ from any other model's ratings. In fact, the ground truth condition ranks in the lower half across all four rating dimensions and yields the lowest probability of receiving label correctness ratings as shown in Figure 2a. Similarly, Narang et al. (2020) note that in their experiment the WT5-11B model reaches a 12%-higher explanation correctness rating than the ground truths. This indicates that e-SNLI might not be suitable to distinguish performances of today's high-performing models. While it remains valuable for training, models should be scored on specifically designed evaluation sets, for example an explainable extension of the NLI stress test dataset.

Limitations and Future Directions. Although we evaluated a total of 11 different model architectures and various different sources of external



(a) Label correctness.

(b) Explanation correctness.

(c) Grammatical correctness.

(d) Commonsense correctness.

Figure 2: Effect displays for user ratings of label, explanation, grammatical and commonsense correctness depending on *model type* following Fox (2003). The rating probability is the probabality that a prediction of a respective model type is perceived to be correct by a human considering fixed effects. Error bars mark 95% confidence limits.

knowledge, this clearly does not exhaust all possible knowledge sources or architectures. While our analysis provides insight into the most common knowledge sources integrated into representative model architectures, future work should confirm our findings for additional sources and architectures. Although our user study already is the largest and most fine-grained evaluation of explainable NLI, future work should further expand the set of dependent variables to potentially reveal effects that are not visible through the lens of our experimental setup. While our work addresses the task of explainable NLI, we expect that the observed disconnect between automatic and human evaluation applies to further tasks and requires to re-assess model evaluation across explainability tasks.

7 Conclusion

In this paper, we addressed three research questions: whether integrating external knowledge can improve explainability for NLI, how effective knowledge implicitly stored in language models is for reasoning, and how humans perceive explanation quality of state-of-the-art natural language inference models. To answer these questions, we proposed different methods of integrating various knowledge sources into deep learning models. We found that fine-tuned language models reach the highest performance on e-SNLI as well as the highest average accuracy within the NLI stress test evaluation. However, their performance can break down on numerical reasoning and negations. In addition to automatic evaluation, we conducted a large-scale human crowdsourcing evaluation and found that high differences in accuracy, BLEU or BLEURT scores do not reflect in significant differences in human ratings of explanation correctness, commonsense inclusion, grammar or label correctness. This highlights an alarming disconnect between automatic evaluation scores and human ratings, that puts the real-world utility of recent model improvements into question and requires to re-think automatic evaluation across the field of explainable AI.

Acknowledgement

We thank the members of the BCAI NLP&KRR research group and the anonymous reviewers for their helpful comments. Ngoc Thang Vu is funded by Carl Zeiss Foundation.

References

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating fact checking explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7352–7364, Online. Association for Computational Linguistics.

Lisa Bauer, Lingjia Deng, and Mohit Bansal. 2021. ERNIE-NLI: Analyzing the Impact of Domain-Specific External Knowledge on Enhanced Representations for NLI. In *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 58–69, Online. Association for Computational Linguistics.

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 *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

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.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 9560–9572.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Diana Inkpen, and Si Wei. 2018. Neural natural language inference models enhanced with external knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2406–2417, Melbourne, Australia. Association for Computational Linguistics.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668, Vancouver, Canada. Association for Computational Linguistics.

Miruna-Adriana Clinciu, Arash Eshghi, and Helen Hastie. 2021. A study of automatic metrics for the evaluation of natural language explanations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2376–2387, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Keyur Faldu, Amit P. Sheth, Prashant Kikani, and Hemang Akabari. 2021. KI-BERT: infusing knowledge context for better language and domain understanding. *CoRR*, abs/2104.08145.

John Fox. 2003. Effect Displays in R for Generalised Linear Models. *Journal of Statistical Software*, 8(15).

Valentin Jijkoun and Maarten de Rijke. 2005. Recognizing textual entailment using lexical similarity. In *Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment*, pages 73–76.

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.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas. Association for Computational Linguistics.

Alexander Hanbo Li and Abhinav Sethy. 2019. Knowledge enhanced attention for robust natural language inference. *CoRR*, abs/1909.00102.

Tao Li and Vivek Srikumar. 2019. Augmenting neural networks with first-order logic. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 292–302, Florence, Italy. Association for Computational Linguistics.

Tianda Li, Xiaodan Zhu, Quan Liu, Qian Chen, Zhigang Chen, and Si Wei. 2019. Several experiments on investigating pretraining and knowledge-enhanced models for natural language inference. *CoRR*, abs/1904.12104.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. KagNet: Knowledge-Aware Graph Networks for Commonsense Reasoning. 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 2829–2839, Hong Kong, China. Association for Computational Linguistics.

Scott Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. volume abs/1705.07874.

Bill MacCartney, Michel Galley, and Christopher D. Manning. 2008. A phrase-based alignment model for natural language inference. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 802–811, Honolulu, Hawaii. Association for Computational Linguistics.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. 2020. Wt5?! training text-to-text models to explain their predictions. *CoRR*, abs/2004.14546.

Xiaoman Pan, Kai Sun, Dian Yu, Jianshu Chen, Heng Ji, Claire Cardie, and Dong Yu. 2019. Improving question answering with external knowledge. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering, MRQA@EMNLP 2019, Hong Kong, China, November 4, 2019*, pages 27–37. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, (8):9.

Colin Raffel, 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. *J. Mach. Learn. Res.*, 21:140:1–140:67.

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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144. ACM.

Hendrik Schuff, Heike Adel, and Ngoc Thang Vu. 2020. F1 is Not Enough! Models and Evaluation Towards User-Centered Explainable Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7076–7095, Online. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Dominic Seyler, Tatiana Dembelova, Luciano Del Corro, Johannes Hoffart, and Gerhard Weikum. 2018. A study of the importance of external knowledge in the named entity recognition task. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 241–246, Melbourne, Australia. Association for Computational Linguistics.

Chen Shi, Shujie Liu, Shuo Ren, Shi Feng, Mu Li, Ming Zhou, Xu Sun, and Houfeng Wang. 2016. Knowledge-based semantic embedding for machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*

(Volume 1: Long Papers), pages 2245–2254, Berlin, Germany. Association for Computational Linguistics.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings*.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451. AAAI Press.

Kyle Swanson, Lili Yu, and Tao Lei. 2020. Rationalizing text matching: Learning sparse alignments via optimal transport. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5609–5626, Online. Association for Computational Linguistics.

Xiaoyan Wang, Pavan Kapanipathi, Ryan Musa, Mo Yu, Kartik Talamadupula, Ibrahim Abdelaziz, Maria Chang, Achille Fokoue, Bassem Makni, Nicholas Mattei, and Michael Witbrock. 2019. Improving Natural Language Inference Using External Knowledge in the Science Questions Domain. In *AAAI*, pages 7208–7215.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Omar Zaidan and Jason Eisner. 2008. Modeling annotators: A generative approach to learning from annotator rationales. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 31–40, Honolulu, Hawaii. Association for Computational Linguistics.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced Language Representation with Informative Entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.

A Ensemble Architectures

Figure 3 depicts the ensemble architectures discussed in Section 3.3.

B Study Interface

Figure 4 shows an example of the study interface used to collect human ratings as discussed in Section 5.

C Knowledge Requirement Annotation

Table 4 lists the annotation guidelines used to decide on low/high levels of required external knowledge as discussed in Section 5.3. Table 5 shows example annotations.

D Examples of Human Ratings

Table 6 displays various model prediction examples and corresponding examples of human ratings.

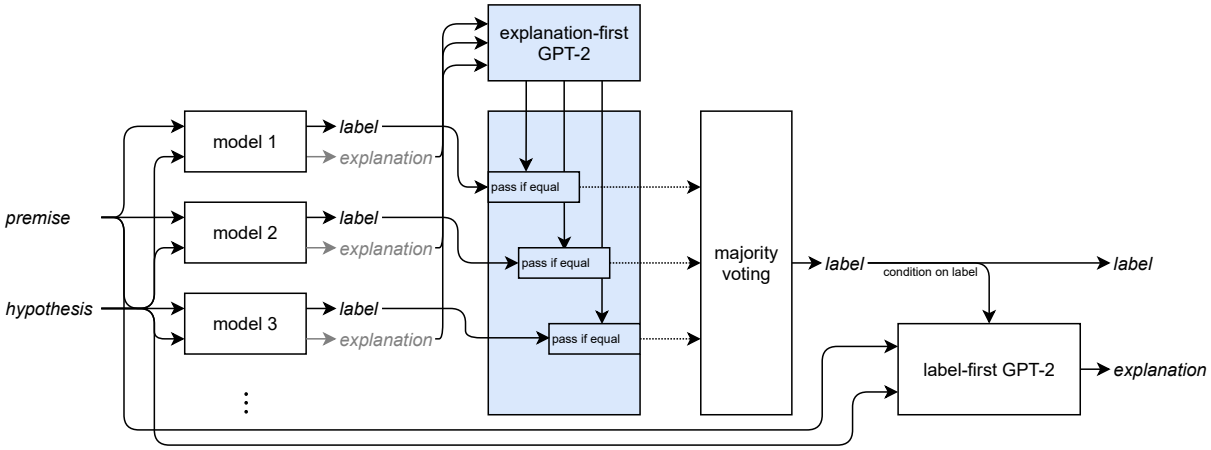


Figure 3: Ensemble architectures. The blue boxes show the consistency-filter extension.

Instance 1

Sentence1: A young girl in glasses observes something in the distance.

Sentence2: A girl sleeping on the ground.

Does the second sentence entail / contradict / is neutral to the first sentence?

Predicted answer: contradiction

Predicted explanation: The girl cannot be sleeping and observing something at the same time.

Q1. The predicted answer is correct:

○ Yes ○ No

Q2. The predicted explanation supports the model's answer prediction:

○ Yes ○ No

Q3. The explanation text is grammatically correct:

○ Yes ○ No

Q4. The explanation includes common sense knowledge required to answer the question:

○ Yes ○ No ○ No need for common sense knowledge

Figure 4: Screenshot of the study interface presented to crowdworkers on Mechanical Turk.

Low Level	Pattern Matching	The entailment can be decided by matching identical parts in the premise and the hypothesis. Premise: *A water scene with a sunset in the background.* Hypothesis:*There is a water scene with the sunset in the back.*
	Unrelated Negation	The entailment can be decided by identifying an unrelated negation. Premise: *Children bathe in water from large drums.* Hypothesis: *The kids are not reading.*
	Rephrasing	The entailment can be decided by simple rephrasing (e.g. replacing a word with a synonym). Premise: *A boy dressed in an orange shirt and a helmet is riding a dirt bike in the woods.* Hypothesis: *A boy in orange rides his dirt bike.*
	Easily-Distinguishable Concepts	The entailment can be decided by identifying unrelated concepts that have no semantic relation. Premise: *Firefighters in full gear are walking up a ladder.* Hypothesis: *The firefighters are eating lunch.*
High Level	Complex Reasoning	The entailment can be decided by resolving more complex relations and reasoning using common sense knowledge. Premise: *Soccer players are playing a night game and the ball is in the air, while the two teams fight for it.* Hypothesis: *The sun was shining during the soccer match.*
	Abstract Concepts	The entailment can be decided using common sense knowledge about abstractions of concepts. Premise: *A girl reaches up to kiss a cat, which is sitting on the counter.* Hypothesis: *A girl is showing affection towards a cat.*

Table 4: Annotation guidelines used during the annotation of low/high levels of required external knowledge with examples.

Low Level	Premise: There is a group of children getting their picture taken with presents. Hypothesis: Two men carry a Christmas tree.
	Premise: A woman looks at a plate filled with steam. Hypothesis: The woman is out shopping at the mall.
	Premise: Man sitting on bench with a suitcase in front of PADDINGTON sign. Hypothesis: A man sitting with a sign.
	Premise: A man grilling a hamburger. Hypothesis: The man is swimming at the bottom of the ocean.
	Premise: The African American man protests against unlawful sex. Hypothesis: The man protests.
High Level	Premise: A boy in a red jacket and black hat sliding on his knees down a snowy hill Hypothesis: A child is playing outside.
	Premise: A man playing a piano. Hypothesis: The man's hands are on the keys of a piano.
	Premise: 3 girls chatting and laughing on the stairwell. Hypothesis: Girls are not having a good time.
	Premise: A man visiting a friend in the hospital. Hypothesis: A man and a patient in a hospital room.
	Premise: Two girls pose along a tree-lined path and blow kisses towards the camera. Hypothesis: Two girls are taking pictures outside.

Table 5: Pairs from the low/high external knowledge requirement annotations sampled from pairs for which annotators agreed.

Premise	Hypothesis	Common Sense	Model	Predicted Label	Predicted Explanation	Human Ratings			
						Label	Expl.	Grammar	Common Sense
By a table with several boxes and two coffee cups on it, a woman in a green sweater is opening a small box and another person (whose head cannot be seen) is operating a small handheld electronic device.	Members of congress opening care packages.	high	WT5-11B	neutral	Not all women are members of congress. Not all boxes are care packages.	✓	✓	✓	✓
In front is a guy singing and playing guitar and a white shirted drummer in the back.	Some people perform at a concert.	high	GPT-LF	neutral	Just because a guy singing and playing guitar and a white shirted drummer in the back does not mean they are performing at a concert.	✓	✓	✓	✓
A man in a black coat and hat and a woman in a blue coat with pink hat snowshoeing on the side of a road with a stoplight in the background .	A man and woman are outside in the snow .	high	COMET +CONT	entailment	in the snow is outside	✓	✓	✓	✓
Two individuals dressed up like animals are posing for the camera .	Two young boys dressed like cows are posing .	low	COMET +CONT	neutral	not all people are young	✗	✗	✓	✓
A man playing a piano	The man's hands are on the keys of a piano.	high	VANILLA	entailment	playing a piano requires one to be on the hands	✓	✓	✗	✓
Two women are hugging on a path through a grassy area with a cow visible past them.	There are women outside in this picture	high	WT5-11B	entailment	A grassy area is outside.	✓	✓	✓	✓
A woman in a spaghetti strap tank wearing a flower in her hair is staring towards the right .	The woman is sleeping on a pillow .	low	COMET +CONT	contradiction	the woman can not be looking towards a sleeping she is sleeping	✓	✗	✗	✓

Table 6: Examples of model predictions and human ratings of the predicted labels and explanations for various model types.