A Dataset and Benchmark on Extraction of Novel Concepts on Trust in AI from Scientific Literature

Melanie McGrath and **Harrison Bailey** and **Necva Bölücü**
Xiang Dai and **Sarvnaz Karimi** and **Andreas Duenser**[*] and **Cecile Paris**
CSIRO Data61, Sydney, Australia
firstname.lastname@csiro.au

Abstract

Information extraction from the scientific literature is a long-standing technique for transforming unstructured knowledge hidden in text into structured data, which can then be used for further analytics and decision-making in downstream tasks. A large body of scientific literature discusses *Trust in AI*, where factors contributing to human trust in artificial intelligence (AI) applications and technology are studied. It explores questions such as why people may or may not trust a self-driving car, and what factors influence such trust. The relationships of these factors with *human trust* in AI applications are complex. We explore this space through the lens of information extraction. That is, we investigate how to extract these factors from the literature that studies them. The outcome could inform technology developers to improve the acceptance rate of their products. Our results indicate that (1) while Named Entity Recognition (NER) is largely considered a solved problem in many domains, it is far from solved in extracting factors of human trust in AI from the relevant scientific literature; and, (2) supervised learning is more effective for this task than prompt-based LLMs.

1 Introduction

The rapid rate at which *Artificial Intelligence* (AI) is developing and becoming integrated into human life requires a thorough understanding of the dynamics of human trust in AI technology (Glikson and Woolley, 2020; Teaming, 2022). Addressing questions about the factors, or *antecedents*, influencing trust in specific AI systems and thresholds for excessive or insufficient trust is crucial for developing AI responsibly and preventing potential misuse (Parasuraman and Riley, 1997; Lockey et al., 2021; McGrath et al., 2025). It would also support its adoption and acceptance rate among users (Henrique and Santos, 2024).

Studies published in psychology and behavioural science, management, and computer science offer extensive insight into this domain (e.g., Glikson and Woolley, 2020; Kaplan et al., 2021; Saßmannshausen et al., 2023). However, this literature constantly expands, making it increasingly difficult for researchers to review and extract relevant knowledge, and for decision makers to understand the extent of contributing factors in the uptake of a given technology. To address this challenge, we explore *whether factors influencing human trust in AI can be identified automatically from the scientific literature?* This is particularly challenging as identifying these factors from a scientific article requires relevant expertise, such as that of a social psychologist specialised in technology uptake. To answer this question, we create the `Trust in AI` dataset, where the factors influencing trust are captured in a structured dataset, making it more accessible and easier for domain experts to navigate. For example, in the sentence on self-driving cars, "..Undesired consequences such as anxiety, competency loss and risk are deemed to foster resistance.." The underlined factors need to be extracted. However, it is not that simple. The AI and the factors fall into different categories. That is, we need to extract the type of AI *application* (e.g., self-driving car), the *factor* (trust antecedent), and the type of factor (*human*, *technological*, *context*). The extracted information has practical applications in research, industry, and commercial AI production. To build this resource, we frame the task as named entity recognition (NER).

Trust in AI as an NER task requires entity annotations. However, annotating a large number of documents can be expensive, time-consuming, and requires extensive human expertise. Limited annotated training data makes it harder to train models that rely on large datasets. To tackle this, one approach is distant supervision to automatically generate labels (Shang et al., 2018; Liang et al., 2020; Xu

[*] Corresponding author

et al., 2023). In distant supervision, the labelling procedure involves automatically detecting entity candidates using knowledge bases with syntactic or semantic similarity. Distant-annotations have both pros and cons: on the one hand, distant-annotated datasets can complement human-annotated ones, potentially mitigating biases or limitations inherent in the human-annotated dataset, and they are easy and low-cost. On the other hand, the labels generated might suffer from noisy and incomplete annotation, as human expertise was not involved.

In this study, we create and benchmark a dataset using both distant and human annotations. It is a substantial training set that incorporates both distant- and human-annotated data, enriching the training set and enhancing the generalisation capacity of supervised models. The effectiveness of distant supervision methods underscores the utility of distant-annotated datasets in training NER models, particularly in domains requiring specialised entity recognition (Jiang et al., 2021), such as Trust in AI.

Our contributions are as follows: (1) We formulate the challenging problem of information extraction (IE) for trust in AI, an area previously unexplored in the NLP domain (§3); (2) We investigate if using LLM-guided annotations as a part of the annotation process is feasible and effective, drawing inspiration from studies demonstrating the capabilities of large language models (LLMs) in simulating human-annotation (Bansal and Sharma, 2023; Goel et al., 2023; Zhang et al., 2023b); (3) We construct a human- and distant-annotated dataset of factors shown to influence trust development named `Trust in AI` (§4)[1], through an extensive process of creating re-usable guidelines with domain experts; and, (5) We provide benchmark results for the NER task with a detailed error analysis (§6).

2 Related Work

Trust in AI Trust is critical to the human willingness to adopt AI technology in a safe and productive way (Jacovi et al., 2021; Schaefer et al., 2021). Consequently, it is important to know what factors contribute to the development of an appropriate level of trust in an AI application. Over 450 distinct factors influencing trust development have been identified in the scientific literature (Saßmannshausen et al., 2023). Decades of research investigating trust in both humans and machines indicate that the antecedents of trust can be reliably classified as: (1) properties of the trustor or *human* factors (e.g., experience, risk aversion); (2) properties of the trustee or *technological* factors (e.g., performance, transparency); or, (3) properties of the task or interaction *context* (e.g., time pressure, task difficulty) (Hancock et al., 2011; Kaplan et al., 2021; Schaefer et al., 2016).

Which of these hundreds of antecedents influences trust in a particular AI application is highly variable. For example, the factors that contribute to trust in an embodied robot (Hancock et al., 2011) may be quite different to those in an algorithmic decision aid (Kaplan et al., 2021). As a result, researchers interested in trust development are increasingly seeking approaches to specifying idiosyncratic models of trust in individual applications. The `Trust in AI` dataset will provide these domain experts with a resource to identify the most relevant factors for their application based on the existing literature. To our knowledge, this is the first such resource created to be utilised by researchers and developers of both NLP and trust in AI.

Annotation of IE Dataset on Scientific Domain Annotating scientific IE datasets can be approached in two key ways: (a) annotating a small amount of data with the help of domain experts and carefully designed annotation guidelines (Kim et al., 2003; Karimi et al., 2015; Friedrich et al., 2020); and, (b) leveraging existing resources including LLMs to automatically annotate a large amount of data with no or little human intervention (Agrawal et al., 2019; Jain et al., 2020; Liu et al., 2022; Ding et al., 2023; Goel et al., 2023).

Each approach has its advantages and disadvantages, with trade-offs in terms of cost, scale, and precision in annotations. Our study fits into both categories, as the concepts of interest are complex expert annotations for this first attempt to create such a resource, along with an additional distantly annotated resource using existing knowledge (Saßmannshausen et al., 2023).

IE using LLMs IE using LLMs has gained prominence in the literature due to its potential advantages, particularly in scenarios with limited annotated data or in domains where traditional supervised approaches face challenges (Brown et al., 2020; Bubeck et al., 2023). LLMs show the capability of recognising novel entities following natural language instructions (Sainz et al., 2024;

[1] Data available at CSIRO Data Portal.

1	Trust in automation and AI query	(*artificial intelligence* **OR** *robot** **OR** *automation* **OR** *machine intelligence* **OR** *autonomy*) **AND** (*trust** **OR** *trust models* **OR** *trustworthiness* **OR** *trust calibration* **OR** *trust repair* **OR** *trust propensity* **OR** *trust development*)
2	Trust in collaboration with AI query	(*human-robot collaboration* **OR** *hybrid intelligence* **OR** *collaborative intelligence* **OR** *robot** **OR** (*collaboration* **AND** *artificial intelligence*) **OR** *human-AI collaboration* **OR** *human-robot team** **OR** *human-autonomy team** **OR** *augmented intelligence* **OR** *human-machine team**) **AND** (*trust** **OR** *trust models* **OR** *trustworthiness* **OR** *trust calibration* **OR** *trust repair* **OR** *trust propensity* **OR** *trust development*)
3	Meta-analysis query	(*trust* **OR** *trustworth**) **AND** (*technolog** **OR** *robot* **OR** *machine** **OR** *automation** **OR** *autonomy** **OR** *agent** **OR** *IT system** **OR** *IT artifact** **OR** *artificial intelligence* **OR** *machine learning*) **AND** (*trust** **OR** *trust models* **OR** *trustworthiness* **OR** *trust calibration* **OR** *trust repair* **OR** *trust propensity* **OR** *trust development*)

Table 1: Queries by domain experts that are used in dataset curation.

Munnangi et al., 2024), or via few-shot learning, useful when annotated data is scarce or expensive to obtain (Agrawal et al., 2022; Li et al., 2023; Zhang et al., 2023a). By employing prompt techniques, LLMs provide a consistent approach to various IE tasks through a single model (Wang et al., 2022). However, studies on fine-tuning LLMs called supervised fine-tuning (SFT) show that supervised models, and consequently annotated datasets, remain essential for achieving high performance in IE tasks with LLMs (Zhou et al., 2024; Gui et al., 2024; Xu et al., 2024).

3 Problem Formulation

The `Trust in AI` dataset can be conceptualised as $D = \{S_i, P_i, L_i\}_{i=1}^{N}$, where N is the total number of sentences in the dataset. For each sentence, S_i, P_i represents its context, which is the paragraph where the sentence S_i is located, and L_i is the set of entity mentions. Each element (entity) in L_i is represented as a triplet, consisting of the start index of a span, the end index, and the entity category (i.e., human factor, technology factor, context factor, and application name).

The dataset can be used for benchmarking the Named Entity Recognition task: for a given sentence S_i and the P_i (context information), the objective is to recognise all elements in L_i.

4 Trust in AI Dataset

We outline the process of curating the dataset, design and adjustments of the annotation guidelines, as well as the annotation process of the `Trust in AI` dataset below.

4.1 Dataset Curation

The dataset was built in two stages, yielding two complementary sets of articles.

Article Set 1 The first set of articles, obtained in Stage 1, was initially collected by a researcher (one of the annotators) for an unrelated literature review on trust in automation and AI, with a focus on human-AI collaboration. That researcher holds a PhD in social psychology. Articles were sourced using two searches, one focused on (i) *Trust in automation and AI* and (ii) *Trust in collaboration with AI*; see 1^{st} and 2^{nd} queries given in Table 1.

Article Set 2 The second set of articles, for Stage 2, was drawn from a 2023 meta-analysis of the antecedents of trust in AI (Saßmannshausen et al., 2023). The authors of the meta-analysis conducted an electronic search using the query (3^{rd} query given in Table 1). Of the 178 articles included in the meta-analysis, we removed papers that did not report empirical findings following a manual inspection by the domain expert.

	Number of articles		
Field	Set 1	Set 2	Total
Cognitive science	0	2	2
Computing	15	51	66
Defence	0	1	1
Economics	0	4	4
Education	0	1	1
Engineering	1	6	7
Ergonomics	2	30	32
Health and medicine	0	3	3
Law	0	2	2
Management	1	10	11
Neuroscience	0	1	1
Operations research	0	1	1
Psychology	1	9	10
Robotics	10	39	49

Table 2: Fields of research represented in the `Trust in AI` dataset.

Databases Databases searched included the ACM Digital Library, EBSCO[2], Emerald Insight[3], the IEEE Xplore Digital Library, JSTOR[4], ProQuest, PsycARTICLES[5], the Psychology and Behavioural Science Collection[6], PsycINFO[7], PSYNEX[8], ScienceDirect[9], and Web of Science[10].

Composition of Final Dataset Upon combining sets 1 and 2, eight articles were duplicated, leaving a total of 186 articles in the final `Trust in AI` dataset. The articles come from a wide range of fields, including *computing*, *robotics*, *psychology*, *economics*, and *management*, with each article potentially belonging to one or more of these fields. The largest number of articles is from the fields of computing (66 articles) and robotics (49 articles), with sub-fields represented, including human-computer interaction, control systems, communications, and information systems. The full list of fields of research is given in Table 2.

4.2 Annotation Process

Two annotators (one researcher holding a PhD in social psychology and one final year student majoring in computer science and politics) completed the annotation task using the Prodigy annotation tool (Montani and Honnibal, 2018). Details about the annotation interface are available in Appendix B.1. The annotation was conducted in five phases:

i. Preparation of the guideline: The annotation guideline was developed through a small pilot annotation by one annotator (the same annotator who conducted the literature search) using 5 articles.

ii. Initial annotation: Both annotators annotated the same 5 articles used in the first phase (233 sentences, total 535 entities).

iii. Resolution: Annotators discussed their annotations on these 5 articles to resolve discrepancies, leading to updates in the annotation guideline. The annotation guideline is presented in Appendix B.2.

iv. Test-set annotation: We conducted the annotations of the test set (in total 16 articles) with two annotators in two parts:

- *Manual annotation:* Annotators annotated the same 5 articles (used in phase iii) using the updated guidelines and then annotated an additional 5 randomly chosen articles manually.
- *LLM guided annotation:* Inspired by studies in dataset annotation (Bansal and Sharma, 2023; Goel et al., 2023; Zhang et al., 2023b), we utilised LLMs as guidance for annotators in the annotation of 6 new randomly chosen articles. We displayed the predictions of LLMs (details in Appendix B.3) with the aim to assist annotators in the process and reduce annotation time. Annotators subsequently rectified any errors made by LLM, allowing us to compare the effectiveness of LLM with manual annotation. The Cohen's Kappa score between LLM-agent and human annotators was low (0.129) on these 6 articles, highlighting the complexity of the task for LLMs and the necessity of a human-annotated dataset. Annotators also noted that LLMs tend to over-annotate a given sentence, and, therefore, guidance in annotation increases the annotation time due to the rectification of errors compared to manual annotation.

Test-set Inter-Annotator Agreement After phase ii, inter-annotator Cohen's kappa (κ) score (Cohen, 1960) is 0.395 on the 5 articles for 233 sentences and 535 entities. Upon resolution

[2] https://www.ebsco.com

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

[4] https://www.jstor.org

[5] www.apa.org/pubs/databases/psycarticles

[6] https://www.ebsco.com/products/research-databases/psychology-behavioral-sciences-collection

[7] https://www.apa.org/pubs/databases/psycinfo

[8] https://www.medicinesfaq.com/brand/psynex

[9] https://www.sciencedirect.com

[10] https://www.webofscience.com/wos

in phase iii, it is observed that the main disagreement is the annotation of *application* and *technology*. Following the resolution, a substantial overall agreement of 0.933 is achieved on these 5 articles. Kappa values over 0.9 are considered near perfect agreement (Cohen, 1960; McHugh, 2012), possibly reflecting the high coverage of annotation guidelines and the training of the annotators. The agreement score of test-set annotation (phase iv) is 0.818 for 16 articles.

v. Training-set annotation: The annotator with a PhD in social psychology annotated 34 articles based on the updated guidelines (6 articles with LLM-guided annotation and 28 articles manually).

4.3 Distant Annotation

To construct the distant-annotated training data, we utilised the meta-analysis of the antecedents of trust in AI and their corresponding dictionary of factors (*human*, *technological*, *context*), consisting of 483 factors in total (Saßmannshausen et al., 2023). [11] First, we extracted noun phrases, such as "Pro-social virtual AI's behaviors" and "the average human rating", from each article using the SpaCy library[12] (Honnibal and Montani, 2017). Then, we measured similarities between these extracted noun phrases against each factor item in the above-mentioned dictionary. We employed two approaches based on string similarity and embedding similarity, respectively. For the string similarity-based approach, we calculated the longest common character subsequence between the candidate phrase and the dictionary item. For the method based on embedding (vector) similarity, we encoded all noun phrases and dictionary items using the same sentence-transformers model[13] and calculated cosine similarity between the obtained vectors. Finally, each candidate noun phrase was labelled with the corresponding entity category based on its most similar factor item from the dictionary.

4.4 Dataset Details

Descriptive statistics of the human and distant annotation (string+vector) datasets are given in Table 3. Note that we split our training set into training and development sets. The distribution of *application* and factor types (*context, human, technology*) in the human-annotated dataset is presented in Figure 1.

	Human Annotated				
Statistic	**Train**	**Dev**	**Test**	**Distant**	**Overall**
# articles	–34–		16	136	186
# paragraphs	340	41	95	7,087	7,563
# sentences	1,833	184	548	35,173	37,738
avg len sentences	146.53	157.83	153.22	152.12	151.89
# tokens	47,829	5,101	15,229	971,127	1,039,286
# entities	4,140	286	1,142	880,112	5,568
avg len entities	12.56	9.83	13.66	8.04	12.65

Table 3: Descriptive statistics of `Trust in AI` dataset including both human- and distant-annotations.

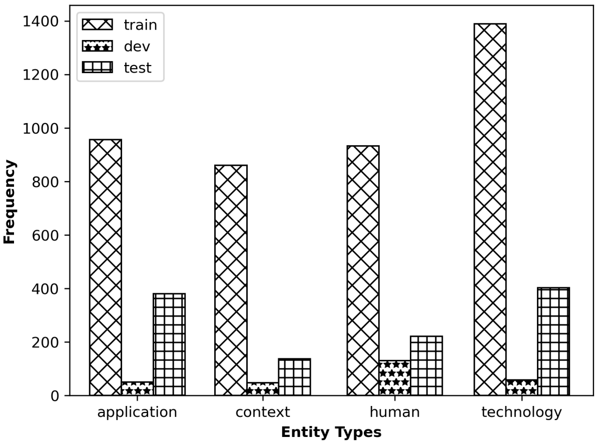


Figure 1: Distribution of factors and application of our human-annotated dataset.

5 Experimental Setup

Task Description Our task involves the identification and classification of factors (e.g., human, context) within a given sentence. Since a span can have more than one-factor category or nested entities, the task is formulated as a span-based NER task. The problem can be formulated as the token-level classification task for a sequence of words $\{w_1, \cdots, w_n\}$, wherein entity labels $\{y_1, \cdots, y_n\}$ are assigned.

NER models We benchmark the effectiveness of several models that belong to three groups: (1) **Fully supervised:** We investigate three supervised methods; (1.1) *PLM*, which consists of a PLM encoder (Liu et al., 2019) and a span-based classifier on top of the encoder (Zhong and Chen, 2021); (1.2) *Seq2seq-BERT*[14] (Straková et al., 2019), a sequence-to-sequence model consisting of an encoder-decoder with LSTM; and (1.3) *BiaffineNER*[15] (Yu et al., 2020), formulated as a graph-based parsing task composed of a BiLSTM encoder with a biaffine classifier. (2) **Few-shot**

[11] The dictionary of factors can be found at `tandfonline.com/doi/full/10.1080/00140139.2022.2120634`.

[12] https://spacy.io

[13] all-mpnet-base-v2 (accessed Oct 2025)

[14] https://github.com/ufal/acl2019_nested_ner

[15] https://github.com/LindgeW/BiaffineNER

	Category				Overall	
Method	Application	Human fac	Technology fac	Context fac	Micro F_1	Macro F_1
PLM	**63.71**$_{\pm 1.8}$	30.74$_{\pm 3.4}$	22.04$_{\pm 2.9}$	**54.36**$_{\pm 1.2}$	**49.38**$_{\pm 1.1}$	**42.71**$_{\pm 1.9}$
Seq2seq-BERT	57.48$_{\pm 1.2}$	28.14$_{\pm 2.5}$	19.45$_{\pm 3.0}$	41.47$_{\pm 1.1}$	43.52$_{\pm 0.9}$	35.16$_{\pm 2.0}$
BiaffineNER	61.12$_{\pm 1.4}$	**32.17**$_{\pm 3.4}$	20.17$_{\pm 2.5}$	46.17$_{\pm 2.0}$	46.43$_{\pm 1.5}$	37.23$_{\pm 2.1}$
Few-shot Learning	33.04$_{\pm 0.2}$	17.91$_{\pm 0.1}$	40.52$_{\pm 0.4}$	18.31$_{\pm 0.4}$	26.42$_{\pm 0.3}$	26.20$_{\pm 0.3}$
Distant Supervision	62.30$_{\pm 1.6}$	23.12$_{\pm 2.7}$	**53.92**$_{\pm 0.5}$	30.08$_{\pm 1.7}$	48.51$_{\pm 1.3}$	42.36$_{\pm 1.0}$

Table 4: Comparison of models in terms of entity-level F_1 for NER task. The best results are **boldfaced**, while the second-best is underlined. 'fac' stands for factor. We report the best results for PLM (RoBERTa-Large), Few-shot Learning (GPT-OSS BM25 5-shot) and Distant Supervision (RoBERTa-Large string+vector), and the full results can be found in Appendix A.

learning: Leveraging in-context learning (ICL) methods as formulated by Bölücü et al. (2023). (3) **Distant supervision:** We use PLM with RoBERTa encoder (Base or Large) (Liu et al., 2019) and a span-based classifier on top of the encoder, with a weighted loss where the labels of weights are the similarity scores used in the distant annotation. In the rest of the paper, we refer to the PLM with the encoder (supervised setting) as PLM (encoder name) and to the distant annotation model simply as RoBERTa Base/Large to avoid confusion. The training consists of two steps: training with distant-annotated data, followed by continual learning with the human-annotated training set.

Training Configuration

- **Fully supervised:** All hyper-parameters used in the supervised baseline models for the NER task are tuned on the development set. For PLM, the hyperparameters are the learning rate of 5e-4, max length of 128, a context window of 200 tokens and a batch size of 16, and models are trained for max epochs of 30. For Seq2seq-BERT and BiaffineNER, we use the default hyperparameters suggested by the authors except for the learning rate (5e-3).

- **Few-shot learning:** For zero- and few-shot learning, we adopt the prompt template provided by the EasyInstruct library[16] (Ou et al., 2024) for ICL. We use to models: GPT3.5-Turbo and GPT-OSS (20B) (OpenAI et al., 2025) (temperature: 0.1). We follow the study of Bölücü et al. (2023) to select ICL samples.

- **Distant Supervision:** We fine-tune all hyper-parameters of the method using the development set. The hyperparameters are the learning rate of 5e-4, max length of 128, a batch size of 16, and models are trained for a maximum of 30.

All experiments are repeated three times, and mean values and standard deviations are reported.

Evaluation Metrics We use entity-level Macro F_1 score (Nakayama, 2018) for the NER task.

6 Results and Analysis

The results are shown in Table 4. We observe that the supervised models outperform those using LLM in zero and few-shot settings, consistent with the studies of Jimenez Gutierrez et al. (2022) and Bölücü et al. (2023). The PLM (RoBERTa-Large) model performs the best in most categories, except for *technology*. Additionally, both the best and second-best models rely on human-annotated datasets for their performance. In few-shot learning, ICL also requires a very small amount of human-annotated data. While this approach is powerful in its ability to adapt to new tasks with limited annotated data, the performance of this model is still behind that of the supervised models. This highlights how important human-annotated data is for these models to do well in NER tasks.

Another observation is the substantial improvement observed in recognising the *technology* factor through distant supervision. Considering that this factor has the highest distribution within the distant-annotated dataset, it suggests that achieving better results with distant supervision may require a larger distant-annotated dataset during the training process.

In the annotation of factor types and *application*, one word can refer to one or more factor types. For instance, the word *adaptability* in *user and robot adaptability* refers to both *human* and *technology* factors. Moreover, the mentioned factor may span several words, not all of which are included in the

[16] https://github.com/zjunlp/EasyInstruct

Human Annotated	Extracted by NER	Error Type
—	displaying information (*technology*)	Complete False Positives
care context (*context*)	—	Complete False Negatives
context (*context*)	context (*technology*)	Wrong Label Right Span
production (*context*)	production robots (*application*)	Wrong Label Overlapping Span
design factors of the robotic interface (*technology*)	robotic (*application*)	Right Label Overlapping Span

Table 5: Selected examples for the error types made by the fully supervised method PLM (RoBERTa-large).

Method	Error Type	Context	Application	Technology	Human	Total
PLM (RoBERTa-Large)	Complete False Negatives	100	135	204	189	628
	Complete False Positives	51	95	148	100	394
	Wrong Label Right Span	14	23	61	62	160
	Wrong Label Overlapping Span	3	26	48	57	134
	Right Label Overlapping Span	83	102	19	40	244
RoBERTa-Large string+vector	Complete False Negatives	96	170	209	187	662
	Complete False Positives	115	76	128	76	395
	Wrong Label Right Span	8	3	38	70	119
	Wrong Label Overlapping Span	3	2	37	75	117
	Right Label Overlapping Span	32	69	62	34	197

Table 6: Statistical details of error types observed in the NER methods.

same factor. For instance, an article might mention *training of communication and trust calibration*, where *training of communication* is a *technology* factor while *training · · · of trust calibration* is *human* factor. This complexity makes the NER task challenging. Even though span-based models are applied to extract factors and applications, the results of NER models on the annotated dataset remain relatively poor, except for *application*. It aligns with the resolution phase of the annotation, where annotators find that *human* and *technology* factors are the most confusing, prompting an update to the annotation guidelines to provide clarity in distinguishing between the annotation of these factors. Supervised models still struggle to extract factors, a task that is challenging even for human annotators with domain expertise. Finally, *application* is expected to be used to label entities that may contain the AI technology or the studied use case, potentially contributing to lower results for the *technology* factor.

6.1 Error Analysis

Entity-level F_1 score, the most common metric for NER models, only credits a prediction when both the span and the label precisely match the annotation. We investigate the predictions of the NER methods to elucidate the common errors made by these methods for the newly designed problem of Trust in AI and the newly defined entity types.

We analyse the predictions of the two methods from our baseline methods: (1) **Fully supervised:** PLM (RoBERTa-Large) and (2) **Distant supervision:** RoBERTa-Large string+vector.

Following Nejadgholi et al. (2020), we analyse the errors in five categories:

- **Complete False Positive:** An entity is predicted by the NER model, but it is not annotated in the human-annotated text.
- **Complete False Negative:** A human-annotated entity is not predicted by the NER model.
- **Wrong Label Right Span:** A human-annotated entity and a predicted entity by the NER model share the same spans but different entity types.
- **Wrong label Overlapping Span:** A human-annotated entity and a predicted entity have overlapping spans but different entity types.
- **Right label Overlapping Span:** A human-annotated entity and a predicted entity have overlapping spans and the same entity types.

Samples of error types made by the fully supervised method can be found in Table 5.

The error analysis of the NER methods, as detailed in Table 6, provides insights into the performance and challenges faced by different approaches. For the PLM (RoBERTa-Large) model (fully supervised), the most common error type is *Complete False Negatives*, particularly with the *technology* and *human* entities, indicating that many entities annotated by humans were missed by the models. It is observed that the PLM (RoBERTa-Large) (fully supervised) method detects the correct span in the sentence but mislabels it (*Wrong Label Right Span*).

For the RoBERTa-Large string+vector model (distant supervision), the errors are distributed similarly to the PLM (RoBERTa-Large) model, with *Complete False Negatives* being the most common error type. However, this model shows improvement in identifying entities with overlapping spans, as indicated by a higher count of *Right Label Overlapping Span* errors compared to the PLM (RoBERTa-Large) model.

We also analysed the entities for each error type and observed that entities are typically single tokens in the *Wrong Label Right Span* error type, whereas entities are often multiple tokens in the *Wrong Label Overlapping Span* and *Right Label Overlapping Span* error types.

The occurrence of the Right Label Overlapping Span error type may be attributed to the nature of span-based annotation. In span-based annotation, a span may encompass more than one entity type or contain nested entities (e.g., *synchronous and co-located teamwork* and *co-located teamwork* are annotated as *context* in the dataset).

We observe that fully supervised approaches still struggle with newly designed entity types (context, human, technology) of the newly designed problem. Distant supervision methods, while improving in some aspects such as overlapping span identification, still exhibit similar error distributions. Additionally, few-shot learning using LLMs struggles with NER tasks for the newly designed problem with new entity types. These findings underscore again the importance of the annotated dataset for extracting trust-related factors in AI literature. Moreover, while the distant-annotated dataset is noisy, it is still a valuable resource for training NER models in our context of trust in AI. Future work should focus on refining distant annotation techniques and exploring hybrid approaches that combine the strengths of both fully supervised and distant supervision methods to improve the accuracy and reliability of entity recognition in the domain of trust in AI.

7 Conclusion and Future Work

Identifying antecedents of human trust in AI from scientific literature has been largely explored only via manual inspection of relevant literature. This manual process, crucial in the development of new AI applications, has hardly been automated itself using techniques developed in artificial intelligence domains, including NLP. We are the first study to tackle this problem.

We investigated whether information extraction techniques, and in particular named entity recognition (NER), can be developed in this space to extract factors of trust in AI. While NER is considered a largely solved problem in many domains, using it to obtain factors related to trust in AI in scientific literature is unexplored, requiring the careful creation of a dataset with expert annotators. We also explore distant annotation. These led to a novel dataset, which we named `Trust in AI`. We then benchmarked state-of-the-art NER techniques such as those using in-context learning and LLMs. We showed that the existing LLMs, such as GPT, are not effective in extracting concepts of interest in Trust in AI. Our dataset is one important step in opening an avenue for further research in this space.

In the future, we plan to extend the dataset to include relations between *factors* and *trust* and entity resolution to identify and link entities that refer to the same entity, providing a more cohesive and accurate representation.

Ethics Statement

As we create a dataset, there are ethical considerations about the use of the data. The dataset used in our work is collected from scientific articles that are publicly available. However, some may require subscriptions to the journals for their users. We make links to the articles available so as not to redistribute those without their publishers' permission. The annotations were conducted by two of the authors as part of their research duties.

Limitations

Language. This dataset only uses English scientific literature, which may limit its usage for other languages.

Subjectivity and Background Knowledge. The dataset annotation is done by two human annotators with different background knowledge, with one expert in the *Trust in AI* domain with a psychology background and another in computer science and politics.

References

Kritika Agrawal, Aakash Mittal, and Vikram Pudi. 2019. Scalable, semi-supervised extraction of structured information from scientific literature. In *Proceedings of the Workshop on Extracting Structured Knowledge from Scientific Publications*, pages 11–20.

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Parikshit Bansal and Amit Sharma. 2023. Large language models as annotators: Enhancing generalization of nlp models at minimal cost. *arXiv preprint arXiv:2306.15766*.

Necva Bölücü, Maciej Rybinski, and Stephen Wan. 2023. impact of sample selection on in-context learning for entity extraction from scientific writing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5090–5107, Singapore. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. 33:1877–1901.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. 2023. Is GPT-3 a Good Data Annotator? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11173–11195, Toronto, Canada. Association for Computational Linguistics.

Annemarie Friedrich, Heike Adel, Federico Tomazic, Johannes Hingerl, Renou Benteau, Anika Maruscyk, and Lukas Lange. 2020. The SOFC-Exp Corpus and Neural Approaches to Information Extraction in the Materials Science Domain. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Ella Glikson and Anita Williams Woolley. 2020. Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, 14(2):627–660.

Akshay Goel, Almog Gueta, Omry Gilon, Chang Liu, Sofia Erell, Lan Huong Nguyen, Xiaohong Hao, Bolous Jaber, Shashir Reddy, Rupesh Kartha, Jean Steiner, Itay Laish, and Amir Feder. 2023. LLMs Accelerate Annotation for Medical Information Extraction. In *Proceedings of the 3rd Machine Learning for Health Symposium*, volume 225 of *Proceedings of Machine Learning Research*, pages 82–100. PMLR.

Honghao Gui, Shuofei Qiao, Jintian Zhang, Hongbin Ye, Mengshu Sun, Lei Liang, Jeff Z. Pan, Huajun Chen, and Ningyu Zhang. 2024. InstructIE: A Bilingual Instruction-based Information Extraction Dataset.

Peter A. Hancock, Deborah R. Billings, Kristin E. Schaefer, Jessie Y.C. Chen, Ewart J. De Visser, and Raja Parasuraman. 2011. A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors*, 53(5):517–527.

Bruno Miranda Henrique and Eugene Santos. 2024. Trust in artificial intelligence: Literature review and main path analysis. *Computers in Human Behavior: Artificial Humans*, 2(1):100043.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1):411–420.

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*, FAccT '21, page 624–635, New York, NY, USA. Association for Computing Machinery.

Sarthak Jain, Madeleine van Zuylen, Hannaneh Hajishirzi, and Iz Beltagy. 2020. SciREX: A challenge dataset for document-level information extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7506–7516, Online. Association for Computational Linguistics.

Haoming Jiang, Danqing Zhang, Tianyu Cao, Bing Yin, and Tuo Zhao. 2021. Named Entity Recognition with Small Strongly Labeled and Large Weakly Labeled Data. 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 1775–1789.

Bernal Jimenez Gutierrez, Nikolas McNeal, Clayton Washington, You Chen, Lang Li, Huan Sun, and Yu Su. 2022. Thinking about GPT-3 In-Context Learning for Biomedical IE? Think Again. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4497–4512, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alexandra D. Kaplan, Theresa T. Kessler, J. Christopher Brill, and P. A. Hancock. 2021. Trust in artificial intelligence: Meta-analytic findings. *Human Factors*, page 001872082110139.

Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. 2015. CADEC: A corpus of adverse drug event annotations. *JBI*, 55.

J.-D. Kim, T Ohta, Y Tateisi, and J Tsujii. 2003. GENIA corpus–a semantically annotated corpus for biotextmining. *Bioinformatics*, 19.

Peng Li, Tianxiang Sun, Qiong Tang, Hang Yan, Yuanbin Wu, Xuanjing Huang, and Xipeng Qiu. 2023. CodeIE: Large Code Generation Models are Better Few-Shot Information Extractors. pages 15339–15353.

Chen Liang, Yue Yu, Haoming Jiang, Siawpeng Er, Ruijia Wang, Tuo Zhao, and Chao Zhang. 2020. BOND: BERT-Assisted Open-Domain Named Entity Recognition with Distant Supervision. New York, NY, USA. Association for Computing Machinery.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. WANLI: Worker and AI Collaboration for Natural Language Inference Dataset Creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Steven Lockey, Nicole Gillespie, Daniel Holm, and Ida Asadi Someh. 2021. A review of trust in artificial intelligence: Challenges, vulnerabilities and future directions. In *Proceedings of the Annual Hawaii International Conference on System Sciences*, volume 2020-January, pages 5463–5472. IEEE Computer Society. ISSN: 15301605.

Melanie J. McGrath, Andreas Duenser, Justine Lacey, and Cécile Paris. 2025. Collaborative human-AI trust (CHAI-T): A process framework for active management of trust in human-AI collaboration. *Computers in Human Behavior: Artificial Humans*, 6:100200.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Ines Montani and Matthew Honnibal. 2018. Prodigy: A new annotation tool for radically efficient machine teaching. *Artificial Intelligence*.

Monica Munnangi, Sergey Feldman, Byron Wallace, Silvio Amir, Tom Hope, and Aakanksha Naik. 2024. On-the-fly Definition Augmentation of LLMs for Biomedical NER. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*.

Hiroki Nakayama. 2018. seqeval: A python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

Isar Nejadgholi, Kathleen C. Fraser, and Berry de Bruijn. 2020. Extensive Error Analysis and a Learning-Based Evaluation of Medical Entity Recognition Systems to Approximate User Experience. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 177–186, Online. Association for Computational Linguistics.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, Kai Chen, Mark Chen, Enoch Cheung, Aidan Clark, Dan Cook, Marat Dukhan, Casey Dvorak, Kevin Fives, Vlad Fomenko, Timur Garipov, Kristian Georgiev, Mia Glaese, Tarun Gogineni, Adam Goucher, Lukas Gross, Katia Gil Guzman, John Hallman, Jackie Hehir, Johannes Heidecke, Alec Helyar, Haitang Hu, Romain Huet, Jacob Huh, Saachi Jain, Zach Johnson, Chris Koch, Irina Kofman, Dominik Kundel, Jason Kwon, Volodymyr Kyrylov, Elaine Ya Le, Guillaume Leclerc, James Park Lennon, Scott Lessans, Mario Lezcano-Casado, Yuanzhi Li, Zhuohan Li, Ji Lin, Jordan Liss, Lily, Liu, Jiancheng Liu, Kevin Lu, Chris Lu, Zoran Martinovic, Lindsay McCallum, Josh McGrath, Scott McKinney, Aidan McLaughlin, Song Mei, Steve Mostovoy, Tong Mu, Gideon Myles, Alexander Neitz, Alex Nichol, Jakub Pachocki, Alex Paino, Dana Palmie, Ashley Pantuliano, Giambattista Parascandolo, Jongsoo Park, Leher Pathak, Carolina Paz, Ludovic Peran, Dmitry Pimenov, Michelle Pokrass, Elizabeth Proehl, Huida Qiu, Gaby Raila, Filippo Raso, Hongyu Ren, Kimmy Richardson, David Robinson, Bob Rotsted, Hadi Salman, Suvansh Sanjeev, Max Schwarzer, D. Sculley, Harshit Sikchi, Kendal Simon, Karan Singhal, Yang Song, Dane Stuckey, Zhiqing Sun, Philippe Tillet, Sam Toizer, Foivos Tsimpourlas, Nikhil Vyas, Eric Wallace, Xin Wang, Miles Wang, Olivia Watkins, Kevin Weil, Amy Wendling, Kevin Whinnery, Cedric Whitney, Hannah Wong, Lin Yang, Yu Yang, Michihiro Yasunaga, Kristen Ying, Wojciech Zaremba, Wenting Zhan, Cyril Zhang, Brian Zhang, Eddie Zhang, and Shengjia Zhao. 2025. gpt-oss-120b gpt-oss-20b Model Card.

Yixin Ou, Ningyu Zhang, Honghao Gui, Ziwen Xu, Shuofei Qiao, Runnan Fang, Lei Li, Zhen Bi, Guozhou Zheng, and Huajun Chen. 2024. EasyInstruct: An Easy-to-use Instruction Processing Framework for Large Language Models. pages 94–106.

Raja Parasuraman and Victor Riley. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2):230–253. Publication Title: HUMAN FACTORS Volume: 39 Issue: 2.

Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, and Eneko Agirre. 2024. Gollie: Annotation guidelines improve zero-shot information-extraction. In *International Conference on Learning Representations*.

Till Saßmannshausen, Peter Burggräf, Marc Hassenzahl, and Johannes Wagner. 2023. Human trust in otherware–a systematic literature review bringing all antecedents together. *Ergonomics*, 66(7):976–998.

Kristin E. Schaefer, Jessie Y.C. Chen, James L. Szalma, and P. A. Hancock. 2016. A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human Factors*, 58(3):377–400. Publisher: SAGE Publications Inc.

Kristin E. Schaefer, Brandon S. Perelman, Gregory M. Gremillion, Amar R. Marathe, and Jason S. Metcalfe. 2021. A roadmap for developing team trust metrics for human-autonomy teams. In *Trust in Human-Robot Interaction*, pages 261–300. Elsevier.

Jingbo Shang, Liyuan Liu, Xiaotao Gu, Xiang Ren, Teng Ren, and Jiawei Han. 2018. Learning Named Entity Tagger using Domain-Specific Dictionary. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2054–2064, Brussels, Belgium. Association for Computational Linguistics.

Jana Straková, Milan Straka, and Jan Hajic. 2019. Neural Architectures for Nested NER through Linearization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5326–5331, Florence, Italy. Association for Computational Linguistics.

Human-AI Teaming. 2022. *Human-AI State-of-the-Art and Research Needs*. Washington, DC: National Academies Press. Pages: 26355.

Chenguang Wang, Xiao Liu, Zui Chen, Haoyun Hong, Jie Tang, and Dawn Song. 2022. DeepStruct: Pretraining of language models for structure prediction. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 803–823, Dublin, Ireland. Association for Computational Linguistics.

Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, Yang Wang, and Enhong Chen. 2024. Large language models for generative information extraction: A survey. *Frontiers of Computer Science*, 18(6):186357.

Lu Xu, Lidong Bing, and Wei Lu. 2023. Sampling Better Negatives for Distantly Supervised Named Entity Recognition. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4874–4882, Toronto, Canada. Association for Computational Linguistics.

Juntao Yu, Bernd Bohnet, and Massimo Poesio. 2020. Named Entity Recognition as Dependency Parsing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6470–6476, Online. Association for Computational Linguistics.

Kai Zhang, Bernal Jimenez Gutierrez, and Yu Su. 2023a. Aligning Instruction Tasks Unlocks Large Language Models as Zero-Shot Relation Extractors. pages 794–812.

Ruoyu Zhang, Yanzeng Li, Yongliang Ma, Ming Zhou, and Lei Zou. 2023b. LLMaAA: Making Large Language Models as Active Annotators. pages 13088–13103.

Zexuan Zhong and Danqi Chen. 2021. A Frustratingly Easy Approach for Entity and Relation Extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 50–61, Online. Association for Computational Linguistics.

Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. 2024. UniversalNER: Targeted Distillation from Large Language Models for Open Named Entity Recognition.

A All results

We applied several models to benchmark the newly proposed problem with new entity types. We used the pattern-based method, which was employed to label the distantly annotated dataset, to label the test set. By including the results of the pattern-based method, all the results belonging to the four groups are listed in this section.

Fully supervised For the PLM fully supervised model, we applied several PLMs as encoders of the model. The results are given in Table 7.

Few-shot learning We applied the ICL methods (5 shot) explained in the study of Bölücü et al. (2023). The results of each ICL sample selection method are given in Table 8.

Distant supervision We applied the distant supervision method with RoBERTa-Base and RoBERTa-Large as encoders using string (syntactic), vector (semantic), and combined string+vector datasets to assess the individual impacts of string,

Method	Category				Overall	
	Application	Human fac	Technology fac	Context fac	Micro F$_1$	Macro F$_1$
BERT-Base	60.45$_{\pm 1.2}$	29.13$_{\pm 4.7}$	19.45$_{\pm 4.2}$	46.14$_{\pm 0.7}$	43.10$_{\pm 2.09}$	38.30$_{\pm 2.3}$
RoBERTa-Base	64.25$_{\pm 1.1}$	33.24$_{\pm 1.1}$	17.61$_{\pm 1.4}$	51.43$_{\pm 1.8}$	48.10$_{\pm 0.9}$	41.63$_{\pm 0.6}$
RoBERTa-Large	63.71$_{\pm 1.8}$	30.74$_{\pm 3.4}$	22.04$_{\pm 2.9}$	54.36$_{\pm 1.2}$	**49.38**$_{\pm 1.1}$	**42.71**$_{\pm 1.9}$

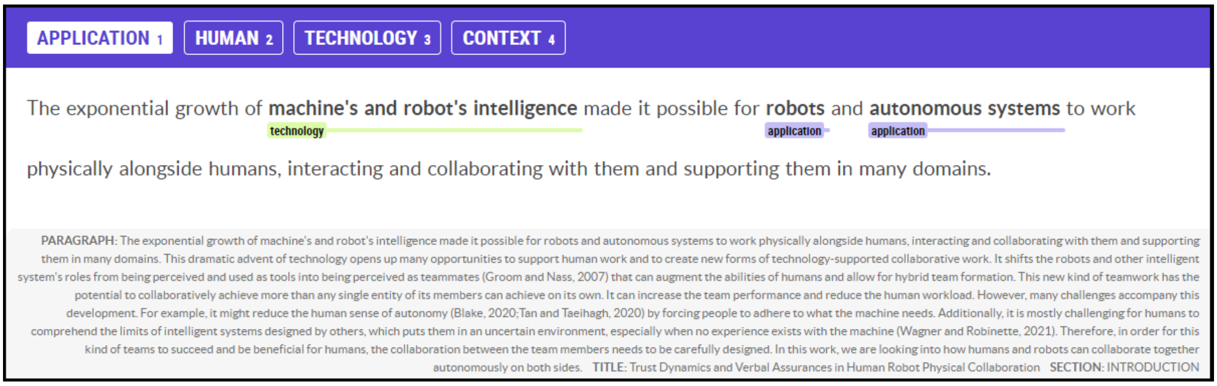
Table 7: PLM fully supervised model with several PLMs as an encoder for the NER task. The best results are **boldfaced**. 'fac' stands for factor.

LLM	Category				Overall	
	Application	Human fac	Technology fac	Context fac	Micro F$_1$	Macro F$_1$
Zero-shot						
GPT 3.5	7.81$_{\pm 0.7}$	6.50$_{\pm 0.5}$	8.64$_{\pm 1.3}$	10.92$_{\pm 0.0}$	7.61$_{\pm 0.2}$	6.91$_{\pm 0.1}$
GPT-OSS	18.60$_{\pm 0.2}$	14.54$_{\pm 0.1}$	22.00$_{\pm 0.4}$	14.11$_{\pm 0.0}$	14.32$_{\pm 0.2}$	13.81$_{\pm 0.1}$
Random Sampling						
GPT 3.5	21.31$_{\pm 2.1}$	11.21$_{\pm 1.7}$	13.33$_{\pm 1.8}$	11.50$_{\pm 2.0}$	15.22$_{\pm 2.4}$	15.51$_{\pm 2.0}$
GPT-OSS	34.10$_{\pm 0.5}$	19.72$_{\pm 0.6}$	36.20$_{\pm 0.9}$	**19.81**$_{\pm 1.1}$	24.84$_{\pm 0.9}$	24.24$_{\pm 0.8}$
KATE						
GPT 3.5	22.03$_{\pm 0.8}$	12.87$_{\pm 0.8}$	18.08$_{\pm 0.9}$	10.76$_{\pm 0.6}$	15.62$_{\pm 0.8}$	15.94$_{\pm 0.7}$
GPT-OSS	**35.20**$_{\pm 0.4}$	**20.02**$_{\pm 0.4}$	39.41$_{\pm 0.5}$	17.44$_{\pm 0.3}$	25.42$_{\pm 0.4}$	25.02$_{\pm 0.3}$
BM25						
GPT 3.5	18.22$_{\pm 0.4}$	10.53$_{\pm 0.8}$	25.62$_{\pm 1.2}$	9.14$_{\pm 1.3}$	16.13$_{\pm 1.0}$	15.88$_{\pm 1.1}$
GPT-OSS	33.04$_{\pm 0.2}$	17.91$_{\pm 0.1}$	**40.52**$_{\pm 0.4}$	18.31$_{\pm 0.4}$	**26.42**$_{\pm 0.3}$	**26.20**$_{\pm 0.3}$

Table 8: Zero-shot, and 5-shot results of each ICL method for the NER task. The best results are **boldfaced**. 'fac' for factor.

Method	Category				Overall	
	Application	Human fac	Technology fac	Context fac	Micro F$_1$	Macro F$_1$
RoBERTa-Base string	60.84$_{\pm 2.2}$	17.24 $_{\pm 1.1}$	48.85$_{\pm 2.3}$	31.36$_{\pm 2.6}$	46.08$_{\pm 0.7}$	39.57$_{\pm 0.3}$
RoBERTa-Large string	59.81$_{\pm 2.1}$	22.42$_{\pm 1.4}$	54.01$_{\pm 1.2}$	29.84$_{\pm 0.4}$	46.91$_{\pm 1.4}$	41.51$_{\pm 1.6}$
RoBERTa-Base vector	64.22$_{\pm 0.8}$	17.81$_{\pm 1.5}$	51.32$_{\pm 0.9}$	31.83$_{\pm 1.8}$	**48.51**$_{\pm 0.9}$	41.30$_{\pm 1.3}$
RoBERTa-Large vector	60.62$_{\pm 2.0}$	21.29$_{\pm 2.5}$	54.07$_{\pm 0.4}$	26.57$_{\pm 3.0}$	48.12$_{\pm 1.3}$	40.64$_{\pm 0.8}$
RoBERTa-Base string+vector	64.88$_{\pm 0.2}$	17.49$_{\pm 1.0}$	50.40$_{\pm 2.8}$	32.33$_{\pm 1.1}$	48.38$_{\pm 0.7}$	41.27$_{\pm 0.8}$
RoBERTa-Large string+vector	62.30$_{\pm 1.6}$	23.12$_{\pm 2.7}$	53.92$_{\pm 0.5}$	30.08$_{\pm 1.7}$	**48.51**$_{\pm 1.3}$	**42.36**$_{\pm 1.0}$

Table 9: Distant supervision experiments. The best results are **boldfaced**.



The exponential growth of **machine's and robot's intelligence** made it possible for **robots** and **autonomous systems** to work physically alongside humans, interacting and collaborating with them and supporting them in many domains.

PARAGRAPH: The exponential growth of machine's and robot's intelligence made it possible for robots and autonomous systems to work physically alongside humans, interacting and collaborating with them and supporting them in many domains. This dramatic advent of technology opens up many opportunities to support human work and to create new forms of technology-supported collaborative work. It shifts the robots and other intelligent system's roles from being perceived and used as tools into being perceived as teammates (Groom and Nass, 2007) that can augment the abilities of humans and allow for hybrid team formation. This new kind of teamwork has the potential to collaboratively achieve more than any single entity of its members can achieve on its own. It can increase the team performance and reduce the human workload. However, many challenges accompany this development. For example, it might reduce the human sense of autonomy (Blake, 2020;Tan and Taeihagh, 2020) by forcing people to adhere to what the machine needs. Additionally, it is mostly challenging for humans to comprehend the limits of intelligent systems designed by others, which puts them in an uncertain environment, especially when no experience exists with the machine (Wagner and Robinette, 2021). Therefore, in order for this kind of teams to succeed and be beneficial for humans, the collaboration between the team members needs to be carefully designed. In this work, we are looking into how humans and robots can collaborate together autonomously on both sides. **TITLE**: Trust Dynamics and Verbal Assurances in Human Robot Physical Collaboration **SECTION**: INTRODUCTION

Figure 2: Interface for manual annotation of the NER task.

vector, and their combination. The results are given in Table 9. The results indicate that using vector similarity alone or in combination with string similarity generally yields better performance com-

	String			Vector			String+Vector		
Entity	**Precision**	**Recall**	**F_1**	**Precision**	**Recall**	**F_1**	**Precision**	**Recall**	**F_1**
Application	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Human	0.7	5.4	1.3	4.7	3.6	4.1	1.1	8.8	1.9
Context	1.3	22.0	2.4	1.0	56.2	1.8	1.0	64.1	9.5
Technology	5.1	60.5	9.5	0.0	0.0	0.0	5.1	60.5	0.0
Micro	3.2	25.0	5.7	1.0	7.4	1.7	2.1	31.5	3.9
Macro	1.8	22.0	3.3	1.4	15.0	1.5	1.8	33.3	3.3

Table 10: Entity-type-wise results of the rule-based method using string, vector and string+vector similarity for NER task.

pared to using string similarity alone.

Pattern-based We applied the pattern-based method to label the test data. Table 10 shows precision, recall and F_1 scores for different entity types. As the knowledge dictionary contains entities for factors (*human*, *technological*, *context*), the precision, recall, and F_1 scores of *application* are both reported as 0.0. The results indicate that the rule-based method effectively captures certain relevant instances for specific entity types. The string similarity method performs exceptionally well in identifying *technology* factors, achieving a recall of 60.5% and a F_1 score of 9.5. In contrast, the vector similarity method excels with *context* factors, with a recall of 56.2% and an F_1 score of 1.8.

B Details of Dataset

B.1 Annotation Interface

The Prodigy annotation tool is utilised for annotation. We design a web page that integrates the Prodigy annotation tool, allowing annotators to input their names and select the article and specific sections before initiating the annotation process. The interface for NER is illustrated in Figure 2, demonstrating the annotation process. As depicted in the figure, context information, including the paragraph containing the sentence, as well as the title and section names of the article, is presented for each sentence.

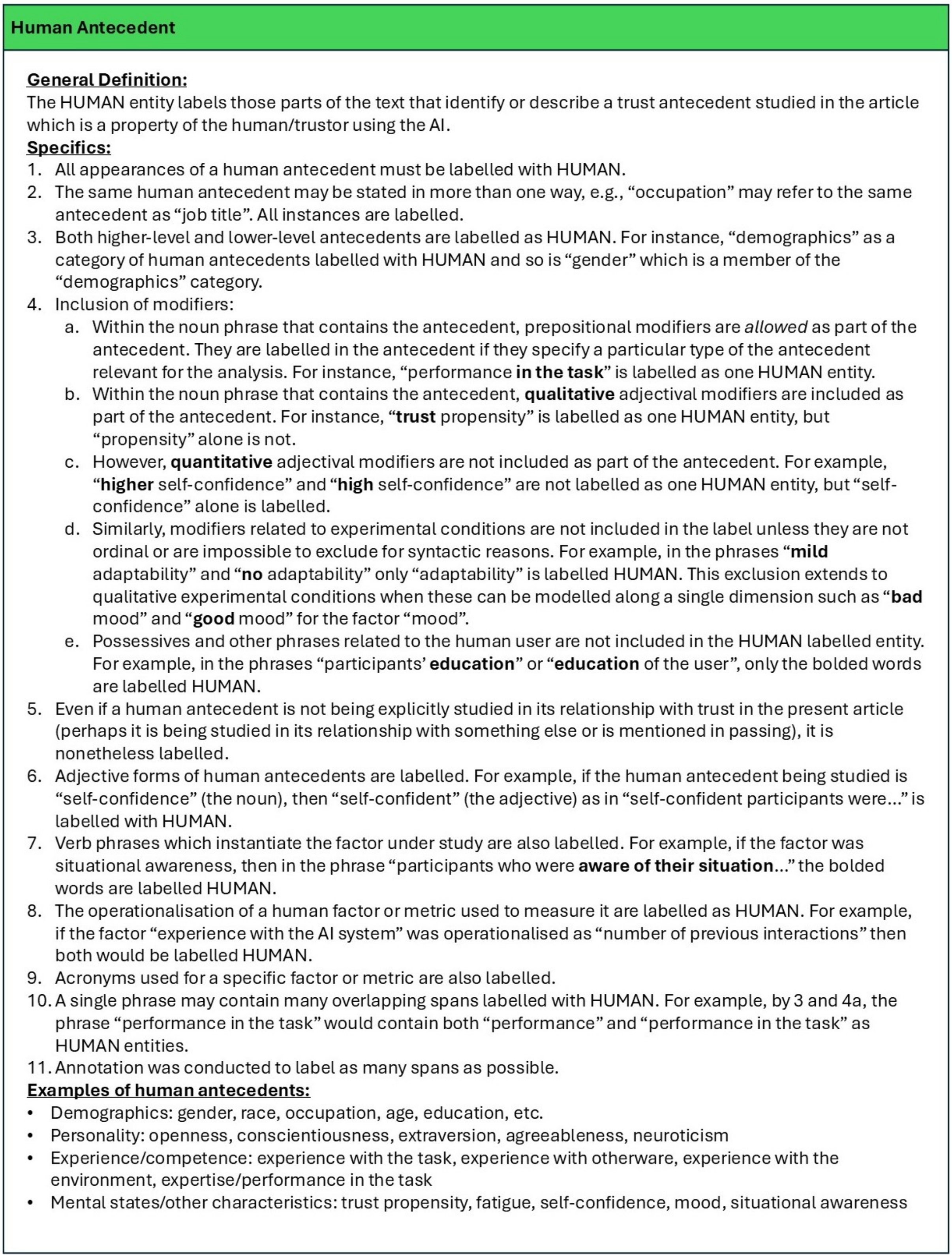
B.2 Annotation Guideline

The model extracts a variety of information from scientific articles that study the relationship between certain factors (or antecedents) and human trust in an artificial intelligence (AI) system. The model extracts the Application of the AI under consideration as well as the antecedents, each of which belongs to one of three categories: Human, Technology, and Context. Each article may contain more than one factor and more than one application. The aim is to create a large, searchable database that contains the key information of the existing research on the antecedents of human trust in AI. The following guideline defines the specific annotation criteria for each piece of information of the model (Figures 3, 4, 5, and 6).

B.3 LLM Guidance Annotation

For LLM guidance, we called the `gpt-3.5-turbo` version of ChatGPT integrated into the spaCy library. We adopted a temperature of 0.3 and used ICL (random sampling- 3-shot) with `spacy.SpanCat.v2` and `spacy.TextCat.v2` components of spaCy and prompt defined by the library to pre-annotate the dataset for NER and RE tasks, respectively (Figure 7).

The LLM-guided annotation interface for NER is illustrated in Figure 8. As depicted in the figure, in addition to manual annotation, LLM guidance is presented for each sentence.



Human Antecedent

General Definition:

The HUMAN entity labels those parts of the text that identify or describe a trust antecedent studied in the article which is a property of the human/trustor using the AI.

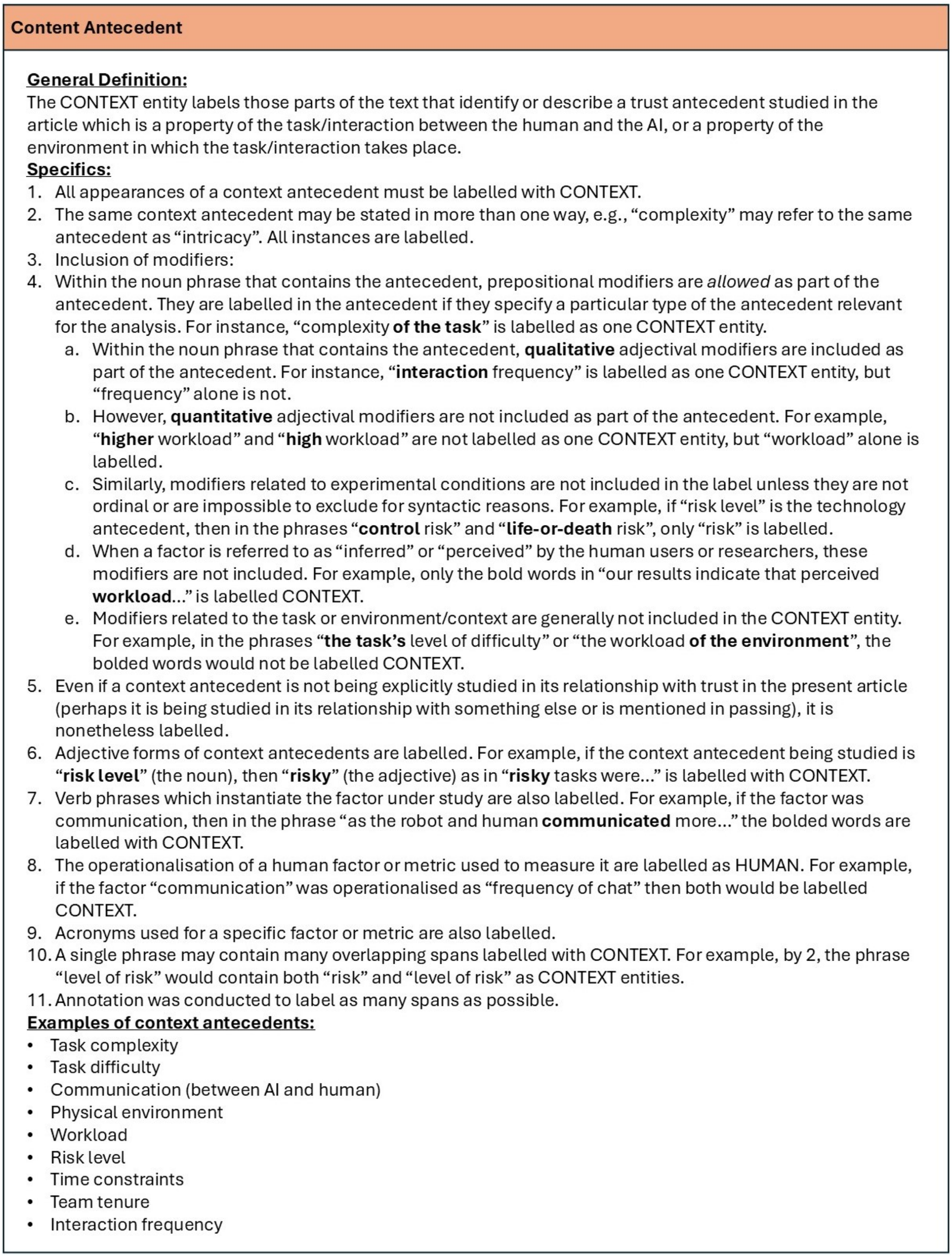
Specifics:

1. All appearances of a human antecedent must be labelled with HUMAN.
2. The same human antecedent may be stated in more than one way, e.g., "occupation" may refer to the same antecedent as "job title". All instances are labelled.
3. Both higher-level and lower-level antecedents are labelled as HUMAN. For instance, "demographics" as a category of human antecedents labelled with HUMAN and so is "gender" which is a member of the "demographics" category.
4. Inclusion of modifiers:
   a. Within the noun phrase that contains the antecedent, prepositional modifiers are *allowed* as part of the antecedent. They are labelled in the antecedent if they specify a particular type of the antecedent relevant for the analysis. For instance, "performance **in the task**" is labelled as one HUMAN entity.
   b. Within the noun phrase that contains the antecedent, **qualitative** adjectival modifiers are included as part of the antecedent. For instance, "**trust** propensity" is labelled as one HUMAN entity, but "propensity" alone is not.
   c. However, **quantitative** adjectival modifiers are not included as part of the antecedent. For example, "**higher** self-confidence" and "**high** self-confidence" are not labelled as one HUMAN entity, but "self-confidence" alone is labelled.
   d. Similarly, modifiers related to experimental conditions are not included in the label unless they are not ordinal or are impossible to exclude for syntactic reasons. For example, in the phrases "**mild** adaptability" and "**no** adaptability" only "adaptability" is labelled HUMAN. This exclusion extends to qualitative experimental conditions when these can be modelled along a single dimension such as "**bad** mood" and "**good** mood" for the factor "mood".
   e. Possessives and other phrases related to the human user are not included in the HUMAN labelled entity. For example, in the phrases "participants' **education**" or "**education** of the user", only the bolded words are labelled HUMAN.
5. Even if a human antecedent is not being explicitly studied in its relationship with trust in the present article (perhaps it is being studied in its relationship with something else or is mentioned in passing), it is nonetheless labelled.
6. Adjective forms of human antecedents are labelled. For example, if the human antecedent being studied is "self-confidence" (the noun), then "self-confident" (the adjective) as in "self-confident participants were…" is labelled with HUMAN.
7. Verb phrases which instantiate the factor under study are also labelled. For example, if the factor was situational awareness, then in the phrase "participants who were **aware of their situation**…" the bolded words are labelled HUMAN.
8. The operationalisation of a human factor or metric used to measure it are labelled as HUMAN. For example, if the factor "experience with the AI system" was operationalised as "number of previous interactions" then both would be labelled HUMAN.
9. Acronyms used for a specific factor or metric are also labelled.
10. A single phrase may contain many overlapping spans labelled with HUMAN. For example, by 3 and 4a, the phrase "performance in the task" would contain both "performance" and "performance in the task" as HUMAN entities.
11. Annotation was conducted to label as many spans as possible.

Examples of human antecedents:

- Demographics: gender, race, occupation, age, education, etc.
- Personality: openness, conscientiousness, extraversion, agreeableness, neuroticism
- Experience/competence: experience with the task, experience with otherware, experience with the environment, expertise/performance in the task
- Mental states/other characteristics: trust propensity, fatigue, self-confidence, mood, situational awareness

Figure 3: Human.



Content Antecedent

General Definition:

The CONTEXT entity labels those parts of the text that identify or describe a trust antecedent studied in the article which is a property of the task/interaction between the human and the AI, or a property of the environment in which the task/interaction takes place.

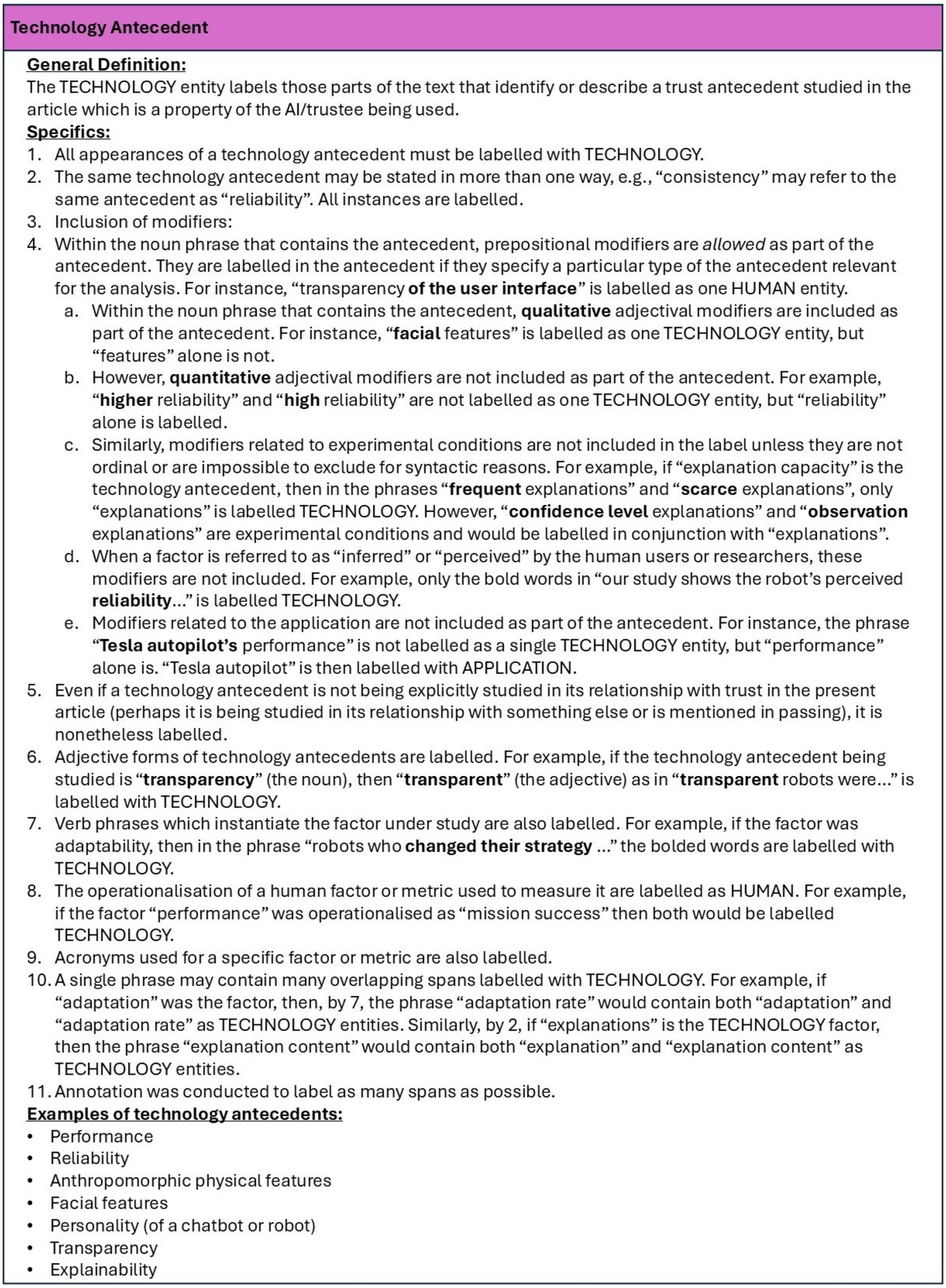
Specifics:

1. All appearances of a context antecedent must be labelled with CONTEXT.
2. The same context antecedent may be stated in more than one way, e.g., "complexity" may refer to the same antecedent as "intricacy". All instances are labelled.
3. Inclusion of modifiers:
4. Within the noun phrase that contains the antecedent, prepositional modifiers are *allowed* as part of the antecedent. They are labelled in the antecedent if they specify a particular type of the antecedent relevant for the analysis. For instance, "complexity **of the task**" is labelled as one CONTEXT entity.
   a. Within the noun phrase that contains the antecedent, **qualitative** adjectival modifiers are included as part of the antecedent. For instance, "**interaction** frequency" is labelled as one CONTEXT entity, but "frequency" alone is not.
   b. However, **quantitative** adjectival modifiers are not included as part of the antecedent. For example, "**higher** workload" and "**high** workload" are not labelled as one CONTEXT entity, but "workload" alone is labelled.
   c. Similarly, modifiers related to experimental conditions are not included in the label unless they are not ordinal or are impossible to exclude for syntactic reasons. For example, if "risk level" is the technology antecedent, then in the phrases "**control** risk" and "**life-or-death** risk", only "risk" is labelled.
   d. When a factor is referred to as "inferred" or "perceived" by the human users or researchers, these modifiers are not included. For example, only the bold words in "our results indicate that perceived **workload**..." is labelled CONTEXT.
   e. Modifiers related to the task or environment/context are generally not included in the CONTEXT entity. For example, in the phrases "**the task's** level of difficulty" or "the workload **of the environment**", the bolded words would not be labelled CONTEXT.
5. Even if a context antecedent is not being explicitly studied in its relationship with trust in the present article (perhaps it is being studied in its relationship with something else or is mentioned in passing), it is nonetheless labelled.
6. Adjective forms of context antecedents are labelled. For example, if the context antecedent being studied is "**risk level**" (the noun), then "**risky**" (the adjective) as in "**risky** tasks were..." is labelled with CONTEXT.
7. Verb phrases which instantiate the factor under study are also labelled. For example, if the factor was communication, then in the phrase "as the robot and human **communicated** more..." the bolded words are labelled with CONTEXT.
8. The operationalisation of a human factor or metric used to measure it are labelled as HUMAN. For example, if the factor "communication" was operationalised as "frequency of chat" then both would be labelled CONTEXT.
9. Acronyms used for a specific factor or metric are also labelled.
10. A single phrase may contain many overlapping spans labelled with CONTEXT. For example, by 2, the phrase "level of risk" would contain both "risk" and "level of risk" as CONTEXT entities.
11. Annotation was conducted to label as many spans as possible.

Examples of context antecedents:

- Task complexity
- Task difficulty
- Communication (between AI and human)
- Physical environment
- Workload
- Risk level
- Time constraints
- Team tenure
- Interaction frequency

Figure 4: Context.



Technology Antecedent

General Definition:

The TECHNOLOGY entity labels those parts of the text that identify or describe a trust antecedent studied in the article which is a property of the AI/trustee being used.

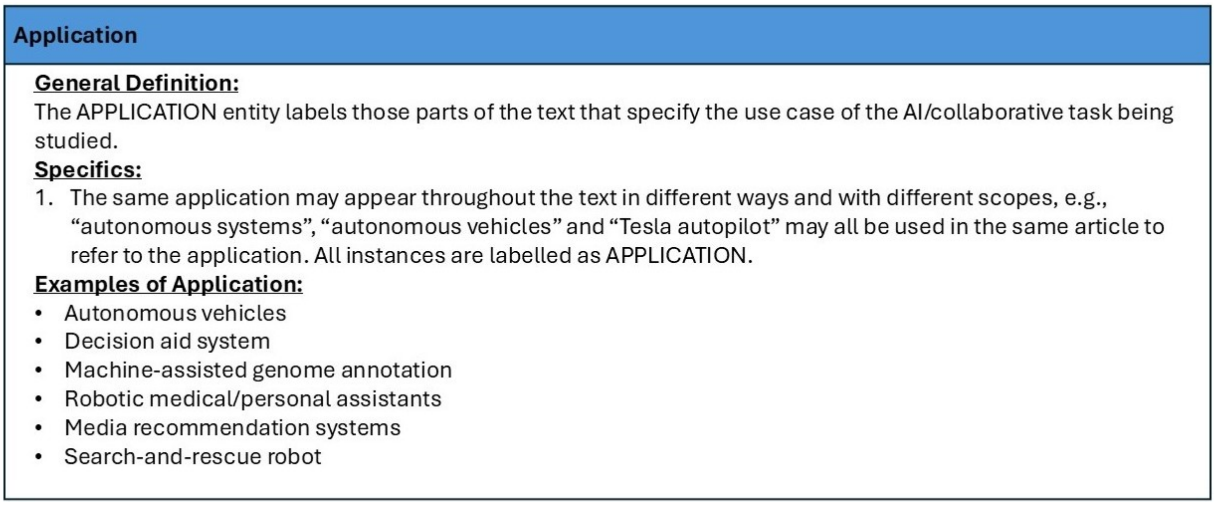
Specifics:

1. All appearances of a technology antecedent must be labelled with TECHNOLOGY.
2. The same technology antecedent may be stated in more than one way, e.g., "consistency" may refer to the same antecedent as "reliability". All instances are labelled.
3. Inclusion of modifiers:
4. Within the noun phrase that contains the antecedent, prepositional modifiers are *allowed* as part of the antecedent. They are labelled in the antecedent if they specify a particular type of the antecedent relevant for the analysis. For instance, "transparency **of the user interface**" is labelled as one HUMAN entity.
    a. Within the noun phrase that contains the antecedent, **qualitative** adjectival modifiers are included as part of the antecedent. For instance, "**facial** features" is labelled as one TECHNOLOGY entity, but "features" alone is not.
    b. However, **quantitative** adjectival modifiers are not included as part of the antecedent. For example, "**higher** reliability" and "**high** reliability" are not labelled as one TECHNOLOGY entity, but "reliability" alone is labelled.
    c. Similarly, modifiers related to experimental conditions are not included in the label unless they are not ordinal or are impossible to exclude for syntactic reasons. For example, if "explanation capacity" is the technology antecedent, then in the phrases "**frequent** explanations" and "**scarce** explanations", only "explanations" is labelled TECHNOLOGY. However, "**confidence level** explanations" and "**observation** explanations" are experimental conditions and would be labelled in conjunction with "explanations".
    d. When a factor is referred to as "inferred" or "perceived" by the human users or researchers, these modifiers are not included. For example, only the bold words in "our study shows the robot's perceived **reliability**..." is labelled TECHNOLOGY.
    e. Modifiers related to the application are not included as part of the antecedent. For instance, the phrase "**Tesla autopilot's** performance" is not labelled as a single TECHNOLOGY entity, but "performance" alone is. "Tesla autopilot" is then labelled with APPLICATION.
5. Even if a technology antecedent is not being explicitly studied in its relationship with trust in the present article (perhaps it is being studied in its relationship with something else or is mentioned in passing), it is nonetheless labelled.
6. Adjective forms of technology antecedents are labelled. For example, if the technology antecedent being studied is "**transparency**" (the noun), then "**transparent**" (the adjective) as in "**transparent** robots were..." is labelled with TECHNOLOGY.
7. Verb phrases which instantiate the factor under study are also labelled. For example, if the factor was adaptability, then in the phrase "robots who **changed their strategy** ..." the bolded words are labelled with TECHNOLOGY.
8. The operationalisation of a human factor or metric used to measure it are labelled as HUMAN. For example, if the factor "performance" was operationalised as "mission success" then both would be labelled TECHNOLOGY.
9. Acronyms used for a specific factor or metric are also labelled.
10. A single phrase may contain many overlapping spans labelled with TECHNOLOGY. For example, if "adaptation" was the factor, then, by 7, the phrase "adaptation rate" would contain both "adaptation" and "adaptation rate" as TECHNOLOGY entities. Similarly, by 2, if "explanations" is the TECHNOLOGY factor, then the phrase "explanation content" would contain both "explanation" and "explanation content" as TECHNOLOGY entities.
11. Annotation was conducted to label as many spans as possible.

Examples of technology antecedents:

- Performance
- Reliability
- Anthropomorphic physical features
- Facial features
- Personality (of a chatbot or robot)
- Transparency
- Explainability

Figure 5: Technology.



Application

General Definition:
The APPLICATION entity labels those parts of the text that specify the use case of the AI/collaborative task being studied.

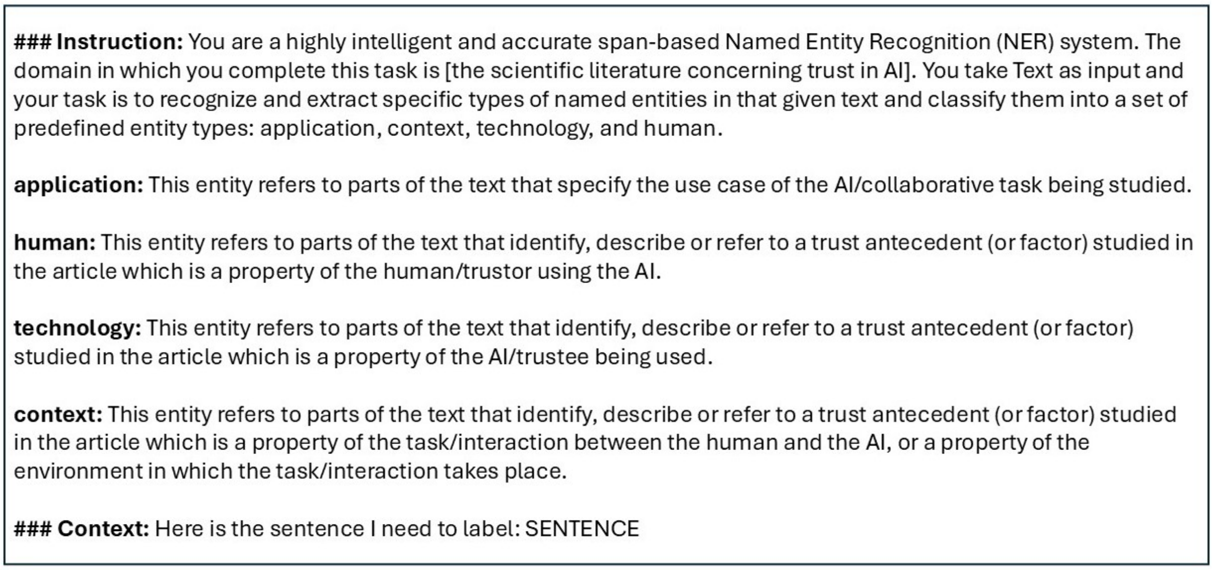
Specifics:

1. The same application may appear throughout the text in different ways and with different scopes, e.g., “autonomous systems”, “autonomous vehicles” and “Tesla autopilot” may all be used in the same article to refer to the application. All instances are labelled as APPLICATION.

Examples of Application:

- Autonomous vehicles
- Decision aid system
- Machine-assisted genome annotation
- Robotic medical/personal assistants
- Media recommendation systems
- Search-and-rescue robot

Figure 6: Application.



### Instruction: You are a highly intelligent and accurate span-based Named Entity Recognition (NER) system. The domain in which you complete this task is [the scientific literature concerning trust in AI]. You take Text as input and your task is to recognize and extract specific types of named entities in that given text and classify them into a set of predefined entity types: application, context, technology, and human.

application: This entity refers to parts of the text that specify the use case of the AI/collaborative task being studied.

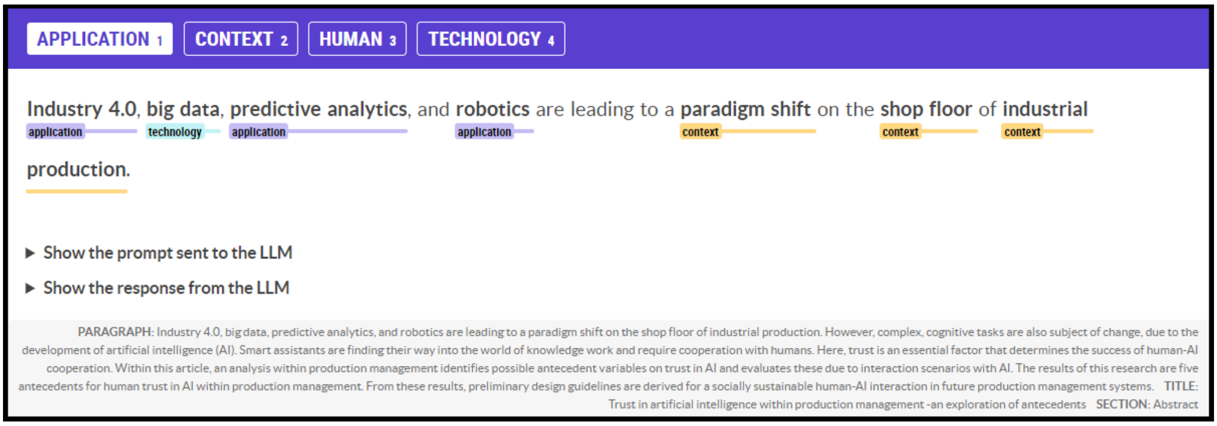
human: This entity refers to parts of the text that identify, describe or refer to a trust antecedent (or factor) studied in the article which is a property of the human/trustor using the AI.

technology: This entity refers to parts of the text that identify, describe or refer to a trust antecedent (or factor) studied in the article which is a property of the AI/trustee being used.

context: This entity refers to parts of the text that identify, describe or refer to a trust antecedent (or factor) studied in the article which is a property of the task/interaction between the human and the AI, or a property of the environment in which the task/interaction takes place.

### Context: Here is the sentence I need to label: SENTENCE

Figure 7: LLM guidance annotation guideline.



Industry 4.0, big data, predictive analytics, and robotics are leading to a paradigm shift on the shop floor of industrial production.

▶ Show the prompt sent to the LLM

▶ Show the response from the LLM

PARAGRAPH: Industry 4.0, big data, predictive analytics, and robotics are leading to a paradigm shift on the shop floor of industrial production. However, complex, cognitive tasks are also subject of change, due to the development of artificial intelligence (AI). Smart assistants are finding their way into the world of knowledge work and require cooperation with humans. Here, trust is an essential factor that determines the success of human-AI cooperation. Within this article, an analysis within production management identifies possible antecedent variables on trust in AI and evaluates these due to interaction scenarios with AI. The results of this research are five antecedents for human trust in AI within production management. From these results, preliminary design guidelines are derived for a socially sustainable human-AI interaction in future production management systems. TITLE: Trust in artificial intelligence within production management -an exploration of antecedents SECTION: Abstract

Figure 8: Interface for LLM-guided annotation of the NER task.