Explanations for explainability: towards an annotated corpus

Anonymous ACL submission

Abstract

Providing good explanations plays a pivotal role in enhancing human understanding. First, we organize explanations into categories based on a framework inspired by scientific and philosophical discussions on the nature of explanations. We then focus on developing retrieval techniques for single-sentence explanations, aiming to lay the groundwork for creating an open-source corpus of scientific articles containing annotations of explanations. A user study was conducted to label 100 sentences according to our classification categories. This collection of annotated examples, balanced with topic-related non-explanatory sentences, was used to refine three large language models (LLMs) via the Cohere API, enabling them to perform (a) semantic search, (b) binary classification and (c) single-label classification. Models (b) and (c) presented results superior to base Llama 3 8B and on par with GPT-4, with model (b) showing balanced results and outperforming GPT-4 by 12% accuracy.

1 Introduction

As generative models become more sophisticated and a standard tool, and as Large Language Models (LLMs) are employed for text generation, a notable use of this technology is its combination with ML models that emphasize explainability. Explanations for machine learning decisions, crucial in sectors like healthcare (Ribeiro et al., 2016; Linardatos et al., 2020; Ghassemi et al., 2021), need to be impactful and human-like (Kulesza et al., 2015; Ali et al., 2023). Addressing the challenge of creating explanations prioritizing proximal over procedural aspects remains a key issue (Tan, 2022). The scarcity of large-scale datasets containing human-generated explanations poses a challenge yet offers a potential solution (Wiegreffe and Marasović, 2021). This research aims to develop a corpus of high-quality scientific explanatory data and explores the performance of easily accessible LLMs for classifying explanatory sentences. Our approach begins with the joint annotation of 100 sentences, from which we derive classification categories based on the data. This document will designate the annotated sentence collection as the Annotated Explanatory Dataset. We provide the Annotated Explanatory Dataset and supplementary materials to the research community via a designated GitHub repository (anonymized for the purpose of the review) (git).

The remainder of this paper is organized as follows: Section 2 puts our work in context, Section 3 explains the construction of the Annotated Explanatory Dataset and presents the techniques for extracting explanations from corpora. The subsequent sections, 4 and 5, explore and analyze the findings. The paper ends with conclusions and an outline of future research directions.

2 Related Work

The quest to understand the essence of explanations spans a wide array of scientific research, with significant contributions from both the social sciences and, more recently, the fields of ML and AI. In social sciences, many definitions and research directions have been drawn from the seminal efforts of philosophers like Aristotle, John Stuart Mill, and Hume, among others. The scholarly contributions of Miller (2019); Mill (2012); Thagard (2012); Lombrozo (2006); Halpern and Pearl (2005); Lewis (1986) have explored the multifaceted nature of explanations. These works examine various critical aspects, such as *causality*, which delves into the cause-and-effect relationships; *contrast*, exploring the distinctions between differing scenarios; *relevance*, focusing on the importance and applicability of explanations; and *truth*, evaluating the accuracy and verifiability of explanations. Meanwhile, ML and NLP focus on operational definitions and the importance of constructing datasets, as seen in studies by Tan (2022); Wiegreffe and Marasović (2021); Hartmann and Sonntag (2022).

Additionally, NLP researchers have crafted highly accurate methods for identifying relationships between concepts. The efforts to derive causal connections from textual content are particularly relevant to our study. A review of these efforts is Yang et al. (2021). The Penn Discourse Treebank (PDTB) version 2.0 is a relevant dataset for causal relations. It was introduced in Prasad et al. (2008) and is the most extensive collection of annotated discourse relations to date, featuring 72,135 non-causal and 9,190 causal examples derived from 2,312 articles from The Wall Street Journal. Focusing instead on the relations dictated by comprehension and common sense, the ECQA and TriviaQA datasets are an example of the most recent direction in the field, a question-answer-evidence approach (Aggarwal et al., 2021; Joshi et al., 2017).

Particularly noteworthy is the work by Overton (Overton, 2012), which bridges the gap between the philosophical discourse on explanations and the practical concerns of analyzing scientific texts; it does so by exploring the diverse philosophical accounts of scientific explanation by analyzing Science journal articles using text mining. Overton identifies the prevalent use of terms like "explain" and "cause," noting that "explain" words are especially common and often used with qualifiers or negations, offering new insights into the practice of scientific explanation beyond traditional analyses. Our work aims to connect the dots between the social science perspective and AI explainability, utilizing the latest LLMs techniques to distil explanatory sentences from scientific articles corpora.

3 Methods

This section outlines the creation of the Annotated Explanatory Dataset, validated through user studies, and discusses the LLMs employed for the automated classification of explanations.

3.1 Sentence Selection

The focus of our search for valid corpora for explanation extraction was narrowed to scientific text for three main reasons. The first reason was the need for concise explanations that would present information in a direct manner; the second was to avoid the potential semantic complications emerging from the social aspect of non-scientific discourse; the third was to limit the subjectivity of the information presented as much as possible.

The PMC Open Access Subset selected as our corpus facilitates easy replication of results and contains plenty of scientific explanations, and offers millions of freely usable journal articles and preprints under licenses like Creative Commons. It's a key part of PubMed Central's effort to enhance access to scientific research for text mining and reuse. This subset enables broader distribution and use than typical copyrighted works, supporting NIH's open access goals through services like cloud and FTP for efficient retrieval and analysis of biotechnology-related literature. Specifically, we employed:

1. The content from `txt` format documents located in the `PMC008` split within the `oa_bulk/oa_comm/txt` directory. This split contains approximately 530,000 documents of varying lengths and formats, from abstracts to full papers, spanning multiple specializations such as chemistry, medicine, and physics, all unified under the primary topic of *biotechnology*.

2. Every document was processed using the NLTK (Bird and Loper, 2004) library, particularly the `nltk.tokenize.sent_tokenize` function, with 'English' as the chosen language and any trailing white spaces removed before processing. The sentences extracted through this process were then stored in a local database and categorized as described in the introduction to the following section.

Drawing upon the research detailed in Overton (2012), which links explanatory sentences to prevalent scientific literature keywords, we initially sorted the data. We assigned each sentence one or more identifiers grounded in specific categories, delineated by their pertinent keywords. These categories and their respective keywords are outlined as follows:

- **because**: associated with the keyword *because*.
- **cause**: linked to keywords such as *cause** and *due to*.
- **confirm**: corresponding to *confirm**.
- **contrast**: encompasses *although, contrast*, despite, however,* and *while*.

- **effects**: pertains to *effect* and *effects*.
- **evidence**: involves *eviden**.
- **explain**: includes *expla** and *unexpla**.
- **indicate**: related to *indicat**, *point*, and *direct*.
- **negation**: identified by *not*.
- **show**: involves *show** and *illustrate**.
- **suggest**: associated with *sugges**.

Each keyword pattern (denoted with an asterisk) represents a wildcard, indicating any extension of the root word.

The differences in keywords and categories between this research and that of Overton (2012) stem from the varied thematic realms explored in the datasets of each study. After organizing the dataset and selecting a representative sample of 1200 sentences that mirror the overall keyword distribution, a preliminary qualitative review was conducted by hand. This review pinpointed around 430 sentences with potential for explanatory significance.

3.2 Annotated Explanatory Dataset

Refining around 430 potential explanations led to a concise set of seed sentences through manual evaluation and categorization, focusing on identifying core characteristics that define each group. This categorization process, driven by the dataset, differentiated explanatory from non-explanatory content, aiming to understand the commonalities and differences within the explanations. This method avoided pre-set criteria, instead exploring the intrinsic connections between categories and the dataset's subject, informed by existing discussions in philosophical and scientific discourse.100 single-sentence explanations deemed appropriate for acting as foundational sentence seeds have been chosen. This is our Annotated Explanatory Dataset from which we derived the explanation categories.

Causation. Explanations in this category identify and describe the relationship between cause and effect, emphasizing that one event or condition leads to another. These explanations connect the cause and outcome without exploring the detailed mechanisms between them. For foundational insights on causation, see Mackie (1974).

Example: "A deficiency of vitamin D in the body causes weakened bones and the onset of osteoporosis."

Mechanistic causation. This category delves into the processes or mechanisms by which a cause leads to an effect, offering a deeper understanding than simple causation. It describes the intermediate steps or biological processes that elucidate how and why the cause effects the outcome, as discussed in Machamer et al. (2000).

Example: "Treatment at an early stage when cancer cells are confined in the organ significantly increases the curative rate."

Contrastive. Contrastive explanations focus on comparing scenarios to explain why a particular outcome occurred in one case but not in another, emphasizing divergent outcomes. This approach is explored in Jacovi et al. (2021).

Example: "The temperature of a large objective lens was higher than that of a small one due to stronger light concentration at higher magnification."

Correlation. These explanations detail relationships between variables where changes in one are associated with changes in another but without establishing causality. It highlights observed patterns or trends indicating simultaneous changes in variables.

Example: "Greater improvements in DXA-based BMD are associated with a greater reduction in fracture risk, especially for spine and hip fractures."

Functional Functional explanations describe the evolution or maintenance of traits due to their utility or role. They focus on the function of a trait in relation to its form and effectiveness, particularly in biology, as discussed in Mayr (1988).

Example: "The owl's wing feathers have evolved for silent flight, aiding in stealthy hunting."

Pragmatic approach. This category emphasizes practicality in choices or actions, focusing on real-world applicability. It explains the selection of methods or models based on convenience or effectiveness, further elaborated in Morgan and Morrison (1999).

Example: "Liquid formulations are preferred in paediatrics for their ease of administration."

3.3 User study and annotator consensus

To reduce the impact of any possible biases from the authors on how sentences were categorized, we conducted a study involving a total of fifteen volunteers who graduated from diverse academic fields (i.e., computer science, linguistics, psychology and robotics) that were not represented in the topic domain of the sentences. This method was chosen to help prevent knowledge bias by forcing the analysis of unfamiliar data purely on a sentence-structure level, without precognitions. The sentences were divided into three equal parts, each containing 33 or 34 sentences. These groups were then utilized in a survey, which included a learning section, referred to as *tutorial*, and a task where participants categorized sentences, referred to as *classification*. The survey was administered using Google Forms, which were divided into two macro-sections. In the tutorial, for each category, the following were provided:

(a) An example sentence,

(b) A written definition,

(c) A graphical representation illustrating the definition.

After the tutorial, participants were tasked with a classification activity structured as a multiple-choice questionnaire. Each of the three questionnaires was delivered to five different annotators, with no annotator being exposed to more than one questionnaire to avoid carry-on knowledge bias; the form was filled in one sitting by each of the users, and no interaction between annotators was allowed to preserve the quality of the results. The average per-sentence consensus between users resulted in a score of 3.57; to further confirm the robustness of the consensus, we computed the Fleiss kappa (Fleiss, 1971) for the set, resulting in a score of 0.303. At first glance, such a score might not seem to indicate quality agreement, but Fleiss' kappa score uses a peculiar agreement scale and it is known to produce lower results with the scaling of categories and annotators (McHugh, 2012). Therefore, considering the kappa score being categorized as "fair agreement" (Landis and Koch, 1977) and the consensus score having a potential range from 1 to 5, the quality test was deemed satisfactory for the seed and the definitions.

While the size of the sentence seed might seem too small for the number of categories available (100 to 6), we believe that the limitations on language imposed by the topic domain and the source of the original data can mitigate the semantic biases that would naturally appear. The annotated sentence seed is available as a `csv` file at the anonymized GitHub repository (git).

3.4 Approaches to explanation classification

Since vector embeddings from large text corpora effectively maintain the semantic connections between sentences (Guha et al., 2003; Bast et al., 2016; Uren et al., 2007), our first approach used semantic search to extract explanations.

The Cohere API (coh) offers developers access to advanced natural language processing capabilities, enabling easy text generation, classification, and analysis integration into applications. It's designed to make cutting-edge language AI technologies accessible for various uses, from automating tasks to enhancing user interactions and extracting insights from data.

The 'embed-english-v3.0‘ model was fed with a seed sentence and approximately 50,000 sentences from the dataset. By tweaking the input configurations, the process was enhanced to rerank and cluster the sentences based on their vector cosine similarity. This methodology allowed us to pinpoint and collect the 20 sentences closely aligned with each seed sentence from its specific cluster.

However, a different approach was adopted after it was found that the initial method did not produce the desired results; less than 30% of the retrieved sentences were actual explanations, with many simply mirroring the seed sentences. The following sections introduce two classification-focused methods tested on a randomly selected subset of around 3,700 sentences from our dataset.

Considering the selected seed sentences did not provide a sufficiently large dataset for full model training, a decision was made to fine-tune a pre-existing large language model (LLM) trained on English text for classification and embedding tasks. Two models were experimented with, starting with `embed-english-v3.0` from Cohere (coh), and the following fine-tuning steps were undertaken:

1. For the binary classification task:
   1.1 Label the chosen explanatory sentences from the biotechnology domain as *positive*.
   1.2 Collect and label a set of 95 non-explanatory sentences from related topics (wik) as *negative*.
   1.3 Create the fine-tuning dataset by combining the positive and negative sets.
   1.4 Adapt the base LLM into a binary classification model.
2. For the multi-class classification task:

2.1 Individually label the sentences from the explanatory seed according to their specific explanation category.

2.2 Select and label a set of 20 non-explanatory sentences from the previously collected ones as *non-explanatory*.

2.3 Produce the fine-tuning dataset by merging these sets.

2.4 Refine the base LLM into a multi-class classification model.

Two types of models were designed and evaluated: a binary classifier and a multi-class classifier. The binary classifier determines whether a sentence is an explanation. The multi-class classifier categorizes sentences into one of the explanatory categories from the Annotated Explanatory Dataset or labels them as non-explanatory.

The datasets for fine-tuning these models are accessible at the anonymized GitHub repository (git), stored in `tsv` format. The model IDs for the Cohere API are provided in the same repository and can be called through the Cohere API.

3.5 Baseline LLMs and comparative evaluation

Given the advancements in OpenAI's GPT architecture, particularly with the introduction of GPT-4, it was logical to employ this architecture for the research. Similarly, the most recent architecture by MetaAI, Llama-3 (lla), was integrated. To ensure lightweight solutions for ease of reproducibility, scalability, and general use, the 8B version of Llama-3 was chosen.

Three templates (*t0, t1, t2*) were developed to aid the models in their classification tasks and determine the optimal amount of information to include in the prompt. The first template exemplifies zero-shot learning, while the next two exemplify few-shot learning. The information was distributed in the following ways:

(a) Executing multi-class classification on any given English sentence, allocating the sentence to predefined categories, *(t0, t1, t2)*.

(b) Integrating a comprehensive list of these categories, each accompanied by definitions, *(t0, t1, t2)*

(c) Accompanying the definitions with three illustrative sentences, *(t1, t2)*

(d) Adding additional illustrative sentences (min 0, max 7) to mimic the proportions in the original seed, *(t2)*

(e) Presenting the analyzed sentence alongside a prompt for the appropriate category label (t0, t1, t2).

These templates applied consistently across our dataset, offering clear examples and directives for the classification task. Examples of these templates are available at our anonymized GitHub repository (git), where they have been uploaded in a `txt` format.

The testing was done using Google Colab notebooks, with the baseline Llama 3 8B run on L4 GPUs and GPT-4 through API; the overall cost for operating Llama 3 and GPT-4 was $\simeq$ 80 euros. However, the Llama model required more than 3 hours compared to GPT-4, which needed just a few minutes.

4 Results

For a thorough comparison, 300 sentences ranging from 50 to 500 characters in length were randomly selected from the test set and manually annotated to serve as a *golden standard* for assessment. This subset did not include *functional* explanations, highlighting their rarity in the larger dataset due to the domain's specific nature. Since the absence of the functional category had a negligible effect on the baseline models and no effect on the fine-tuned ones, it was excluded when evaluating the multiclass performance of the models.

Table 1 offers a side-by-side general performance evaluation of all models tested: the fine-tuned Cohere binary classifier and multi-class classifiers, GPT-4, Llama 3 8B. The *t0/t1/t2* mark represents the template used to prompt the generative model. The fine-tuned models demonstrated slightly superior accuracy when compared to GPT-4, with the performance of base Llama 3 8B being inferior to both models independently of the prompt template used to run the tests.

An important finding was the repetition of high recall scores achieved by GPT-4's and Llama 3's binary classification, largely due to the tendency of both models to broadly label sentences as explanations. This approach correctly identified all positive instances while mistakenly categorizing a large amount of the non-explanatory sentences. The class-by-class comparison for the fine-tuned

model	precision	recall	accuracy	F1-score
finetuned binary	.63 / —	.70 / —	.76 / —	.66 / —
finetuned multi	— / .60	— / .44	— / .70	— / .51
GPT - 4 (t0)	.41 / .32	.99 / .42	.51 / .31	.58 / .36
GPT - 4 (t1)	.46 / .47	.99 / .58	.61 / .49	.63 / .52
GPT - 4 (t2)	.56 / .45	.93 / .49	.73 / .58	.70 / .47
Llama 3 8B (t0)	.34 / .22	.98 / .21	.35 / .11	.50 / .22
Llama 3 8B (t1)	.35 / .14	.87 / .15	.40 / .17	.49 / .15
Llama 3 8B (t2)	.34 / .14	.94 / .21	.35 / .13	.50 / .17

Table 1: Evaluation metrics of the fine-tuned classifiers, base GPT-4 and base Llama 3 8B. The values presented are *binary score / multiclass score.*

multi finetuned			
	precision	**recall**	**F1**
causation	0.40	0.47	0.44
contrastive	0.73	0.57	0.64
correlation	0.38	0.28	0.32
mech. caus.	0.83	0.33	0.48
prag. app.	0.50	0.07	0.13
non-expl	0.78	0.88	0.83

GPT-4 (t2)			
	precision	**recall**	**F1**
causation	0.28	0.57	0.37
contrastive	0.50	0.21	0.30
correlation	0.27	0.46	0.34
mech. caus.	0.36	0.27	0.31
prag. app	0.31	0.79	0.45
non-expl	0.95	0.63	0.76

Table 2: Performance comparison of the two best-performing models by class label.

multi-class model and the GPT-4 with the best performing template is depicted in Table 2.

5 Discussion

With the results provided in the previous section, it is possible to extract useful information regarding the performance of the two fine-tuned LLM classifiers, the baseline models and the possible pitfalls and issues within the procedures. Firstly, the random sampling of the test set (300 sentences out of 3600+) and its subsequent manual annotation as the golden standard has led to the non-representation of the *functional* category of explanations, as it can be seen missing from Table 2. While this might seem counterproductive for the testing process, it is also important to note that the *functional* category is related to the *biology* specific niche of the topic macro-domain. This representation could be a fairly accurate approximation when scaled to real corpora.

Second, as shown in Table 1, even fine-tuning with just 200 sentences enabled a binary classification model to achieve slightly better accuracy than a sophisticated system like GPT-4. This model demonstrated more balanced precision and recall values and avoided the overclassification of *positive* labels, a problem observed with GPT-4 in Tabel 2. Although the 0.76 accuracy may not entail a fully automated classification process, it suggests the feasibility of employing binary classification models for accurately compiling large collections of *explanatory* sentences. This approach could be executed semi-supervised, with future progress leading to unsupervised approaches.

Third, although the multi-class classifier failed to recall the majority of *pragmatic approach* explanations within the test sample, its performance across the remaining categories was strong enough to surpass the best-prompted GPT-4 model in terms of overall accuracy and precision and scores. Despite the results not being revolutionary for LLM or GPT-4 architectures, the potential for improvement with additional high-quality data is evident and significant. This allows combining a fine-tuned binary classifier for preliminary screening with a prompted GPT model for more nuanced classification tasks.

As an aside, the inferior performance of baseline Llama 3 8B was surprising but not entirely so. An interesting finding was the difference in performance depending on the template complexity, achieving slightly better results with a medium-complexity zero-shot template (t1) compared to both the simpler and more complex templates (t0, t2). Perhaps a comparison between the larger 70B

Llama 3 and the other models used in the paper might have been more appropriate considering the parameter size; alternatively, using a fine-tuned version of the 8B model could have led to better results. Nonetheless, the base 8B model was a good enough compromise between size and effectiveness to be used as a baseline, given the previously mentioned constraints.

6 Conclusion and future work

This study was initiated to establish a foundation for creating a corpus of explanatory sentences to pinpoint effective data-gathering and categorization methods. We have introduced a framework for identifying explanatory sentences within biotechnology-related topics and reported findings from experiments with the fine-tuned Cohere LLM, base Llama 3 8B and GPT-4, demonstrating over 0.7 accuracy in binary classification of explanatory content. Considering the Cohere API's performance with a relatively small qualitative dataset against a system like GPT-4, combined with its user-friendly nature and minimal resource demands, this suggests promising avenues for further exploration. This lays the basis for AI-aided user annotations for a wider sentence seed, further refining of the model, and even better corpus-building capabilities to be achieved.

Future research directions involve more extensive comparisons between tunable LLMs to help expand the qualitative sentence seed from this project and investigate potential avenues to develop a classification system capable of handling explanations that span multiple sentences. We believe that by assembling vast collections of human-generated explanations, we can refine the annotated explanatory dataset with improved annotations for more efficient model tuning, which would not require specific pairs of explanations and "added theory" to extract explanatory sentences from textual data. Furthermore, this could enable the conversational outputs of XAI generative models to more accurately reflect human conversation and produce explanatory text; this could pair well with effective counterfactual frameworks in providing understandable AI outputs for both laymen and outsiders of the machine-learning field.

Acknowledgments

A sincere *thank you* to our colleagues for their creative input, endless coffee supplies, and unwavering support. We extend our special gratitude to the volunteering dataset annotators for their dedication and insightful feedback.

Ethical implications and risks

All the work done with annotators has been carried on within the best ethical constraints, with voluntary work being paid for in kindness and treats, correct and compliant use of the licensing provided by the datasets used in the paper and the fair and correct use of the models deployed. While volunteering annotators were not allowed to complete the survey in multiple tranches, the time required was between 30 and 45 minutes and thus did not endanger, harm or strain the annotators mentally or physically.

Since the information contained in the datasets, the sentence seed, and the test set are obtained from academically trusted scientific resources, the risk of spreading misinformation or biased production of results should be minimal and non-threatening for the scientific community. Moreover, since the dataset focuses on explanatory single sentences related to the biotechnology domain, the risk of bias towards marginalised communities is almost nonexistent. We did not personally read the entirety of the PMC corpus, so we cannot say that the risk is zero, but there is a strong assumption of safety.

While future work down the line could provide materials that could be used with malicious intent, such as applying convincing explanatory output to biased or faulty models, we believe that the current risk is not heightened by the publication of this work.

Reproducibility

To provide as much reproducibility of the results presented in this paper as possible, all the test data, the tuning data and the templates to correctly prompt the GPT-4 and Llama 3 8B models have been included in the currently anonymized GitHub repository (git) previously mentioned in the paper. The folder is organized to provide an easily understandable division of all the materials relevant to this paper, and in addition to the aforementioned data, contains the executable Python files derived from the Colab notebooks used to run the GPT-4 and Llama 3 8B models. The exact split of the test set randomly selected to evaluate the models is also freely available, along with the Cohere model IDs to allow for reproducible API calls and the original sentence seed with the annotator consensus score. For the purpose of the review, the data and software used will also be uploaded in the respective sections of the ARR form.

Limitations

Time and computational constraints were not the main limitations of this work since using lightweight, fast-to-deploy architectures was a reasoned choice to avoid gatekeeping materials and procedures from anybody without easy access to powerful cloud computing structures. However, extensive testing and template engineering could not be performed to assess the best possible version of GPT-4 and baseline Llama 3 against the Cohere LLMs; three templates are certainly enough, but perhaps not extensively so, since it is known that slight modification in a prompt for generative LLMs can produce a wide array of unexpected results.

Certainly, the number of annotators can be addressed as a limitation in the scope of the presented work, alongside the narrow domain topic chosen for the dataset. Future work will consider both of these limitations to produce more robust claims and strive for a higher annotator consensus, aiming for wider-reaching studies and clearer definitions. Similarly, the reduced sample test set of 300 sentences out of 3600+ could have skewed the results in favour of one model or another; the development of a bigger golden-standard test set is planned for future refinement of the dataset.

References

Explanations for explainability GitHub repository. https://anonymous.4open.science/r/Explanation_Datasets-2A51/.

Introducing embed v3. https://txt.cohere.com/introducing-embed-v3/.

Introducing meta llama 3: The most capable openly available llm to date. https://ai.meta.com/blog/meta-llama-3/.

Wikipedia: The Free Encyclopedia. wikimedia foundation, inc, 22 july 2004. https://en.wikipedia.org/.

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.

Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. 2023. Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence. *Information fusion*, 99:101805.

Hannah Bast, Björn Buchhold, Elmar Haussmann, et al. 2016. Semantic search on text and knowledge bases. *Foundations and Trends® in Information Retrieval*, 10(2-3):119–271.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

JL Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378—382.

Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L Beam. 2021. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 3(11):e745–e750.

Ramanathan Guha, Rob McCool, and Eric Miller. 2003. Semantic search. In *Proceedings of the 12th international conference on World Wide Web*, pages 700–709.

Joseph Y Halpern and Judea Pearl. 2005. Causes and explanations: A structural-model approach. part ii: Explanations. *The British journal for the philosophy of science*.

Mareike Hartmann and Daniel Sonntag. 2022. A survey on improving NLP models with human explanations. In *Proceedings of the First Workshop on Learning with Natural Language Supervision*, pages 40–47, Dublin, Ireland. Association for Computational Linguistics.

Alon Jacovi, Swabha Swayamdipta, Shauli Ravfogel, Yanai Elazar, Yejin Choi, and Yoav Goldberg. 2021. Contrastive explanations for model interpretability. *CoRR*, abs/2103.01378.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *Preprint*, arXiv:1705.03551.

Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces*, pages 126–137.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

David Lewis. 1986. Causal explanation.

Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. 2020. Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1):18.

Tania Lombrozo. 2006. The structure and function of explanations. *Trends in cognitive sciences*, 10(10):464–470.

Peter Machamer, Lindley Darden, and Carl F. Craver. 2000. Thinking about mechanisms. *Philosophy of Science*, 67(1):1–25.

John Leslie Mackie. 1974. *The Cement of the Universe*. Clarendon Press, Oxford,.

Ernst Mayr. 1988. *Toward a New Philosophy of Biology: Observations of an Evolutionist*. Harvard University Press, Cambridge, MA.

Mary McHugh. 2012. Interrater reliability: The kappa statistic. *Biochemia medica : časopis Hrvatskoga društva medicinskih biokemičara / HDMB*, 22:276–82.

John Stuart Mill. 2012. A system of logic. In *Arguing About Science*, pages 243–267. Routledge.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38.

Mary S. Morgan and Margaret Morrison, editors. 1999. *Models as Mediators: Perspectives on Natural and Social Science*. Cambridge University Press.

James A. Overton. 2012. Explanation in science.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Chenhao Tan. 2022. On the diversity and limits of human explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2173–2188, Seattle, United States. Association for Computational Linguistics.

Paul Thagard. 2012. *The cognitive science of science: Explanation, discovery, and conceptual change*. Mit Press.

Victoria Uren, Yuangui Lei, Vanessa Lopez, Haiming Liu, Enrico Motta, and Marina Giordanino. 2007. The usability of semantic search tools: a review. *The Knowledge Engineering Review*, 22(4):361–377.

Sarah Wiegreffe and Ana Marasović. 2021. Teach me to explain: A review of datasets for explainable natural language processing. *Preprint*, arXiv:2102.12060.

Jie Yang, Soyeon Caren Han, and Josiah Poon. 2021. A survey on extraction of causal relations from natural language text. *CoRR*, abs/2101.06426.