On Evaluating Explanation Utility for Human-AI Decision-Making in NLP

Fateme Hashemi Chaleshtori **Atreya Ghosal** **Ana Marasović**
Kahlert School of Computing
University of Utah
{fateme.hashemi,atreya.ghosal,ana.marasovic}@utah.edu

Abstract

Is explainability a false promise? This debate has emerged from the lack of consistent evidence that explanations help in situations they are introduced for. In NLP, the evidence is not only inconsistent but also scarce. While there is a clear need for more human-centered, application-grounded evaluations, it is less clear where NLP researchers should begin if they want to conduct them. To address this, we introduce evaluation guidelines established through an extensive review and meta-analysis of related work.

1 Introduction

Decision-makers can make use of imperfect AI models if they can detect when these models are correct. Explanations of individual predictions are proposed to this end as they are expected to reveal useful signals about the model's reasoning process (Jacovi et al., 2021). Before undertaking realistic evaluations involving people, NLP researchers aspired to first implement working methods. Thus, prior NLP explainability work has mostly focused on overcoming technical challenges and used proxy evaluations. Consequently, human-centered evaluations of explanations grounded in real NLP applications are scarce (see Table 1). There is a prevailing perspective that this now needs to change since explainability methods passed proof-of-concept tests. However, given that this is a nascent NLP research space and the notable variation among prior studies evident from Table 1, determining an experimental setup can be challenging. This paper aims to alleviate this difficulty by providing guidelines.

An existing resource for the development and evaluation of explanations in NLP already includes over 50 datasets. Can these be used for application-grounded evaluations of explanations? To address this, in §3, we establish criteria to assess each dataset's suitability in computing explanation usefulness with measurements overviewed in §2. We discover that 17/51 datasets are apt for studying appropriate reliance and complementary human-AI team performance, but only involve low risk. 4/51 are suitable for these measurements, involve higher risk, and do not have quality concerns. We recommend prioritizing these 4 tasks, as high stakes necessitate proper explanations more.

The model performance on these datasets should not be reaching the upper bound. If so, the chance of hazards, and hence, the risk, is low. There is also no room to improve complementary performance, and the issue of overreliance becomes irrelevant. In §4, we demonstrate the importance of reassessing backbone model performance in a rapidly evolving field like NLP.

Finally, we review the experimental design for human subject evaluations from previous studies. We distill prevailing trends and noteworthy deviations, as well as refine the protocol proposed by Schemmer et al. (2023). This refined protocol better isolates the impact of explanations on human reliance and human-AI team performance. Future research could employ our proposed

XAI in Action: Past, Present, and Future Applications @ NeurIPS 2023.

	Task	Usefulness Measurements	Explanations Evaluated	Models Explained	Baseline Expl.	Helpful
W1	D1; D28	Human accuracy (human init. wrong, no AI advice, with explanation)	Free-text explanations	T5-large (FT-full); T5-3B (FT-128); davinci-instruct-beta (ICL-6)	None	No
W2	D2	Reliance	Manually extracted evidence independent of the model†	DPR (Karpukhin et al., 2020)	Post-hoc calibrated model confidence	Yes
W3	D3; D4	Complementary team performance	LIME input attribution of top-1 or 2 predictions or *human* free-text explanations	RoBERTA-Base (FT)	Post-hoc calibrated model confidence	No
W4	D5	Regression analysis est. how each condition influences player accuracy	Manually extracted evidence‡	TF-IDF to find & return the label of the most similar doc or previously seen question	Similarity score between a question & a retrieved doc	Yes
W5	D6	Complementary team performance**	Input attribution from SVM's weights; k-NN train examples	Linear SVM with BoW features	Providing the accuracy of the SVM	Yes
W6	D6	Reliance	LIME input attribution	SVM	None	Yes

Table 1: Overview of prior application-grounded explanation usefulness evaluations in NLP. W1 (Joshi et al., 2023); W2 (González et al., 2021); W3 (Bansal et al., 2021); W4 (Feng and Boyd-Graber, 2019); W5 (Lai and Tan, 2019); W6 (Schemmer et al., 2023). FT stands for "finetuned" and ICL for "in-context learning". **They omit $\mathrm{acc}(y_p|x)$ from Eq (5) in App. A.

approach to assess the usefulness of different explanations on human reliance on models we trained, as well as complementary performance, on the tasks spotlighted by our meta-analysis.

2 Background: Application-Grounded Explanation Evaluation

Using Doshi-Velez and Kim (2017)'s taxonomy, an evaluation of explanations can be (1) *proxy* (no humans, proxy tasks; e.g., the level of sparsity), (2) *human-grounded* (with humans, simplified tasks; e.g., simulatability), or (3) *application-grounded* (with humans, realistic tasks; e.g., human-AI decision making). The third category is the focus of this paper. Forward and counterfactual simulatability (Xie et al., 2022; Arora et al., 2022) are human-grounded, but not application-grounded: Buçinca et al. (2020) show that the effects of explanations on simulatability differ from their effects on human-AI decision-making within the same experimental setup. Liao et al. (2022) outline six usage contexts within explainable AI, one of which is decision making which is the focus of our work.

Application-grounded evaluations of explanations have predominantly been executed for applications with interpretable features such as people's age or income (Liao and Varshney, 2022). Explaining tasks that involve text has unique challenges: features are a sequence of high-dimensional non-interpretable vectors; an arbitrary number of features; continuous representations of discrete inputs; explaining models with billions of parameters; pretrained models; and inherently interpretable models (e.g., linear models, short decision trees) performing nowhere close to large language models. Prior NLP explainability work has mostly focused on overcoming these challenges. Moreover, numerous realistic applications of language technology have become evident and possible only with recent advances. Thus, to conclusively establish — or disprove — the value of explanations for human-AI decision-making in NLP, more research is needed together with a more meticulous evaluation protocol designed to collectively guide us towards settling this matter. To this end, we start with an overview of explanation usefulness measurements. We provide the equations for calculating each measurement in Appendix A.

Reliance. It is often asserted that explanations can deter people from rejecting correct predictions, i.e., **underreliance**. This expectation stems from assuming that the model is correct for the right reasons, and explanations are anticipated to unveil this. To measure the usefulness of explanations in mitigating underreliance, it has been proposed to compare the average rate at which people reject correct predictions with and without the provision of explanations (Wang and Yin, 2021). If explanations are helpful this rate should lower with showing them. Explanations could also aid people in rejecting incorrect predictions, thereby countering **overreliance** (Vasconcelos et al., 2023). This becomes possible when explanations present information that appears illogical, self-contradictory, or inconsistent with what the person already knows. Explanations are useful if the average rate at which people accept incorrect predictions lowers with the provision of explanations. The ultimate goal is **appropriate reliance**—have people accept correct predictions and dismiss erroneous ones (Wang and Yin, 2021). A gain in the average rate at which people do so upon seeing explanations quantifies their usefulness. While some require annotators to only detect model errors (Wang and Yin, 2021; González et al., 2021), others require that the person provides the final label (Schemmer et al., 2023; Joshi et al., 2023). Schemmer et al. (2023) urge to first ask participants to guess the answer before showing additional information and propose reporting the fraction of times a person (1) flips their initial, wrong judgment of model correctness after seeing a correct model prediction, and (2) sticks with their initial, correct judgment after seeing a wrong model prediction. Joshi et al. (2023)'s measurement is similar to (1) but a person needs to flip their initial, wrong answer to the correct answer upon seeing AI's explanation but not its prediction.[1] Another related measurement is the switch percentage (Zhang et al., 2020b). Fok and Weld (2023) define a desired reliance behavior based on the expected performance of a person and a model. Specifically, they state that it is acceptable if people always accept predictions of a "super-human" model (or always reject predictions made by a "sub-human" model), even if in certain cases predictions are not correct (wrong).

Complementary Performance. It is argued that explanations can boost human-AI team performance for the same reasons they can support reliance. For this to even be possible, the state-of-the-art model or people alone should not already reach the upper bound on performance. Instead of measuring the difference in accuracy, Feng and Boyd-Graber (2019) perform a regression analysis that determines the influence of AI's advice with explanations on players' accuracy. When measuring reliance or complementary performance, it is common to also ask annotators to self-report their **confidence** in decisions they made and **trust** in the AI model on a case-by-case basis or as a post-task survey.

3 Analysis of Task Appropriateness

In this section, we present criteria that can be used to determine the suitability of tasks for application-grounded human evaluations of explanations (§3.1) and analyze 51 existing datasets introduced for developing and evaluating explanations in NLP (§3.2). We refer to a task as a realization of that task in the data.

3.1 Task Criteria

We determine that the following criteria must be fulfilled to ensure that evaluations are rooted in genuine human-AI interactions:

c_1: The task has a meaningful connection to a real-world application, involving people who seek model outputs and act on them.
c_2: The dataset inputs must be realistic.
c_3: Handling task instances requires a notable effort from people, or people are bad at it.

For example, CommonsenseQA (Talmor et al., 2019) has no associated application as people do not need answers to questions such as "At the end of your meal what will a waiter do? serve food, eat, set table, serve meal, or present bill". PubHealth (Kotonya and Toni, 2020) has actionable outputs but lacks realistic task inputs. The task is to verify a claim based on a professional fact-checking report on the same claim that won't be available for a new claim a model gets post-deployment. Finally, while there might be a use for sentiment classification of laptop reviews (Pontiki et al., 2014), their brief average length of only 15 words allows people to correctly and confidently gauge sentiment without assistance. Hence, concerns about under- or overreliance do not arise in this context because people never end up really relying on anything.

[1] Joshi et al. (2023) also study whether automated model explanations support the human ability to reason about new situations where the same logic applies, like human-authored explanations (Blanchard et al., 2018; Vasilyeva and Lombrozo, 2022). Specifically, they measure the accuracy of people who incorrectly answered questions that require the same reasoning as a control question (not just a paraphrase), and are asked to answer again after seeing an explanation of the control question. Their procedure cannot be applied post-deployment, as control questions are not accessible for new inquiries, and is thus not application-grounded, although it is human-grounded.

The three criteria above are sufficient if our sole focus is on reliance and complementary performance. However, the definition of human trust in AI (Jacovi et al., 2021) implies that we cannot talk about trust with no risk involved, as one cannot accept vulnerability when none exists. Thus, studying human trust in AI demands an extra condition:

c_4: There is some undesirable event that can possibly (but not certainly) occur when collaborating with models for the task.

Although risk is not pivotal to defining studies of reliance and human-AI teams well, **we urge giving precedence to tasks involving higher risk because under- and overreliance have more pronounced consequences for them.** It is more valuable to develop explanations that boost appropriate reliance for them, and this is how the need for explanations is often motivated.

3.2 Categorization of ExNLP Tasks

We analyze 51 datasets that are reported on the website that collects datasets for explainable NLP (Wiegreffe and Marasović, 2021) according to how they satisfy the criteria in §3.1. [2] In Appendix E, we report details of our decisions for each task and provide an overview in Table 2. We use ■ if a benchmark criterion is satisfied, and □ otherwise. A suitable dataset for application-grounded human-subject evaluations of explanations should have an application (c_1) and realistic inputs (c_2) as well as either require notable effort, or be a difficult task for people (c_3), and ideally more than low levels of risk (c_4). We mark tasks that satisfy $c_{\{1,2,3\}}$, i.e., those suitable for studying reliance with 👍 and those that satisfy all criteria and that should be prioritized with 🌟.

Are ExNLP tasks connected to real-world applications beyond debugging? We first determine that we can imagine people using the outputs of a model trained on dataset instances. E.g., sentiment predictions of reviews can be used to decide whether to make a purchase. We then assess that task instances resemble what models can realistically access to make their predictions in the future (unlike the fact-checking example in §3.1). If both of these two conditions are met, we deem that a task is connected to real-world application, and not otherwise. We find that 29/51 (56.9%) datasets have an associated application as well as realistic inputs, i.e., fulfill the central requirement for *application*-grounded evaluations, but 22/51 (43.1%) do not.

Do ExNLP tasks require notable human effort? Are people skilled at solving these tasks? We estimate effort using the average length of task inputs, anticipating that longer inputs demand more effort. The maximum average length that we decide does not require notable effort is 272 words, taking approximately a minute to read (Rayner et al., 2016). Future work could check whether datasets with short examples still require notable cognitive load (e.g., math problems). We estimate human ability with the reported human performance if

		$c_1 \land c_2$	c_3	c_4	
W1	D1	■	■	□	👍
W2	D2	■	■	□	👍
W3	D3	□	■	-	
	D4	■	■	□	👍
W4	D5	■	■	□	👍
W5,W6	D6	■	■	□	👍
	D7	■	■	□	👍
	D8	■	□	□	
	D9	□	■	-	
	D10	□	■	-	
	D11	□	-	-	
	D12	■	-	-	
	D13	□	□	-	
	D14	□	□	-	
	D15	□	■	-	
	D16	□	■	-	
	D17	■	■	■	🌟*👍
	D18	■	■	■	🌟*👍
	D19	□	□	-	
	D20	□	□	-	
	D21	□	□	-	
	D22	■	■	□	👍
	D23	■	-	-	
	D24a	■	□	□	
	D24b	■	□	□	
	D25	□	□	-	
	D26	□	□	-	
	D27	□	□	-	
W1	D28	□	□	-	
	D29	■	■	□	👍
	D30	□	□	-	
	D31	■	■	□	👍
	D32	■	■	□	👍
	D33	□	□	-	
	D34	■	■	□	👍
	D35	■	□	□	
	D36	■	□	□	
	D37	■	□	□	
	D38	■	■	■	🌟👍
	D39	□	□	-	
	D40	■	□	□	
	D41	■	-	-	
	D42	□	■	-	
	D43	□	□	-	
	D44	■	■	■	🌟👍
	D45	□	-	-	
	D46	□	■	-	
	D47	■	-	□	
	D48	■	-	-	
	D49	■	■	■	🌟👍
	D50	■	■	■	🌟👍

Table 2: Appropriateness of ExNLP datasets. See §3.2 for more.

[2] https://exnlpdatasets.github.io/

available. We find that 24/51 (47%) tasks either require notable effort or people do not excel at it, the data is not available for 2/51 (3.9%), we are not able to estimate the human ability for 5/51 (9.8%), and for 20/51 (39.21%) inputs are too short while people do the task well. Of 24 requiring notable effort or people are not good at them, 17/51 (33.3%) also have associated applications and realistic inputs. That is, 33.3% of ExNLP datasets are suitable for studying appropriate reliance and complementary team performance.

Are ExNLP tasks associated with high-risk situations? Motivated by Suresh et al. (2021), we approach answering this question from the perspective of two stakeholder types: (i) people who act on the model output (e.g., doctors) and (ii) decision subjects (e.g., patients). We first determine possible hazards. We decide what a hazard's level of risk is — low, moderate, or high — based on its severity and likelihood. We estimate the likelihood based on the performance of the state-of-the-art model, expecting that the higher the performance is, the lower the likelihood. We subjectively determine their worst-case severity. We find that among the 17 remaining datasets, only 6 cause hazards that are not benign. Upon manual inspection of examples of this data, we discovered problems with D17 and D18 (see Appendix B for more information). Thus we exclude them and recommend prioritizing 4 datasets for application-grounded evaluation of explanations in NLP: EvidenceInference v2 *with document retrieval* (D38), SciFact-Open (D44), ContractNLI (D49), and Indian Legal Documents Corpus (D50).

4 Analysis of Model Appropriateness

We check whether we can train models for 🌟 tasks that fulfill all the criteria in §3.1 and quality checks, that are already on par with the estimated upper bound: performance of domain experts working without time constraints. If a model rarely makes mistakes, a hazard's likelihood is low, so the risk is not as high. These are not the tasks we recommend prioritizing. Moreover, Fok and Weld (2023) argue that a viable strategy, in this case, is to always use the model's predictions, so only underreliance is of interest. We develop a baseline model for EvidenceInference v2 *with document retrieval* (D38), SciFact-Open (D44), ContractNLI (D49), and Indian Legal Documents Corpus (ILDC; D50).

As a backbone model, we use Flan-T5-3B (Wei et al., 2022) due to its larger size and versatility stemming from instruction finetuning with data of 1.8K tasks. For each task, we finetune it with all task training data; details are given in Appendix D. In the same Appendix, we provide examples of task input that are given to the model (Tables 7–10). While larger variants might be even better, unlike them, finetuning a 3B model with long inputs (4.2K tokens) can be done on a single GPU. Larger variants may perform better, but, unlike them, a 3B model with long inputs (4.2K tokens) can be finetuned using a single GPU. However, quantized LLMs are altering this landscape.

Table 3 underscores the importance of reassessing baselines. We obtain a 22.8 macro-F1 point increase on SciFact-Open and a 24.5 point improvement in the average contradiction and entailment F1 scores on Contract-NLI. In the original EvidenceInference task setup, where models are provided with relevant documents and do not require retrieval, we obtrain a 36.7-point macro-F1 improvement (see Table 6 in the Appendix). However, Flan-T5-3B only slightly improves ILDC's state-of-the-art results, despite being a substantially larger, and more pretrained, model. This implies that we cannot assume improvements similar to the former two datasets when incorporating recent advances and dismiss a dataset on that basis: the baseline results must be computed. We find that the baselines for ILDC and SciFact-Open — the two datasets with the reported human performance — have not reached peak performance. Yet, the results on SciFact-Open, and Contract-NLI, are more promising than once believed, suggesting they could assist laypeople, or experts working under time constraints.

5 Discussion: Study Procedure

Almost all prior studies (Table 1) involve a condition where participants receive a model prediction, but they are not consistent with the display of model confidence. Bansal et al. (2021) note that "predictions with confidence" is a simple, yet stronger baseline compared to predictions only.

	P	R	F1
Flan-T5-3B (our)			
Accept	81.7	72.6	76.9
Reject	75.1	83.6	79.1
Micro Avg.	78.4	78.0	78.0
Macro Avg.	78.4	78.1	**78.0**
XLNet + BiGRU			
Macro Avg.	76.8	76.3	76.5
Human Est. Accuracy			93.9

(a) ILDC

	P	R	F1
Flan-T5-3B (our)			
Increase	52.7	64.4	58.8
No Diff	54.7	29.2	38.1
Decrease	41.1	59.1	48.5
Micro Avg.	50.7	49.1	47.6
Macro Avg.	49.5	51.6	48.5
NN + pretrain cond. attn.			
Macro Avg.	53.0	51.8	52.1

(b) EvidenceInference v2 with retrieval

	P	R	F1
Flan-T5-3B (our)			
Support	79.0	79.0	79.0
No Info	71.8	70.5	71.2
Contradict	74.2	76.6	75.4
Micro Avg.	75.3	75.3	75.3
Macro Avg.	75.0	75.4	**75.2**
MultiVerS			
Macro Avg.	73.6	40.7	52.4
Human Est.	94.8	84.1	89.1

(c) SciFact-Open

	P	R	F1
Flan-T5-3B (our)			
Entail	92.5	93.7	93.1
No Mention	93.0	87.0	89.9
Contradict	68.7	82.7	75.0
Micro Avg.	90.2	89.7	89.8
Macro Avg.	84.7	87.7	86.0
Macro (E,C)	80.6	88.2	**84.1**
BERT-Large			
Entail	-	-	83.4
No mention (*not reported*)			
Contradict	-	-	35.7
Macro (E,C)	-	-	59.6

(d) ContractNLI

Table 3: Finetuned Flan-T5-3B and the state-of-the-art reported results. XLNet+BiGRU (Malik et al., 2021); Neural network + Pretrain conditional attention (Lehman et al., 2019); MultiVerS (Wadden et al., 2022b); BERT-Large (Koreeda and Manning, 2021). Wadden et al. (2022b) estimate the human performance in the "abstract-provided" setting.

For a more nuanced analysis, Schemmer et al. (2023) propose a sequential decision process. Participants first make a guess unassisted, then reevaluate upon viewing the model's prediction and explanation. They also report their self-confidence. This robust protocol can be further improved (besides displaying model confidence) by dividing the second step as follows: reveal the prediction, have the participant reassess, provide the explanation, and ask for the final decision. This approach better isolates the effects of explanations; e.g., if participants switch to wrong AI predictions despite making correct guesses initially, this might be because they are blindly following the AI's advice while ignoring the explanation. If these participants persist with the wrong AI prediction but their self-confidence lowers upon receiving explanations, it suggests that explanations may be discouraging overreliance. Our breakdown of all possibilities is in Table 4.

Experts are the targeted audience of applications associated with tasks identified by our meta-analysis, as task instances are too specialized (see Tables 7–10). These tasks require notable effort, but experts are skilled (e.g., ILDC experts' average accuracy is 94%). Thus, if participants make the initial guess without time constraints, human-AI teams likely won't outdo experts alone. Therefore, unlike almost all prior studies that involve only laypeople, application-grounded evaluation with our highlighted tasks should focus on time-constrained experts. Evaluations with experts are more expensive, so the number of instances and participants must also be reevaluated. When measuring complementary team performance, it is essential to estimate human time-constrained performance rather than unconstrained, as team performance must be better than the former but not the latter.

$y_h^{(1)}$	y_p	$y_h^{(2)}$ (after $y_h^{(1)} \wedge y_p \wedge c_p$)		$y_h^{(3)}$ (after $y_h^{(2)} \wedge e_p$)	
✓	✓	✓	Confirmation	✓	Effects of e_p undetermined but also not interesting
				✗	Unlikely (Spammers?)
		✗	Unlikely (Spammers?)		(Don't give 3rd chance)
					(Don't give 3rd chance)
	✗	✓	Correct Self-Reliance (CSR)	✓	e_p could be reinforcing CSR (good), doing nothing, or deterring from CSR
				✗	e_p causing OR
		✗	Overreliance (OR)	✓	e_p fixing OR
				✗	e_p could be reinforcing OR (bad), doing nothing, or deterring from OR (good)
✗	✓	✓	Correct Reliance (CR)	✓	e_p could be reinforcing CR (good), doing nothing, or deterring from CR (bad)
				✗	e_p causing UR
		✗	Underreliance (UR)	✓	e_p fixing UR
				✗	e_p could be reinforcing UR (bad), doing nothing, or deterring from UR (good)
	✗	✓	Unlikely (Spammers?)		(Don't give 3rd chance)
					(Don't give 3rd chance)
		✗	Confirmation	✓	Unlikely (Spammers?)
				✗	Effects of e_p undetermined but also interesting

Table 4: Our extension of Schemmer et al. (2023)'s study. Show a prediction, y_p, and confidence, c_p, and only then an explanation, e_p. $y_h^{(1)}$ is a human's initial guess, $y_h^{(2)}$ is the 2nd guess upon seeing y_p, and $y_h^{(3)}$ is the 3rd guess after seeing e_p. ✓ (✗) is the correct (wrong) guess.

Finally, Table 1 highlights that prior studies primarily concentrate on one or two types of explanations, with input highlights being a common choice. Yet, explainable AI has more to offer (Madsen et al., 2023). Liao et al. (2022) stress that "explainability is not a monolithic concept, and what users need to be explained varies across different types of systems and user tasks". This paper identifies NLP tasks suitable for developing explanations, but we do not suggest that one type of explanation should fit all tasks, nor should explanations be developed without considering specific tasks.

These insights, coupled with the models we trained, can be put together to analyze various explanations for tasks spotlighted by our meta-analysis.

Acknowledgments

We thank Dana EeJae Ahn for her help in calculating the average length of instances in datasets we examine, Jason Wiese and Kyle Lo for helpful discussions on the study procedure, Kyle Lo for his feedback on the framing of this work, UtahNLP group and anonymous reviewers for their feedback, everyone who responded to our requests to share the data, and Q. Vera Liao and Alison Smith-Renner whose presentation at the NAACL'22 tutorial on "Human-centered Evaluations of Explanations" helped us write the background section.

References

Shourya Aggarwal, Divyanshu Mandowara, Vishwajeet Agrawal, Dinesh Khandelwal, Parag Singla, and Dinesh Garg. 2021. Explanations for CommonsenseQA: New Dataset and Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3050–3065, Online. Association for Computational Linguistics.

Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 85–90, Brussels, Belgium. Association for Computational Linguistics.

Siddhant Arora, Danish Pruthi, Norman M. Sadeh, William W. Cohen, Zachary C. Lipton, and Graham Neubig. 2022. Explain, edit, and understand: Rethinking user study design for evaluating model explanations. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI*

2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022, pages 5277–5285. AAAI Press.

David Atkinson, Kumar Bhargav Srinivasan, and Chenhao Tan. 2019. What gets echoed? understanding the “pointers” in explanations of persuasive arguments. 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 2911–2921, Hong Kong, China. Association for Computational Linguistics.

Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21. Association for Computing Machinery.

David Bender. 2015. Establishing a human baseline for the winograd schema challenge. In *Proceedings of the 26th Modern AI and Cognitive Science Conference*, pages 39–45, Greensboro, NC, USA. CEUR Workshop Proceedings.

Thomas Blanchard, Nadya Vasilyeva, and Tania Lombrozo. 2018. Stability, breadth and guidance. *Philosophical Studies*, 175:2263–2283.

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.

Faeze Brahman, Vered Shwartz, Rachel Rudinger, and Yejin Choi. 2021. Learning to rationalize for nonmonotonic reasoning with distant supervision. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 12592–12601. AAAI Press.

Ana Brassard, Benjamin Heinzerling, Pride Kavumba, and Kentaro Inui. 2022. COPA-SSE: Semi-structured explanations for commonsense reasoning. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3994–4000, Marseille, France. European Language Resources Association.

Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, IUI '20, page 454–464, New York, NY, USA. Association for Computing Machinery.

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*, volume 31. Curran Associates, Inc.

Samuel Carton, Qiaozhu Mei, and Paul Resnick. 2018. Extractive adversarial networks: High-recall explanations for identifying personal attacks in social media posts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3497–3507, Brussels, Belgium. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, Dimitrios Tsarapatsanis, Nikolaos Aletras, Ion Androutsopoulos, and Prodromos Malakasiotis. 2021. Paragraph-level rationale extraction through regularization: A case study on European court of human rights cases. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 226–241, Online. Association for Computational Linguistics.

George Chrysostomou and Nikolaos Aletras. 2022. Flexible instance-specific rationalization of NLP models. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 10545–10553. AAAI Press.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. 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 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Mitchell DeHaven and Stephen Scott. 2023. Bevers: A general, simple, and performant framework for automatic fact verification.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020a. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Jay DeYoung, Eric Lehman, Benjamin Nye, Iain Marshall, and Byron C. Wallace. 2020b. Evidence inference 2.0: More data, better models. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 123–132, Online. Association for Computational Linguistics.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning.

Shi Feng and Jordan L. Boyd-Graber. 2019. What can AI do for me?: evaluating machine learning interpretations in cooperative play. In *Proceedings of the 24th International Conference on Intelligent User Interfaces, IUI 2019, Marina del Ray, CA, USA, March 17-20, 2019*, pages 229–239. ACM.

Raymond Fok and Daniel S. Weld. 2023. In search of verifiability: Explanations rarely enable complementary performance in ai-advised decision making.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Ana Valeria González, Gagan Bansal, Angela Fan, Yashar Mehdad, Robin Jia, and Srinivasan Iyer. 2021. Do explanations help users detect errors in open-domain QA? an evaluation of spoken vs. visual explanations. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1103–1116, Online. Association for Computational Linguistics.

Braden Hancock, Paroma Varma, Stephanie Wang, Martin Bringmann, Percy Liang, and Christopher Ré. 2018. Training classifiers with natural language explanations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1884–1895, Melbourne, Australia. Association for Computational Linguistics.

Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. 2019. A richly annotated corpus for different tasks in automated fact-checking. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 493–503, Hong Kong, China. Association for Computational Linguistics.

Shirley Anugrah Hayati, Dongyeop Kang, and Lyle Ungar. 2021. Does BERT learn as humans perceive? understanding linguistic styles through lexica. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6323–6331, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ruining He and Julian J. McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016*, pages 507–517. ACM.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2009. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, pages 94–99, Boulder, Colorado. Association for Computational Linguistics.

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.

Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing trust in Artificial Intelligence: Prerequisites, causes and goals of human trust in ai. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, page 624–635. Association for Computing Machinery.

Peter Jansen, Niranjan Balasubramanian, Mihai Surdeanu, and Peter Clark. 2016. What's in an explanation? characterizing knowledge and inference requirements for elementary science exams. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2956–2965, Osaka, Japan. The COLING 2016 Organizing Committee.

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.

Brihi Joshi, Ziyi Liu, Sahana Ramnath, Aaron Chan, Zhewei Tong, Shaoliang Nie, Qifan Wang, Yejin Choi, and Xiang Ren. 2023. Are machine rationales (not) useful to humans? measuring and improving human utility of free-text rationales. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7103–7128, Toronto, Canada. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, 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.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. QASC: A dataset for question answering via sentence composition. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8082–8090. AAAI Press.

Yuta Koreeda and Christopher Manning. 2021. ContractNLI: A dataset for document-level natural language inference for contracts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1907–1919, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754, Online. Association for Computational Linguistics.

Mücahid Kutlu, Tyler McDonnell, Tamer Elsayed, and Matthew Lease. 2020. Annotator rationales for labeling tasks in crowdsourcing. *J. Artif. Intell. Res.*, 69:143–189.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Vivian Lai and Chenhao Tan. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 29–38, New York, NY, USA. Association for Computing Machinery.

Matthew Lamm, Jennimaria Palomaki, Chris Alberti, Daniel Andor, Eunsol Choi, Livio Baldini Soares, and Michael Collins. 2021. QED: A framework and dataset for explanations in question answering. *Transactions of the Association for Computational Linguistics*, 9:790–806.

Eric Lehman, Jay DeYoung, Regina Barzilay, and Byron C. Wallace. 2019. Inferring which medical treatments work from reports of clinical trials. 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 3705–3717, Minneapolis, Minnesota. Association for Computational Linguistics.

Q. Vera Liao and Kush R. Varshney. 2022. Human-centered explainable ai (xai): From algorithms to user experiences.

Q. Vera Liao, Yunfeng Zhang, Ronny Luss, Finale Doshi-Velez, and Amit Dhurandhar. 2022. Connecting algorithmic research and usage contexts: A perspective of contextualized evaluation for explainable ai. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 10(1):147–159.

Bill Yuchen Lin, Dong-Ho Lee, Ming Shen, Ryan Moreno, Xiao Huang, Prashant Shiralkar, and Xiang Ren. 2020. TriggerNER: Learning with entity triggers as explanations for named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8503–8511, Online. Association for Computational Linguistics.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 158–167, Vancouver, Canada. Association for Computational Linguistics.

Andreas Madsen, Siva Reddy, and Sarath Chandar. 2023. Post-hoc interpretability for neural NLP: A survey. *ACM Comput. Surv.*, 55(8):155:1–155:42.

Vijit Malik, Rishabh Sanjay, Shubham Kumar Nigam, Kripabandhu Ghosh, Shouvik Kumar Guha, Arnab Bhattacharya, and Ashutosh Modi. 2021. ILDC for CJPE: Indian legal documents corpus for court judgment prediction and explanation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4046–4062, Online. Association for Computational Linguistics.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14867–14875. AAAI Press.

Julian J. McAuley, Jure Leskovec, and Dan Jurafsky. 2012. Learning attitudes and attributes from multi-aspect reviews. In *12th IEEE International Conference on Data Mining, ICDM 2012, Brussels, Belgium, December 10-13, 2012*, pages 1020–1025. IEEE Computer Society.

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, Brussels, Belgium. Association for Computational Linguistics.

Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document ranking with a pretrained sequence-to-sequence model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 708–718, Online. Association for Computational Linguistics.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 309–319, Portland, Oregon, USA. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

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.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Keith Rayner, Elizabeth R Schotter, Michael E J Masson, Mary Potter, and Rebecca Treiman. 2016. So much to read, so little time. *Psychological Science in the Public Interest*, 17:34 – 4.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Pedro Rodriguez, Shi Feng, Mohit Iyyer, He He, and Jordan L. Boyd-Graber. 2019. Quizbowl: The case for incremental question answering. *CoRR*, abs/1904.04792.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *Logical Formalizations of Commonsense Reasoning, Papers from the 2011 AAAI Spring Symposium, Technical Report SS-11-06, Stanford, California, USA, March 21-23, 2011*. AAAI.

Rachel Rudinger, Vered Shwartz, Jena D. Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A. Smith, and Yejin Choi. 2020. Thinking like a skeptic: Defeasible inference in natural language. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4661–4675, Online. Association for Computational Linguistics.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Max Schemmer, Niklas Kühl, Carina Benz, Andrea Bartos, and Gerhard Satzger. 2023. Appropriate reliance on AI advice: Conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces, IUI 2023, Sydney, NSW, Australia, March 27-31, 2023*, pages 410–422. ACM.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Shashank Srivastava, Igor Labutov, and Tom Mitchell. 2017. Joint concept learning and semantic parsing from natural language explanations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1527–1536, Copenhagen, Denmark. Association for Computational Linguistics.

Harini Suresh, Steven R. Gomez, Kevin K. Nam, and Arvind Satyanarayan. 2021. Beyond expertise and roles: A framework to characterize the stakeholders of interpretable machine learning and their needs. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. Association for Computing Machinery.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. 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 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. 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 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Andrew Trotman, Antti Puurula, and Blake Burgess. 2014. Improvements to BM25 and language models examined. In *Proceedings of the 2014 Australasian Document Computing Symposium, ADCS 2014, Melbourne, VIC, Australia, November 27-28, 2014*, page 58. ACM.

Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S. Bernstein, and Ranjay Krishna. 2023. Explanations can reduce overreliance on AI systems during decision-making. *Proc. ACM Hum. Comput. Interact.*, 7(CSCW1):1–38.

Nadya Vasilyeva and Tania Lombrozo. 2022. Explanations and causal judgments are differentially sensitive to covariation and mechanism information. *Frontiers in Psychology*, 13.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. 2022a. SciFact-open: Towards open-domain scientific claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4719–4734, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

David Wadden, Kyle Lo, Lucy Lu Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. 2022b. MultiVerS: Improving scientific claim verification with weak supervision and full-document context. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 61–76, Seattle, United States. Association for Computational Linguistics.

Cunxiang Wang, Shuailong Liang, Yue Zhang, Xiaonan Li, and Tian Gao. 2019. Does it make sense? and why? a pilot study for sense making and explanation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4020–4026, Florence, Italy. Association for Computational Linguistics.

Xinru Wang and Ming Yin. 2021. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *26th International Conference on Intelligent User Interfaces*, IUI '21, page 318–328, New York, NY, USA. Association for Computing Machinery.

Ziqi Wang, Yujia Qin, Wenxuan Zhou, Jun Yan, Qinyuan Ye, Leonardo Neves, Zhiyuan Liu, and Xiang Ren. 2020. Learning from explanations with neural execution tree. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Sarah Wiegreffe and Ana Marasović. 2021. Teach me to explain: A review of datasets for explainable natural language processing. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*.

Kaige Xie, Sarah Wiegreffe, and Mark Riedl. 2022. Calibrating trust of multi-hop question answering systems with decompositional probes. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2888–2902, Abu Dhabi, United Arab Emirates. 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 Twelfth Language Resources and Evaluation Conference*, pages 5456–5473, Marseille, France. European Language Resources Association.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018, Lisbon, Portugal. Association for Computational Linguistics.

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.

Zhiwei Yang, Jing Ma, Hechang Chen, Hongzhan Lin, Ziyang Luo, and Yi Chang. 2022. A coarse-to-fine cascaded evidence-distillation neural network for explainable fake news detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2608–2621, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Qinyuan Ye, Xiao Huang, Elizabeth Boschee, and Xiang Ren. 2020. Teaching machine comprehension with compositional explanations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1599–1615, Online. Association for Computational Linguistics.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. Reclor: A reading comprehension dataset requiring logical reasoning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Omar Zaidan, Jason Eisner, and Christine Piatko. 2007. Using "annotator rationales" to improve machine learning for text categorization. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 260–267, Rochester, New York. Association for Computational Linguistics.

Hongming Zhang, Xinran Zhao, and Yangqiu Song. 2020a. WinoWhy: A deep diagnosis of essential commonsense knowledge for answering Winograd schema challenge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5736–5745, Online. Association for Computational Linguistics.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark. Association for Computational Linguistics.

Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020b. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 295–305, New York, NY, USA. Association for Computing Machinery.

Wenxuan Zhou and Muhao Chen. 2022. An improved baseline for sentence-level relation extraction. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 161–168, Online only. Association for Computational Linguistics.

A Explanation Usefulness Measurements Equations

We use this notation: x for input, y_p for a predicted label, e_p for an explanation of y_p, e_p^b for a baseline explanation (e.g., model confidence) or no explanation, y for the gold label, and y_h is the final decision a human makes upon seeing y_p.

Underreliance:

$$\begin{aligned}
&\bar{r}_u(x, y_p, e_p) := \frac{1}{|\mathcal{C}|} \sum_{x \in \mathcal{C}} \mathbb{1}_{\{\text{reject}(x, y_p, e_p)\}} \\
&\mathcal{C} := \{x : y_p = y\} \\
&u(e_p) := \bar{r}_u(x, y_p, e_p^b) - \bar{r}_u(x, y_p, e_p^b, e_p)
\end{aligned} \tag{1}$$

Overreliance:

$$\begin{aligned}
&\bar{r}_o(x, y_p, e_p) := \frac{1}{|\mathcal{W}|} \sum_{x \in \mathcal{W}} \mathbb{1}_{\{\text{accept}(x, y_p, e_p)\}} \\
&\mathcal{W} := \{x : y_p \neq y\} \\
&u(e_p) := \bar{r}_o(x, y_p, e_p^b) - \bar{r}_o(x, y_p, e_p^b, e_p)
\end{aligned} \tag{2}$$

Appropriate Reliance (weaker):

$$\begin{aligned}
&r_a(x, y_p, e_p) := \begin{cases} 1, & \big(y_p = y \land \text{accept}(x, y_p, e_p)\big) \lor \\ & \big(y_p \neq y \land \text{reject}(x, y_p, e_p)\big) \\ 0, & \text{otherwise} \end{cases} \\
&\bar{r}_a := \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} r_a(x, y_p, e_p) \\
&u(e_p) := \bar{r}_a(x, y_p, e_p^b, e_p) - \bar{r}_a(x, y_p, e_p^b)
\end{aligned} \tag{3}$$

Appropriate Reliance (stricter):

$$\begin{aligned}
&r_a(x, y_p, e_p) := \begin{cases} 1, & \big(y_p = y \land \text{accept}(x, y_p, e_p)\big) \lor \\ & \big(y_p \neq y \land \text{reject}(x, y_p, e_p) \land \boldsymbol{y_h = y}\big) \\ 0, & \text{otherwise} \end{cases} \\
&\bar{r}_a := \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} r_a(x, y_p, e_p) \\
&u(e_p) := \bar{r}_a(x, y_p, e_p^b, e_p) - \bar{r}_a(x, y_p, e_p^b)
\end{aligned} \tag{4}$$

Complementary Performance:

$$\begin{aligned}
&u(e_p) = \text{acc}(y_h | x, e_p, y_p) - \\
&\quad \max\{\text{acc}(y_h | x, y_p, e_p^b), \text{acc}(y_h | x), \text{acc}(y_p | x)\},
\end{aligned} \tag{5}$$

B Analysis of LIAR-RAW

The main purpose of this task is to assess the veracity of statements about a diverse range of topics, using a handful of reports as references. Upon conducting a manual examination of a randomly selected sample from the dataset, a few issues related to data quality became apparent. Notably, it appeared that during some data processing stages, all instances of “to be” verbs had been replaced with “be”, sentences and phrases had been truncated, and other grammatical problems were identified. In order to evaluate the quality of the data, we conducted an analysis on a sample of 97 data points randomly selected from the dataset. For this analysis, a single annotator carefully reviewed each data point to determine its acceptability based on the claim and the accompanying reports extracted from the relevant articles. This assessment encompassed confirming the coherence and alignment of the reports with the claim. Out of the 97 data points reviewed, 36 (37.1%) were deemed acceptable. Additionally, we calculate the perplexity scores for these data points using the GPT2-XL (1.5B) model (Radford et al., 2019). Our goal was to find a link between data acceptability and perplexity for potential data filtering. However, our analysis showed no such correlation, making perplexity unsuitable for filtering low-quality data. In light of this, we will not include this dataset in the subsequent phases of our study.

C Analysis of UKPSnopes

In this task, the objective is to assess the truthfulness of claims spanning various domains. Each claim is paired with an associated article sourced from the fact-checking website Snopes[3]. We recognize that this setup does not reflect real-world scenarios as there may not be a corresponding fact-checked article available for every new claim. Therefore, we primarily rely on articles procured from non-fact-checking sources. To enable retrieval, we compile all articles from non-fact-checking resources in the original dataset to build a corpus of size 13.1K. We observed that around 15.24% of claims in the training set had contrary labels when matched with different sections of the same Snopes article. As we are excluding Snopes articles from our analysis, we cannot use these claims due to the uncertainty of their gold labels. For the remaining claims, we retrieved the most relevant documents from the corpus of non-fact-checking articles and finetuned the Flan-T5-3B model using these documents. While the retrieval recall in the train set is 51.03%, the experiment had poor results (Table 5a). To investigate the issue, we conducted two more experiments. First, we fine-tuned the same model using the gold non-fact-checking articles linked in the Snopes articles that are paired with the claims, and second, using the gold Snopes articles. We found poor performance with non-fact-checking articles (Table 5b) but good performance with Snopes articles (Table 5c). This indicates that the corpus of articles lacks the required information to solve the task.

	P	R	F1	P	R	F1	P	R	F1
SUPPORT	42.3	100	59.4	44	84.3	57.8	81.6	75.3	78.3
NO INFO.	80	1.6	3.1	49.6	21.2	29.7	76	82.6	79.1
CONTRADICT	0	0	0	0	0	0	67.1	62.82	64.9
MICRO AVG.	40.8	33.9	20.9	31.2	35.2	29.2	74.9	73.6	**74.1**
MACRO AVG.	52.4	42.7	26.3	40	44.6	37.2	79.7	77.8	78.6

(a) Retrieved articles

(b) Gold articles linked in the Snopes articles

(c) Gold Snopes articles

Table 5: Three setups for fine-tuning the Flan-T5-3B model for the UKPSnoeps task using different sources for verifying the claims.

D Details of Model Finetuning

In Tables 7–10, we provide an illustrative instance demonstrating how we craft the input for each baseline model we develop following the recommended templates.[4]

EvidenceInference v2 (D38) The task aims to compare the effect of treatment A relative to treatment B on a specified outcome within a scientific article. In a real-world scenario, the ideal scientific article to look into might not always be readily available. Hence, we formulate the task to involve document retrieval and thereby, we aggregate all articles within the dataset to establish a corpus for the retrieval of apposite articles. Our approach consists of these two steps:

[3] https://www.snopes.com

[4] https://github.com/google-research/FLAN/blob/main/flan/v2/templates.py

- We use the BM25Plus algorithm (Trotman et al., 2014) to get the top 100 relevant documents for each query, after which, we rank them with the method introduced by Nogueira et al. (2020), and finally select the top 10 ranked.
- We finetune Flan-T5-3B using the query and top 10 documents as input.

	P	**R**	**F1**
Increase	87.77	91.27	89.49
No Diff.	90.75	87.81	89.26
Decrease	87.54	87.85	87.69
Micro avg.	89	88.96	**88.96**
Micro avg.	88.69	88.98	88.81

Table 6: ERASER EvidenceInference task performance with finetuned Flan-T5-3B when gold documents are provided to the model.

Table 6 shows how well our model performs when we finetune it with the true relevant document. Our retrieval module, however, has a 3% recall rate, i.e., it retrieves the true relevant document for only 3% of the queries. Note that we use the same retrieval procedure for other datasets where we get the recall of 50%. The significant difference in F1 scores between using true relevant documents and retrieved documents (see Table 3) underscores the retrieval challenge, indicating the need for more comprehensive data collections and stronger retrieval models.

SciFact-Open (D44) This is another fact-checking task, but the claims are limited to the scientific domain. To train the Flan-T5-3B model, we follow the same steps in previous tasks. We extract the top 10 most pertinent documents related to each claim from 500K research abstracts.

ContractNLI (D49) Given a contract and a set of hypotheses, the objective of this task is to determine for each hypothesis whether it implies, contradicts, or remains neutral in relation to the contract. This is a three-class classification task, with "Yes" signifying hypothesis entailment to the contract, "No" denoting contradiction with the contract, and "Cannot say" indicating a neutral relationship. To prepare data for finetuning, we concatenate the contract and hypothesis (see Table 9), ensuring the hypothesis remains in the input by truncating the left side.

Indian Legal Documents Corpus (D50) This task involves predicting whether claims presented by an appellant/petitioner against a respondent should be accepted or rejected using a case proceeding document sourced from the Supreme Court of India (Malik et al., 2021). Following the proposed approach accompanying the dataset, we use as many final tokens of ILDC$_{\text{single}}$ instances as we can for training our model. The later tokens are expected to encapsulate the key information and reasoning underpinning the judgment. Malik et al. (2021) could fit only 512 tokens, but Flan-T5 does not have restrictions on the input size. The number of input tokens it can process is determined by memory capacity, hence we could fit 4200.

Energetic 3.20 ± 0.10 3.28 ± 0.10 < 0.05\nParticipants reported being significantly more relaxed, calmer, more energetic, less tired, less sluggish, and felt a higher overall sense of well-being during the intervention period compared to the control.\n**Based on the above text, what's the best answer to this question**: Does administering the treatment 'sit-stand desks (SSDs)' significantly change the Energetic compared to the baseline treatment?\n\n **Options:**\n**Significantly increase**\n, **No significantly difference**\n, **Significantly decrease**\n\n**Answer:**

Table 7: A representative input sample for the ERASER EvidenceInference V2.0 task. In this sample, there is only one document, but through retrieval, additional documents should be added. The template is: "{**text**}\nBased on the above text, what is the best answer to this question: {**question**}\n\n**Options**:\nsignificantly increase\nno significantly difference\nsignificantly decrease\n\n**Answer**: "

Determine if the claim is true based on the text below:\n **Claim:** A high microerythrocyte count raises vulnerability to severe anemia in homozygous alpha (+)- thalassemia trait subjects.\n\n**Options: True, False, Not enough information**\n\nBACKGROUND The heritable haemoglobinopathy alpha(+)-thalassaemia is caused by the reduced synthesis of alpha-globin chains that form part of normal adult haemoglobin (Hb). \nIndividuals homozygous for alpha(+)-thalassaemia have microcytosis and an increased erythrocyte count.\nAlpha(+)-thalassaemia homozygosity confers considerable protection against severe malaria, including severe malarial anaemia (SMA) (Hb concentration < 50 g/l), but does not influence parasite count. \nWe tested the hypothesis that the erythrocyte indices associated with alpha(+)-thalassaemia homozygosity provide a haematological benefit during acute malaria. \nThis model predicted that children homozygous for alpha(+)-thalassaemia lose less Hb than children of normal genotype for a reduction in erythrocyte count of >1.1 x 10(12)/l as a result of the reduced mean cell Hb in homozygous alpha(+)-thalassaemia.\nIn addition, children homozygous for alpha(+)-thalassaemia require a 10% greater reduction in erythrocyte count than children of normal genotype (p = 0.02) for Hb concentration to fall to 50 g/l, the cutoff for SMA. \nWe estimated that the haematological profile in children homozygous for alpha(+)-thalassaemia reduces the risk of SMA during acute malaria compared to children of normal genotype (relative risk 0.52; 95% confidence interval [CI] 0.24-1.12, p = 0.09).\nCONCLUSIONS The increased erythrocyte count and microcytosis in children homozygous for alpha(+)-thalassaemia may contribute substantially to their protection against SMA.\nOther host polymorphisms that induce an increased erythrocyte count and microcytosis may confer a similar advantage.\nRBC counts, Hb, Hct, MCH, MCHC values were significantly higher in b- thalassemia minor comparing with IDA patients but MCV showed no significant difference in these two groups. \nThis point sometimes leads misdiagnosis particularly in coincident IDA and β-thalassemia minor.\nTherefore in suspicious cases of β-thalassemia trait in IDA background, it is better to do hemoglobin electrophoresis after treatment of IDA. \nHowever, the Hb F level was significantly higher in patients with S/Thal having two XmnI sites carrying Arab-Indian and Senegal haplotypes as compared to Bantu, Benin and Cameroon haplotypes. \nThalassemia trait (TT)-related anemia is a common hematologic problem in Mediterranean region. \nThis type of anemia may be frequently confused with iron deficiency anemia (IDA).\nAnemia becomes more severe in case of co-existence of both anemia types. \nThalassemia is a congenital hemolytic disorder caused by a partial or complete deficiency of α- or β-globin chain synthesis.\nHomozygous carriers of β-globin gene defects suffer from severe anemia and other serious complications from early childhood.\nThe disease is treated by chronic blood transfusion.\nSome forms of α thalassemia are also associated with a similar clinical picture. \nAs a consequence, additional previously undescribed, complications are now being recognized. \nBackground-Alpha-thalassemia is one of the most prevalent hemoglobin disorders in the world.\nAs a result, a considerable number of patients with microcytic, hypochromic anemia and normal Hb A2 levels might be misdiagnosed as silent β-thalassemia.\nThey were tested for the 2 most frequent α -thalassemia deletions (-α 3.7, -α 4.2).\nResults of CBC, hemoglobin analysis, and average annual frequencies of severe pain episodes and numbers of transfused red cell units were documented.\nSickle2̆013thalassemia association resulted in higher hemoglobin, hematocrit, and erythrocyte counts with reduced MCV and reticulocytes. \n

Table 8: A representative input sample for the SciFact-Open task. The template is: Determine if the claim is true based on the text below:\n Claim: **claim**\n\n**Options**: True, False, Not enough information\n\n**text**\nanswer:

AGREEMENT ON THE NON-DISCLOSURE OF CONFIDENTIAL INFORMATION (TWO-WAY)\nThis Agreement is made with an effective date of the day of 200_ between The University of Bristol whose registered address is Senate House, Tyndall Ave, Bristol, BS8 1TH, and ____ whose registered address is ________________________________ (the “Parties”)\nTHE PURPOSE of this Agreement is to regulate the exchange and subsequent treatment of confidential information to be received by or disclosed to the signatories to this Agreement, in the field of , so as to protect the proper interests of the disclosing party whilst this confidential information is in the possession or control of the receiving party. For the purposes of this Agreement the term confidential information includes proprietary materials and information relating thereto including without limitation specifications, drawings, designs, computer software and knowhow. In general the receiving party must afford disclosed confidential information the same degree of protection as it would afford its own.\nNOW IT IS HEREBY AGREED:-\n1. The disclosure of confidential information is for the specific purpose of evaluating technology in the field described above in the first instance, and will normally be between of the University of Bristol and of . Any specific documents or materials which are necessarily provided on loan for the above purposes will be specified in a schedule to this Agreement and the receiving party will return these and any other documents or materials subsequently provided to the disclosing party on request.\n2. The parties will mark or otherwise designate confidential information to show expressly or by necessary implication that it is imparted in confidence.\n3. The receiving party will receive all confidential information (whether recorded in writing or by other means or given orally without record) which is disclosed in connection with this Agreement subject to the following conditions:\na) it will take all proper and reasonable measures to ensure that the confidentiality of such information is maintained;\nb) it will not use the information for any commercial purpose or manufacture without obtaining a written licence or other agreement from the disclosing party;\nc) it will not disclose the information to any third party without written permission;\nd) it will not disclose the information to employees other than those above except to the extent necessary to fulfil the purposes set out above and all such other employees to which it will disclose it will be made aware of the confidential nature of the information, and the conditions of disclosure herein;\ne) it will not make any copy of or abstract of the information without specific written permission from the disclosing party;\nf) it will acknowledge the source (i.e. one of the organisations signatory to this Agreement) of, and will mark “Confidential”, any drawing, document or software incorporating the information.\n4. Under the terms of this Agreement there is no explicit or implied transfer of ownership to the receiving party of any drawings, documents or software, or the copyright subsisting in them. PROVIDED that the obligations herein undertaken will not apply to:\na) information which at the time of disclosure is in the public domain or which after disclosure becomes part of the public domain through no fault of the recipient, or\nb) information which the recipient party can show was in its possession at the time of disclosure or which is independently developed by the recipient and was not acquired directly or indirectly from the disclosing party, or\nc) information which is made public at any time by the disclosing party, or by others with the permission of the disclosing party, or\nd) information which is received by the receiving party from a third party without similar restriction and without breach of this Agreement.\ne) information which is required to be disclosed by legal process, law or regulatory authority.\n5. Both parties agree that at all times, during and after the current discussions, and thereafter for a period of ten (10) years, starting from the effective date of this Agreement, not to communicate or to divulge to third parties confidential information received from the other party.\n6. This Agreement is to be construed and enforced in accordance with English Law and is subject to the exclusive jurisdiction of the English courts to which the parties hereto submit. This clause shall not prevent a party from seeking interim relief in any court of competent jurisdiction. Signed for and on behalf of the University of Bristol\n____________________________________ Date_________________\nName in block letters ____\nSigned for and on behalf of\n____________________________ Date_________________\n\n\n**Does this contract follow that** Receiving Party may create a copy of some Confidential Information in some circumstances?\n**Options: Yes, No, Cannot say**

Table 9: A representative input sample for the Contract-NLI task. The temples is: “{**premise**}\n\nDoes this contract follow that “{**hypothesis**}”?\nOptions: Yes, No, Cannot say \n{answer}”

civil appellate jurisdiction civil appeal number 1415 of\n1981.\nappeal by special leave from the judgment and order\ndated the 7th january 1981 of the allahabad high companyrt in\ncivil misc. application number 113 of 1981 in second appeal number\n1484 of 1973.\n\np. rana m. qamaruddin and mrs. m. qamaruddin for the\nappellants. k. sanghi for respondent number 1.\nthe judgment of the companyrt was delivered by\ndesai j. special leave granted. we have heard mr. o. p. rana learned companynsel for the\nappellant and mr. a.k. sanghi learned companynsel for the\nrespondent. the high companyrt disposed of the appeal preferred\nby the present appellant in the absence of the learned\ncounsel for the appellant. when the appellant became aware\nof the fact that his appeal had been disposed of in the\nabsence of his advocate he moved an application in the high\ncourt to recall the order dismissing his appeal and permit\nhim to participate in the hearing of the appeal. this\napplication was rejected by the high companyrt on the ground\nthat though the application was prepared and drafted and an\naffidavit was sworn on 29th october 1980 the same was number\npresented to the companyrt till numberember 12 1980 and that there\nis numbersatisfactory explanation for this slackness on the\npart of the learned advocate who was requested to file the\napplication. the disturbing feature of the case is that under our\npresent adversary legal system where the parties generally\nappear through their advocates the obligation of the\nparties is to select his advocate brief him pay the fees\ndemanded by him and then trust the learned advocate to do\nthe rest of the things. the party may be a villager or may\nbelong to a rural area and may have numberknumberledge of the\ncourts procedure. after engaging a lawyer the party may\nremain supremely companyfident that the lawyer will look after\nhis interest. at the time of the hearing of the appeal the\npersonal appearance of the party is number only number required\nbut hardly useful. therefore the party having done\neverything in his power to effectively participate in the\nproceedings can rest assured that he has neither to go to\nthe high companyrt to inquire as to what is happening in the\nhigh companyrt with regard to his appeal number is he to act as a\nwatchdog of the advocate that the latter appears in the\nmatter when it is listed. it is numberpart of his job. mr. a.k. sanghi stated that a practice has grown up in the high companyrt\nof allahabad amongst the lawyers that they remain absent\nwhen they do number like a particular bench. maybe he is better\ninformed on this matter. ignumberance in this behalf is our\nbliss. even if we do number put our seal of imprimatur on the\nalleged practice by dismissing this matter which may\ndiscourage such a tendency would it number bring justice\ndelivery system into disrepute. what is the fault of the\nparty who having done everything in his\npower and expected of him would suffer because of the\ndefault of his advocate. if we reject this appeal as mr.\n\nk. sanghi invited us to do the only one who would suffer\nwould number be the lawyer who did number appear but the party\nwhose interest he represented. the problem that agitates us\nis whether it is proper that the party should suffer for the\ninaction deliberate omission or misdemeanumberr of his agent. the answer obviously is in the negative. maybe that the\nlearned advocate absented himself deliberately or\nintentionally. we have numbermaterial for ascertaining that\naspect of the matter. we say numberhing more on that aspect of\nthe matter. however we cannumber be a party to an innumberent\nparty suffering injustice merely because his chosen advocate\ndefaulted. therefore we allow this appeal set aside the\norder of the high companyrt both dismissing the appeal and\nrefusing to recall that order. we direct that the appeal be\nrestored to its original number in the high companyrt and be\ndisposed of according to law. if there is a stay of\ndispossession it will companytinue till the disposal of the\nmatter by the high companyrt. there remains the question as to\nwho shall pay the companyts of the respondent here.\n\n**Multi-choice problem: Determine whether this petition should be accepted or not.**\n**Options: Accept, Reject**\n**Answer:**

Table 10: A representative input sample for the ILDC$_{\text{single}}$ task. The template is: **petition**\n\nMulti-choice problem: Decide whether this petition should be accepted or not.\nOptions: Accept, Reject\nAnswer: answer

E Categorization of ExNLP Tasks

[D1] StrategyQA (Geva et al., 2021)

Prediction Task: Open-ended QA (1) without any additional context or (2) in the context of retrieved Wikipedia paragraphs

Average Input Length: 960[5] [context] + 46 [question] = 1003 words

Human Ability: Reported human accuracy is 87% ("given access to Wikipedia articles and an option to reveal the decomposition for every question")

Application: Information search

Hazard from Immediate Usage:

- *Who:* Anyone
- *Hazard:* Propagating misinformation from accepting the wrong answer
- *Probability:* Moderate (currently 12% model-human accuracy gap)[6]
- *Severity:* Low (questions are not about critical information such as health, law, etc.)
- *Risk:* Low

Hazard from Downstream Impact: Same as for the immediate usage.

[D2] NaturalQuestions (Kwiatkowski et al., 2019)

Prediction Task: Identifying a span in a Wikipedia article that answers an open-ended question (originally asked in Google Search)

Average Input Length: 5197 [document] + 9 [question] = 5206 words

Human Ability: Reported human F1 score is 87% (long answers), 76% (short answers)

Application: Information search

Hazard from Immediate Usage:

- *Who:* Anyone
- *Hazard:* Propagating misinformation from accepting the wrong answer
- *Probability:* Moderate ($\exists$ model-human performance gap)[7]
- *Severity:* Low (questions are not about critical information such as health, law, etc.)
- *Risk:* Low

Hazard from Downstream Impact: Same as for the immediate usage.

[D3] ReClor (Yu et al., 2020)

Prediction Task: Multiple-choice reading comprehension targeting logical reasoning

Average Input Length: 65 [context] + 15 [question] + 75 words [choices] = 155 words

Human Ability: Although it can be 100%, Bansal et al. (2021) report 67%

[5] The models are set to retrieve 10 Wikipedia paragraphs from the corpus and the average paragraph length is 96.

[6] `https://leaderboard.allenai.org/strategyqa/submissions/public`; `https://paperswithcode.com/sota/strategyqa-on-big-bench`

[7] `https://ai.google.com/research/NaturalQuestions/leaderboard`; `https://paperswithcode.com/sota/question-answering-on-natural-questions`

Application: No. Models trained on this data could be used to practice for law school admissions if new exams with multiple choices are available, but not correct solutions. However, practice exams come with solutions.

Hazard from Immediate Usage: None as there is no realistic application.

Hazard from Downstream Impact: None as there is no realistic application.

[D4] BeerAdvocate (McAuley et al., 2012)

Prediction Task: Sentiment classification of beer reviews

Average Input Length: 88 words [review]

Human Ability: 87%; Although this is already high, Bansal et al. (2021) show this is not the upper bound

Application: Deciding whether to buy a beer

Hazard from Immediate Usage:

- *Who:* Beer buyers
- *Hazard:* Buying a beer they do not like
- *Probability:* Assuming that beers that are positively reviewed are liked by new customers, we expect the probability to be low since today's models accurately classify the sentiment of reviews in other domains.
- *Severity:* Low since the cost of a bottle/can of beer is generally low
- *Risk:* Low

Hazard from Downstream Impact: Nothing noteworthy.

[D5] QuizBowl (Feng and Boyd-Graber, 2019)

Prediction Task: Quizbowl (answering questions from all areas of knowledge with as few clues as possible)

Average Input Length: ∼20 words [question] based on the similar data (Rodriguez et al., 2019)

Human Ability: An average player "buzzes with 65% of the question shown with 60% accuracy" (Rodriguez et al., 2019)

Application: Playing Quizbowl as a cooperation with a machine. This version does not exist yet but could happen.

Hazard from Immediate Usage:

- *Who:* Quizbowl player
- *Hazard:* Loosing a game
- *Probability:* Depends on the player
- *Severity:* Low
- *Risk:* Low

Hazard from Downstream Impact: If a player loses they are affecting only themselves.

[D6] Ott et al. (2011)

Prediction Task: Finding deceptive opinion spam ("fictitious opinions that have been deliberately written to sound authentic, in order to deceive the reader") in the context of hotel reviews

Average Input Length: 146 words [review]

Human Ability: 53-62% (majority baseline 58%)

Application: Deciding whether to engage with a hotel review and book the hotel

Hazard from Immediate Usage:

- *Who:* A person booking a hotel
- *Hazard:* Booking a disappointing hotel
- *Probability:* Low. People take multiple factors, not only a few reviews, when booking a hotel, especially if more expensive/important. However, if we assume that they looked at only reviews, we still expect the probability to be low since today's models accurately classify the sentiment of reviews in other domains.
- *Severity:* Depends on personal circumstances and expense, but generally low.
- *Risk:* Low

Hazard from Downstream Impact:

- *Who:* Hotel management
- *Hazard:* Public complaints that the room was not as described; A customer with the right expectations does not get a room
- *Probability:* Low, since the probability from the immediate usage is low
- *Severity:* Moderate, since repeatedly getting public complaints and missing the right customers can hurt the business to some degree
- *Risk:* Low

[D7] HotPotQA (Yang et al., 2018)

Prediction Task: Reading comprehension targeting multi-hop reasoning

Average Input Length: 4633 [context] + 15 [question] = 4648 words

Human Ability: 98.8 F1

Application: Information search

Hazard from Immediate Usage:

- *Who:* Anyone
- *Hazard:* Propagating misinformation from accepting the wrong answer
- *Probability:* Moderate ($\exists$ model-human performance gap)[8]
- *Severity:* Low (questions are not about critical information such as health, law, etc.)
- *Risk:* Low

Hazard from Downstream Impact: Same as for the immediate usage.

[D8] Amazon Book Reviews (He and McAuley, 2016)

Prediction Task: Sentiment classification of book reviews

Average Input Length: 105 words [review]

Human Ability: Not reported, but we expect people to be good at this task

Application: Deciding whether to buy a book

Hazard from Immediate Usage:

- *Who:* Book buyers
- *Hazard:* Buying a book they do not like
- *Probability:* Assuming that books that are positively reviewed are liked by new customers, we expect the probability to be low since today's models accurately classify the sentiment of reviews in other domains.

[8] https://hotpotqa.github.io/; https://paperswithcode.com/sota/question-answering-on-hotpotqa

- *Severity:* Low since the cost of a book is generally low
- *Risk:* Low

Hazard from Downstream Impact: Nothing noteworthy.

[D9] Jansen et al. (2016)

Prediction Task: Multiple-choice science exam QA

Average Input Length: 20 [question] + 20 [choices] = 40 words

Human Ability: Depends, but can be 100%

Application: No. Models trained on this data could be used that students in 3rd- to 5-th grade to practice for science exams if exams are available, but not correct solutions. However, practice exams come with solutions.

Hazard from Immediate Usage: None as there is no realistic application.

Hazard from Downstream Impact: None as there is no realistic application.

[D10] Ling et al. (2017)

Prediction Task: Solving multiple-choice algebraic word problems

Average Input Length: 31 [question] + 10 [choices] = 41 words

Human Ability: Depends, but can be 100%

Application: No. Models trained on this data could be used by college students to practice for GMAT/GRE if exams are available, but not correct solutions. However, practice exams come with solutions.

Hazard from Immediate Usage: None as there is no realistic application.

Hazard from Downstream Impact: None as there is no realistic application.

[D11] Srivastava et al. (2017)

Prediction Task: Classification of the purpose of an email (including an email to oneself) into 7 categories: "personally keep note of a person contact", "requesting something to be done [from an employee]", "asking [a friend] to meet up at some event", sharing "something humorous from the Internet" to a friend, "request a meeting about something", "announcement of some new policy", "reminder to do something"

Average Input Length: Data not available

Human Ability: Not reported

Application: No. These are personal reminders and we expect that people do not want them to be categorized automatically in these specific 7 categories.

Hazard from Immediate Usage: None as there is no realistic application.

Hazard from Downstream Impact: None as there is no realistic application.

[D12] Hancock et al. (2018)

Prediction Task: Given a sentence with highlighted (1) names of two people, predict whether they are spouses, (2) a chemical and a disease, predict whether the disease is chemical-induced, and (3) a protein and a kinase, predict “whether or not the kinase influences the protein in terms of a physical interaction or phosphorylation”

Average Input Length: (1) 23 [sentence with a spouse relationship], (2) 10 words [a sentence with a chemical-disease pair], (3) The protein data is not available.

Human Ability: Not reported

Application: (1) No, we expect there is no interest in a tool that only predicts whether two people named in a given sentence are spouses. (2) Automatic completion of bioinformatics databases based on new biomedical publications. (3) Hancock et al. say that predicting a relation between a given protein and kinase can be useful for “targeting biological pathways of Parkinson’s disease”.

Hazard from Immediate Usage: We focus on (2) that has an application and its data is available.

- *Who:* Biocurator
- *Hazard:* Accepting a wrong prediction and consequently (1) adding to a database a wrong relation or (2) not adding a correct relation. These can result in the biocurator’s job performance problems if done repeatedly and propagating misinformation.
- *Probability:* Undetermined, as the recent models’ performance for this application is not known
- *Severity:* Moderate
- *Risk:* Depends on the probability, but could be moderate

Hazard from Downstream Impact:

- *Who:* Scientist/biologist; Database owner
- *Hazard:* Getting the wrong information about a relation; Providing wrong or missing information to their customers based on their biocurators’ final decisions
- *Probability:* Undetermined, as the recent models’ performance for this application is not known
- *Severity:* Moderate
- *Risk:* Depends on the probability, but could be moderate

[D13] e-SNLI (Camburu et al., 2018)

Prediction Task: Natural language inference

Average Input Length: 13 [premise] + 7 [hypothesis] = 20 words

Human Ability: 89% (Bowman et al., 2015)

Application: No. SNLI is introduced to probe models’ understanding of entailment and contradiction.[9]

Hazard from Immediate Usage: None as there is no realistic application.

Hazard from Downstream Impact: None as there is no realistic application.

[D14] e-δ-SNLI (Brahman et al., 2021)

Prediction Task: Defeasible natural language inference (Rudinger et al., 2020)

The remaining information is the same as for E-SNLI *above.*

---

[9] There are application-grounded versions of NLI such as EvidenceInference v2 (D38).

[D15] LIAR-PLUS (Alhindi et al., 2018)

Prediction Task: Verification of claims about a broad range of topics based on (1) metadata, or (2) metadata and a summary of a report written by a fact checker that discusses the veracity of a claim

Average Input Length: (1) 17 [claim] + 50 [metadata] = 67 words; (2) 17 [claim] + 50 [metadata] + 69 [summary] = 136 words

Human Ability: Not reported. (1) We expect that fact checking a claim based on metadata, without any reports on the claim, is hard. (2) We expect that is easy to fact check a claim based on a short report written by a professional fact-checker that summarizes their research on the veracity of the claim.

Application: No. (1) Fact-checking without reading any reports on the claim is not realistic. (2) A summary written by professionals to fact-check a claim already clearly indicates the author's decision of veracity. The LIAR-RAW version (see [D17]) where the input is the statement and a few reports, some of which are unreliable, is a reasonable application.

Hazard from Immediate Usage: None as there is no realistic application.

Hazard from Downstream Impact: None as there is no realistic application.

[D16] PubHealth (Kotonya and Toni, 2020)

Prediction Task: Verification of claims about public health from a fact-checking/news article discussing the claim written by a professional

Average Input Length: 14 [claim] + 707 [article] = 721 words

Human Ability: Not reported, but we expect that is easy to fact check a claim based on a report written by a professional fact-checker that summarizes their research on the veracity of the claim.

Application: No. A summary written by professionals to fact-check a claim already clearly indicates the author's decision of veracity. The LIAR-RAW version (see [D17]) where the input is the statement and a few documents, some of which are unreliable, is a reasonable application.

Hazard from Immediate Usage: None as there is no realistic application.

Hazard from Downstream Impact: None as there is no realistic application.

[D17] LIAR-RAW (Yang et al., 2022)

Prediction Task: Verification of claims about a broad range of topics, given a few reports (media reports, user comments, blogs, etc.), some of which are unreliable.

Average Input Length: 17 [claim] + 1568 [reports] = 1585 words

Human Ability: Not reported

Application: The task setup is realistic because people will first find related articles (some of which are unreliable) to go about verifying a claim.

Hazard from Immediate Usage:

- *Who:* Fact checker; Anyone
- *Hazard:* Job performance problems; Propagating misinformation
- *Probability:* Moderate, models' performance is not high (Yang et al., 2022)
- *Severity:* Can be high (e.g. if someone was defamed); Moderate (the statements are about more important information than in open-ended QA datasets, but not all are about vital information such as health)
- *Risk:* High; Moderate

Hazard from Downstream Impact:

- *Who:* An entity that false statements were made about and that a fact checker falsely confirmed; Anyone
- *Hazard:* Defamation; Propagating misinformation
- *Probability:* Moderate (same as above)
- *Severity:* High; Moderate (same as above)
- *Risk:* High; Moderate

[D18] RAWFC (Yang et al., 2022)

Prediction Task: Verification of short statements on a broad range of topics based on a few reports (media reports, user comments, blogs, etc.), some of which are unreliable.

Average Input Length: 18 [claim] + 4075 [reports] = 4093 words

The remaining information is the same as for LIAR-RAW above.

[D19] ECQA (Aggarwal et al., 2021)[10]

Prediction Task: Multiple-choice QA targeting commonsense

Average Input Length: 13 [question] + 13 [choices] = 26 words

Human Ability: 88.9% (Talmor et al., 2019)

Application: No. CQA is introduced to test models' commonsense understanding. People do not need answers to commonsense-probing questions.

Hazard from Immediate Usage: None as there is no realistic application.

Hazard from Downstream Impact: None as there is no realistic application.

[D20] Sensemaking (Wang et al., 2019)

Prediction Task: Given two sentences, predict which one is nonsensical

Average Input Length: 17 [sentence1 + sentence2] = 17 words

Human Ability: 99.1%

Application: No. SENSEMAKING is introduced to test models' commonsense understanding. People do not need predictions of which of two sentences is nonsensical.

Hazard from Immediate Usage: None as there is no realistic application.

Hazard from Downstream Impact: None as there is no realistic application.

[D21] WinoWhy (Zhang et al., 2020a)

Prediction Task: Winograd Schema Challenge (pronoun coreference resolution)

Average Input Length: 16 [sentence 1] + 24 [sentence 2] = 40 words

Human Ability: 92.1% accuracy (Bender, 2015)

Application: No. WSC is introduced to test models' commonsense understanding. People do not need such pronouns resolved (in isolation).

Hazard from Immediate Usage: None as there is no realistic application.

[10] ECQA explanation annotations replace CoS-E's (Rajani et al., 2019) that are too nosiy.

Hazard from Downstream Impact: None as there is no realistic application.

[D22] ChangeMyView (Atkinson et al., 2019)

Prediction Task: Predicting is a forum counter-argument to someone's opinion persuasive

Average Input Length: 351 [opinion] + 215 [counterargument] = 566 words

Human Ability: Not reported

Application: Assistant writing when the goal is to write a convincing, but not deceptive, counter-argument by showing the writer if their current response is predicted to change someone's mind.

Hazard from Immediate Usage:

- *Who:* Anyone
- *Hazard:* Failing to change someone's opinion
- *Probability:* Undetermined, as the models' performance for this application is not known
- *Severity:* Low
- *Risk:* Low

Hazard from Downstream Impact:

- *Who:* Person who changes their mind
- *Hazard:* Not changing their opinion. This does not lead to propagating misinformation as original posts are presented as opinions, not facts.
- *Probability:* Undetermined, as recent models' performance for this application is not known
- *Severity:* Low
- *Risk:* Low

[D23] SBIC (Sap et al., 2020)

Prediction Task: Classify a social media or forum post as offensive or not

Average Input Length: 19 words [post]

Human Ability: Not reported

Application: Content moderation

Hazard from Immediate Usage:

- *Who:* Content moderator
- *Hazard:* Job performance problems from repeatedly not flagging attacking comments or flagging non-attacking comments
- *Probability:* Undetermined, as recent models' and human performance are not known
- *Severity:* Moderate
- *Risk:* Depends on the probability, but can be moderate

Hazard from Downstream Impact:

- *Who:* Someone who is targeted (in-group or personally) by an attacking comment; A poster of an inoffensive post that is flagged
- *Hazard:* Mental health harms
- *Probability:* Undetermined, as recent models' and human performance are not known
- *Severity:* Depends on personal circumstances, but can be moderate
- *Risk:* Depends on the probability, but can be moderate

[D24a] Wang et al. (2020); relation extraction

Prediction Task: Relation extraction between people and organizations (TACRED; Zhang et al., 2017) or relations that are chosen because they have broad coverage (SEMEVAL; Hendrickx et al., 2009)

Average Input Length: 36 words [sentence] (TACRED) / 19 words [sentence] (SEMEVAL)

Human Ability: Not reported, but we expect good human abilities for the task

Application: Extraction of TACRED relations will be requested by people in form of open-ended QA. SemEval relations are too generic and we do not see a specific application for them.

Hazard from Immediate Usage:

- *Who:* Anyone
- *Hazard:* Propagating misinformation about relations between certain people and organizations.
- *Probability:* Low, a RoBERTa-based model gets a 91+ F1-score (Zhou and Chen, 2022).
- *Severity:* Low (relations are not about critical information such as health, law, etc.).
- Low

Hazard from Downstream Impact: Same as for the immediate usage.

[D24b] Wang et al. (2020); sentiment analysis

Prediction Task: Sentiment classification of laptop and restaurant reviews

Average Input Length: 15 words [laptop reviews]; 13 words [restaurant reviews]

Human Ability: Not reported

Application: Deciding whether to buy a laptop / visit a restaurant

Hazard from Immediate Usage:

- *Who:* Laptop buyers, restaurant-goers
- *Hazard:* Dissatisfying laptop/restaurant
- *Probability:* Low. People take multiple factors, not only a few reviews when buying a laptop or booking a restaurant, especially if more expensive/important. However, if we assume that they looked at only reviews, we still expect the probability to be low since today's models accurately classify the sentiment of reviews in other domains.
- *Severity:* Depends on personal circumstances and expense, but generally low.
- *Risk:* Low

Hazard from Downstream Impact: Nothing noteworthy.

[D25] COPA-SSE (Brassard et al., 2022)

Prediction Task: Given a premise and two choices, select the choice that more plausibly has a causal relation with the premise

Average Input Length: 6 [premise] + 12 [choices] = 18 words

Human Ability: "We have established that human raters can perform extremely well on this task, with near perfect agreement." (Roemmele et al., 2011)

Application: No. COPA is introduced to test models' commonsense causal reasoning that people possess.

Hazard from Immediate Usage: None, as there is no realistic application.

Hazard from Downstream Impact: None, as there is no realistic application.

[D26] WorldTree v1 (Jansen et al., 2018)

Prediction Task: Multi-choice middle-school level science exam QA

Average Input Length: 23 [question] + 20 [options] = 43 (v1) words

Human Ability: Depends, but can be 100%

Application: No. Models trained on this data could be used that students in 3rd- through 5-th to practice for science exams if exams are available, but not correct solutions. However, practice exams come with solutions.

Hazard from Immediate Usage: None, as there is no realistic application.

Hazard from Downstream Impact: None, as there is no realistic application.

[D27] WorldTree V2 (Xie et al., 2020)

Prediction Task: Multi-choice middle-school level science exam QA

Average Input Length: 19 [question] + 15 [options] = 34 (v2) words

Human Ability: Depends, but can be 100%

Application: No. Models trained on this data could be used that students in 3rd- through 9-th grade to practice for science exams if exams are available, but not correct solutions. However, practice exams come with solutions.

Hazard from Immediate Usage: None, as there is no realistic application.

Hazard from Downstream Impact: None, as there is no realistic application.

[D28] OpenBookQA (Mihaylov et al., 2018); e-OBQA (Jhamtani and Clark, 2020)

Prediction Task: Multi-choice middle-school level science exam QA

Average Input Length: 12 [question] + 11 [options] = 23 words

Human Ability: Reported human performance is 92%, but it could be anything from 0 to 100% depending on a person's knowledge

Application: No. Models trained on this data could be used for students in 3rd through 9th grade to practice for science exams if exams are available, but not correct solutions. However, practice exams come with solutions.

Hazard from Immediate Usage: None, as there is no realistic application.

Hazard from Downstream Impact: None, as there is no realistic application.

[D29] QED (Lamm et al., 2021)

Extended NaturalQuestions with their explanation annotations. See [D2].

[D30] QASC (Khot et al., 2020) / e-QASC (Jhamtani and Clark, 2020)

Prediction Task: Multi-choice middle-school level science exam QA

Average Input Length: 8 [question] + 13 [options] = 21 words

Human Ability: Reported human performance is 93%

Application: No. Models trained on this data could be used that middle-school students to practice for science exams if exams are available, but not correct solutions. However, practice exams come with solutions.

Hazard from Immediate Usage: None, as there is no realistic application.

Hazard from Downstream Impact: None, as there is no realistic application.

[D31] Ye et al. (2020)

Extended NaturalQuestions and SQuAD (Rajpurkar et al., 2016) with their explanation annotations. See [D2].

[D32] R4C (Inoue et al., 2020)

Extended HotPotQA with their explanation annotations. See [D8].

[D33] TriggerNER (Lin et al., 2020)

Prediction Task: Named entity recognition

Average Input Length: 14 words [sentence]

Human Ability: Not reported, but we expect good human abilities for this task

Application: While NER is a useful component of larger systems (automatic tag generation, information retrieval, content recommendation, etc.), it is not realistic to expect that a person will check each labeled entity manually for another purpose.

Hazard from Immediate Usage: None, as there is no realistic application.

Hazard from Downstream Impact: None, as there is no realistic application.

[D34] Zaidan et al. (2007) / ERASER Movie Reviews (DeYoung et al., 2020a)

Prediction Task: Sentiment classification of movie reviews

Average Input Length: 648 words [reviews]

Human Ability: Reported human performance ranges from 92–97%

Application: Deciding whether to go see or rent a movie

Hazard from Immediate Usage:

- *Who:* Movie watchers
- *Hazard:* Buying a cinema ticket or renting a movie they do not like
- *Probability:* Low since sentiment classifiers of movie reviews work well[11]
- *Severity:* Low since the cost of renting or seeing a movie is generally low
- *Risk:* Low

Hazard from Downstream Impact: Nothing noteworthy.

---

[11] https://paperswithcode.com/sota/text-classification-on-imdb

[D35] Stanford Sentiment Treebank (Socher et al., 2013)

Prediction Task: Sentiment classification of movie reviews

Average Input Length: 16 words [review]

Human Ability: Not reported

The rest of the information is the same as for the dataset above ([D35]).

[D36] ERASER BoolQ (DeYoung et al., 2020a)

Prediction Task: Answering yes/no questions from a Wikipedia passage

Average Input Length: 9 [question] + 93 [passage] = 102 words

Human Ability: Reported human accuracy is 90% (Clark et al., 2019)

Application: Information search

Hazard from Immediate Usage:

- *Who:* Anyone
- *Hazard:* Propagating misinformation from accepting the wrong answer
- *Probability:* Low since models achieve accuracy above the estimated human accuracy[12]
- *Severity:* Low (questions are not about critical information such as health, law, etc.)
- *Risk:* Low

Hazard from Downstream Impact: Same as for the immediate usage.

[D37] FEVER (Thorne et al., 2018)

Prediction Task: Verification of claims "containing a single piece of information, focusing on the entity that its original Wikipedia page was about", given Wikipedia articles

Average Input Length: 8 [claim] + 227 [article] = 235 words

Human Ability: Not reported

Application: Open-ended QA with Wikipedia articles. FEVER claims are simple facts about entities and the task in the real world resembles open-ended QA more than a task that a professional fact-checker does.[13]

Hazard from Immediate Usage:

- *Who:* Anyone
- *Hazard:* Propagating misinformation from accepting the wrong answer
- *Probability:* Moderate; SOTA achieves ∼80% accuracy (DeHaven and Scott, 2023)
- *Severity:* Low (questions are not about critical information such as health, law, etc.)
- *Risk:* Low

Hazard from Downstream Impact: Same as for the immediate usage.

[D38] ERASER EvidenceInference (DeYoung et al., 2020a; Lehman et al., 2019) / EvidenceInference v2 (DeYoung et al., 2020b)

Prediction Task: Given a scientific article, predict whether administering treatment A will significantly increase/decrease some outcome compared to treatment B, or make no difference.

[12] `https://paperswithcode.com/sota/question-answering-on-boolq`

[13] An example of a claim in FEVER is: "Berlin is the capital of Germany."

The original task setup asks that a model predict the effect of a treatment given a relevant publication. Having the relevant publication provided is not realistic, but the available data allows a more realistic setup where it must be retrieved first.

Average Input Length: 11 [query (treatment A | treatment B | outcome)] + 3627 [passage] = 3638 words

Human Ability: Not reported

Application: Treatment recommendations

Hazard from Immediate Usage:

- *Who:* Clinicians
- *Hazard:* Cause physical harm to patients; Job performance problems
- *Probability:* Moderate; Chrysostomou and Aletras (2022) report 83% F1 score
- *Severity:* Can be high (for both hazards)
- *Risk:* High (for both hazards)

Hazard from Downstream Impact:

- *Who:* Patients
- *Hazard:* Getting hurt
- *Probability:* Moderate; Chrysostomou and Aletras (2022) report 83% F1 score
- *Severity:* Can be high
- *Risk:* High

[D39] ERASER MultiRC (DeYoung et al., 2020a; Khashabi et al., 2018)

Prediction Task: Multiple-choice QA from a few passages

Average Input Length: 15 [question] + 43 [passage] = 58 words

Human Ability: 84.3 F1-score

Application: No. MultiRC is introduced to probe models' multiple-choice reading comprehension abilities when they need to take "into account information from multiple sentences". If we imagine a version without answer choices, we still deem that there is no realistic application because the source documents are not broad enough for open-ended QA (search engines) but also not specific enough (e.g., healthcare-related questions).

Hazard from Immediate Usage: None, as there is no realistic application.

Hazard from Downstream Impact: None, as there is no realistic application.

[D40] WikiQA (Yang et al., 2015)

Prediction Task: Identifying a span in a Wikipedia article that answers an open-ended question (originally asked in Bing)

Average Input Length: 234 [Wikipedia summary] + 7 [question] = 241 words

Human Ability: Not reported

Application: Information search

Hazard from Immediate Usage:

- *Who:* Anyone
- *Hazard:* Propagating misinformation from accepting the wrong answer
- *Probability:* Low since models achieve high performance[14]
- *Severity:* Low (questions are not about critical information such as health, law, etc.)

[14] https://paperswithcode.com/sota/question-answering-on-wikiqa

- *Risk:* Low

Hazard from Downstream Impact: Same as for the immediate usage.

[D41] WikiAttack (Carton et al., 2018)

Prediction Task: Predict is a Wikipedia revision comment a personal attack

Average Input Length: 65 words [comment]

Human Ability: Not reported

Application: Content moderation

Hazard from Immediate Usage:

- *Who:* Content moderator
- *Hazard:* Job performance problems from repeatedly not flagging attacking comments or flagging non-attacking comments
- *Probability:* Undetermined, as recent models' and human performance are not known
- *Severity:* Moderate
- *Risk:* Depends on the probability, but can be moderate

Hazard from Downstream Impact:

- *Who:* Someone who is targeted (in-group or personally) by an attacking comment; A poster of an inoffensive post that is flagged
- *Hazard:* Mental health harms
- *Probability:* Undetermined, as recent models' and human performance are not known
- *Severity:* Depends on personal circumstances, but can be moderate
- *Risk:* Depends on the probability, but can be moderate

[D42] UKPSnopes (Hanselowski et al., 2019)

Prediction Task: Verification of claims about a broad range of topics, given an article from Snopes fact-checking website, which is not a realistic application setup. However, the available data could possibly allow a more realistic setup where relevant documents (that are not fact-checking reports) must be retrieved first. After running various experiments, it became clear that these documents were insufficient for solving the task (refer to §C for more details) and there is a need for constructing a more comprehensive and suitable document corpus to retrieve relevant articles from.

Average Input Length: 15 [claim] + 947 [documents] = 962 words

Human Ability: 80.2% F1-score

Application: No. The dataset does not represent a realistic task setup (similar to PubHlealth ([D16]). The veracity of the claims is assessed based on an article that specifically discusses the target claim, which does not exist in real-world situations.

Hazard from Immediate Usage: None, as there is no realistic application.

Hazard from Downstream Impact: None, as there is no realistic application.

[D43] CoQA (Reddy et al., 2019)

Prediction Task: "Given a passage and a conversation so far, the task is to answer the next question in the conversation."

Average Input Length: 264 [passage] + 5 [question] + 3 [answer] = 272 words

Human Ability: Reported human F1 score is 88.8

Application: No. Resembles conversational information search, but the first question in CoQA conversations is not standalone (without the passage), e.g., "Who had a birthday", so unlike StrategyQA (D1) and NaturalQuestions (D2) we cannot re-purpose CoQA such that for the first question in the conversation, we retrieve the relevant article then the most relevant passage in it, (i.e., for conversational information search).

Hazard from Immediate Usage: None as there is no realistic application.

Hazard from Downstream Impact: None as there is no realistic application.

[D44] SciFact-Open (Wadden et al., 2022a); SciFact (Wadden et al., 2020)

Prediction Task: Given a claim and a set of abstracts, the *open* scientific claim verification task asks a model to first retrieve abstracts that are relevant for verifying a given claim, and then for each retrieved abstract, predict whether it provides the evidence that supports or refutes the claim.

Average Input Length: 11 [claim] + 12 [title] + 1860 [retrieved abstracts] = 1883 words[15]

Human Ability: Wadden et al. (2022b) estimate human performance in the setting where relevant abstracts are provided to be 89.1% F1 score

Application: Scientific claim verification

Hazard from Immediate Usage:

- *Who:* Clinicians; Researchers/readers of the relevant journals; Anyone
- *Hazard:* Cause physical harm to patients; Publishing new articles based on wrong answers; Defamation; Job performance problems; Propagating misinformation from accepting the wrong answer
- *Probability:* Moderate–High (models do not achieve very high F1 score in the more realistic setup with 500K abstracts)
- *Severity:* Can be high (for all hazards)
- *Risk:* High (for all hazards)

Hazard from Downstream Impact:

- *Who:* Patients; Anyone
- *Hazard:* Getting hurt; Propagating misinformation from accepting the wrong answer from a person who was misinformed by the model
- *Probability:* Moderate–High (see immediate impact)
- *Severity:* Can be high (for both hazards)
- *Risk:* High (for both hazards)

[D45] Kutlu et al. (2020)

Prediction Task: Rating the relevance of Web pages for different search topics

Average Input Length: Data (documents/webpages and search topics/queries) are not available.

Human Ability: Reported human accuracy is 65%

Application: Information search

Hazard from Immediate Usage:

- *Who:* Anyone
- *Hazard:* Propagating misinformation from accepting the wrong answer

[15] The models are set to retrieve 10 relevant abstracts from the corpus and the average paragraph length is 186.

- *Probability:* N/A
- *Severity:* Low (questions are not about critical information such as health, law, etc.)[16]
- *Risk:* Low

Hazard from Downstream Impact: Same as for the immediate usage.

[D46] ECtHR (Chalkidis et al., 2021)

Prediction Task: "Given a set of paragraphs that refer to the facts of each case [...] in judgments of the European Court of Human Rights (ECtHR), [...] predict the allegedly violated articles of the European Convention of Human Rights (ECHR)."

Average Input Length: 1579 words [facts sequence]

Human Ability: Not reported

Application: No. The facts of a case are explicitly provided by legal professionals while in real-world situations, they are not. This is similar to PubHealth ([D16]) where a claim is verified based on a report about this claim written by a professional fact checker. The ILDC version (see [D50]) with unstructured/unannotated case proceedings is more realistic.

Hazard from Immediate Usage: None, as there is no realistic application.

Hazard from Downstream Impact: None, as there is no realistic application.

[D47] Hummingbird (Hayati et al., 2021)

Prediction Task: Classifying text if it has the following styles: politeness, sentiment, offensiveness, and five emotion types.

Average Input Length: 184 words [sentence]

Human Ability: Inter-annotator agreement ranges from ≈63 (politeness) to ≈83 (joy)

Application: Assistant writing when the goal is to write text with one of the styles above

Hazard from Immediate Usage:

- *Who:* Anyone
- *Hazard:* Writing text in undesired style, e.g., not sufficiently polite or sad
- *Probability:* Low-Moderate (based on the 2021 model performance; Hayati et al., 2021)
- *Severity:* Depends who is the text written for, but generally low
- *Risk:* Low

Hazard from Downstream Impact: Nothing noteworthy.

[D48] HateXplain (Mathew et al., 2021)

Prediction Task: Hate speech detection

Average Input Length: 23 words [sentence]

Human Ability: Not reported

Application: Content moderation

Hazard from Immediate Usage:

- *Who:* Content moderator

[16] https://trec.nist.gov/data/web/09/wt09.topics.queries-only

- *Hazard:* Job performance problems from repeatedly not flagging attacking comments or flagging non-attacking comments
- *Probability:* Undetermined, as recent models' and human performance are not known
- *Severity:* Moderate
- *Risk:* Depends on the probability, but can be moderate

Hazard from Downstream Impact:

- *Who:* Someone who is targeted (in-group or personally) by an attacking comment; A poster of an inoffensive post that is flagged
- *Hazard:* Mental health harms
- *Probability:* Undetermined, as recent models' and human performance are not known
- *Severity:* Depends on personal circumstances, but can be moderate
- *Risk:* Depends on the probability, but can be moderate

[D49] ContractNLI (Koreeda and Manning, 2021)

Prediction Task: "Given a contract and a set of hypotheses (each being a sentence), classify whether each hypothesis is entailed by, contradicting to or not mentioned by (neutral to) the contract"

Average Input Length: 1631 [contract] + 13 [hypothesis] = 1644 words

Human Ability: Not reported

Application: Reviewing a contract

Hazard from Immediate Usage:

- *Who:* Business owner; Person working for a company that reviews contracts
- *Hazard:* Incorrectly reviewing the contract leading to business damages/liability; Job performance problems
- *Probability:* High (based on the model's performance for the contradiction label; Koreeda and Manning, 2021)
- *Severity:* High
- *Risk:* High

Hazard from Downstream Impact:

- *Who:* A company hired someone to review their contracts
- *Hazard:* Getting an incorrectly reviewed contract leading to business damages/liability
- *Probability:* High (based on the model's performance for the contradiction label; Koreeda and Manning, 2021)
- *Severity:* High
- *Risk:* High

[D50] ILDC (Malik et al., 2021)

Prediction Task: Based on a case proceeding document from the Supreme Court of India, predict "whether the claim(s) filed by the appellant/petitioner against the respondent should be accepted or rejected".

Average Input Length: 3731 words [petition] (ILDC$Multi$), 3731 words [petition] (ILDC$Single$)

Human Ability: Reported average expert accuracy is 94%

Application: AI-assisted judicial decision making

Hazard from Immediate Usage:

- *Who:* SCI legal professionals
- *Hazard:* Accepting a claim that should be rejected or rejecting a claim that should be accepted

- *Probability:* Moderate (based on the 2021 model's performance; Malik et al., 2021)
- *Severity:* High
- *Risk:* High

Hazard from Downstream Impact:

- *Who:* Appellants/petitioners; Respondents
- *Hazard:* Getting a wrong decision for their claim; Wrongful accusation/defamation
- *Probability:* Moderate (based on the 2021 model's performance; Malik et al., 2021)
- *Severity:* High (for both hazards)
- *Risk:* High (for both hazards)