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

Fateme Hashemi Chaleshtori **Atreya Ghosal** **Alexander Gill**
Purbid Bambroo **Ana Marasović**
Kahlert School of Computing
University of Utah
fateme.hashemi@utah.edu

Abstract

Is explainability a false promise? This debate has emerged from the insufficient evidence that explanations help people in situations they are introduced for. More human-centered, *application-grounded* evaluations of explanations are needed to settle this. Yet, with no established guidelines for such studies in NLP, researchers accustomed to standardized proxy evaluations must discover appropriate measurements, tasks, datasets, and sensible models for human-AI teams in their studies.

To aid with this, we first review existing metrics suitable for application-grounded evaluation. We then establish criteria to select appropriate datasets, and using them, we find that only 4 out of over 50 datasets available for explainability research in NLP meet them. We then demonstrate the importance of reassessing the state of the art to form and study human-AI teams: teaming people with models for certain tasks might only now start to make sense, and for others, it remains unsound. Finally, we present the exemplar studies of human-AI decision-making for one of the identified tasks — verifying the correctness of a legal claim given a contract. Our results show that providing AI predictions, with or without explanations, does not cause decision makers to speed up their work without compromising performance. We argue for revisiting the setup of human-AI teams and improving automatic deferral of instances to AI, where explanations could play a useful role[1].

1 Introduction

Decision makers can make use of imperfect models if they can detect when they 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. There is a prevailing perspective that this now needs to change since explainability methods passed proof-of-concept tests (see Human-centered Evaluations of Explanations Tutorial). However, given that this is a nascent NLP research space and the notable variation among prior studies (see Table 5), choosing explanation evaluation measurements, tasks, models to explain, baseline conditions, and many other study design choices is not straightforward. 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 answer this, in §3, we establish criteria to assess each dataset's suitability for computing explanation usefulness with the measurements overviewed in §2. We discover that 17/53 datasets are apt for studying appropriate reliance and complementary human-AI team performance but only involve low risk. 4/53 additionally involve higher risk and do not have quality concerns. We recommend using these four, as high-stakes scenarios benefit more from effective explanations.

We introduce the final criterion for dataset selection based on how the likelihood of hazards, and therefore risk, changes if the performance of a state-of-the-art model peaks or is so low the model cannot be collaborated with. We show that this criterion requires continuous assessment of model performance in a rapidly evolving field like NLP.

Finally, we present explanation usefulness studies for a task identified by our meta-analysis: verifying a legal claim given a contract. These serve as exemplars to NLP researchers planning similar

[1] Our code and human study templates: https://github.com/utahnlp/nlp-explanation-utility-guideline/

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 7456–7504

studies. We use both the common human-AI setup, where people make all final decisions with AI assistance, and an overlooked setup, where people decide only for those referred to them by a deferral model. We isolate the effect of explanations, use strong baseline conditions, deploy multiple cognitive forcing functions, integrate real-world situations, and implement attention checks tailored to the application. We show that input highlights and influential training examples do not improve human decision making assisted with the model's predictions and confidence. However, we find that major advances are needed for functional human-AI teams in NLP, *regardless of explanations*. We discover that successful deferral is a promising direction to this end.

2 Review of Application-Grounded Explanation Evaluation

In Table 5 (Appendix), we overview prior human-centered application-grounded evaluations of explanations of NLP models (Lai and Tan, 2019; Feng and Boyd-Graber, 2019; González et al., 2021; Bansal et al., 2021; Schemmer et al., 2023; Mozannar et al., 2023; Joshi et al., 2023; Si et al., 2024). There is a notable variation in the choice of models explained, explanations used, evaluation measurements, baseline conditions, datasets, and outcomes among them. To conclusively establish, or disprove, the value of explanations for human-AI decision-making in NLP, more research is needed together with a more rigorous evaluation protocol designed to collectively guide us towards settling this matter. To assist with running such studies, we start with an overview of suitable explanation usefulness measurements as they are not yet widely adopted in NLP research.

Taxonomy of explanation evaluation. Doshi-Velez and Kim (2017) categorize evaluations of explanations as: (1) *proxy* (no humans, proxy tasks; e.g., the proportion of all features selected as important), (2) *human-grounded* (with humans, but simplified tasks; e.g., simulatability), or (3) *application-grounded* (with humans, realistic tasks; e.g., human-AI decision making). Human-AI decision-making, which is the focus of this paper, is one of the six usage contexts within explainable AI (Liao et al., 2022). Forward and counterfactual simulatability, that are common in NLP (Xie et al., 2022; Arora et al., 2022), are human- but not application-grounded; Buçinca et al. (2020) show that explanations affect simulatability and human-AI decision-making differently.

Reliance definitions. 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. Explanations could also aid people in rejecting incorrect predictions, thereby countering **overreliance**. This becomes possible when explanations present information that appears illogical, self-contradictory, or inconsistent with what the person already knows. The ultimate goal is **appropriate reliance** — have people accept correct predictions and dismiss erroneous ones. A gain in the average rate at which people do so upon seeing explanations quantifies their usefulness.

Fok and Weld (2023) define desired reliance behavior based on expected performance, which implies it is fine to accept a "super-human" model's predictions (or reject a "sub-human" model's predictions) even if some are wrong (correct).

Measuring reliance. Researchers rarely ask people to accept/reject predictions to measure reliance, except González et al. (2021). Instead, people are often shown model predictions and asked to make the final decision.[2] Overreliance is measured by how often the final decision agrees with the model's when it is wrong (Vasconcelos et al., 2023). A possible confounder is that people might make the same wrong decisions as the model, not because they are blindly following it, but because they genuinely find the same wrong answer to be correct.

Schemmer et al. (2023) thus propose that participants first make a guess unassisted, then reevaluate upon viewing the model's prediction. They propose reporting the fraction of times a person (1) flips their initial, wrong judgment after seeing a correct model prediction, and (2) sticks with their initial, correct judgment after seeing a wrong model prediction.[3] Wang and Yin (2021) aim for a similar procedure but allow scrolling to the model prediction while making the first guess, potentially influencing the standalone guess. Joshi et al. (2023)'s approach is similar to (1), but a person needs to flip their initial, wrong answer to the correct one upon

[2] The accept/reject setup might seem a step removed from actual decision making, thus less grounded in application.

[3] The switch percentage (Zhang et al., 2020b) is related.

seeing AI's explanation but not its prediction.[4]

This approach to measuring reliance where people make all final decisions with AI assistance is also how **complementary human-AI team performance** is typically measured. Human-AI teams should surpass the accuracy of both the AI alone and the human alone (Bansal et al., 2021), and explanations could provide a boost. For this to even be possible, the performance of a state-of-the-art model or time-constrained people alone should not already peak. It should not be too low either, because then collaborating with such a model would not be advisable. Instead of reporting the gain (if any) of teaming up and explanations, Feng and Boyd-Graber (2019) perform a regression analysis.

Providing a model prediction with its confidence is a simple, yet stronger **baseline condition** for explanation usefulness compared to predictions only (Bansal et al., 2021). However, not all prior studies test the explanations relative to displaying model confidence; see Table 5 (Appendix). When measuring reliance or complementary performance, it is common to ask annotators to self-report their **confidence** in their decisions and **trust** in the AI model on a case-by-case basis or as a post-task survey.

Deferral. Integrating a deferral model (Dvijotham et al., 2023), $\mathcal{M}_D$, that decides whether an instance can be correctly processed by a prediction model, $\mathcal{M}_P$, presents an alternative to having people make all final decisions with AI in the loop. Explanation usefulness has not been studied in this human-AI team setup. Feeding explanations to $\mathcal{M}_D$ could enhance its correctness, if explanations indeed indicate when $\mathcal{M}_P$ is correct or wrong, as commonly assumed. They could also assist the human reviewer that gets a small fraction of instances deemed hard for $\mathcal{M}_P$. Knowing that a highly accurate model likely made a mistake on a given example, along with its reasoning for it, can nudge the reviewer to consider why the model erred and preempt them from making the same mistake. Surprisingly, we find that these explanations point to the correct evidence (§5).

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 53 existing datasets introduced for developing and evaluating explanations in NLP (§3.2). We refer to a task as its realization 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: Task instances require a notable effort from people, or people are bad at them.

An application must exist if the goal is to evaluate the usefulness of explanations within the context of human-AI interaction for that application. For example, COMMONSENSEQA (Talmor et al., 2019), one of the most popular datasets in NLP explainability research, 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". A dataset might be related to an application, but the task could be narrowed down to take inputs that would never realistically occur. For instance, 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 an unverified claim post-deployment. Finally, if a task instance is easy and quick to handle, collaboration with AI is unnecessary: the person already knows what to do, making the reliance irrelevant. For example, 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.

These three criteria are sufficient if the sole focus is on reliance/complementary performance. However, the definition of human trust in AI (Jacovi et al., 2021) implies that trust inherently involves risk, as one cannot accept vulnerability when none exists. Thus, studying human trust in AI demands an extra criterion:

[4] Joshi et al. (2023) also study whether model explanations support the human ability to reason about new situations where the same logic applies, like human explanations do (Blanchard et al., 2018; Vasilyeva and Lombrozo, 2022). However, their approach cannot be applied once the model is deployed.

> 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 sound studies of reliance and human-AI teams, **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. The necessity of the data selection criterion c_4 might be questioned if risk can only be simulated. We expect simulations of risk to provide more meaningful insights about high-risk scenarios than using a task that inherently lacks any moderate risk.

3.2 Categorization of ExNLP Tasks

We analyze 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.[5] In Appendix F, we report details of our decisions for each task and provide an overview in Table 1. We use ■ if a benchmark criterion is satisfied, and □ otherwise. A suitable dataset for application-grounded 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 30/53 (56.6%) datasets have an associated application and realistic inputs, i.e., fulfill the central requirement for *application*-grounded evaluations, but 23/53 (43.4%) do not.

		$c_1 \wedge c_2$	c_3	c_4	
W1,6	D1	■	■	□	✓
W2	D2	■	■	□	✓
W3	D3	□	■	-	
	D4	■	□	□	
W4	D5	■	■	□	✓
W5	D6	■	■	□	✓
	D7	□	□	-	
W7	D8	□	■	□	
W8	D9	■	□	□	
	D10	■	□	□	
	D11	□	■	-	
	D12	■	□	□	
	D13	□	□	-	
	D14	□	□	-	
	D15	□	■	-	
	D16	□	■	-	
	D17	■	■	■	⋆*
	D18	■	■	■	⋆*
	D19	□	□	-	
	D20	□	□	-	
	D21	□	□	-	
	D22	■	■	□	✓
	D23	■	-	-	
	D24a	■	□	□	
	D24b	■	□	□	
	D25	□	□	-	
	D26	□	□	-	
	D27	□	□	-	
	D28	■	■	□	
	D29	■	■	□	✓
	D30	□	□	-	
	D31	■	■	□	✓
	D32	■	■	□	✓
	D33	□	□	-	
	D34	■	■	□	✓
	D35	■	□	□	
	D36	■	□	□	
	D37	■	-	-	
	D38	■	■	■	⋆
	D39	□	□	-	
	D40	■	□	□	
	D41	■	-	-	
	D42	■	■	■	⋆
	D43	□	□	-	
	D44	■	■	■	⋆
	D45	□	-	-	
	D46	□	■	-	
	D47	■	-	□	
	D48	■	-	-	
	D49	■	■	■	⋆
	D50	■	■	■	⋆
	D51	□	■	-	
	D52	□	-	-	

Table 1: Categorization of ExNLP datasets. See §3.2 for a description of symbols.

[5] https://exnlpdatasets.github.io/; We excluded datasets in non-English languages and multimodal datasets.

Do ExNLP tasks require notable human effort? Are people skilled at solving these tasks? We estimate effort using the avg. length of task inputs, anticipating that longer inputs demand more effort. The maximum average length that we decide does not need notable effort is 272 words, taking around a minute to read (Rayner et al., 2016). We estimate human ability using reported human performance when available. We find that 25/53 (47.2%) tasks either require notable effort or people do not excel at it, the data is not available for 2/53 (3.8%), we are not able to estimate the human ability for 5/53 (9.4%), and for 21/53 (39.6%) inputs are too short while people do the task well. Of 25 requiring notable effort or people are not good at them, 17/53 (32.1%) also have associated applications and realistic inputs. That is, 32.1% 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 answer this question from the perspective of 2 stakeholders: (i) people acting 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.1). 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 (ILDC; D50).

3.3 Task Checks with Model Performance

The final check for a dataset's suitability is based on a chosen model performance:

> c_5: The model performance should be high enough to warrant collaboration, but not so high that it can operate effectively on its own without human oversight.

If a model rarely makes mistakes, the likelihood and thus risk of hazards are typically low. These are tasks we recommend deprioritizing. Fok and Weld (2023) argue that using the predictions is viable in this case. However, a worse model should not be chosen when a better-performing one is available and resource-appropriate.

	Precision	Recall	F1 score
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
CONTRADICT	-	-	35.7
MACRO (E,C)	-	-	59.6

(a) ContractNLI

	Precision	Recall	F1 score
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

(b) EvidenceInference v2 *with retrieval*

Table 2: Finetuned Flan-T5-3B and the state-of-the-art reported results. Koreeda and Manning (2021) do not report finetuned BERT-Large F1 for the "No-mention" class. EvidenceInference is proposed without retrieval, so there is no prior model performance to report. Results for SciFact and ILDC are in Table 7 (Appendix). Macro F1 averages F1 scores across classes, while Micro F1 calculates F1 from aggregated results.

Finetuning large language models (LLMs) is an effective method for specializing a model to a task. We do so for each ⋆ task that fulfills all the criteria in §3.1 and quality checks. We use Flan-T5-3B (Wei et al., 2022) due to its size and versatility stemming from instruction finetuning with data of 1.8K tasks.[6] Details and examples of task inputs to the model (Tables 11–14) are in Appendix C.

Tables 2 and 7 (Appendix) underscore the importance of reassessing baselines. We obtain a

[6]We did not explore prompting because we did not expect notable improvements from it compared to finetuning Flan-T5-3B at the time of writing this paper. We did not train Flan-T5 to self-rationalize with free-text explanations, as human-authored free-text explanations are not available for supervision.

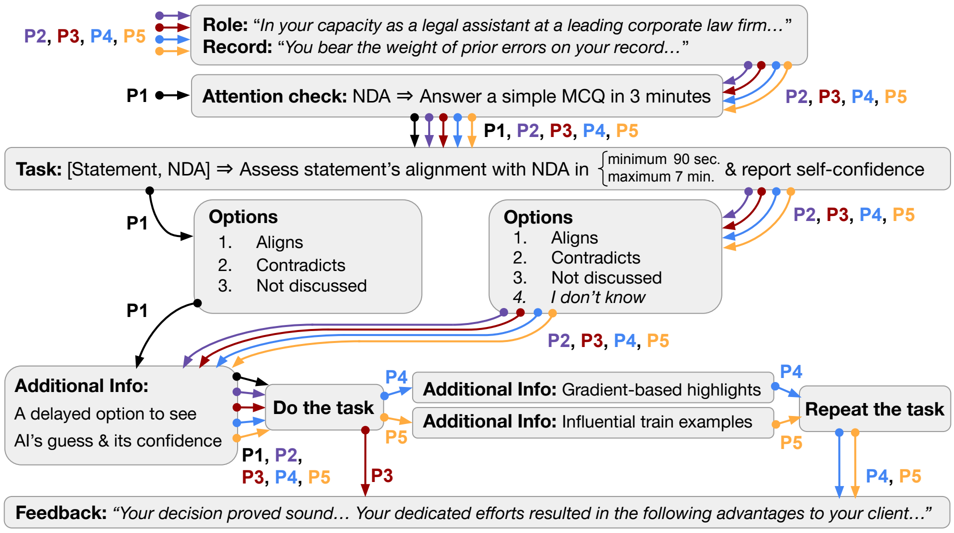


Figure 1: An overview of the five task presentations, P1–P5, used for Study I (see §4.2).

24.5 point improvement in the average contradiction and entailment F1 scores on Contract-NLI (Table 2), and a 22.8 macro-F1 point increase on SciFact-Open (Table 7). Without reassessing performance on these datasets, we would not realize that it now makes sense to team up people with models for these tasks. Moreover, we find that the baselines for ILDC and SciFact-Open — the two datasets with the reported human performance — have not reached peak performance. Thus, human-AI teams might provide benefits over using AI alone. The performance of EvidenceInference-v2 with retrieval remains low, and studying human-AI teams on this task is not justified without a stronger model in the loop.

4 Study I: People Make All Decisions

In §2, we overview two strategies for human-AI decision-making: (1) people make all final decisions with AI assistance, and (2) a deferral model refers only a fraction to people. We aim to provide an exemplar for user study design and initial insights into the usefulness of explanations in improving human-AI decision making in both of these setups for one of the tasks identified in our meta-analysis: verifying claims based on the ContractNLI Non-Disclosure Agreements (NDAs). In this section, we use (1), and in §5, we look at (2).

4.1 Study Design

We overview various design choices we consider that we recommend integrating in future studies.

To isolate the effects of explanations, in the first step, a participant may reveal the prediction and make their first guess. Only then we provide the explanation and ask for their final decision.[7] To encourage thoughtful engagement with the model predictions in the first step, we use three **cognitive forcing functions** (Buçinca et al., 2021). (1) Let participants choose whether to reveal the model's predictions. (2) Delay when the option to reveal the prediction becomes available.[8] (3) Disable the option to move to the next step for a minimum amount of time considered necessary to reasonably attempt the task; in our case, for 90 seconds.

Making hypothetical decisions in a questionnaire differs from making real decisions, where people might reassess their trust and opt for more cautious actions, or not act at all. Consider claim verification. People may temporarily perceive false information to be true, but do not disseminate all such misconceptions. When considering making decisions based on them, one might reflect more deeply and consider their confidence more carefully.

To bring participants **closer to real-world scenarios**, we have them adopt a specific **hypothetical role**, as shown in the example in Fig. 8. Henceforth, all figures except Fig. 1 appear in the Appendix. Additionally, they should have the **don't know option**, which allows them to refrain from making decisions when they normally would not, thereby avoiding an overestimation of overreliance. To prevent using this option unrealistically often, we

[7] In Appendix D.1, we discuss why we do not ask people to make three guesses to isolate the effects of the predictions.

[8] We randomly select the delay time between 50–70 or 30–45 seconds, depending on the study. Randomly to avoid waiting out a fixed amount of time.

		F1($\mathcal{M}$)-F1($\mathcal{H}$) ↓	$\frac{\#\ \mathcal{H}\text{ correct}}{\#\ \mathcal{M}\text{ correct}}$ ↑	$\frac{\#\ \mathcal{H}\text{ correct}}{\#\ \mathcal{M}\text{ wrong}}$ ↑	Avg. # Reveal	Avg. $\mathcal{H}$ Conf.	% IDK
P1	AI + confidence	17.0	76.0	28.1	3.8	3.8	-
P2	↳ + IDK, role, record	18.5	77.3	14.3	3.3	3.8	3.3
P3	↳ + feedback	10.7	84.2	26.9	3.6	3.6	1.7
P4	↳ + highlights	9.6	87.0	14.8	-	3.9	0.6
P5	↳ + influential examples	10.0	86.3	15.4	-	3.8	0.0

Table 3: Human performance for ContractNLI claim verification across conditions in Study I (§4). $\mathcal{M}$ denotes the finetuned Flan-T5-3B, $\mathcal{H}$ human participants, # Reveal for how many of 6 instances annotators reveal the model's prediction, $\mathcal{H}$ Conf. their self-confidence, and % IDK the rate at which they chose the I don't know option.

show them "their" **prior record** that occasionally mentions they previously refrained from making judgments and that they are asked by their superior to decrease that rate; see Fig. 9. Another method to discourage excessive use of the don't know option is to provide **feedback** after a decision is made; see Fig. 12. Moreover, people typically receive some feedback when deciding for others, which in turn influences their future decision making.

We create roles, records, and feedback by prompting `gpt-4-0613` (OpenAI et al., 2024) and revising its generations. We also use it to create a simple multiple-choice question (MCQ) for each NDA serving as an **attention check** in the main annotation tasks. We provide details of these processes in Appendices D.2 and D.3.

Finally, in a small experiment, we check that we recruit annotators who can perform this task alone. We ask 5 participants to annotate 2 legal claims: one with a short NDA and one with a long. Note that in the main study (§4) we have ≈40 participants. The first claim should be accurately assessed given the short NDA, and all annotators meet our expectations. See details in Appendix D.4. For the claims with longer NDAs, 4/5 participants give the correct response. This suggests that annotators with a degree in administration and law who work in legal functions perform well on this task. Thus, they should be **time-constrained** when deciding for all instances; otherwise, the benefit of providing AI's predictions should not be expected.

4.2 Task Presentations

Annotators' task is to assess whether a statement aligns with a given NDA. We test the following five task presentations, outlined in Fig. 1:

- **`(P1;Baseline)`** A delayed option to see Flan-T5's prediction and its **calibrated confidence**.[9]
- **`(P2)`** `P1` information with the don't know option and priming with roles and records.
- **`(P3)`** `P2` information with providing feedback.
- **`(P4)`** `P2` information, then input highlights in the next step, and finally feedback.[10]
- **`(P5)`** `P2` information, followed by influential train examples, and at the end, feedback.[11]

Figures 2–4 show the instructions given to the annotators. In each task setting, participants are given 6 instances. For each, they first answer an attention-check MCQ about an NDA within 3 minutes. If time runs out, they are moved to the next instance.[12] Following the MCQ, they should evaluate a statement based on the same NDA in 7 minutes (Fig. 10) and then report their self-confidence (Fig. 11). In the final two settings, after participants make the first guess, they are shown the model explanation and asked to guess again. In `P4`, participants may see the NDA with top 5%, 10%, or 20% of the important words highlighted (Fig. 13), and in `P5`, the top 3 most influential labeled train examples with `Input×Gradient` highlights incorporated—participants should not have to *fully* review three different NDAs to *quickly* verify if the AI's guess is correct. Appendix D.5 provides more info.

For each study, we aim for 80 examples for each of the three labels. We gather 35-40 participant responses per study, with each participant annotating up to 6 samples. We collect 1108 annotations across all conditions. Participant recruitment is done through Prolific. More on participants in D.6.

4.3 Results

We provide our findings in Table 3. Across all conditions, time-constrained humans ($\mathcal{H}$) collaborating with the model perform at least 9.5 F1 score points worse than the model alone ($\mathcal{M}$). This underscores the need for major improvements in the effectiveness of this type of human-AI teams, regardless

[9]Temperature scaling (Guo et al., 2017) on the dev set.

[10]`Input×Gradient` highlights (Shrikumar et al., 2016) obtained with inseq (Sarti et al., 2023).

[11]EK-FAC influences (George et al., 2018; Grosse et al., 2023) obtained with kronfluence.

[12]Thus, the total num. of examples varies across conditions.

of explanations. We notice that the rate at which experts are correct when the model is also correct is far from 100%. This hesitation among experts to rely on the model is also evident from their decision to reveal the model's prediction in just over half of the six provided examples. P3 (feedback) significantly improves this rate.[13] Despite seemingly being more risk-averse, participants choose "I don't know" with a low rate of at most 3%.

The rate at which people are correct when the model is wrong never exceeds 28%. When both the model and people are wrong, they concur in at least 80% of such cases (not shown in Table 3). Future human-AI teaming should focus on strategies that could notably improve this. On average, people's confidence ranges 3–4 (moderately to highly confident), even when they make wrong decisions.

In Table 9 (Appendix), we present a breakdown of the impact of highlights. As evident by the number of different situations that arise from the people's first guess (1st col.), asking people to reasses their decisions and self-confidence with highlights helps better understand their effects. This analysis confirms that highlights do not consistently meet expectations to mitigate under- and overreliance; in some cases, they may even contribute to it.

Our main takeaway is that the key challenge in this human-AI team setup is achieving that experts spend just enough time on an instance to develop an informed opinion and then effectively use the AI's guess, confidence, and explanation to *quickly* transition from initial opinion to a confident, accurate decision. Only in this way can they speed up their work without sacrificing performance. However, this raises the question: what could help achieve this delicate balance? We suspect that adding more information to the AI's guess is not the solution.

5 Study II: People Decide Only for Deferred Instances

This human-AI decision making setup defers a fraction of all instances to experts. We investigate explanation usefulness for finetuning/prompting LLMs to defer, and to human decision makers.

5.1 Usefulness to Deferral Models

We finetune Llama-2-13B-Chat and gpt-3.5--turbo-1106, and prompt gpt-4o-2024-05-13. We use 1.4K model's dev-set predictions for finetuning, 371 predictions for testing deferral models, and 30 for in-context learning.[14] Wrong predictions are positive (defer) examples and correct ones are negative (don't defer) examples for deferral models. We provide all models with a ContractNLI claim+NDA to defer or not, and the model's label for the claim.[15] Some models also get a few demonstrations and/or Input×Gradient highlights.

Model	Train Setup	Input	Recall
Llama-2-Chat-13B	Finetuned	P+T+H	5.0
		P+T	10.0
gpt-4o-2024-05-13	Zero-shot	P+T	10.0
	30-shot	P+T	17.5
		P+T+H	15.0
gpt-3.5-turbo-1106	Finetuned	P+T	**30.0**

Table 4: Recall of deferral models for the defer-to-expert (AI wrong) class (§5). P denotes the AI's prediction, T denotes ContractNLI instance text, and H highlights.

The main challenge we observe with all models is a low recall for the deferral class, as seen in Table 4, with more results in Table 19 (Appendix). Prompting gpt-4o-2024-05-13 improves recall compared to the finetuned Llama-2-13B-Chat, if a few examples are provided. Finetuning gpt-3.5--turbo-1106 notably improves recall, but remains low, reaching only 30%. This shows that our best deferral model is not functional. Thus, building effective deferral models remains a challenge, and human-AI decision-making in this setup needs notable improvements, much like the previous setup.

Including highlights (details on how we do that in Appendix E) reduces recall, contrary to the belief they might provide additional signal about when the model is correct/wrong.

5.2 Usefulness to Experts

We still aim to provide initial insights into whether explanations for likely-wrong predictions help decision makers, as we hypothesize in §2. Without workable deferral models, we do not study this in a realistic setup with deferral mistakes; all participants get instances that the model mishandles. Given the lack of usefulness of explanations so far, we opt for a small-scale study and include a free-form question about how explanations are helpful, if at all. We use the setups in P4 (w/ highlights) and P5 (w/ highlighted influential examples) in §4, but

[13] A two-proportion z-test yields p-value of 0.03 for P3.

[14] 10 positive and 20 negative as fewer examples should be deferred to experts.

[15] We exclude confidence because we find no correlation between it and prediction accuracy (see Fig. 15).

modify the instructions to warn participants that AI most likely mislabeled examples they will review (see Fig. 6). We end with 23 participants for the former and 20 for the latter, each annotating 2 instances. The final number of annotated instances is 73 because some annotators reach the time limit.

Participants, on average, rate the impact of explanations on their decision making with 2.56 (slightly to moderately). For 55/73 statements participants explain whether and how explanations are helpful. We categorize these insights in Table 18 and provide examples. While explanations of a likely-wrong model can encourage different reasoning approaches as we hypothesize, highlights are more often helpful in finding relevant information in an NDA despite wrong predictions.[16] These results suggest that the model finds necessary information, but uses it in a way that leads it to a wrong decision. Thus, rather than suggesting to annotators that these explanations may reflect likely flawed reasoning, we should inform them that the model could identify the correct evidence yet still arrive at the wrong conclusion.

We notice a low F1 score of only 34.2% in this study (Table 20, Appendix), which is a stark contrast to the previous small-scale study of people operating alone (§4.1). Five participants state they disagree with feedback mentioning the correct label (see Appendix E). Therefore, one author of this paper has checked 29 examples for which participants respond about their use of highlights. The gold label is changed if the author's reasoning matches the participant's reasoning and they disagree with the gold label. This occurs in 11/29 cases, suggesting that examples that are challenging to the model might also have noisy labels. Table 17 shows one such sample. Using the revised gold labels, participants' F1 score increases from 34.9% to 79.4% on this reviewed subset (29 samples), and AI's F1 score from 0% to 27.3%. Future work should be cautious with deferred instances (where AI is wrong) because these likely present the tail-end distribution of the dataset.

Finally, we evaluate the human-AI team with the default gold labels. 237/2091 ContractNLI test instances (11.3%) are mislabeled by the model and would be deferred to experts by a perfect deferral model. We assume that the experts' performance on all 237 deferred samples is similar to the 34.2% F1 score they achieve on the 73 samples they annotated. 88.7% (AI correct) * 100% (AI F1) + 11.3% (AI wrong) * 34.2% (experts F1) gives a joint performance of 92.6% under the perfect deferral model. This is 13.9 points higher than the best team in §4, and would additionally increase with cleaner labels. Based on the average time needed for reviewing a statement-NDA pair in §4 (212 seconds), we determine that the human decision makers' time needed for deferred instances (11.3% of the data) would be 14 hours, while for all instances it would be 123 hours.

These results suggest that achieving workable deferral is a promising way to advance human-AI teams. A new goal for explanations could be to help a deferral model to identify when the predictor is likely accurate. Moreover, given the challenges of the other human-AI team setup discussed in §4.3), we find the deferral setup more feasible. Here, the primary challenge is technical: developing accurate automatic deferral. Once this is done, domain experts can use their time to accurately assess a small set of instances, which we know they can do.

6 Conclusions

We provide guidelines for specifying a sound experimental setup for application-grounded evaluation of explanations in NLP. The dataset selection criteria we set should be used to determine a dataset's suitability for such evaluations. Future studies should consider 4 datasets we identify by validating over 50 datasets against our criteria, or validate/create new datasets guided by them. We show that the performance of a resource-appropriate state-of-the-art model should be continuously reassessed: it should neither peak nor be low, otherwise teaming the model with people is unsound. We recommend following our user study design that isolates the effect of explanations, uses strong baseline conditions, deploys multiple cognitive forcing functions, integrates real-world situations, implements attention checks tailored to the application, among other things.

We believe that the main reason AI assistance is unhelpful to specialists, regardless of explanations, is the difficulty of making the right decision without investing enough effort into a given problem. Yet, more effort makes teaming with AI unnecessary. We argue that integrating deferral could be a more viable approach to speeding up work without sacrificing overall performance.

[16]Another strong baseline could be offering "model-free" highlights based on the statement-NDA lexical overlap.

7 Limitations

The dataset selection criteria we set (§3.1) could have been used to validate some promising datasets beyond the resource with existing explainability-research datasets we used. The application of the criteria could also be enhanced. For example, we approximate effort with length, but shorter examples people can solve could also require effort, e.g., certain math problems. On the other hand, finding an answer quickly in a longer task instance might be possible with a keyword search. Future work should continue to improve methods for quantifying human effort and ability. We subjectively determine hazard severity, and sometimes likelihood, which leaves room for disagreement. However, acknowledging this, we provide higher risk when less confident, such that a dataset is not unfairly filtered. While newer models than Flan-T5 could have been finetuned in §3.3, and consequently not all of the 4 datasets might remain as suitable, the main takeaway of this section remains: reassess the performance of a model you plan to team people with. Finally, our user studies face challenges all user studies do. Specifically, they would be stronger with more participants, examples annotated, explanation types, and datasets evaluated, among other things, which is not possible due to financial restrictions. Despite our best efforts to provide reliable outcomes, as in most studies with human subjects, we cannot guarantee there are no confounders.

Acknowledgments

We thank anonymous reviewers for their useful feedback, 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, Nazanin Hashemi Chaleshtori for her assistance with understanding examples in medical domain, UtahNLP group, 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 non-monotonic 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.

Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To trust or to think: Cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1).

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.

Krishnamurthy Dvijotham, Jim Winkens, Melih Barsbey, Sumedh Ghaisas, Robert Stanforth, Nick Pawlowski, Patricia Strachan, Zahra Ahmed, Shekoofeh Azizi, Yoram Bachrach, Laura Culp, Mayank Daswani, Jana von Freyberg, Christopher J. Kelly, Atilla P. Kiraly, Timo Kohlberger, Scott Mayer McKinney, Basil Mustafa, Vivek Natarajan, Krzysztof J. Geras, Jan Sylwester Witowski, Zhi Zhen Qin, Jacob Creswell, Shravya Shetty, Marcin Sieniek, Terry Spitz, Greg C. Corrado, Pushmeet Kohli, taylan. cemgil, and Alan Karthikesalingam. 2023. Enhancing the reliability and accuracy of ai-enabled diagnosis via complementarity-driven deferral to clinicians. *Nature Medicine*, 29:1814–1820.

Julian Eisenschlos, Bhuwan Dhingra, Jannis Bulian, Benjamin Börschinger, and Jordan Boyd-Graber. 2021. Fool me twice: Entailment from Wikipedia gamification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 352–365, Online. Association for Computational Linguistics.

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.

Thomas George, César Laurent, Xavier Bouthillier, Nicolas Ballas, and Pascal Vincent. 2018. Fast approximate natural gradient descent in a kronecker factored eigenbasis. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 9573–9583.

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.

Roger B. Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, Evan Hubinger, Kamile Lukosiute, Karina Nguyen, Nicholas Joseph, Sam McCandlish, Jared Kaplan, and Samuel R. Bowman. 2023. Studying large language model generalization with influence functions. *CoRR*, abs/2308.03296.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

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.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

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.

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.

Hussein Mozannar, Jimin J. Lee, Dennis Wei, Prasanna Sattigeri, Subhro Das, and David A. Sontag. 2023. Effective human-ai teams via learned natural language rules and onboarding. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, and Yinfei Yang. 2022. Large dual encoders are generalizable retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9844–9855, Abu Dhabi, United Arab Emirates. 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.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

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.

Gabriele Sarti, Nils Feldhus, Ludwig Sickert, and Oskar van der Wal. 2023. Inseq: An interpretability toolkit for sequence generation models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 421–435, Toronto, Canada. 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.

Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. 2016. Not just a black box: Learning important features through propagating activation differences. *CoRR*, abs/1605.01713.

Chenglei Si, Navita Goyal, Sherry Tongshuang Wu, Chen Zhao, Shi Feng, Hal Daumé III, and Jordan L. Boyd-Graber. 2024. Large language models help humans verify truthfulness - except when they are convincingly wrong. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

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.

Appendix Overview

In this supplementary material, we provide:

- **Appendix A.** Discussion on the lack of explanation evaluations in NLP that are application-grounded.
- **Appendix B.** Additional analyses of the quality and suitability of three datasets highlighted by our meta-analysis.
- **Appendix C**. Details about the data processing and training of Flan-T5-3B on four datasets.
- **Appendix D.** Information on various components of our human studies and additional results complementing Study I (§4).
- **Appendix E.** Details on preparing the input data for deferral (§5) and participants' feedback on the study's verdict for instances they annotate.
- **Appendix F.** Details of our reasoning behind the categorization of each dataset.

A Discussion: Scarcity of Application-Grounded Explanation Evaluations in NLP

Why are application-grounded evaluations of explanations currently limited in NLP? Such evaluations of explanations have predominantly been done 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 LLMs outperform inherently interpretable models (e.g., linear models, short decision trees). Prior NLP explainability work has mostly focused on overcoming these challenges. Moreover, many realistic language technology applications have become evident and possible with LLMs in the past two years.

B Quality Analysis of Filtered Datasets

In this section, we explain why we exclude three datasets, although they meet the necessary criteria for studying the impact of explanations on human-AI teams.

B.1 Analysis of LIAR-RAW (D17) and RAWFC (D18)

The goal of this task is to assess the veracity of statements about a diverse range of topics, using a handful of reports as references. A few issues related to data quality became apparent after conducting a manual examination of a randomly selected sample from the dataset. Notably, it appears that perhaps during data processing stages, all instances of "to be" verbs are replaced with "be", sentences and phrases are truncated, unnecessary repetitions of sentences, and other grammatical problems are identified. We evaluate the quality of the data by analyzing a sample of 100 data points randomly selected from the dataset.

A single author carefully reviews 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 utterances with the claim. Out of the 100 data points reviewed, 38 are deemed acceptable.[17] We find similar issues with RAWFC (D18). We thus exclude these datasets.

B.2 Analysis of UKPSnopes (D42)

In this task, the objective is to evaluate the veracity of claims across various domains. Each claim is accompanied by a fact-checking report sourced from the fact-checking website, Snopes[18]. This configuration, however, does not mirror real-world situations since a fact-checking report may not always be available for every new claim. Instead, one needs to retrieve relevant documents. Hanselowski et al. (2019) note that relevant articles can be obtained from links within the Snopes fact-checking reports. To this end, we compile all linked articles into a corpus of size 13K. We retrieve the most relevant documents from this corpus for each claim and then finetune the Flan-T5-3B model for claim verification using this information.[19] For finetuning details, see how we approach this for other retrieval tasks in Appendix C.

Our retrieval model achieves a recall of 51% for claims in the training set. The micro-F1 score

[17] We additionally calculate the perplexity for these data points using the `gpt-2-XL` (1.5B) model (Radford et al., 2019). We could use perplexity for data filtering if it is correlated with data acceptability. We get a low correlation between them.

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

[19] Each claim in the dataset is matched with multiple snippets from the associated Snopes article, with each (claim, snippet) pair receiving a fact-checking label. We observe that approximately 15.2% of claims in the training set have different inconsistent labels when matched with different snippets of the same Snopes article. We exclude these claims due to the uncertainty in determining the gold label in the absence of the accompanying Snopes report and pairing them with other gathered articles.

	Task	Usefulness Measurements	Explanations Evaluated	Models Explained	Baseline Condition	Helpful
W1	D1	Increased team performance	Input attribution from SVM's weights; k-NN train examples	Linear SVM with BoW features	Providing the accuracy of the SVM	Yes
W2	D2	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
W3	D3; D4	Increased 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	Increase in % accept correct; Decrease in % accept wrong	Manually extracted evidence independent of the model†	DPR (Karpukhin et al., 2020)	Post-hoc calibrated model confidence	Yes
W5	D6; D7	% initially wrong, correct after the explanation	Free-text explanations	T5-large (FT-full); T5-3B (FT-128); `davinci-instruct-beta` (ICL-6)	None	No
W6	D6	% initially wrong, flip after the model's guess; % stick with initial correct guess	LIME input attribution	SVM	None	Yes
W7	D8	Increased team performance Decreased task time	Free-text explanations	Flan-T5-XLarge	None	No
W8	D9	Increased team performance Decreased task time	Free-text explanation; Contrastive explanations; Retrieved passages	ChatGPT + GTR-XXL (Ni et al., 2022) retriever	None	No

Table 5: Overview of prior application-grounded explanation usefulness evaluations involving a text-based task. W1 (Lai and Tan, 2019); W2 (Feng and Boyd-Graber, 2019); W3 (Bansal et al., 2021); W4 (González et al., 2021); W5 (Joshi et al., 2023); W6 (Schemmer et al., 2023); W7 (Mozannar et al., 2023); W8 (Si et al., 2024). FT stands for "finetuned" and ICL for "in-context learning".

	Precision	Recall	F1 score
SUPPORT	42.3	100.0	59.4
NO-INFO.	80.0	1.6	3.1
CONTRADICT	0.0	0.0	0.0
MICRO AVG.	40.8	33.9	20.9
MACRO AVG.	52.4	42.7	26.3

(a) Using automatically retrieved articles

Precision	Recall	F1 score
44.0	84.3	57.8
49.6	21.2	29.7
0.0	0.0	0.0
31.2	35.2	29.2
40.0	44.6	37.2

(b) Using articles linked in an associated Snopes fact-checking report

Precision	Recall	F1 score
81.6	75.3	78.3
76.0	82.6	79.1
67.1	62.82	64.9
74.9	73.6	**74.1**
79.7	77.8	78.6

(c) Using a Snopes fact-checking report

Table 6: Finetuned Flan-T5-3B results for claim verification in the UKPSnopes dataset across three input setups: with automatically retrieved articles, with articles linked in a fact-checking report for the claim, and with a fact-checking report for the claim.

for claim verification using Flan-T5-3B finetuned with retrieved documents is only 20.9 (Table 6a). We conduct additional experiments to investigate what is responsible for this: the compiled corpus of documents for retrieval is insufficient and therefore the dataset is not suitable, or retrieval is hard for this task. We omit the retrieval step and finetune the Flan-T5-3B model using (1) the gold articles linked within the Snopes fact-checking report about the claim, and (2) the gold Snopes fact-checking reports. We get the micro-F1 score of only 29.2 with the articles linked in the fact-checking report (Table 6b), but 74.1 F1 with the report (Table 6c). This indicates that the corpus of linked articles lacks the necessary information to address the task effectively, contrary to Hanselowski et al. (2019) hypothesis.

C Details of Model Finetuning

In Tables 11–14, we provide illustrative instances demonstrating how we craft the input for each baseline model we develop following the recommended templates.[20] Below, we describe the process of

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

	Precision	Recall	F1 score
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

	Precision	Recall	F1 score
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

(b) SciFact-Open

Table 7: Finetuned Flan-T5-3B and the state-of-the-art reported results. XLNet+BiGRU (Malik et al., 2021); MultiVerS (Wadden et al., 2022b). Wadden et al. (2022b) estimate the human performance in the "abstract-provided" setting. ContractNLI and Evidence-Inference v2 (w/ retrieval) results are in Table 2.

	Precision	Recall	F1 score
INCREASE	87.8	91.3	89.5
NO-DIFF.	90.8	87.8	89.3
DECREASE	87.5	87.9	87.7
MICRO AVG.	89.0	89.0	**89.0**
MICRO AVG.	88.7	89.0	88.8

Table 8: ERASER EVIDENCEINFERENCE task performance with finetuned Flan-T5-3B when gold documents are provided to the model.

finetuning Flan-T5-3B for each of the four datasets highlighted in our meta-analysis.

EvidenceInference v2 (D38) This 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 articles 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 of articles. Our approach consists of these two steps:

- Use the BM25Plus algorithm (Trotman et al., 2014) to get the top 100 relevant documents for each query, after which rerank those 100 with the method introduced by Nogueira et al. (2020), and finally select the top 10.
- Finetune Flan-T5-3B using the query and the top 10 documents obtained as input.

Our retrieval module has a low recall rate of 3%, 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, and we get a recall score as high as 51%. Table 2b presents the results of finetuning the Flan-T5-3B model with retrieved documents, and Table 8 shows how well our model performs when finetuned with the true relevant document instead of retrieved ones. The significant difference in F1 scores between using the true relevant documents and the retrieved documents underscores the retrieval challenge, indicating the need for stronger retrieval models.

SciFact-Open (D44) This is another fact-checking task, but the claims are limited to the scientific domain. To finetune the Flan-T5-3B model, we follow the same two-step approach described in the previous task. We extract the top 10 most pertinent documents related to each claim from a corpus of 500K research abstracts. See Table 7b for results.

ContractNLI (D49) Given a contract and a set of legal statements, the objective is to determine whether each statement implies, contradicts, or remains neutral in relation to the contract. This is a three-class classification task, with "Yes" signifying the statement entailment to the contract, "No" denoting contradiction with the contract, and "Cannot say" indicating the statement is undiscussed within the contract. To prepare data for finetuning, we concatenate the contract and statement (example in Table 12), ensuring the statement remains in the input by truncating the left side. Table 2a presents the model performance.

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 $\text{ILDC}_{\text{single}}$ instances as we

$\mathcal{M}$-$\mathcal{H}_1$	$\mathcal{H}_2$	# Conf. Δ	Effect of h
c-w [46]	c [11]	-	corrects underreliance
	w [36]	4 ↑ 5 ↓ 27*	reinforces underreliance discourages underreliance no effect
w-c [8]	c [5]	1 ↓ 4*	discourages self-reliance no effect
	w [3]	-	causes overreliance
c-c [323]	w [15]	-	causes underreliance

Table 9: Effects of highlights (h). c / w stand for correct/wrong decisions. $\mathcal{M}$-$\mathcal{H}_1$ shows the correctness of a model prediction and the 1st human guess; $\mathcal{H}_2$ the correctness of the 2nd guess upon seeing h. ↑ / ↓ / * means the human confidence increased / decreased / stayed the same after seeing h. The numbers in brackets show counts for each scenario.

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 tokens. Results are presented in Table 2b.

D Additional Details and Results of ContractNLI (D49) User Study I (§4)

In this section, we provide additional details that complete §4 and §5. We design our studies using the Qualtrics online survey maker.[21] In Table 9, we provide a breakdown of the impact of highlights that is discussed in 4.3.

D.1 On Isolating Effects of Both AI's Predictions and Explanations

Schemmer et al. (2023)'s two-step approach to measuring reliance (§2) accurately isolates the impact of model predictions on final human decisions. However, applying this approach to tasks identified by our meta-analysis requires extra considerations.

The users of models trained for these tasks are experts, not laypeople, as task instances are highly specialized (see examples in Tables 11–14 such as the SciFact statement "*A high microerythrocyte count raises vulnerability to severe anemia in homozygous alpha (+)- thalassemia trait subjects.*"). These tasks require notable effort, but experts are skilled (e.g., ILDC experts' average accuracy is 94%). Thus, if experts are asked to make the initial guess without time constraints, human-AI teams likely will not outdo experts alone, making human-AI teaming unwarranted. However, if time constraints are imposed then AI can be helpful by assisting experts to make accurate decisions more quickly and confidently, i.e., teaming makes sense. Therefore, unlike almost all prior studies in NLP that involve only laypeople, application-grounded evaluation with our highlighted tasks should focus on *time-constrained* decision-making by experts. Experts are more expensive, so the number of instances and participants may be different from studies with laypeople.

Schemmer et al. (2023)'s robust protocol should be further extended (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 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. The breakdown of all possibilities is in Table 10. However, by adding this step, annotators need to make three guesses for the same examples, which is burdensome and makes it less likely that they will change their decision by the third guess. For evaluating explanation usefulness, we believe it is more important to reliably isolate the effect of showing the explanation. We thus recommend first asking annotators to make a guess, giving them an option to reveal the model's prediction, and then asking them to guess again upon seeing the explanation.

D.2 Generating Roles, Records, and Feedback

We aim to encourage participants to envision themselves in a realistic scenario, shift away from a typical crowdsourcing task, and approach it as they would in real life. To this end, we investigate the effect of presenting the participants with a hypothetical role and the consequences of "their" past errors in that role. Here, we explain how each piece of the aforementioned information is generated.

Roles (Condition `P2`). The role is exemplified by scenarios depicting a paralegal assistant's tasks

[21] https://www.qualtrics.com/

within a corporate law firm, particularly involving work on NDAs. We add cautionary notes regarding the potential negative consequences of both errors and inaction. We request `gpt-3.5-turbo` to refine and expand a manually crafted draft of the described scenario and warnings using more formal and technical language. Subsequently, we instruct `gpt-3.5-turbo` to reword the text, resulting in four variations. Fig. 8 shows an instance.

Record (Condition P2). We construct the record information by integrating three components:

- We simply mention that the participant has prior errors. We use `gpt-3.5-turbo` to articulate this segment and generate four alternative phrasings to avoid redundancy.
- Providing one of the hypothetical situations described above to `gpt-4-0613`, we ask it to generate a list of potential consequences of making mistakes in this context. We end up with fourteen different consequences. Here is an example:
  `Subjecting the client to legal liabilities and the possibility of facing lawsuits, increasing their financial burdens and legal complications.`
  We present one or two randomly chosen consequences from the aforementioned list of 14, allocating two consequences for 50% of the samples and one consequence for the remainder.
- We underscore the importance of avoiding indecision — equivalent to selecting the don't know option — as a common practice since it is deemed unacceptable. We randomly select a percentage between 20 and 35 to represent the frequency with which the participant has abstained from making decisions in the past, resulting in reprimands from their supervisor. This last part is added to 40% of the samples.

In Fig. 9, we show an example containing two consequences and an inaction warning.

Feedback (Condition P3). We generate feedback based solely on the statement. Given three labels, there are 9 potential combinations of the statement's true label and what a participant might decide the label is. For each of the 17 unique statements in the ContractNLI dataset, we create feedback corresponding to each of the nine label combinations, resulting in a total of 153 unique feedback instances. To generate feedback, we prompt `gpt-4-0613` in multiple iterations. First, we provide each of the 17 statements paired with each of the 3 true labels. We ask `gpt-4-0613` to imagine it is a client who owns a company entering into a non-disclosure agreement with another company, and ask it to generate actions the client should take to adhere to the provided pair of statement and label. Table 15 illustrates one example per gold label. In the next iteration, we provide `gpt-4-0613` with the triplet of the statement, each of the three possible true labels, and each of the three labels a participant might choose, along with the actions provided in the previous step. We inform `gpt-4-0613` whether the actions were fulfilled or violated based on the match between the true label and the participant label, and we ask it to suggest a legal penalty or potential benefit based on the participant's mistakes or sound judgment. Additionally, we use `gpt-4-0613` to rephrase feedback according to the frequency of (statement, true label) pairs in the data. Table 16 provides two examples of instructing `gpt-4-0613` to generate positive (for entail–entail combination) and negative (for undiscussed–entail combination) feedback. For other combinations, we slightly adjust the prompt to reflect the response accuracy compared to the true label.

We use a generic declaration unrelated to task data to provide feedback on choosing the don't know option. Again, we present a draft to `gpt-4-0613`, tasking it with elaborating and producing five distinct versions. For instance:

`Your response did not adequately attend to your client's concerns. The lack of clarity is causing delays within the company and adversely affecting its business operations.`

An example of the feedback we present to the participants is shown in Fig. 12.

D.3 Constructing Qualification Exam and Attention Checks

To ensure participant engagement and prevent spam, we create a qualification exam. We create 30 multiple-choice questions (MCQs) based on 30 distinct NDAs from the training dataset; this prevents qualified participants from encountering NDAs used in the main tasks since they are picked from the test set. Below is the prompt we use to instruct `gpt-4-0613` to generate an MCQ based on a text excerpt:

`You are a helpful assistant specializing in the legal domain. You want to check whether someone read a text. Construct a simple 4-choice question about this text so that anyone who reads the text is able to answer it. The correct answer must be explicitly mentioned in the text. Make sure only one answer is correct. Define the correct`

```
answer at the end. Try to balance the options in length (number of words). Format your question as follows:
  <Question text>
  1) option 1
  2) option 2
  3) option 3
  4) option 4
  Correct answer: [option_number]
```

To verify its generation accuracy, we present the same text excerpt along with the question and options to `gpt-3.5-turbo` and check if the generated answer matches the response from `gpt-4-0613`. If they align, we assume the question is valid. However, in cases where the text contains negation or exclusions, the generated questions are generally inaccurate, requiring manual verification or replacement. From this pool of 30 questions, each participant is presented with 6 randomly chosen questions and the corresponding NDA. Passing the exam requires correctly answering at least 5 of these 6 questions. Around 74% of participants passed the 10-minute exam. All participants receive $2 compensation. We compile a list of verified users and employ it in Prolific to regulate participation in the main studies. Fig. 7 shows an example.

We use the same method to generate MCQs for each sample assessed by participants in the main human studies. Participants must answer the MCQ correctly to proceed to the next step. We monitor the number of attempts as an indicator of participants' attention. The average is 1.2 attempts.

D.4 Estimating Expert Performance on ContractNLI

To approximate experts' performance, we conduct a small study consisting of an attention check MCQ followed by a two-step statement assessment: labeling and open-ended reasoning description. See Fig. 5 for the instructions provided to the participants. Each participant is given a unique, longer, potentially more challenging NDA with roughly 1650 words (close to the average length of NDAs in the ContractNLI dataset) and one smaller NDA of about 650 words, which is the same for all participants. The smaller NDA is paired with a statement explicitly mentioned in the NDA for easy labeling; this serves as an additional check for participants' attentiveness, complementing the MCQ, and all of them answered it correctly. We pair each of the 5 longer NDAs with a randomly selected unique statement[22]. Participants are recruited through the Prolific platform and meet the qualifications.

D.5 Additional Details on Producing and Presenting Explanations

We present three types of explanations to participants across various studies: model confidence scores, input highlights, and influential training examples. In the following, we detail how we generate and present them to the participants.

Confidence score. We use the model's prediction probabilities calibrated with temperature scaling (Guo et al., 2017) [23] over the development dataset of ContractNLI to calculate a confidence score for each sample, ranging from 0 to 100. The confidence score is shown alongside the AI's guess when presented to participants. Examples can be seen in Fig. 13 and 14.

Input highlights. Highlights show which parts of the input are important for the model to make its prediction. We use the `Input×Gradient` (Shrikumar et al., 2016) scores obtained with `inseq` (Sarti et al., 2023). We display the highlight scores by varying the color intensity of the important words. Higher scores correspond to more intense highlight colors presented to participants. To display the Top-N% of highlights, we retain the highest N% of highlight scores and set the remainder to zero. Each word's highlight intensity is determined by the average non-zero highlight scores of its tokens. To prevent overwhelming the participants while maintaining NDA integrity, we give them the option to expand/collapse sections with no highlights. An example is shown in Fig. 13.

Influential training examples. We identify influential training examples with EK-FAC (George et al., 2018; Grosse et al., 2023). As previously mentioned, the maximum number of input tokens Flan-T5-3B can process is determined by GPU memory capacity, which for us results in 4.2K tokens. Since NDAs in ContractNLI can be even 8K tokens long, truncation is necessary. Some statements, like those defining confidential information, are typically discussed at the beginning of contracts, while others, such as termination conditions, are addressed towards the end. To prevent loss of vital information in lengthier contracts, we produce

[22] All NDAs are paired with all 17 statements in the ContractNLI data.

[23] https://github.com/gpleiss/temperature_scaling

two input versions — one where the statement is appended and another where it is prepended to the NDA (as shown in Table 12) — and truncate input from the opposite side. We then calculate influence scores for both versions of inputs for each NDA-statement pair in an evaluation sample and NDA-statement in the training data, average them, and select the top three most influential training examples. For these three influential examples, we additionally incorporate `Input×Gradient` highlights to streamline the verification process so that participants do not have to review three more NDAs to swiftly verify if the AI's guess is correct. See Fig. 14 for an example.

D.6 Recruiting, Examining, and Paying Participants

All our studies begin with informed consent. We filter for individuals who have a degree in administration and law, work in legal functions, and are fluent in English. Unfortunately, even some domain experts on Prolific behaved suspiciously during our preliminary studies. Therefore, we create a qualification exam to deter spammers; details on its creation are provided in §D.3. No annotator sees the same example more than once across conditions. However, ensuring no overlap across studies for the contradicting label is not possible due to insufficient samples. Nonetheless, we ensure that no sample is repeated across two consecutive studies. We pay annotators $12.45 per hour.

E Details on Deferral Models Input and Deferral Results for User Study II (§5)

Data preprocessing. We use the development set of the ContractNLI dataset, dividing it into 1,400 samples for training, 320 for evaluation, and 371 for testing, to develop the deferral models. We finetune `Llama-2-13B-Chat` using all the training samples. For finetuning `gpt-3.5-turbo-1106`, we balance the training set by preserving all 156 positive samples and randomly selecting an equal number of negative samples.

To incorporate highlights in the input, we wrap the top 5% of highlighted words with an HTML-like tag: `<important></important>`.

End-of-study reflections. Below is the feedback we get from some participants in Study II, mainly regarding the accuracy of gold labels presented to them when they received the feedback on their assessment claiming they made a wrong decision.

Participant-1: "i disagree with your final feedback. additionally it is difficult to distinguish whether something is absent, or whether it contradicts, when there is a contradicting statement within the NDA"

Participant-2: "Interpretation of agreements is at times difficult. Lawyers by their training try to interpept agreements in many different ways. There was one example I disagreed with the ending decision. "

Participant-3: "I felt like some of the answers were wrong - especially the one relating to verbal information. The NDA discussed orally conveyed information, but the answer still said that the NDA did not discuss it. "

Participant-4: "I disagree with the outcomes (final feedback) based on the reasoning I gave."

Participant-5: "I feel some of these answers were not correct."

Participant-6: "It seemed more difficult than the previous studies!"

$y_h^{(1)}$	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 10: Extending 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.

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.\nSickle2̌013thalassemia association resulted in higher hemoglobin, hematocrit, and erythrocyte counts with reduced MCV and reticulocytes. \n

Table 11: 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:

CONFIDENTIALITY AND NON-DISCLOSURE AGREEMENT\nBUSINESS: ____________________\nADDRESS: ____________________\nDESCRIPTION:\nASKING PRICE: .\nThis is intended to be a legally binding document. This agreement shall be governed by and enforced in accordance with the laws of the State of California, USA as applicable to contracts to be performed therein. The undersigned (hereinafter, collectively, 'Buyer') acknowledges its/his/her desire to receive from Epsteen & Associates ('Broker') and from the owner of the Business, described above ('Seller') certain information pertaining to the Business, the Seller and/or the possible sale of the Business (the 'Transaction'). For purpose of this Agreement, (a) the term 'Buyer' means all of undersigned, including both the potential buyer interested in the Transaction, and such buyer's broker, and both such buyer and buyer's broker are bound by the provisions of this agreement; and (b) any information provided to Buyer, or otherwise learned by Buyer, concerning the Business, Seller or Transaction shall collectively be referred to herein as 'Confidential Information'. In consideration of Broker providing Confidential Information to Buyer, Buyer agrees to the following:\nKEEP INFORMATION CONFIDENTIAL\nBuyer acknowledges that any Confidential Information disclosed to others may be damaging to the Business and the Seller. Buyer understands that Confidential Information includes, without limitation: the fact that Business is for sale; financial details; identity of suppliers and customers; and any information not generally known by public. Buyer agrees not to disclose Confidential Information to anyone other than its/his/her advisors and affiliates who both (a) have a need to know the information in connection with the Transaction; and (b) have agreed by signing a copy of this agreement to be bound by the terms of this agreement. Buyer agrees that all copies of materials and data provided to Buyer (and any information derivative of such information) shall also be 'Confidential Information'; and all Confidential Information shall be returned to Broker in the event that Buyer decides not to pursue the Transaction. Buyer shall be legally responsible for the actions and omissions of Buyer's advisors and affiliates.\nDIRECT ALL CONTACT THROUGH BROKER\nBuyer shall not contact the Seller or any other individual or entity associated with Seller or the Business including, without limitation, landlords, employees, suppliers or customers except upon the prior written consent of Seller. All correspondence, inquiries, and offers to purchase, and other documents relating to the Business or Transaction (all of which is 'Confidential Information) will be delivered solely through Broker, and all negotiations relating to the Business or Transaction will be conducted exclusively through Broker.\nUSE INFORMATION FOR EVALUATION PURPOSE ONLY\nWithout limiting the other restrictions in this agreement, Buyer agrees to use Confidential Information solely to internally evaluate the Business for the possible Transaction and not for any other purposes whatsoever.\nDO NOT CIRCUMVENT SELLER AND/OR BROKER Buyer will not circumvent Seller and/or Broker by contacting any person or persons involved with the Business including, without limitation, landlords, employees, suppliers or customers. CONFIDENTIAL INFORMATION IS PROVIDED BY SELLER\nAll information about the Business is provided by the Seller and is not verified by Broker. Buyer understands that purchasing any business represents investment risks and that Buyer should obtain professional assistance from independent accounting, legal, and financial advisors to verify all information prior to consummating an agreement to purchase the Business. Buyer will not rely on the information provided by Broker or Seller, including the Confidential Information, but shall conduct its own independent due diligence. Seller (and not Broker) is the source of all information and statements about the Business. Broker makes no warranty, guarantee, expressed or implied, as to the accuracy of such information.\nBuyer agrees to defend, indemnify, protect and hold harmless Broker, and release Broker, in connection with all information provided to Buyer, including all Confidential Information, and in connection with any breach by Buyer of any of its obligations under this agreement.\nPROVIDE EVIDENCE OF FINANCIAL ABILITY\nShould Buyer present an offer to purchase the Business, Buyer will provide a financial statement and a personal and business history, and Buyer authorizes Broker and Seller to obtain through standard reporting agencies, financial and credit information about Buyer and/or the companies Buyer represents.\nENFORCEMENT\nBuyer acknowledges and agrees that any breach of any of its/his/her obligations hereunder will cause Seller and the Business irreparable harm for which Seller and the Business have no adequate remedy at law, and that Seller and the Business shall be entitled to injunctive and other equitable relief to prevent a breach or continued breach of this agreement, in addition to any other remedies Seller and Business may have at law or in equity, and that this agreement shall be specifically enforceable in accordance with its terms. Both Broker and Seller are beneficiaries of this agreement and are both entitled to enforce this agreement.\nIn any action or proceeding, whether or not resulting in litigation, between Buyer (or either of them) and Seller, or between Buyer (or either of them) and Broker, including any litigation to enforce any of the terms of this agreement, the prevailing party shall be entitled to recover, in addition to any damages or compensation received, its costs and expenses incurred in connection with such action or proceeding, including any reasonable attorneys' fees, expenses and court costs.\nWe, the undersigned, understand and agree that this agreement is legally binding upon us. We understand that Seller and/or Broker have the right to seek any and all lawful remedies to enforce the terms of this agreement. We acknowledge that we have read and understand the disclosures contained herein.\nBUYER: BUYER'S BROKER/AGENT:\n____________________ ____________________\nSIGNATURE DATE SIGNATURE DATE\n____________________ ____________________\nPRINT NAME PRINT NAME\n \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 12: 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.\nOptions: Accept, Reject\nAnswer:**

Table 13: A representative input sample for the $\text{ILDC}_{\text{single}}$ task. The template is: **petition**\n\nMulti-choice problem: Decide whether this petition should be accepted or not.\nOptions: Accept, Reject\nAnswer: answer

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:\nSignificantly increase\n, No significantly difference\n, Significantly decrease\n\nAnswer:**

Table 14: A representative input sample for the ERASER EvidenceInference V2 task. In this sample, there is only one document, but additional documents should be added through retrieval. 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**: "

Prompt: You are a client who owns a company, entering into a non-disclosure agreement with another company. The following statement *entails* the NDA: Agreement shall not grant Receiving Party any right to Confidential Information. What actions will you take to satisfy the above statement? **GPT-4 Response**: To satisfy the statement, I would ensure that the receiving party does not gain any rights to the confidential information. This means they will not be allowed to use the information for their own benefit or disclose it to third parties. The information shared would be strictly for the purpose of the project or partnership at hand. I would also insist on clear guidelines in the NDA about how the confidential information should be handled, stored, and disposed of after the agreement ends. Additionally, I would ensure there are severe penalties for breach of the agreement to deter any misuse of the confidential information. This could include financial penalties or legal action. Finally, I would monitor compliance with the NDA to ensure the receiving party is respecting the terms of the agreement.
Prompt: You are a client who owns a company, entering into a non-disclosure agreement with another company. The following statement *is not discussed* in the NDA: Receiving Party shall destroy or return some Confidential Information upon the termination of Agreement. Considering that the above statement is undiscussed and thus unclear, what are you allowed to do while still being within the bounds of the NDA? **GPT-4 Response**: Without a clear statement in the NDA regarding the handling of confidential information after termination of the agreement, you are not explicitly required to destroy or return the information. However, maintaining confidentiality remains paramount. You should continue to treat the information as confidential and not disclose it to any third parties. If you're unsure about how to handle the information post-agreement, it would be best to consult with your legal advisor. To maintain good business relations, you could also reach out to the other party for further clarification. It's always advisable to handle such matters carefully to avoid potential legal disputes.
Prompt: You are a client who owns a company, entering into a non-disclosure agreement with another company. The following Statement *contradicts* the NDA: All Confidential Information shall be expressly identified by the Disclosing Party. Considering that the above statement does NOT hold, what are you allowed to do while still being within the bounds of the NDA? **GPT-4 Response**: Even with the statement not holding, you are still required to protect all confidential information that you receive from the other party under the NDA. This means you cannot disclose this information to anyone who is not part of the agreement, use it for any purpose other than what is stipulated in the agreement, or use it to the detriment of the disclosing party. You are also expected to take reasonable steps to protect the information and limit access to those who really need to know it. However, you are not obligated to identify or label the information as confidential since the statement does not hold, and you can treat all information received in the course of the agreement as confidential. It's also important to remember that any obligations under the NDA continue even after the termination of the agreement until a specified time. Always consult with a legal professional to understand your obligations under any legal agreement.

Table 15: Samples of input prompts along with the responses from GPT-4 for generating actions regarding pairs of statements and labels.

Prompt: You are a client who owns a company, entering into a non-disclosure agreement with another company. The statement <Some obligations of Agreement may survive termination of Agreement.> *entails* the NDA, and your lawyer correctly confirms this inference.
Because your lawyer was able to inform you of the conditions, you were able to fulfill the following intended requirements/actions:
"As a client entering into a non-disclosure agreement (NDA), I would be required to continue to uphold certain obligations even after the termination of the agreement, as per the statement. To satisfy these obligations, I would ensure that any confidential information received during the term of the agreement remains confidential and is not disclosed to any third parties. This could involve continued secure storage of the information or its proper destruction if it's not needed. I would also ensure that my employees, contractors or anyone who had access to the information during the partnership understand their ongoing obligations to maintain confidentiality. I would refrain from using the confidential information for any purpose other than what was specified in the agreement, even after its termination. Any breach of these obligations could result in legal consequences, so it's crucial to respect the terms of the NDA."
Can you state some benefits that the above actions give to your company due to your lawyer's correct advice directly specific to the statement above? Phrase your answer in one to two sentences, in a punchy way, as if you were writing a report TL;DR to a supervisor so that later on can be used to provide feedback to the lawyer.
GPT-4 Response: By adhering to our lawyer's advice on the NDA terms, we not only safeguard our company's reputation and maintain trust with external partners, but also mitigate potential legal liabilities, ensuring our ongoing business operations remain uncompromised.

Prompt: You are a client who owns a company, entering into a non-disclosure agreement with another company. Although the statement <Receiving Party may create a copy of some Confidential Information in some circumstances.> *is not discussed* in the NDA, your lawyer made a mistake and reported that the statement *entails* the NDA. As a consequence, your company failed to comply with the following intended actions:
"In the absence of explicit permission in the NDA, we should assume that we are not allowed to make copies of the Confidential Information. NDAs are designed to protect sensitive information, and typically restrict unauthorized copying, sharing, or usage of such information. To stay within the bounds of the NDA, we should only use the Confidential Information as specifically permitted in the agreement and avoid creating any copies. If we find it necessary to create a copy for your business operations, it is advisable to seek written consent from the disclosing party. This ensures that we remain compliant with the NDA and respect the other party's proprietary rights."
Can you mention a concrete legal penalty you might face due to your lawyer's incorrect advice directly specific to the statement above? Phrase your answer in one to two sentences, in a punchy way, as if you were writing a report TL;DR to a supervisor so that later on can be used to provide feedback to the lawyer.
GPT-4 Response: Due to the lawyer's incorrect advice, our company may face severe legal penalties including a breach of contract lawsuit, potentially resulting in substantial financial damages and injunctions prohibiting further use of the confidential information. Additionally, this breach could tarnish our reputation, jeopardizing future business partnerships and opportunities.

Table 16: Two instances of input prompts and GPT-4 responses for generating feedback on the benefits of correctly labeling a specific statement (top example) and the consequences of mislabeling a particular statement (bottom example).

Statement	Receiving Party may retain some Confidential Information even after the return or destruction of Confidential Information.
Gold label	No mention
AI label	Contradict
Expert label	Contradict
Author label	Contradict
Expert comment	"The highlighted phrases show statements that may be in relation to the statement that has been asked. It shows the most accurate statements and paragraphs to search the answer from . the highlighred phrases gave guidance in terms of where to look and how to answer so in conclusion it did reveal that some confidential information must be returned and never to be retained."
Relevant spans of NDA (by author)	5. RETURN OF CONFIDENTIAL INFORMATION Either party may request in writing at any time that any confidential information or any documents containing such confidential information, disclosed pursuant to the terms of this agreement and any copies thereof be returned with a written statement to the effect that upon such return the receiving party has not knowingly retained in its possession or under its control, either directly or indirectly, any such confidential information or any documents containing such confidential information or any copies thereof and the receiving party shall comply with any such request within seven days of receipt of such request.
NDA file	https://contract-nli.s3.us-east-2.amazonaws.com/hit_files/57.pdf

Table 17: One example where we change the gold label after reviewing the NDA and the expert's reasoning.

Theme	Description	Examples
Relevant information highlighted (49.1%)	Participants use the highlights to identify specific clauses or phrases that are crucial to answering the question.	**Example-1:** "the highlighted phrases were very important cause they helped me too lok through the document thoroughly and see what the NDA say about the higlighted word. This made it easier to do a process of elimination. It is easier and quickier to go through a document this way. it ensures accuracy and also saves time."
		Example-2: "It still contradicts the statement because no where in there does it say that confidential information may be shared. It states that they should not disclose confidential information at any time. It is discussed in this NDA. The examples highlight that this is the case. It mentions third-parties, however it specifically states it cannot be shared with them."
Comparison (16.4%)	Examples and highlights verify or adjust the participants' initial assessment.	**Example-1:** "It is not express within the document that the definition of 'Confidential Information' includes verbally conveyed information, but it flows from the highlighted statements logically that one of the 'forms' of transmitting 'Confidential Information' could logically include verbally conveyed information. Thus, it is not on all fours within the document but I would say generally the statement aligns with the document. I was on the fence between identifying the document as aligning, or not including express reference to the statement, so therefore not referenced in the document. AI has highlighted the relevant area, and so I have in this case leant slightly more towards it aligning."
		Example-2: "i did not see any discussion of technical information within the NDA during my own analysis. when looking at the highlighted portions, i still did not see any discussion of technical information. this added to the confidence level of my own analysis and decision that the statement is not included within the NDA."
Irrelevant or misleading information highlighted (10.1%)	Examples sometimes highlight irrelevant or misleading sections, causing confusion for participants and potentially helping them detect AI's mistakes.	**Example-1:** "All the highlighted sentence were not consistent with the statement made by the model. This led me to make a different final decision that the one made by the model. Actually from the highlighted sentence it is evident that the the statement is not discussed in the NDA and therefore the answer provided by the model is not correct."
		Example-2: "the three most influential examples refer to the form of the disclosure of the confidential information and not the form in which the confidential information is provided to the receiving party (written or verbal). that means that the receiving party must refrain from disclosing the confidential information either in written or verbal form, but does not discus the form in which the confidential information was provided to the receiving party"
Not helpful (7.3%)	Some participants found highlights not to be useful in addressing the statement, leading to confusion or wasted time.	**Example-1:** "the highlighted sections provided by the AI are not all that helpful. it highlighted so many sections, and a lot of those sections are not relevant to the question or statement. looking at the highlighted sections wasted a lot of time and in the end was not helpful at all in making my decision."
		Example-2: "The three most influential examples in the training model had rather little impact on my decision. It seems the AI is not looking at the most relevant portion of the agreement to answer the quested posed by the researcher. In fact, some of the information in these examples is actually completely irrelevant."

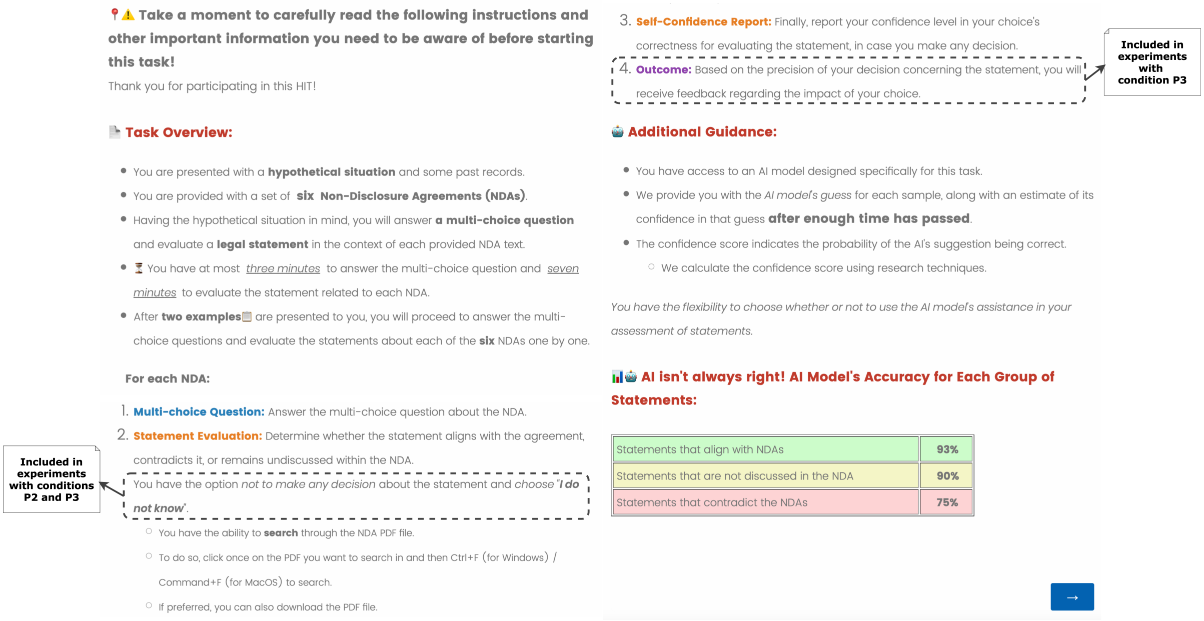
Table 18: Analysis of open-ended responses to the question on the impact of explanations on the participants' decision-making process. This table shows the most common themes in the responses. For each theme, the first example is from responses in the study with highlights, and the second one is for influential examples.

Model	Train Setup	Input	Class	Precision	Recall	F1 score
Llama-2-13B-chat-hf	Finetuned	P+T	DEFER	5.6	10.0	7.1
			DO-NOT-DEFER	75.6	33.2	46.2
			MICRO AVG.	66.9	30.3	41.3
			MACRO AVG.	40.6	21.6	26.7
		P+T+H	DEFER	2.9	5.0	3.6
			DO-NOT-DEFER	54.6	21.4	30.8
			MICRO AVG.	48.1	19.4	27.4
			MACRO AVG.	28.7	13.2	17.2
GPT-4o	Zero-shot	P+T	DEFER	12.5	10.0	11.1
			DO-NOT-DEFER	87.5	90.0	88.7
			MICRO AVG.	78.1	80.0	79.0
			MACRO AVG.	50.0	50.0	49.9
	30-shot	P+T	DEFER	**18.9**	17.5	18.2
			DO-NOT-DEFER	88.3	89.3	88.8
			MICRO AVG.	79.7	80.3	80.0
			MACRO AVG.	53.6	53.4	53.5
		P+T+H	DEFER	17.7	15.0	16.2
			DO-NOT-DEFER	88.1	90.0	89.1
			MICRO AVG.	79.3	80.6	79.9
			MACRO AVG.	52.9	52.5	52.6
GPT-3.5-turbo-1106	Finetuned	P+T	DEFER	12.5	**30.0**	17.7
			DO-NOT-DEFER	87.5	70.0	77.8
			MICRO AVG.	78.1	65.0	70.3
			MACRO AVG.	50.0	50.0	47.7

Table 19: Complete results of deferral models (§5). P denotes the finetuned Flan-T5-3B's prediction, T a ContractNLI statement-NDA pair text, and H `Input×Gradient` highlights.

Decision Maker	Data Sample	Class	Precision	Recall	F1 score	# Examples
Experts	Deferred w/ highlights	ENTAIL	15.4	50.0	23.5	4
		NO-MENTION	77.8	24.1	36.8	29
		CONTRADICT	12.5	40.0	19.0	5
		MICRO AVG.	28.9	28.9	28.9	38
		MACRO AVG.	35.2	38.0	26.5	38
	Deferred w/ highlighted influential examples	ENTAIL	0.0	0.0	0.0	3
		NO-MENTION	75.0	44.4	55.8	27
		CONTRADICT	13.3	40.0	20.0	5
		MICRO AVG.	40.0	40.0	40.0	35
		MACRO AVG.	29.4	28.1	25.3	35
	All examples deferred to experts	ENTAIL	11.8	28.6	16.7	7
		NO-MENTION	76.0	33.9	46.9	56
		CONTRADICT	12.9	40.0	19.5	10
		MICRO AVG.	34.2	34.2	**34.2**	73
		MACRO AVG.	33.6	34.2	27.7	73
Flan-T5-3B	Undeferred samples	ALL	100.0	100.0	100.0	1854

Table 20: Performance of experts and the finetuned Flan-T5-3B on different sets of data in Study II (§5).



📍⚠️ **Take a moment to carefully read the following instructions and other important information you need to be aware of before starting this task!**

Thank you for participating in this HIT!

📄 **Task Overview:**

- You are presented with a **hypothetical situation** and some past records.
- You are provided with a set of **six Non-Disclosure Agreements (NDAs)**.
- Having the hypothetical situation in mind, you will answer **a multi-choice question** and evaluate a **legal statement** in the context of each provided NDA text.
- ⏳ You have at most _three minutes_ to answer the multi-choice question and _seven minutes_ to evaluate the statement related to each NDA.
- After **two examples** are presented to you, you will proceed to answer the multi-choice questions and evaluate the statements about each of the **six** NDAs one by one.

For each NDA:

1. **Multi-choice Question:** Answer the multi-choice question about the NDA.
2. **Statement Evaluation:** Determine whether the statement aligns with the agreement, contradicts it, or remains undiscussed within the NDA.
   - You have the option *not to make any decision* about the statement and choose *'I do not know'*.
     - You have the ability to **search** through the NDA PDF file.
     - To do so, click once on the PDF you want to search in and then Ctrl+F (for Windows) / Command+F (for MacOS) to search.
     - If preferred, you can also download the PDF file.
3. **Self-Confidence Report:** Finally, report your confidence level in your choice's correctness for evaluating the statement, in case you make any decision.
4. **Outcome:** Based on the precision of your decision concerning the statement, you will receive feedback regarding the impact of your choice.

🤖 **Additional Guidance:**

- You have access to an AI model designed specifically for this task.
- We provide you with the *AI model's guess* for each sample, along with an estimate of its confidence in that guess **after enough time has passed**.
- The confidence score indicates the probability of the AI's suggestion being correct.
  - We calculate the confidence score using research techniques.

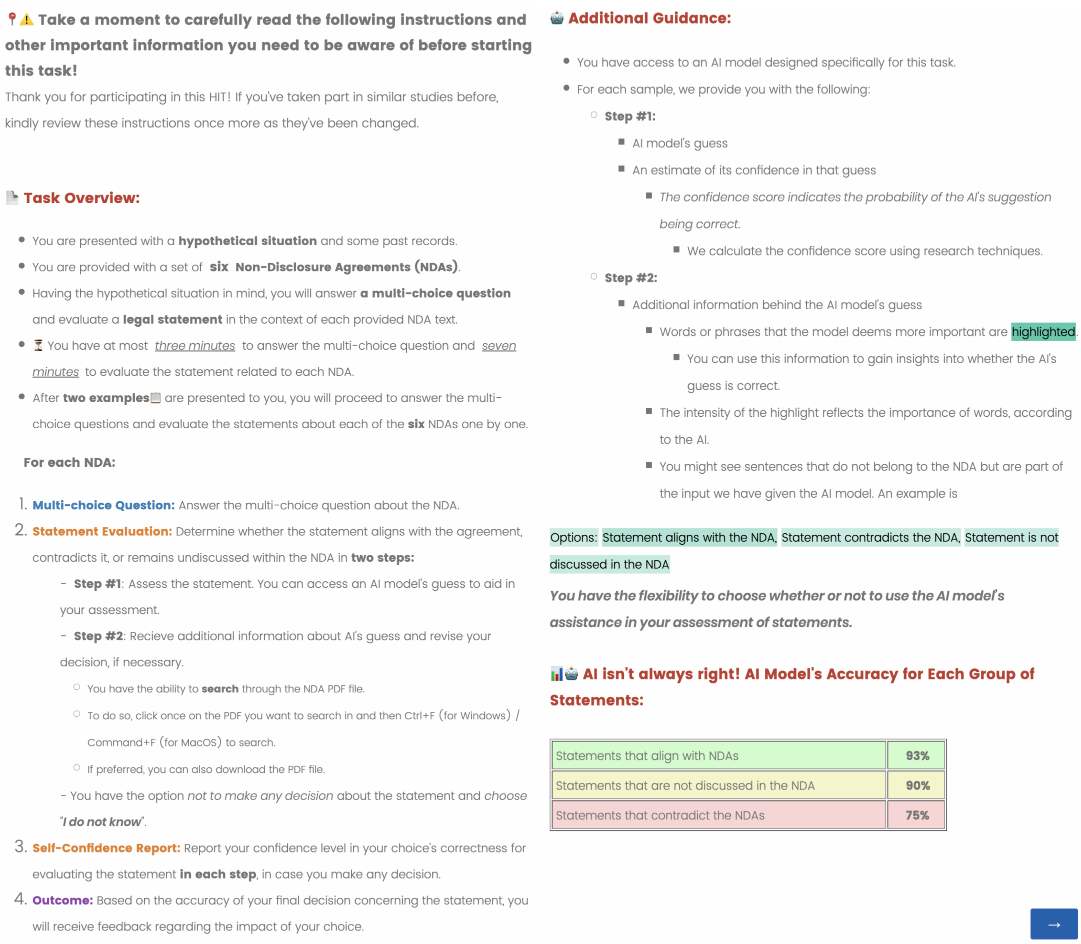
You have the flexibility to choose whether or not to use the AI model's assistance in your assessment of statements.

📊🤖 **AI isn't always right! AI Model's Accuracy for Each Group of Statements:**

Statements that align with NDAs	93%
Statements that are not discussed in the NDA	90%
Statements that contradict the NDAs	75%

→

Figure 2: Instructions for Study I (§4), the experiments with conditions P1 to P3.



📍⚠️ **Take a moment to carefully read the following instructions and other important information you need to be aware of before starting this task!**

Thank you for participating in this HIT! If you've taken part in similar studies before, kindly review these instructions once more as they've been changed.

📄 **Task Overview:**

- You are presented with a **hypothetical situation** and some past records.
- You are provided with a set of **six Non-Disclosure Agreements (NDAs)**.
- Having the hypothetical situation in mind, you will answer **a multi-choice question** and evaluate a **legal statement** in the context of each provided NDA text.
- ⏳ You have at most _three minutes_ to answer the multi-choice question and _seven minutes_ to evaluate the statement related to each NDA.
- After **two examples** are presented to you, you will proceed to answer the multi-choice questions and evaluate the statements about each of the **six** NDAs one by one.

For each NDA:

1. **Multi-choice Question:** Answer the multi-choice question about the NDA.
2. **Statement Evaluation:** Determine whether the statement aligns with the agreement, contradicts it, or remains undiscussed within the NDA in **two steps:**
   - **Step #1**: Assess the statement. You can access an AI model's guess to aid in your assessment.
   - **Step #2**: Recieve additional information about AI's guess and revise your decision, if necessary.
     - You have the ability to **search** through the NDA PDF file.
     - To do so, click once on the PDF you want to search in and then Ctrl+F (for Windows) / Command+F (for MacOS) to search.
     - If preferred, you can also download the PDF file.
   - You have the option *not to make any decision* about the statement and *choose 'I do not know'*.
3. **Self-Confidence Report:** Report your confidence level in your choice's correctness for evaluating the statement **in each step**, in case you make any decision.
4. **Outcome:** Based on the accuracy of your final decision concerning the statement, you will receive feedback regarding the impact of your choice.

🤖 **Additional Guidance:**

- You have access to an AI model designed specifically for this task.
- For each sample, we provide you with the following:
  - **Step #1:**
    - AI model's guess
    - An estimate of its confidence in that guess
      - *The confidence score indicates the probability of the AI's suggestion being correct.*
        - We calculate the confidence score using research techniques.
  - **Step #2:**
    - Additional information behind the AI model's guess
      - Words or phrases that the model deems more important are highlighted.
        - You can use this information to gain insights into whether the AI's guess is correct.
      - The intensity of the highlight reflects the importance of words, according to the AI.
      - You might see sentences that do not belong to the NDA but are part of the input we have given the AI model. An example is

Options: Statement aligns with the NDA, Statement contradicts the NDA, Statement is not discussed in the NDA

You have the flexibility to choose whether or not to use the AI model's assistance in your assessment of statements.

📊🤖 **AI isn't always right! AI Model's Accuracy for Each Group of Statements:**

Statements that align with NDAs	93%
Statements that are not discussed in the NDA	90%
Statements that contradict the NDAs	75%

→

Figure 3: Instructions for Study I (§4), the experiment with conditions P4.

📍⚠️ **Take a moment to carefully read the following instructions and other important information you need to be aware of before starting this task!**

Thank you for participating in this HIT! If you've taken part in similar studies before, kindly review these instructions once more as they've been changed.

📄 Task Overview:

- You are presented with a **hypothetical situation** and some past records.
- You are provided with a set of **six Non-Disclosure Agreements (NDAs)**.
- Having the hypothetical situation in mind, you will answer **a multi-choice question** and evaluate a **legal statement** in the context of each provided NDA text.
- ⏳ You have at most *three minutes* to answer the multi-choice question and *seven minutes* to evaluate the statement related to each NDA.
- After **two examples** 📰 are presented to you, you will proceed to answer the multi-choice questions and evaluate the statements about each of the **six** NDAs one by one.

For each NDA:

1. **Multi-choice Question:** Answer the multi-choice question about the NDA.
2. **Statement Evaluation:** Determine whether the statement aligns with the agreement, contradicts it, or remains undiscussed within the NDA in **two steps:**

- **Step #1**: Assess the statement. You can access an AI model's guess to aid in this assessment.
- **Step #2**: Recieve additional information about AI's guess and revise your decision, if necessary.

  - You have the ability to **search** through the NDA PDF file.
  - To do so, click once on the PDF you want to search in and then Ctrl+F (for Windows) / Command+F (for MacOS) to search.
  - If preferred, you can also download the PDF file.

- You have the option *not to make any decision* about the statement and *choose '**I do not know**'*.

1. **Self-Confidence Report:** Report your confidence level in your choice's correctness for evaluating the statement **in each step**, in case you make any decision.
2. **Outcome:** Based on the accuracy of your final decision concerning the statement, you will receive feedback regarding the impact of your choice.

🤖 Additional Guidance:

- You have access to an AI model designed specifically for this task.
- For each sample, we provide you with the following:
  - **Step #1:**
    - AI model's guess
    - An estimate of its confidence in that guess
      - *The confidence score indicates the probability of the AI's suggestion being correct.*
        - We calculate the confidence score using research techniques.
  - **Step #2:**
    - Additional information behind the AI model's guesses: **Influential Examples**
    - AI learns to perform a task from so-called training examples. Influential training examples for a new <statement-contract> pair are those that influence what the model decides for this pair.
      - You can access the **three most influential examples** for each provided sample that impacted AI's guess the most.
        - For each example, the top 5% of words or phrases the model deems more important are highlighted.
          - The intensity of the highlight reflects the importance of words, according to the AI.
          - You might see sentences that do not belong to the NDA but are part of the input we have given the AI model. An example is Options: Statement aligns with the NDA, Statement contradicts the NDA, Statement is not discussed in the NDA
      - You can use this information to gain insights into whether the AI's guess is correct or not.

You have the flexibility to choose whether or not to use the AI model's assistance in your assessment of statements.

📊🤖 AI isn't always right! AI Model's Accuracy for Each Group of Statements:

Statements that align with NDAs	93%
Statements that are not discussed in the NDA	90%
Statements that contradict the NDAs	75%

Figure 4: Instructions for Study I (§4), the experiment with conditions P5.

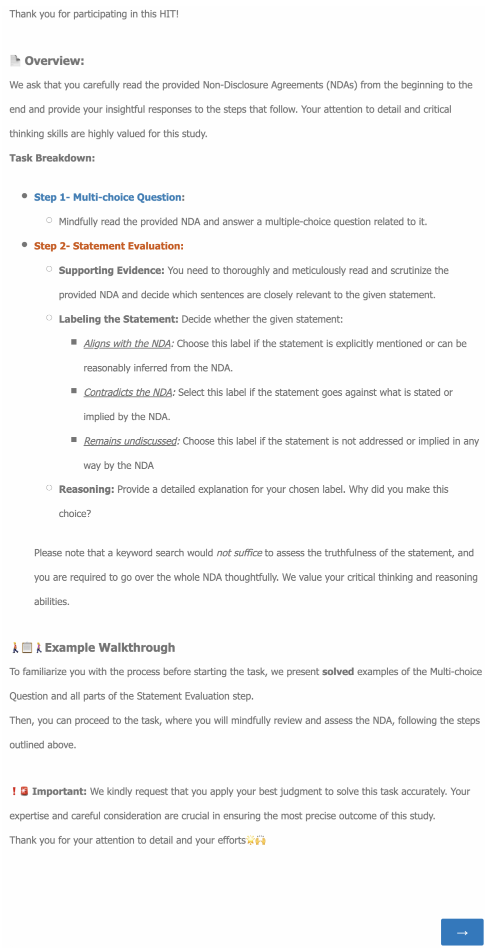


Figure 5: Instructions for expert performance estimation experiment.

Figure 7: Attention check sample. Before solving the main task, participants are asked to solve a multi-choice question regarding the same NDA.

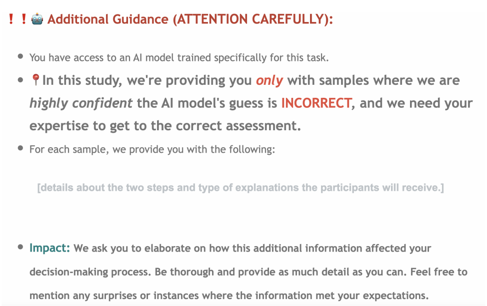


Figure 6: Parts of instructions unique to Study I (§4). The rest is the same as in Fig. 3–4 for the experiment with conditions P4 and P5, respectively.

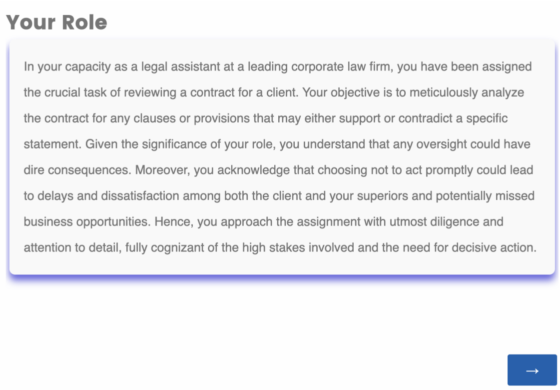


Figure 8: An example of the hypothetical roles we provide participants.

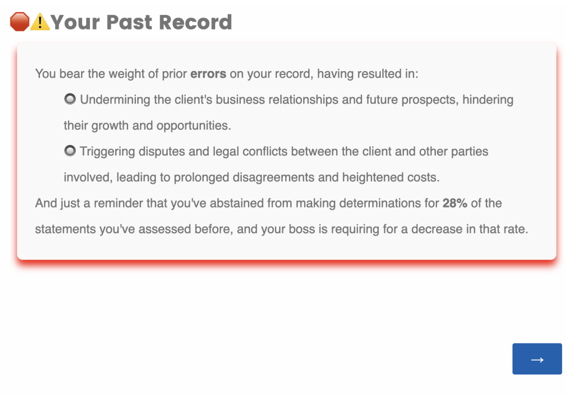


Figure 9: An example of the consequences of past errors we show to participants.

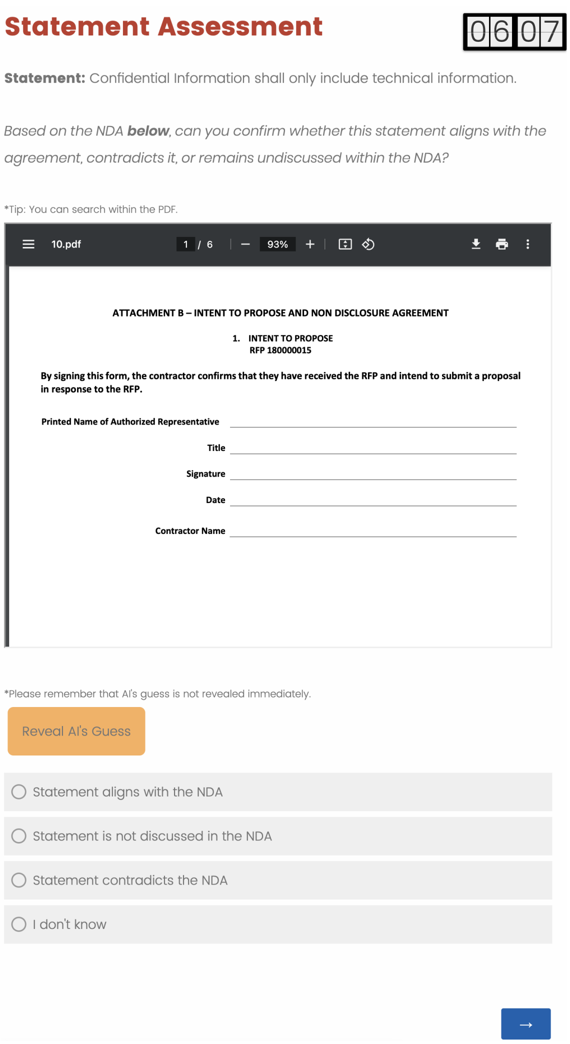


Figure 10: Statement Assessment. Participants can click the "Reveal AI" button to view AI's prediction. Upon doing so, a GIF animation is displayed for approximately 5 seconds, followed by presenting AI's prediction and confidence score. AI's prediction is color-highlighted (green, yellow, or red) according to the label.

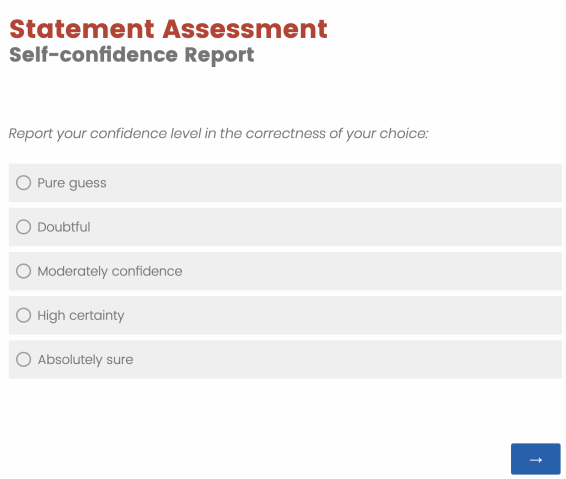


Figure 11: Likert-scale self-confidence report question.

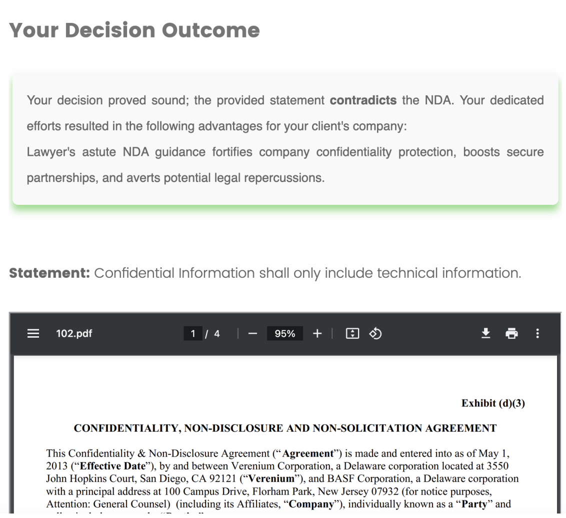


Figure 12: Positive feedback for labeling a specific statement correctly. We provide the statement and NDA again in this step for the participant's review.

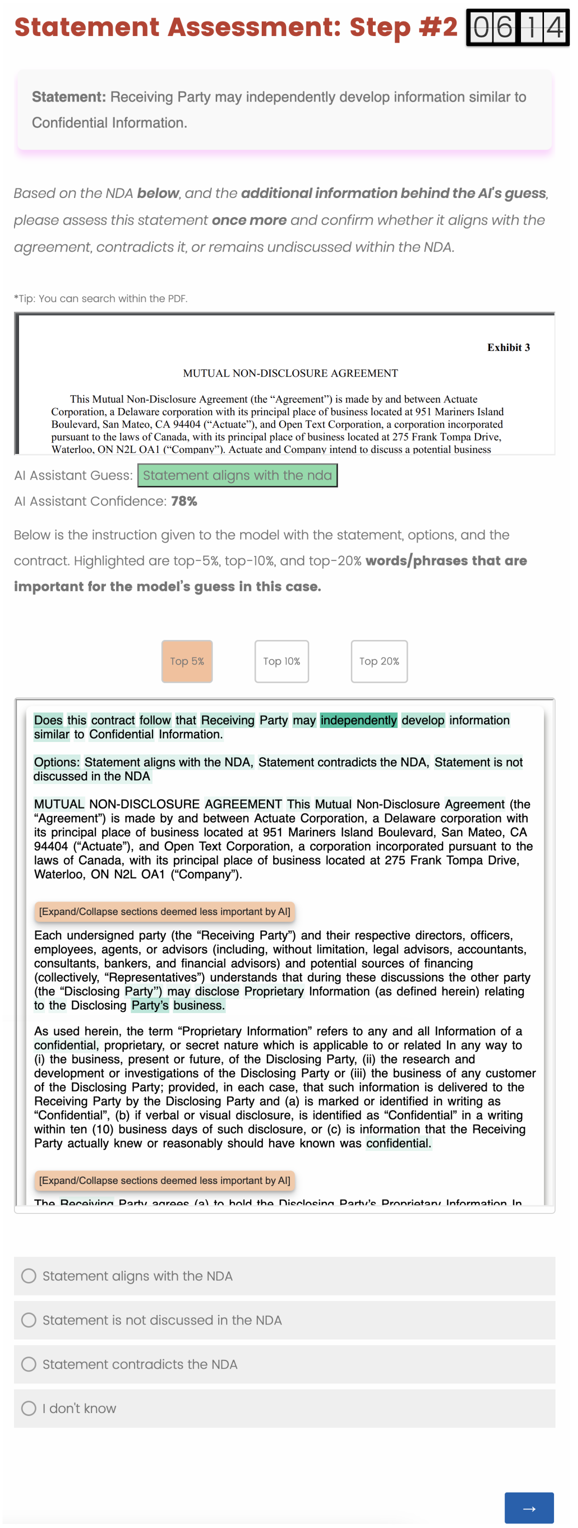


Figure 13: Second iteration of statement evaluation with access to Input$\times$Gradient highlights.

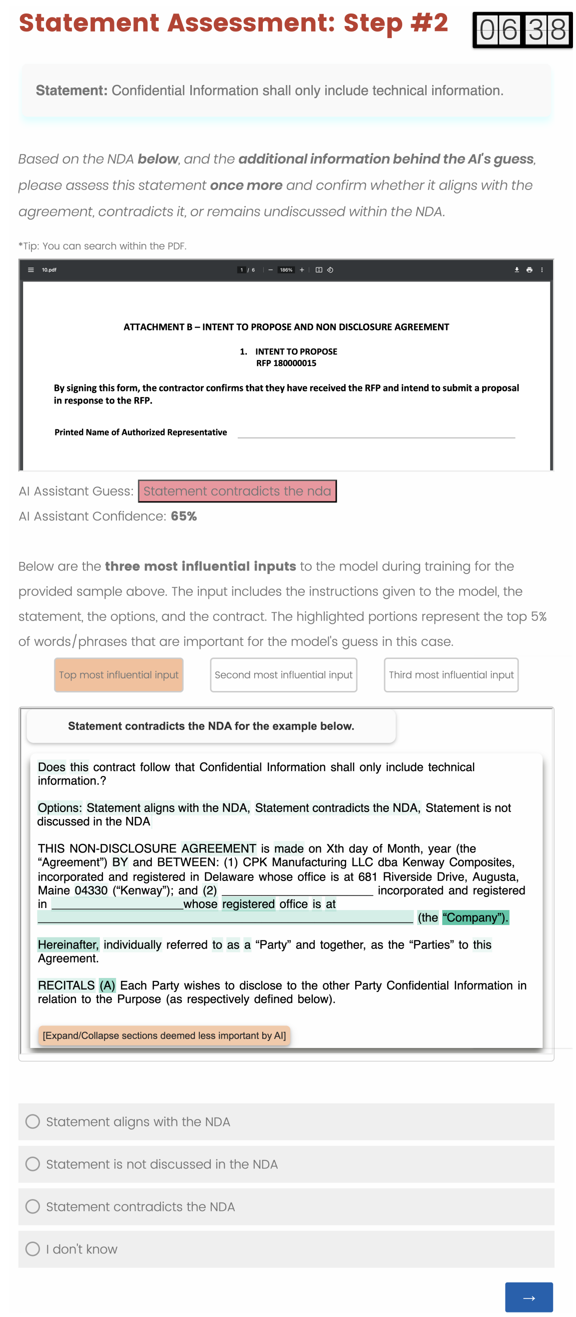


Figure 14: Second iteration of statement evaluation while accessing the EK-FAC influential training examples.

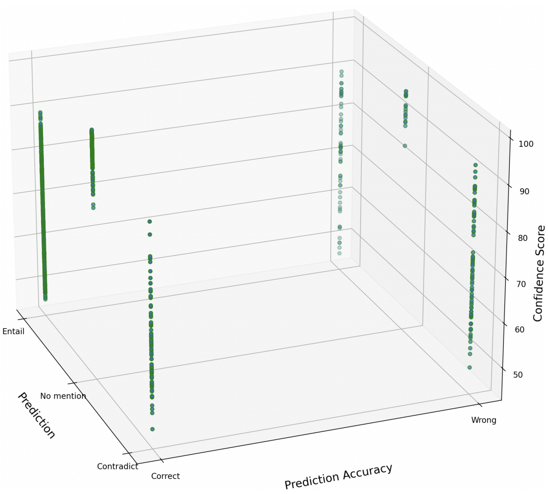


Figure 15: This plot illustrates the relationship between the model's prediction, confidence score, and prediction accuracy. Despite confidence score calibration, there is no clear correlation between the model's confidence scores and prediction accuracy.

F Categorization of ExNLP Tasks

[D1] 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

[D2] 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.

[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%
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] 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 (∃ model-human performance gap)[24]
- *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.

[D6] 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[25] [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)[26]
- *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.

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

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

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

[D7] 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.

[D8] MMLU (Hendrycks et al., 2021)

Prediction Task: Multi-choice QA in multiple domains such as elementary mathematics, US history, and law.
Average Input Length: 50 [question] + 4 [choices] = 54 words
Human Ability: 34.5% (Unspecialized people), but it varies based on the question domains and user expertise.
Application: No. The dataset does not represent a realistic task setup (similar to [D11]).
Hazard from Immediate Usage: None, as there is no realistic application.
Hazard from Downstream Impact: None, as there is no realistic application.

[D9] FoolMeTwice (Eisenschlos et al., 2021)

Prediction Task: Determine whether a given claim, spanning various domains, is supported or refuted by the provided evidence from a Wikipedia page.
Average Input Length: 15 [claim] + 30 [evidence text] = 45 words
Human Ability: Reported average accuracy is 78.1%.
The remaining information is the same as for FEVER *([D12]).*

[D10] 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.
- *Severity:* Low since the cost of a book is generally low
- *Risk:* Low

Hazard from Downstream Impact: Nothing noteworthy.

[D11] 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 for students in 3rd to 5th 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.

[D12] 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.[27]
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.

[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.[28]
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.*

[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

[27] An example of a claim in FEVER is: "Berlin is the capital of Germany."
[28] There are application-grounded versions of NLI such as EvidenceInference v2 (D38).

that it 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 it 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)[29]

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.

---

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

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

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

Prediction Task: Predicting is a forum counterargument 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, counterargument 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 words
Human Ability: Depends, but can be 100%
Application: No. Models trained on this data could be used that students in 3rd through 5th 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 words
Human Ability: Depends, but can be 100%
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.

[D28] 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 (∃ model-human performance gap)[30]
- *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.

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

Extended NATURALQUESTIONS with their explanation annotations. See [D5].

[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 [D5].

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

Extended HOTPOTQA with their explanation annotations. See [D28].

[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[31]
- *Severity:* Low since the cost of renting or seeing a movie is generally low
- *Risk:* Low

Hazard from Downstream Impact: Nothing noteworthy.

[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]).

---

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

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

[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[32]
- *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] 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

[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. The original task setup asks for a model to 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

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

- *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[33]
- *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.

---

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

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

Prediction Task: Predict is a Wikipedia revision comment on 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 §B.2 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 (D6) and NaturalQuestions (D5) 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[34]

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
- *Probability:* N/A
- *Severity:* Low (questions are not about critical information such as health, law, etc.)[35]
- *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

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

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

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
- *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

[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)

[D51] 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.

[D52] 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 anticipate 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.
