On the Challenges of Evaluating Compositional Explanations in Multi-Hop Inference: Relevance, Completeness, and Expert Ratings

Peter A. Jansen and **Kelly Smith** and **Dan Moreno** and **Huitzilin Ortiz**

University of Arizona, USA

`pajansen@arizona.edu`

Abstract

Building compositional explanations requires models to combine two or more facts that, together, describe why the answer to a question is correct. Typically, these "multi-hop" explanations are evaluated relative to one (or a small number of) gold explanations. In this work, we show these evaluations substantially underestimate model performance, both in terms of the *relevance* of included facts, as well as the *completeness* of model-generated explanations, because models regularly discover and produce valid explanations that are different than gold explanations. To address this, we construct a large corpus of 126k domain-expert (science teacher) relevance ratings that augment a corpus of explanations to standardized science exam questions, discovering 80k additional relevant facts not rated as gold. We build three strong models based on different methodologies (generation, ranking, and schemas), and empirically show that while expert-augmented ratings provide better estimates of explanation quality, both original (gold) and expert-augmented automatic evaluations still substantially underestimate performance by *up to 36%* when compared with full manual expert judgements, with different models being disproportionately affected. This poses a significant methodological challenge to accurately evaluating explanations produced by compositional reasoning models.

1 Introduction

Compositional inference is the high-level task of combining two or more pieces of knowledge to perform reasoning. In the context of question answering, compositional (or "multi-hop") inference typically takes the form of combining facts from a knowledge base that allow a given solver to form a complete chain-of-reasoning that moves from question to correct answer. A desirable consequence is that the facts used to assemble this chain-of-reasoning can then be taken as an inter-

Question: When trees are cleared from the land, what will most likely occur?
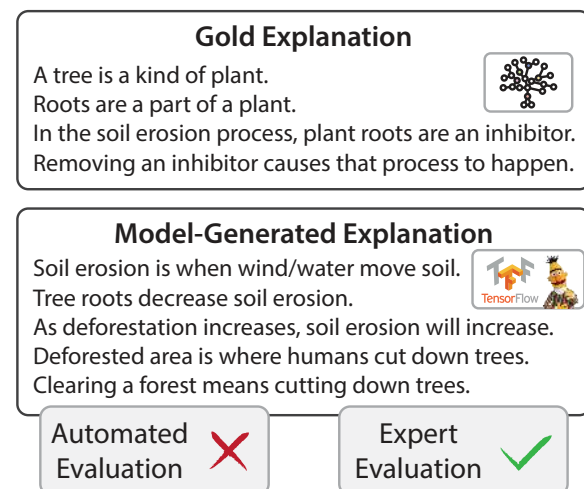**Answer:** Soil Erosion

Figure 1: An example science exam question, its gold explanation from the WorldTree corpus, and a model-generated explanation from one of the models *(Tensorflow-Ranking-BERT)* trained using expert-generated relevance ratings produced in this work. Though the model-generated explanation is strong, it shares no facts in common with the gold explanation, and automatic evaluations rate it neither *relevant* nor *complete*.

pretable record of that reasoning, as well as a human-readable explanation for why the answer is correct.

Compositional inference has seen steady growth in the last three years, in large part due to the recent availability of training and evaluation data for the task (e.g. Yang et al., 2018; Khashabi et al., 2018; Jansen et al., 2018), which has historically been unavailable due to the challenges in annotating explanations, and the expense in generating quality data at scale. To ease these burdens, nearly all datasets have focused on small compositional inference problems that require composing only two representations, typically triples, sentences, or whole paragraphs (see Wiegreffe and Marasović, 2021, for a survey of datasets).

In this work, we focus on the problem of generating and evaluating large explanations to science

exam questions (with the average explanation in this work requiring composing 6 facts). Our evaluation experience in this domain has been uneasy – we have developed seemingly well-reasoned models, only to receive comparatively low evaluation scores relative to baseline models in automatic evaluations. When evaluated manually, as shown in Figure 1, we observe that many models produce compelling explanations, at least in part, but these explanations score poorly because they differ from gold explanations. This parallels the disparity in automatic versus manual evaluations in other fields, such as machine translation (Freitag et al., 2021).

In this work we systematically analyze the difference between automatic and expert manual evaluation, and formalize evaluation of large compositional explanations in two aspects: by examining the **relevance** of each fact to the question and answer, as well as the **completeness** of the entire explanation – that is, whether the collection of facts in the explanation form a complete chain-of-reasoning from question to answer. Of these two metrics, obtaining accurate *relevance* measures is in principle solvable by brute force – by creating an exhaustive corpus of ratings – while exhaustively enumerating possible *n-fact* explanations and rating them for *completeness* is likely less tractable. Here, we focus on generating extensive *relevance* annotation of the facts most likely to be incorporated in explanations, while providing an estimate of the undercounting of *completeness* by manually evaluating the completeness of explanations generated by three state-of-the-art models.

The contributions of this work are:

1. **Resource:** We produce a large set of 126k expert-generated relevance ratings for building explanations to science exam questions from atomic facts. Each rated fact was highly ranked by a large language model trained on gold explanations using distant supervision, complementing those in the WorldTree V2 explanation corpus (Xie et al., 2020). The domain experts (science teachers) discovered 80.6k additional facts, *four times more than provided in gold explanations*, to be relevant for explanation construction.

2. **Models:** We use these ratings to train and evaluate three state-of-the-art exhaustive models, using three different modeling paradigms: explanation-as-ranking, generation, and constraint-based schemas.

3. **Evaluation:** We conduct large automatic and manual evaluations, empirically demonstrating substantial differences in evaluation when using better ratings and judgements. In fully-automatic evaluations, original evaluations underestimate *relevance* by up to 14% compared with a fully-automatic evaluation that includes expert ratings. But, this fully-automatic setting still underestimates *relevance* by up to 29% and *completeness* by up to 36% compared to full manual judgements.

2   Related Work

Compositional explanations: In their survey, Weigreffe and Marasovic (2021) identified 14 structured explanation datasets for compositional reasoning. Due to the challenge in annotating large compositional explanations, nearly all datasets to date (such as QASC (Khot et al., 2020), Open-BookQA (Mihaylov et al., 2018), and R^{4}C (Inoue et al., 2020)) require combining an average of only 2 facts. In this work, to study evaluation challenges with the largest available explanations, we use the WorldTree V2 explanation corpus (Xie et al., 2020), whose explanations require composing an average of 6 (and as many as 16) facts.

Evaluation with multiple gold explanations: Nearly all compositional reasoning datasets annotate (at most) a single gold explanation, with two exceptions. eQASC (Jhamtani and Clark, 2020) generates 10 perturbations of possible 2-fact QASC explanations, and asks crowdworkers to rate these as valid or invalid chains of reasoning. 26% are rated valid, resulting in an average of 2.6 valid 2-fact explanations per question, which Jhamtani and Clark (2020) then use to train a classifier and ranker. Taking a different approach, R^{4}C (Inoue et al., 2020) uses crowdworkers to generate 3 explanations (represented as chains of triples) to select HotpotQA questions (Yang et al., 2018). R^{4}C explanations are short, with 68% containing 2 triples, 23% using 3 triples, and 9% use 4 or more triples. Inoue et al. (2020) then define an alignment procedure between model-generated output triples and the 3 gold explanations, and take the highest scoring alignment as the score of the explanation. In this work, due to the intractability of generating and rating explanatory perturbations with large explanations, we instead use domain-experts to produce relevance ratings for component facts at scale.
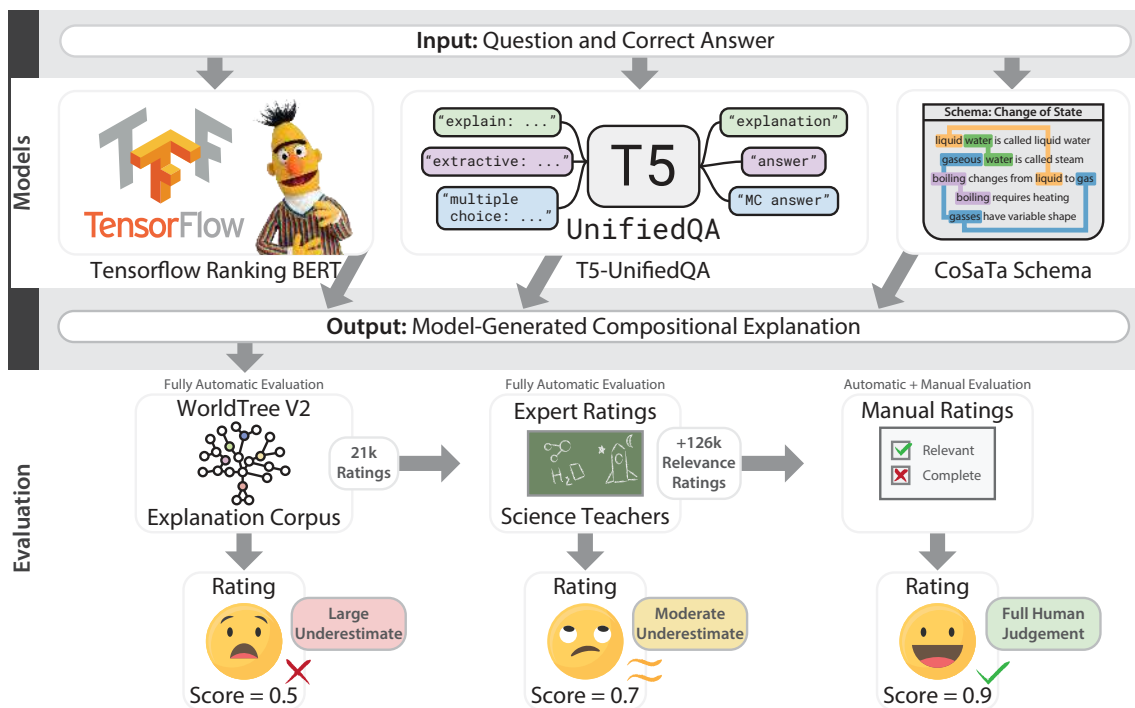
Figure 2: An overview of this work. We generate a large set of relevance ratings for explanatory facts annotated by domain experts (science teachers), that complement the original WorldTree explanation corpus. We use this new annotation for training and evaluating three families of strong models (Tensorflow Ranking BERT, T5-UnifiedQA, and CoSaTa Schemas) on generating explanations. We show through automatic and manual analyses that the current method of using a single gold explanation for evaluation substantially undercounts explanation performance in terms of relevance and completeness, while even expert relevance ratings (when used in fully-automatic evaluations) still moderately undercount true task performance compared to full manual human judgements.

We then rate model explanations for *completeness* manually, as we hypothesize that generating a large database of alternative gold explanations as in Inoue et al. (2020) is likely intractable for explanations longer than two or three facts.

Rating completeness is challenging, with opportunities for bias. For example, Korman et al. (2020) note that "all explanations are incomplete, but reasoners think some explanations are more complete than others", and empirically determined that humans prefer simpler explanations that reduce gaps in causal explanatory steps. Here we consider explanations correct if they are technically correct (as determined by domain experts), and minimize gaps in inferences, without measuring succinctness.

Modeling approaches to generating explanations: A wide variety of approaches have been proposed for building compositional explanations (see Thayaparan et al., 2020, inter alia), including integer linear programming (Khashabi et al., 2016), formal logics and rules (Weber et al., 2019), iterative construction methods (Cartuyvels et al., 2020), and various explanation-as-ranking approaches such as those proposed in the Shared Tasks for Explanation Regeneration (e.g. Jansen and Ustalov, 2020). In this work we explore evaluation challenges grounded in the performance of three strong and methodologically diverse models: reranking exhaustive classifications of large language models (e.g. Das et al., 2019; Li et al., 2020), generative models (e.g. Khashabi et al., 2020), and schema-based models (e.g. Lin et al., 2019; Jansen, 2020)

3 Overview

An overview of our approach is shown in Figure 2. First, in Section 4 we generate a large set of expert relevance ratings that complement those in the WorldTree V2 explanation corpus. We then use these expert relevance ratings to demonstrate existing evaluations substantially undercount relevance in explanation-as-ranking paradigms in Section 5. In Section 6 we implement strong generative, ranking, and schema-based models that produce whole explanations (rather than ranked lists), and show through manual analysis that automated metrics substantially undercount relevance and completeness of whole explanations, even when using better ratings. We conclude with a discussion of implica-

Q: Burning fossil fuels adds pollutants like sulphur into the air. This pollution contributes to: A: acid rain		
TR	**Gold**	**Fact**
3	*	Burning fossil fuels releases sulfur dioxide into the atmosphere.
3	*	Emitting sulfur dioxide causes acid rain.
2		Burning fossil fuels causes pollution.
2	*	Emission is when something is added to. the atmosphere.
2		Gasses from burning oil and coal that dissolve in water in the atmosphere cause acid rain.
2		As the amount of sulphur gas in the atmosphere increases, the PH of rain will decrease.
1		Acid rain negatively impacts water quality.
1		Coal is a kind of fossil fuel.
0		The air contains carbon dioxide.
0		Oil is a kind of pollutant.

Table 1: Example relevance ratings. *(top)* A question and its correct answer. *(bottom)* A subset of the shortlist of facts, teacher-generated relevance ratings (TR) for each fact, and whether a given fact was included in the gold explanation in the original WorldTree V2 explanations. *Note:* for space, only a subset of the shortlist of facts and gold explanation are shown.

4 Dataset Description

To support our experimentation we created a resource that, for a given WorldTree V2 question, provides a set of relevance ratings for the facts most likely to be in an explanation. WorldTree V2 contains 4.4k questions, and its supporting knowledge base contains approximately 9k facts, meaning that exhaustively evaluating the relevance of each fact for each question would require approximately *40 million* ratings, which is intractable even for this comparatively small corpus. Instead, here, annotators rate a shortlist of facts most likely to be relevant to building possible explanations, producing ratings for a total of 126k facts.

Initial Shortlist: To produce the shortlist, we train two large language models, BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), on the task of retrieving relevant sentences from the corpus. To encourage the models to find a broad array of facts that might be relevant to building an explanation, we model this retrieval as a distant supervision classification problem where the gold WorldTree explanations are used as positive examples, while 200 randomly sampled facts from the corpus serve as negative training examples. For each question, we exhaustively score all 9k facts in the knowledge base using both models, take the

top 20 scoring facts from each model, and combine them into a final shortlist. We also add any facts from the gold explanation that were not ranked in the top 20 by either model. Due to significant overlap in the output of both models, the average shortlist per question contains 28.9 facts.[1]

Rating protocol: Each question's shortlist of facts was independently rated by 2 domain-expert annotators (science teachers), using the following 4-point rating scheme:

TR	Label	Description
3	Core	Facts that directly address the core topic the question is testing.
2	Important	Key knowledge supporting the core facts or grounding core knowledge in examples the question uses.
1	Extra Detail	Facts that (a) when included, add extra detail to the explanation, but (b) when missing, do not exclude important details from the explanation.
0	Irrelevant	Facts not relevant to the question.

Table 2: 4-point Relevance Rating Scheme

Central to this rating scheme is a graded notion of relevance, that includes optional facts *(extra detail)* that enrich the explanation when included, but do not cause critical gaps in the inference when not included. Example ratings are shown in Table 1.

Raters: Three graduate research assistants in education served as domain experts, and worked for several months to complete the relevance ratings. Each has between 8 and 20 years of science teaching experience at the elementary, middle-school, or high-school level.

Interannotator Agreement: Even after substantial training, the domain experts found this to be a challenging task. Interannotator agreement (Cohen's Kappa) was $\kappa = 0.46$, which is considered moderate agreement (Landis and Koch, 1977). Raw percent agreement between annotators was 61%, with nearly all disagreements within ± 1 of each other (88% of disagreements). Annotators reported that disagreements tended to be in determining thresholds for the different categories – some annotators tended to err on the side of suggesting more facts were important to generating an explanation, while others preferred generating more minimalistic explanations. To account for individual

tions for evaluation in multi-hop inference.

---

[1] Scoring only the top $\approx$30 facts per question was chosen due to budgetary constraints and timing considerations.

		Teacher Rating			
		0 Irr.	1 Ext.	2 Imp.	3 Core
WT2	Gold	315	2,846	9,063	8,579
	Not Gold	24,795	36,962	36,399	7,245
	Increase	–	1299%	402%	84%

Table 3: Distribution of teacher ratings for shortlisted facts across all explanations, broken down by each fact's original WorldTree V2 (WT2) rating. In total, this annotation procedure discovered 80.6k additional relevant facts in the corpus (approximately 18 facts per question) not originally included in the single gold WorldTree explanation per question.

variation in detail preference, we average the final ratings and round up the final scores.

Comparison with gold explanations: The distribution of relevance ratings is shown in Table 3, broken down by whether a given fact was originally included in the gold WorldTree explanation for a given question. Of the 20.8k facts across all gold WorldTree explanations, the teachers rated 98.5% of these as also relevant (i.e., $TR > 0$), demonstrating strong agreement with the original explanation authors. Teachers rated 17.6k (85%) of these facts as *core* or *important* to the inference, while 2,846 (14%) were rated as *extra detail* facts that the explanations could include or disclude without causing significant gaps in the reasoning.

Most stark is the volume of additional facts rated as relevant by the domain experts. In total, 80.6k additional facts (approximately 18 facts per question) were rated as relevant by teachers, but were not used in a given question's gold explanation, with 43.6k of these facts (10 per question) rated as *core* or *important*. As context, the original WorldTree explanations contain an average of 6 facts – here, the expert ratings found that, on average, four times as many facts are relevant to building an explanation than are annotated in a given question's single explanation. This suggests that *while providing a single example explanation is helpful for training a model, it is insufficient for evaluating that model's capacity for constructing explanations*, as it may be possible to build many different compositional explanations from a given collection of facts.

5 Experiments: Explanation-as-Ranking

To characterize differences in automatic evaluation when using existing gold explanations versus expert relevance ratings, we first explore performance in the explanation-as-ranking paradigm. Rather than directly producing an explanation for a given question, as a stepping-stone task, explanation-as-ranking (Jansen and Ustalov, 2019; Das et al., 2019; Li et al., 2020) is an explanatory retrieval analogue that requires models to exhaustively rank all facts in a knowledge base such that the most relevant facts are selectively ranked to the top of the list.[2]

Models: We include the exhaustive BERT and RoBERTa models trained on the original WorldTree gold explanations, and used in generating the shortlist for the teacher ratings, as described above. We also include a Tensorflow-Ranking-BERT model (Han et al., 2020), which combines BERT embeddings directly in a pointwise learning-to-rank (Pasumarthi et al., 2019a) instead of classification framework, and achieves extremely strong single-model performance for large benchmark ranking tasks such as MS MARCO (Nguyen et al., 2016). Unlike the baseline models, TFR-BERT was trained on the expert-generated relevance ratings. Due to the expense in evaluating this model, here we rerank only the top 100 scoring facts from the exhaustive BERT model.

Evaluation: Explanation-as-ranking performance is reported using Mean Average Precision (MAP). We evaluate in three settings: (1) Using the original gold WorldTree explanations, (2) treating each fact the experts rated as *extra-detail* or higher as gold, or (3) treating each fact the experts rated as *important* or higher as gold. Because the expert ratings are graded (0-3) rather than binary (gold/not gold), we also evaluate using Normalized Discounted Cumulative Gain (NDCG).

Results: The results of the evaluation are shown in Table 4. Using the original gold annotation, the best scoring model, RoBERTa, achieves a MAP of 0.57. When evaluated using the expert relevance ratings, this increases to 0.61 (+4%) when considering only *important* or higher facts as gold, while increasing to 0.68 (+11%) when allowing *extra detail* facts to be considered gold. Conversely, the TFR-BERT shows comparatively low performance using the original gold annotation, at 0.49 MAP.

---

[2]Note that the dataset generated in this paper was used for the Third Shared Task on Multi-Hop Inference for Explanation Regeneration, an explanation-as-ranking task (Thayaparan et al., 2021) that ran concurrently with this submission. The TFR-BERT model described here performs comparably with the winning system (Pan et al., 2021), which reached 0.82 NDCG.

| Model | Ranking Setting | Explanation-as-Ranking Evaluation Method (MAP) | | | | | |
		Baseline WT2	Teacher *(>= Extra (1))*	Δ	Teacher *(>=Important (2))*	Δ	Teacher NDCG
BERT	Exhaustive	0.54	0.66	↑0.11	0.58	↑0.03	0.75
RoBERTa	Exhaustive	**0.57**	**0.68**	↑0.11	0.61	↑0.04	0.78
TFR-BERT	Rerank (K=100)	0.49	0.58	↑0.09	**0.63**	↑0.14	**0.81**

Table 4: Using more extensive expert relevance ratings causes substantially different MAP scores for the same models in fully-automatic evaluation settings. *(Left)* Models, and their ranking setting. All models were exhaustively evaluated on the entire corpus except TFR-BERT, which reranks the top-100 BERT facts per question. *(Right)* Evaluation scores when using original WorldTree (WT2) ratings, or the top-K expert (science teacher) ratings produced in this work. Teacher evaluations are provided in two settings: considering all facts rated "Extra" or above as gold, or all facts rated "Important" or above as gold. All scores represent Mean Average Precision (MAP), except for the last column, which provides Normalized Discounted Cumulative Gain (NDCG) for reference. Delta scores (Δ) represent the difference between a given teacher evaluation and the WT2 baseline. Automatic measures of relevance can differ by up to 14% when using either a single gold explanation or top-K expert relevance ratings as gold.

When evaluated using the exhaustive teacher ratings, this model achieves a MAP of 0.63 (+14%) when considering only *important* or higher facts as gold – becoming the best-performing model – while reaching 0.58 (+9%) when considering *extra detail* or better facts as gold. These results empirically demonstrate that using a single gold explanation as an evaluation standard can dramatically underestimate model performance compared to more extensive relevance annotation. Further, *this performance decrease may not be uniform across models*, as demonstrated by RoBERTa's performance being underestimated by only 4% while TFR-BERT is underestimated by 14% – suggesting that meaningful comparisons between model performance may not be possible with limited relevance annotation.

6 Experiments: Whole Explanations

Here we construct and evaluate a diverse set of generative, top-k ranking, and schema-based models that return whole (short-length) explanations to a user, in place of ranked lists. We evaluate these in terms of both *relevance* and *completeness*, comparing both automatic and manual assessments of these metrics.

6.1 Models

Each model is described briefly below, with details and hyperparameters provided in the `Appendix`.

T5-UnifiedQA (Generative): UnifiedQA (Khashabi et al., 2020) is a variant of T5 (Raffel et al., 2020) that includes multi-task pretraining for 4 different forms of question answering across 20 datasets, including extractive QA (i.e., locating an answer span in a passage), abstractive QA (i.e., generating an answer not supplied in a

passage), multiple-choice QA, and Boolean QA, while achieving state-of-the-art performance on 10 datasets. Here we train T5-UQA-3B to also generate compositional explanations by cueing generation with the question and correct answer candidate, and targeting generation to produce strings of highly-rated facts delimited with an `[AND]` separator token.

The model was trained to produce all facts rated as relevant by expert raters for a given question. This could result in long strings, so we trained two independent subtasks: a CORE subtask that includes all facts rated *important* or greater, and an EXT subtask that includes only *extra-detail* ($0 < TR < 2$) facts. The model was implemented using Huggingface Transformers (Wolf et al., 2020). To simplify evaluation, T5-generated facts are aligned to their best-scoring WorldTree knowledge base fact using ROUGE-1 scores (Lin and Hovy, 2003).

TFR-BERT (Ranking): A Tensorflow-Ranking-BERT model (Han et al., 2020), trained on expert-generated data, as described in Section 5. To move from ranking to explanation generation, we simply take the top-K ranked facts per question as the explanation, using an empirically determined threshold of $K = 8$ (where F1 performance plateaued on the development set).

COSATA (Schema): A schema-based model implemented using the Constraint Satisfaction over Tables (COSATA) solver (Jansen, 2020). Schemas take the form of constraint satisfaction patterns over Worldtree facts represented as semi-structured table rows, where valid solutions of a given schema are only possible if all slots (facts) in a schema can be successfully populated by satisfying their constraints (see Figure 2). We use the 385 science-

| Model | Rel | Automatic Analysis | | | | Average |
		Comp	$F1^{ex}$	$Comp_B$	$F1_B^{ex}$	Expl. Length
Single Models						
Generative and Ranking						
T5-UQA-3B$_{CORE}$	0.62	0.32	0.42	0.20	0.12	8
T5-UQA-3B$_{CORE+EXT}$	0.55	0.39	0.45	0.23	0.14	13
TFR-BERT	0.74	**0.59**	**0.66**	**0.50**	**0.37**	8
Schema-based						
Schema (1 Schema)	**0.82**	0.36	0.50	0.27	0.16	5
Schema (2 Schemas)	0.78	0.42	0.55	0.32	0.20	7
Schema (3 Schemas)	0.75	0.46	0.57	0.36	0.23	8
Ensembles						
T5-UQA-3B$_{C+E}$ + TFR-BERT	0.62	0.70	0.66	0.54	0.48	21
Schema (3S) + TFR-BERT	**0.75**	0.71	**0.73**	0.59	0.49	16
Schema (3S) + T5-UQA-3B$_{C+E}$	0.62	0.61	0.62	0.46	0.36	22
Schema (3S) + TFR-BERT + T5-UQA-3B$_{C+E}$	0.65	**0.77**	0.71	**0.61**	**0.57**	30

Table 5: Performance of all single and ensemble models investigated, using automatic performance metrics. *Relevance* is measured using expert (teacher) ratings, while *completeness* is measured using a combination of orginal WorldTree V2 and teacher ratings (see text).

domain schema included with COSATA, each containing an average of 12 facts (before filtering), and run these over the WorldTree knowledge base, generating a large set of 593k solutions that are pre-cached for speed. Each solution is scored using exhaustive BERT rankings to select 1, 2, or 3 schemas to combine and output to the user. Before output, low-scoring facts are filtered to improve succinctness.

6.2 Automatic Evaluation Metrics

Here we evaluate models on *relevance*, *completeness*, and an F1 analogue combining the two.

Relevance: The proportion of facts in an explanation deemed not-irrelevant to the question (i.e. that received a non-zero expert relevance rating).

Completeness: When evaluated automatically, *completeness* represents the proportion of facts in the gold WorldTree explanation that are also found in the model-generated explanation. We also define $Comp_B$, a binary measure, that is 1 if all facts in the WorldTree gold explanation that were rated as *important* or higher by the experts are included in the model-generated explanation, and 0 otherwise.

Due to anticipated methodological issues – that valid explanations other than the gold annotated explanation are possible – we also evaluate $Comp_B$ manually in Section 6.4, using the criterion that an experienced annotator believes the facts in the explanations form a complete chain-of-reasoning from question to answer without significant gaps.

F1: With *relevance* an analogue of precision, and *completeness* an analogue of recall, to provide a single score that reflects overall explanation performance, we also provide an F1 analogue, defined as the harmonic mean of *relevance* and *completeness*:

$$F1^{ex} = \frac{Relevance \cdot Completeness}{Relevance + Completeness} \quad (1)$$

6.3 Results using Automated Metrics

The performance of the ranking, generative, and schema-based models using the expert-informed automatic evaluation metrics is shown in Table 5, broken down by single models and ensembles. In terms of single models, at 0.82 the Schema-based models perform highest in *relevance* – likely owing to the constraints of each schema ensuring that collections of facts are organized according to a particular theme – followed by TFR-BERT, with the best performing T5-UQA model performing 20 points lower, at 0.62. Conversely, TFR-BERT scores highly in both graded and binary *completeness*, reaching 0.50 $Comp_B$, while T5-UQA reaches less than half this performance, and the schema models reach a middle-ground of 0.36. The best-scoring model, TFR-BERT, reaches 0.66 $F1^{ex}$, or 0.37 $F1_B^{ex}$ using binary completeness.

Given the variety of methodologies used, the three model families have comparatively low overlap, and ensemble models that combine the output of each model substantially improve *completeness* and $F1^{ex}$ scores. The best-scoring model, which

Pattern Scoring Method	Automatic Analysis					Manual Analysis			Underestimate (Δ)		
	Rel	Comp	$F1^{ex}$	$Comp_B$	$F1_B^{ex}$	Rel	$Comp_B$	$F1_B^{ex}$	Rel	$Comp_B$	$F1_B^{ex}$
T5-UQA-3B$_{CORE}$	0.53	0.36	0.43	0.10	0.17	0.82	0.44	0.57	**+0.29**	+0.34	**+0.40**
TFR-BERT	0.72	**0.59**	**0.65**	**0.36**	**0.48**	**0.93**	**0.72**	**0.81**	+0.21	**+0.36**	+0.33
Schema (3 Schemas)	**0.74**	0.46	0.57	0.21	0.33	0.79	0.44	0.57	+0.05	+0.23	+0.24

Table 6: A manual analysis of *relevance* and *completeness* of three length-matched single models on 50 questions from the development set. Here, each model produces an explanation with an average length of 8 facts. *Underestimate* refers to difference scores between manual and automatic measures. *Note:* Automatic analysis numbers differ from those in Table 5 as they represent performance on only the 50 questions included in this analysis.

Question: Which generates waves that are capable of traveling through a vacuum?	
Answer: a light bulb	

Gold Explanation

TR	Fact
3	* A light bulb generates visible light when turned on.
2	Visible light is a kind of light.
3	Light can travel through a vacuum.
2	* Light is a kind of wave.

Schema-Generated Explanation

Schema 1: Light Properties
3	Electromagnetic waves can travel through a vacuum.
3	(Light is a kind of electromagnetic radiation)
1	Light travels fastest through a vacuum.
2	* Light is a kind of wave.

Schema 2: Light Bulb Uses
1	A light bulb is used for seeing in the dark.
3	* A light bulb generates visible light when turned on.

Schema 3: Parts of things
1	A light bulb is a part of a lamp.

Table 7: An example schema-generated explanation rated poorly by the automated analysis, but rated complete by the manual analysis. *TR* signifies expert (teacher) ratings of each fact. * signifies that a fact occurs in both generated and gold explanation. *(brackets)* signify a fact was filtered out using the schema filtering criterion to generate a more succinct explanation, but is included for demonstrative purposes.

combines the output of the generative, ranking, and schema models achieves an $F1_B^{ex}$ of 0.57, reaching 20 points over the best-scoring single model.

6.4 Manual Evaluation of Completeness

The approximate automatic measure of *completeness* used above is still problematic, because it relies on a single gold explanation filtered to include only the most important facts in the expert ratings. To measure the difference between this automatic measure of completeness and actual completeness, we conducted a detailed manual evaluation of single model performance for 50 questions in the development set. To control for explanation length, we chose single models whose average explanation lengths were identical (8 ± 0.5 facts long), while for robustness we evaluated completeness using binary

judgements. In addition, any facts without expert relevance ratings (i.e. facts that were not within the initial top-K list rated by the expert annotators) were provided binary relevance judgements.[3]

Raters: Annotating the completeness of a collection of facts as an explanation can be challenging, particularly when locating gaps in an inference. Due to timing constraints, in this analysis, completeness judgements were initially made by an author (science-domain expert and compositional reasoning expert), then compared against those generated by one of the domain-expert science teachers from Section 4. Percent agreement between the author and teacher was strong, at 89% for binary completeness judgements, and 88% for binary relevance judgements.

Results: The results of this manual analysis are shown in Table 6. This analysis shows that even when supplemented with expert relevance ratings, using a single gold explanation for automatically evaluating completeness still provides a large underestimate of task performance. In particular, manual binary completeness ratings $Comp_B$ exceeded automatic evaluations by +23% to +36%, *and in all cases more than doubled the original estimate of task performance*. To illustrate this, Table 7 shows an example of a gold explanation, and an explanation generated by the Schema model. Both explanations are complete, but the Schema-generated explanation is rated poorly because it includes only half of the highly-rated facts of the gold explanation. Clearly as explanations become large, and composed of increasingly atomic facts, many more paths to generating complete explanations are possible, and alternate methods of accurately estimating completeness are required.

---

[3]A total of 327 facts, approximately 2 per model-generated explanation, required these relevance judgements.

7 Conclusion

Relevance performance is still undercounted:
While the expert-generated relevance ratings produced in this work provide more accurate estimates of performance compared to single gold explanations when used in fully-automatic evaluations, these automatic estimates still undercount overall model performance. In our experiments we show the expert ratings primarily provide a vehicle for training better models, but that automatically evaluating relevance performance still remains a challenge, even with a large targeted increase in relevance annotation. Further, annotators reported that determining relevance of single facts in isolation is challenging because it lacks the broader compositional context of the rest of the candidate explanation, suggesting ultimate limits to the utility of exhaustive annotation.

Measuring completeness is a major challenge:
As explanations become larger, and facts become more atomic, models are afforded more opportunities to build explanations that differ from those annotated as gold. Because of this, automatic completeness judgements substantially undercount true (manual) completeness by at least a factor of two across all models. Alignment approaches (Inoue et al., 2020) and annotating multiple explanations (Jhamtani and Clark, 2020) have been proposed for short explanations, but are unlikely to scale well as compositionality increases. Treatments similar to those that use a formal semantics or theorem proving to evaluate truth (e.g. Weber et al., 2019; Clark et al., 2020) are attractive, but are unlikely to offer generality at scale without substantial development effort. Conversely, streamlined manual evaluation frameworks are becoming increasingly common for simpler generative tasks (e.g. Khashabi et al., 2021), but it is unclear how accurately non-technical crowdworkers would perform on rating compositional completeness – and even if possible, this would raise the time and cost associated with evaluation dramatically.

Automatic model comparisons are inaccurate:
While T5-UQA and the Schema models show large performance differences using automatic measures, our manual analysis shows they actually have similar performance characteristics as performance underestimates disproportionately affect different models. Comparing models to perform hypothesis testing (i.e. *Model A outperforms Model B*) is cur-

rently challenging without substantial manual analysis, and a significant methodological limitation to advancing the science of compositional reasoning for building large explanations.

Open Data

Our corpus and analyses are available at: http://cognitiveai.org/explanationbank/.

Acknowledgements

References

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.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2020. Transformers as soft reasoners over language. *arXiv preprint arXiv:2002.05867*.

Rajarshi Das, Ameya Godbole, Manzil Zaheer, Shehzaad Dhuliawala, and Andrew McCallum. 2019. Chains-of-reasoning at textgraphs 2019 shared task: Reasoning over chains of facts for explainable multi-hop inference. In *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13)*, pages 101–117.

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.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *arXiv preprint arXiv:2104.14478*.

Shuguang Han, Xuanhui Wang, Mike Bendersky, and Marc Najork. 2020. Learning-to-rank with bert in tf-ranking. *arXiv preprint arXiv:2004.08476*.

Naoya Inoue, Pontus Stenetorp, and Kentaro Inui. 2020. R4C: A benchmark for evaluating RC systems to get the right answer for the right reason. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6740–6750, Online. Association for Computational Linguistics.

Peter Jansen. 2020. CoSaTa: A constraint satisfaction solver and interpreted language for semi-structured tables of sentences. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 70–76, Online. Association for Computational Linguistics.

Peter Jansen and Dmitry Ustalov. 2019. TextGraphs 2019 shared task on multi-hop inference for explanation regeneration. In *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13)*, pages 63–77, Hong Kong. Association for Computational Linguistics.

Peter Jansen and Dmitry Ustalov. 2020. TextGraphs 2020 shared task on multi-hop inference for explanation regeneration. In *Proceedings of the Graph-based Methods for Natural Language Processing (TextGraphs)*, pages 85–97, Barcelona, Spain (Online). 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.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262, New Orleans, Louisiana. Association for Computational Linguistics.

Daniel Khashabi, Tushar Khot, Ashish Sabharwal, Peter Clark, Oren Etzioni, and Dan Roth. 2016. Question answering via integer programming over semi-structured knowledge. *arXiv preprint arXiv:1604.06076*.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UNIFIEDQA: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907, Online. Association for Computational Linguistics.

Daniel Khashabi, Gabriel Stanovsky, Jonathan Bragg, Nicholas Lourie, Jungo Kasai, Yejin Choi, Noah A Smith, and Daniel S Weld. 2021. Genie: A leaderboard for human-in-the-loop evaluation of text generation. *arXiv preprint arXiv:2101.06561*.

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

Joanna Korman and Sangeet Khemlani. 2020. Explanatory completeness. *Acta Psychologica*, 209:103139.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Weibin Li, Yuxiang Lu, Zhengjie Huang, Weiyue Su, Jiaxiang Liu, Shikun Feng, and Yu Sun. 2020. Pgl at textgraphs 2020 shared task: Explanation regeneration using language and graph learning methods. In *Proceedings of the Graph-based Methods for Natural Language Processing (TextGraphs)*, pages 98–102.

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.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 150–157.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset. In *CoCo@ NIPS*.

Chunguang Pan, Bingyan Song, and Zhipeng Luo. 2021. Deepblueai at textgraphs 2021 shared task: Treating multi-hop inference explanation regeneration as a ranking problem. In *Proceedings of the Fifteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-15)*, pages 166–170.

Rama Kumar Pasumarthi, Sebastian Bruch, Xuanhui Wang, Cheng Li, Michael Bendersky, Marc Najork, Jan Pfeifer, Nadav Golbandi, Rohan Anil, and Stephan Wolf. 2019a. Tf-ranking: Scalable tensorflow library for learning-to-rank. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2970–2978.

Rama Kumar Pasumarthi, Sebastian Bruch, Xuanhui Wang, Cheng Li, Michael Bendersky, Marc Najork, Jan Pfeifer, Nadav Golbandi, Rohan Anil, and Stephan Wolf. 2019b. Tf-ranking: Scalable tensorflow library for learning-to-rank. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2970–2978.

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. *Journal of Machine Learning Research*, 21(140):1–67.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE.

Mokanarangan Thayaparan, Marco Valentino, and André Freitas. 2020. A survey on explainability in machine reading comprehension. *arXiv preprint arXiv:2010.00389*.

Mokanarangan Thayaparan, Marco Valentino, Peter Jansen, and Dmitry Ustalov. 2021. TextGraphs 2021 shared task on multi-hop inference for explanation regeneration. In *Proceedings of the Fifteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-15)*, pages 156–165, Mexico City, Mexico. Association for Computational Linguistics.

Leon Weber, Pasquale Minervini, Jannes Münchmeyer, Ulf Leser, and Tim Rocktäschel. 2019. NLProlog: Reasoning with weak unification for question answering in natural language. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6151–6161, Florence, Italy. Association for Computational Linguistics.

Sarah Wiegreffe and Ana Marasović. 2021. Teach me to explain: A review of datasets for explainable nlp. *arXiv preprint arXiv:2102.12060*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. 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.

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.

A Appendix

A.1 Data

A.1.1 WorldTree V2 explanation corpus

The WorldTree V2 explanation corpus (Xie et al., 2020) is a set of explanations to standardized science exam questions represented as sets of atomic facts. Each fact in an explanation is connected with either the question, answer, or other facts using lexical overlap (shared words). The average explanation contains 6 facts (range 1-16). The supporting semi-structured knowledge base contains approximately 9k facts distributed across 82 tables, where each table is organized around a particular kind of knowledge relation (e.g. *taxonomic*, *parts-of*, *generic properties*, *sources of things*, *causes*, *changes*, *if-then relationships*, *coupled relationships*, etc).

WorldTree V2 includes a subset of the questions from the Aristo Reasoning Challenge (ARC) corpus (Clark et al., 2018). ARC questions are standardized science exam questions drawn from 12 US states, where each question is a 4-choice multiple choice question. In total, WorldTree V2 contains 2210 training, 496 development, and 1670 test set questions.

A.2 Expert Relevance Ratings

A total of 236k expert judgements were collected for 126k facts. While most facts were rated by 2 annotators, due to scheduling conflicts with the COVID-19 pandemic, approximately 15% of questions are rated by a single teacher. 120 randomly sampled questions (approximately 3.5k facts) were used as training and calibration for all 3 annotators, who iteratively rated facts then discussed and resolved disagreements as a means of calibration before annotating the remainder of questions.

Exclusions: The WorldTree V2 knowledge base contains approximately 1.2k synonymy relations (e.g. *cooler* means *colder*, or *bike* means *bicycle*) that models tend to rate highly even though they have minimal conceptual content. We filtered these synonymy facts from the facts that the expert annotators rated before assembling the shortlists, to ensure that expert time was spent on rating the relevance of core scientific/world knowledge rather than thesaurus-like facts.

A.3 Evaluation Metrics

A.3.1 Ranking Metrics

Mean Average Precision (MAP) and Normalized Discounted Cumulative Gain (NDCG) follow their standard definitions as used in Table 4.

A.3.2 Whole-explanation Metrics

Relevance: Relevance represents the proportion of facts returned by a model that have a non-zero expert-rated relevance score for a given question. More specifically, relevance for a given explanation of length L is defined as:

$$Relevance = \frac{\sum_{i=1}^{L} R_i}{L} \qquad (2)$$

Where $TR(i, q)$ is an expert-annotated relevance rating for fact i for question q, the relevance score

of a given fact in the explanation, R_i, is defined as:

$$R_i = \begin{cases} 1, & \text{if } TR(i, q) \geq 1 \\ 0, & \text{otherwise} \end{cases} \qquad (3)$$

Note that facts without annotated relevance ratings are assumed to have a rating of 0 (irrelevant).

Completeness: Completeness represents the proportion of facts in the gold explanation for a given question that are also present in the model-generated explanation. Given a set of facts representing a gold explanation of length N, $G = \{G_1, G_2, ..., G_N\}$, and set of facts representing the model-generated explanation $M = \{M_1, M_2, ..., M_L\}$, the completeness of M is defined as:

$$Completeness = \frac{|G \cap M|}{|M|} \qquad (4)$$

The binary measure of completeness, Comp_B, is 1 if *Completeness* is 1, and 0 otherwise.

A.4 Additional Model Details and Hyperparameters

A.4.1 T5-UnifiedQA

UnifiedQA (Khashabi et al., 2020) is a variant of T5 (Raffel et al., 2020) that includes multi-task pre-training for 4 different forms of question answering across 20 datasets, including extractive QA (i.e., locating an answer span in a passage), abstractive QA (i.e., generating an answer not supplied in a passage), multiple-choice QA, and Boolean QA, while achieving state-of-the-art performance on 10 datasets. Here we train T5-UQA-3B to also generate compositional explanations by cueing generation with the question and correct answer candidate, and targeting generation to produce long strings of highly-rated facts delimited with an [AND] separator token.

Cueing: For training and evaluation, data was provided in the following format. Source (question) data was provided in the following format:

```
explanation:  <question
text> [ANSWER] <answer text>
[CUETOKEN]
```

Target (explanation) data was provided and generated in the following format:

```
<fact1> [AND] <fact2> [AND]
<fact3> [AND] ...  [EOS]
```

Where `factN` represents the sentence tokens for a given fact in the explanation (e.g. *"water is a kind of liquid"*). 5 permutations of fact orderings were used to discourage reliance on fact ordering and encourage robustness in model generations.

Pretraining: During a pretraining phase, T5-UQA was cued with the question and answer (as above), but provided with only a single fact from the gold explanation to generate.

Model parameters: All experiments were performed with the 3-billion parameter version of T5-UQA. We made use of DeepSpeed ZeRo optimizations (Rajbhandari et al., 2020) to fit the 3B model into the largest GPUs available to us (A100-40GB). Models were trained to 30 epochs, where generation performance (ROUGE-1) plateaued. We use the default hyperparameters for training provided in the Huggingface Transformers library (Wolf et al., 2020). To improve inference quality, at inference time we use a batch size of 1, a beam search over 64 beams, and (given the diversity of generations, and the preference for shorter generations even after considerable training) combine all facts generated in the top 10 beams (after splicing on the fact delimiter) into a candidate list of generated facts.

Model runtime: T5-UQA-3B took approximately 3 days to train on our dataset and another 2 days to evaluate, using 4x A100-40GB GPUs (i.e. approximately 20 A100 GPU days, equivalent to approximately 52 V100 GPU days).

Alignment to WorldTree knowledge base: To enable automatic evaluation and direct comparison with the other models, the output of T5-UQA was aligned to existing WorldTree facts. The output of the model was split on the `[AND]` delimiter, and each fact was exhaustively scored against all 9k facts in the WorldTree tablestore, where the fact with the highest ROUGE-1 (Lin and Hovy, 2003) alignment score was taken to be the appropriate WorldTree fact. We empirically determined that facts whose ROUGE-1 scores were lower than 0.70 tended to be poor alignments, most typically from misgenerations, incorrect generations, or nonsensical generations, though occasionally from correct generations that do not have a corresponding counterpart in the WorldTree knowledge base.

A.4.2 TFR-BERT

A Tensorflow-Ranking-BERT model (Han et al., 2020), which combines large language model embeddings with pointwise ranking (rather than classification) through the Tensorflow-Ranking framework (Pasumarthi et al., 2019b).

Cueing: During training and evaluation, the model was provided both the question and correct answer text. During training, it was provided with relevance rankings for the shortlist (approximately top-30) expert-rated facts. During evaluation, it was provided the top-100 ranked facts from the exhaustive BERT baseline (ranked by their classification scores), and re-ranked these facts.

Model parameters: Due to a large memory dependency (TFR-BERT GPU RAM scales with model parameter size *and* list size), we made use of BERT-base-uncased, a 110M parameter model. Default parameters were used for training and evaluation.

Model runtime: Training took approximately 2 days on a single A100-40GB GPU (multi-GPU training is not currently supported). To evaluate on comparatively large list sizes, evaluation was done using CPUs rather than GPUs, and took approximately 3 days using 32 CPU cores.

Top-K tuning: TFR-BERT results are reported both as a ranking model, where the output is a ranking of the entire knowledge base, as well in a whole-explanation paradigm where the output is a discrete *top-k-fact* explanation. Here we choose $k = 8$, where $F1^{ex}$ performance plateaued at 0.65 on the development set.

A.4.3 CoSaTa Schemas

A schema-based model implemented using the Constraint Satisfaction over Tables (COSATA) solver (Jansen, 2020). Schemas take the form of constraint satisfaction patterns over Worldtree facts represented as semi-structured table rows, where valid solutions of a given schema are only possible if all slots (facts) in a schema can be successfully populated by satisfying their constraints. We use the 385 science-domain schema included with COSATA, each containing an average of 12 facts, and run these over the WorldTree tablestore, generating a large set of 593k solutions that are pre-cached for speed.

Scoring: Scoring schema to create a shortlist of relevant patterns is challenging. Pilot experiments showed the simplest strategy had the best performance: For a given question, we score a given

schema solution by summing the BERT scores of the individual facts used in that solution, while clipping any fact scores below a threshold so as not to heavily penalize a given solution for a small number of irrelevant rows. We report scores for systems that include a single schema solution, or that combine 2 or 3 top-scoring schemas. Reported performance is for a model that post-filters schemas before giving them to the user. Specifically, any facts in a schema with scores below a threshold are filtered,[4] generally significantly increasing relevance while slightly decreasing completeness. *Negative results:* When scoring schema solutions, pilot experiments showed a variety of different scoring methodologies based on top-K scoring, or cued scoring from other methods (e.g. T5, TFR-BERT) performed worse than a simple thresholded sum. We hypothesize TFR-BERT ranking scores performed worse than BERT classification scores due to a comparative robustness of combining multiple classification scores into a single score, versus combining multiple ranks for individual facts.

Model parameters: Scoring for CoSaTa schemas uses the BERT-base-uncased (110M) model. Other hyperparameters reported above.

Model runtime: Initial schema generation and caching took approximately 2 days on 16 CPU cores. Subsequent ranking, scoring, and filtering took approximately 2 hours.

A.4.4 BERT and RoBERTa Baselines

The BERT and RoBERTa exhaustive baselines are trained as classifiers, using a distant supervision paradigm where (for a given question) all gold explanation facts are taken as gold, and N randomly sampled facts from the knowledge base are taken as negative examples. Unlike Das et al. (Das et al., 2019), who train an exhaustive model using all non-gold facts in the knowledge base, here we sub-sample only a subset of the corpus to minimize the likelihood that the model would use a relevant fact not annotated as in the gold explanation for a given question as a negative example.

Model Parameters: We used a vanilla BERT-base-uncased (110M), as well as RoBERTA-Large (355M) pre-trained on the RACE reading comprehension benchmark (Lai et al., 2017).

---

[4]Here, both clipping and filtering thresholds were set at zero.