Understanding Fine-grained Distortions in Reports of Scientific Findings

Amelie Wührl[1,3], **Dustin Wright**[2], **Roman Klinger**[3] and **Isabelle Augenstein**[2]
[1]University of Stuttgart, Germany,
[2]University of Copenhagen, Denmark,
[3]University of Bamberg, Germany
amelie.wuehrl@ims.uni-stuttgart.de
{dw, augenstein}@di.ku.dk
roman.klinger@uni-bamberg.de

Abstract

Distorted science communication harms individuals and society as it can lead to unhealthy behavior changes and decrease trust in scientific institutions. Given the rapidly increasing volume of science communication in recent years, a fine-grained understanding of *how* findings from scientific publications are reported to the general public, and methods to detect distortions from the original work automatically, are crucial. Prior work focused on individual aspects of distortions or worked with unpaired data. In this work, we make three foundational contributions towards addressing this problem: (1) annotating 1,600 instances of scientific findings from academic papers paired with corresponding findings as reported in news articles and tweets with respect to four characteristics: causality, certainty, generality and sensationalism; (2) establishing baselines for automatically detecting these characteristics; and (3) analyzing the prevalence of changes in these characteristics in both human-annotated and large-scale unlabeled data. Our results show that scientific findings frequently undergo subtle distortions when reported. Tweets distort findings more often than science news reports. Detecting fine-grained distortions automatically poses a challenging task. In our experiments, fine-tuned task-specific models consistently outperform few-shot Large Language Model prompting. The dataset is available at https://www.uni-bamberg.de/en/nlproc/resources/sciencecommdistortion/.

1 Introduction

Lay people, i.e., non-experts with limited experience or knowledge of a specific domain, rely on effective science communication to learn about scientific research. In order to make scientific information understandable to a lay audience, science communicators must first simplify the highly technical language of science (Salita, 2015). In doing so, authors may knowingly or unknowingly distort the information conveyed by the original scientific publication to achieve specific rhetorical goals (Ransohoff and Ransohoff, 2001; Dempster et al., 2022; Tichenor et al., 1970; Sumner et al., 2014; Bratton et al., 2019). For example, simplifying findings for a non-expert audience requires balancing accuracy, accessibility and comprehensibility (Kuehne and Olden, 2015) which can lead to information being omitted purposefully. At the same time, the way that science is communicated to the public is crucial as it influences people's behavior and trust in science (Kuru et al., 2021; Fischhoff, 2012; Hart and Feldman, 2016).

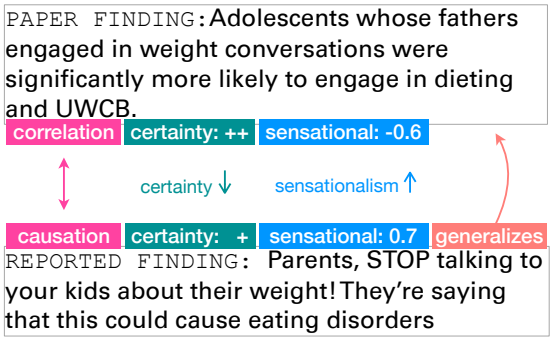


Figure 1: Pair of scientific finding and reported finding with fine-grained labels of distortions.

Consider Fig. 1. The paper finding describes a correlation between "weight conversations" and "dieting and UWCB"[1] while clearly stating to whom the finding applies. In the reported finding those constraints are omitted, it generalizes from "fathers" to all "parents" and "adolescents" to "kids." Further, the reported finding states a causal relationship between "talking to kids about their weight" and "eating disorders." Finally, the reported finding expresses high certainty in the correlational relationship found, while the reported

[1]UWBC stands for unhealthy weight control behaviors.

finding speculates that a causal relationship "could" exist, thus communicating lower certainty. While subtle, the reported finding presents a different picture than that of the original paper, highlighting the need for a fine-grained comparison between paper and reported findings.

Previous work has been limited to either a subset of the distortions studied in this work (Yu et al., 2019; Pei and Jurgens, 2021a; Wright and Augenstein, 2021b) or detecting general information change without fine-grained labels (Wright et al., 2022). We improve on this by making three core contributions: We collect an expert-annotated dataset of 1,600 scientific findings paired across scientific papers and their reported findings in news media and Twitter[2] (**C1**). The data is annotated with fine-grained distortion labels, i.e., causal claim strength, level of certainty, level of generality and sensationalism. Using this labeled data, we train and analyze the performance of benchmark models on the task of fine-grained distortion detection **(C2)**. Finally, using the data from **C1** and models from **C2**, we perform a large scale analysis of the prevalence and types of distortions present in both our expert labeled data and a large-scale automatically labeled dataset of 1,655,570 paper findings, 422,626 news findings and 356,275 tweets **(C3)**. We answer the following research questions:

RQ1 How are scientific findings changed when reported to lay audiences?

RQ2 How reliably can we detect distortions automatically?

For RQ1, we find that scientific findings undergo fine-grained changes when they are reported even when their overall content is well-aligned. This is consistent across scientific disciplines. We find that 54 % of findings are reported with a changed causal relation and 60 % of findings are reported with a changed level of certainty. In 49 % of the paired findings, the reported finding is more general than the paper finding, and reported findings are typically more sensational compared to paper findings. Across all change dimensions, findings reported in tweets are more susceptible to mis-reporting compared to reports in science news. With respect to RQ2, we find that detecting fine-grained distortions automatically poses a challenging task. In our experiments, fine-tuning task-specific models consistently outperform few-shot LLM prompting. Our best models achieve macro F_1 scores of 0.58, 0.56, 0.57 and Pearson correlation of 0.61 for predicting causality, certainty, generalizing and sensationalism, respectively.

[2]Twitter is now called X.

2 Related Work

Science Communication. Science communication is a relatively nascent area of exploration in NLP. The main problems which have been worked on relate to information change in news articles and social media about scientific papers (Wright et al., 2022; Wright and Augenstein, 2021b; Pei and Jurgens, 2021a; Yu et al., 2020), understanding discourse strategies in scientific press releases (August et al., 2020), information loss in medical summaries (Trienes et al., 2024), tasks related to scientific peer review (Kuznetsov et al., 2022), tasks related to scholarly document understanding (Wright and Augenstein, 2021a; Beltagy et al., 2019) and scientific fact checking (Wadden et al., 2020; Mohr et al., 2022). This work is most closely related to those studying information change in science communication, particularly the works of Wright et al. (2022) on general information change and Wright and Augenstein (2021b) and Pei and Jurgens (2021a) on exaggeration and certainty, respectively. These works are limited in a few key aspects, which we address. First, they are concerned with either single, narrow aspects of information change or overarching broad notions of change, missing important types of distortions such as generalizing and sensationalizing results. Additionally, the existing labeled data for exaggeration is limited in size, and the labeled data for certainty are unpaired. We improve on this by augmenting the matched findings in the dataset from Wright et al. (2022) with four specific distortions that are prevalent in science communication: exaggerating causal claim strength, changing the level of certainty, generalizing results, and sensationalizing results.

Misinformation. Inaccurate reports of scientific findings is a form of mis-information. Misinformation detection and fact-checking are established tasks in NLP, both for the general domain and for scientific claims (Guo et al., 2022; Vladika and Matthes, 2023). Scientific fact-checking verifies scientific claims against evidence sources. It is related to our task as it is compares the truthfulness of a statement against a reference document. Technically, a reported finding constitutes a claim about the original which connects our task to claim detection and argument mining (Lawrence and Reed,

2019; Boland et al., 2022). However, compared to both these related tasks, this work requires a more nuanced view of detailed characteristics of the overarching claim.

3 Dimensions of Information Changes

We consider four dimensions found to be notable in the science of science literature (Sumner et al., 2014; Bratton et al., 2019; Fischhoff, 2012; Ransohoff and Ransohoff, 2001) that characterize scientific findings and may undergo change when reported: *Causality*, i.e., the type of causal relation (or its absence) described in finding; *Certainty*, i.e., the level of confidence or certainty that is expressed wrt. a finding; *Generality*, i.e., the level of generalization or specificity of a finding compared to its reporting; *Sensationalism*, i.e., the extent to which a finding is presented in a way to elicit an emotional reaction by using urgent or exaggerated language and descriptions. The divergence between those dimensions in the paper and the reported findings allows us to estimate how accurate a reporting is and which properties may be distorted. More specifically, we consider a reported finding to be mis-reported if the label for a given characteristic changes from the paper finding to the corresponding reported finding. We build a dataset of 1,600 findings from four scientific disciplines (medicine, psychology, biology and computer science). Findings are paired between scientific paper and news and scientific paper and tweet, giving 800 pairs total. In an annotation conducted with crowdworkers, we label each instance/pair with regards to the change type. Annotators rate certainty and sensationalism levels, identify causality relations and check for generalizations in both versions of the finding. Fig. 1 shows an example.

3.1 Change Dimensions

(1) **Causality** describes a cause–effect relationship between two things, variables, agents etc. Correlation describes relationships where two actions relate to each other, but one is not necessarily the effect or outcome of the other. (2) **Certainty** in science communication can be expressed with respect to various aspects (Pei and Jurgens, 2021b). (Un)certainty exists towards specific numbers/quantities ('approximately 50 %'), the extent to which a finding applies ('mainly observed for') or the probability that something applies, occurs or is associated ('possibly associated with'). (3) **Generalizations** claim something is always true, even if it is only valid in certain instances or occasionally. For example, a reporting that generalizes a finding about diabetes type 2 in seniors to all people with diabetes or a generalization from a specified set of medical conditions ('reduced risk of stroke and diabetes') to a general statement ('has health benefits'). (4) Sensational text intends to spark the interest of a reader, make them curious or elicit an emotional reaction. Cues for **sensationalism** can be urgent, exaggerated language or conveyed through the use of informal or colloquial language.

3.2 Dataset Construction

3.2.1 Source Data

To investigate *how* science communication changes scientific findings, we require reports of findings matched with their original counterpart. Therefore, we build on SPICED (Wright et al., 2022) which provides text pairs of scientific findings and associated reports of the finding from news articles or Twitter. Each pair is scored with an *information matching score (IMS)* which indicates how similar the content of the two texts are. It ranges from *5 – completely the same* to *1 – completely different*. We sample instances with a high information matching score (IMS > 4) and filter out instances with a high IMS that the SPICED dataset marks as *easy* cases[3].

The filtering provides us with a total of 837 paired findings across four scientific disciplines: biology (185), computer science (168), medicine (227) and psychology (257). The reported findings stem from science news (515) and Twitter (322).

3.2.2 Annotation Tasks

Considering the substantial differences in change type concepts, we design the data collection as four separate annotation tasks to enable annotators to focus on one concept at a time.[4] This also allows us to operationalize each task independently, which is important to find the optimal annotation method for each task without burdening the annotators with strenuous context switches. We describe the tasks and settings in the following.

[3] SPICED marks instances as *easy* if the reported finding is almost identical to the paper finding.

[4] In a set of pilot studies, we experiment with tasking annotators to label all change types for an instance, instead of focusing on one concept per study. However, inter-annotator agreement in this setup was very low, presumably because of the difficulty of the tasks themselves and the substantial cognitive complexity it takes to understand and switch between multiple concepts during annotation.

Causality. Given a finding, annotators are tasked to identify which type of causal relationship is described. In a classification setting, annotators decide between *No relation stated*, meaning no causal or correlational relation is stated, *Correlation*, *Causation* and *Explicitly states: no relation*, meaning the finding states the absence of a relation.

Certainty. Given a finding, annotators rate the level of certainty with which a finding is being described. This task uses a 4-point rating scale ranging from *Uncertain* to *Certain* with the nuances *somewhat uncertain* and *somewhat certain* in between.

Generalization. Given a paired finding, annotators identify which finding is more general, i.e., the *reported finding*, or the *paper finding*. If they are equally specific/general, annotators can label them as expressing the *same level of generality*.

Sensationalism. Annotators are presented with sets of four findings at a time. In a best-worst-scaling setup (Kiritchenko and Mohammad, 2017), they identify which of the four findings is the *most* and *least sensational*.

3.2.3 Annotation Environment

We use POTATO (Pei et al., 2022) as our annotation environment and recruit crowdworkers using PROLIFIC.[5] To ensure subject expertise, participants must have at least an undergraduate degree in the respective scientific field or a closely related subject (refer to Appendix A.1 for details.) For the change types *causality*, *certainty*, and *generalization*, every annotator works on 12 instances. For *sensationalism*, participants work on 10 instances, i.e., quad-tuples.[6] We provide a detailed description of the annotation setup in Appendix A.1 and screenshots in the supplementary material[7].

3.2.4 Label Aggregation

For *causality*, *certainty* and *generalization*, we aggregate the final labels using MACE (Hovy et al., 2013), a Bayesian model which learns a distribution over labels that takes into account annotator competence. To obtain real-valued scores from best-worst annotation, we calculate the percentage of times an instance was chosen as most sensational minus the percentage of times the term was chosen as least sensational (Kiritchenko and Mohammad, 2017). The score ranges between 1 and 1.

3.3 Analysis

3.3.1 Evaluation Metrics

We evaluate the results of the annotation studies using the following metrics: Average pairwise inter-annotator F_1 (iaF_1), i.e., treating one annotator's labels as gold annotations and consider the other annotator's labels as predictions (Hripcsak and Rothschild, 2005); average pairwise Cohen's ; average[8] pairwise Spearman's correlation ; split-half reliably to evaluate best-worst scaling tasks (Kiritchenko and Mohammad, 2017) for which all annotations for an instance are split into half. For each set, the best-worst scaling score is calculated independently. We report the correlation (Spearman and Pearson) between the sets of scores[9].

3.3.2 Agreement

We report agreement metrics for all tasks in Table 1. For *causality*, we observe an inter-annotator F_1 of 0.38. The average pairwise agreement is 0.21 indicating fair agreement (McHugh, 2012). For *certainty*, the average correlation ( ) between the certainty ratings is 0.44. In the *generalization* task, we observe an inter-annotator F_1 of 0.42 and a of 0.20. We report split-half reliability for the *sensationalism* task and observe an average of 0.44 indicating a positive correlation between the sets of scores.

While we acknowledge that the agreement scores are in parts relatively low, it is crucial to point out that this does not necessarily indicate low annotation quality (Plank, 2022; Reidsma and Carletta, 2008; Sandri et al., 2023). We presume that the scores reflect the difficulty of the tasks as judging scientific findings is not trivial and sensationalism and certainty are to a certain extend subjective. Note that is in line with agreement scores reported for similar tasks such as classifying writing strategies for science communication (August et al., 2020), identifying certainty vs. doubt (Rubin, 2007) and in general, for partially subjective and non-propositional tasks which are known to exhibit stronger label variation as a result of annotators'

[5] https://www.prolific.com/

[6] Quad-tuples contain a mixture of paper findings and reported findings. To avoid biasing the annotators, we do not make it transparent to annotators which source the individual finding originated from.

[7] The supplementary material and dataset is available at https://www.uni-bamberg.de/en/nlproc/resources/sciencecommdistortion/

[8] We average correlations by transforming each correlation coefficient using Fisher's Z, calculating the average of the transformed values, and back-transforming the value.

[9] We use the best-worst-scaling scripts available at https://saifmohammad.com/WebPages/BestWorst.html

	causality		certainty			general.		sensation.	
disc.	iaF$_1$		iaF$_1$			iaF$_1$			*r*
bio	.40	.22	.37	.22	.48	.43	.24	.48	.48
cs	.36	.20	.31	.15	.42	.41	.17	.47	.48
med	.38	.23	.36	.19	.48	.44	.23	.41	.44
psy	.36	.20	.37	.22	.38	.40	.16	.38	.39
avg.	.38	.21	.35	.20	.44	.42	.20	.44	.45

Table 1: Inter-annotator F_1 (iaF$_1$), average pairwise Cohen's , Spearman's correlation ( ) across tasks and disciplines. For sensationalism, and *r* are the correlations from the split-half reliability evaluation.

individual backgrounds (Biester et al., 2022; Sap et al., 2022). Therefore, we consider these findings from the annotation study, i.e., the dataset along with the agreement scores, a research outcome. We provide the individual annotations along with the aggregated labels to enable further research regarding annotator differences.

3.3.3 Results: How are scientific findings changed when they are reported to lay audiences? (RQ1)

We want to understand if and how scientific findings are distorted when they are reported to lay audiences. To this end, we compare for each paired finding how the label for a particular dimension (causality, certainty, generalization, sensationalism) changes going from the paper finding to the reported finding. Fig. 2 visualizes the results. Figures 2 (a) and 2 (b) present distortions in Sankey diagrams, with the left side of each chart depicting the label for the paper finding and the right side depicts the label for the reported finding. Each label is represented by a set of strands going from left to right. Figure 2 (c) plots the distribution of labels for generalization in a bar plot. For each label we plot the number of instances separated by communication outlet. Fig. 2 (d) visualizes the sensationalism scores for the paper finding and the scores for the reported findings (tweets and science news) in a density plot.

Fig. 2 (a) shows changes in causality. Within each relation type (causation, correlation etc.) for the paper findings, we see that the strongest Sankey strand typically leads to its same-label-counterpart on the right side (e.g., *Correlation* to *Correlation*). However, while these strands tend to be the strongest, the sum of the other strands originating from each group, is equally substantial. In fact, overall only 45.5 % of paired findings convey the same relation in the paper and reported finding.

Fig. 2 (b) shows changes in certainty. Overall we observe that both paper and reported findings typically describe the finding with relative certainty. The most frequent distortion is turning a *somewhat certain* paper finding to a *certain* reported finding. In general, we observe that transitions to neighboring labels on the certainty scale are typically more frequent than changes to certainty levels further away on the scale. Collapsing the *somewhat (un)certain* findings, we find that 15 % of paper findings labeled as *certain* are reported in an *uncertain* manner, while 13 % of paper findings labeled as *uncertain* are reported in a *certain* manner.

We visualize changes in generality in Fig. 2 (c). Overall we observe that in the majority of cases reported findings are more general compared to the paper findings. This means findings typically start out being specific and become more general in the reporting. The opposite occurs less frequently.

The density plots in Fig. 2 (d) show the distribution of sensationalism scores across paper and reported findings. Reported findings are separated by communication outlet (science news, tweets). The score on the x-axis ranges from -1 (least sensational) to 1 (most sensational). We see that the majority of findings have a sensationalism score around 0. Notably, the distribution of the reported finding scores is offset more toward +1, indicating that the reportings are typically more sensational compared to the paper findings.

Changes across communication outlets. We want to understand if the communication outlet (science news, Twitter) impacts the types of distortions we observe. For this, we identify four distortions which are potentially most harmful for the news consumer. Critical distortions are:

caus″ Increase in causal claim strength: *Correlation* or *Explicitly States: no relation* in paper finding to *Causation* in reported finding.

gen″ Increase in generality: the reported finding being *more general* than the paper finding.

cert″ Increase in certainty.

sens″ Increase in sensationalism scores from paper to reported finding > 1sd.

Table 2 shows the percentage of finding pairs affected by critical distortions across the two communication outlets, i.e., Twitter and science news. For the categories *Increase in causal claim strength* (caus″) and *Increase in certainty* cert″, there are no major differences between science news and tweets. For *Increase in generality* (gen″) however, reports

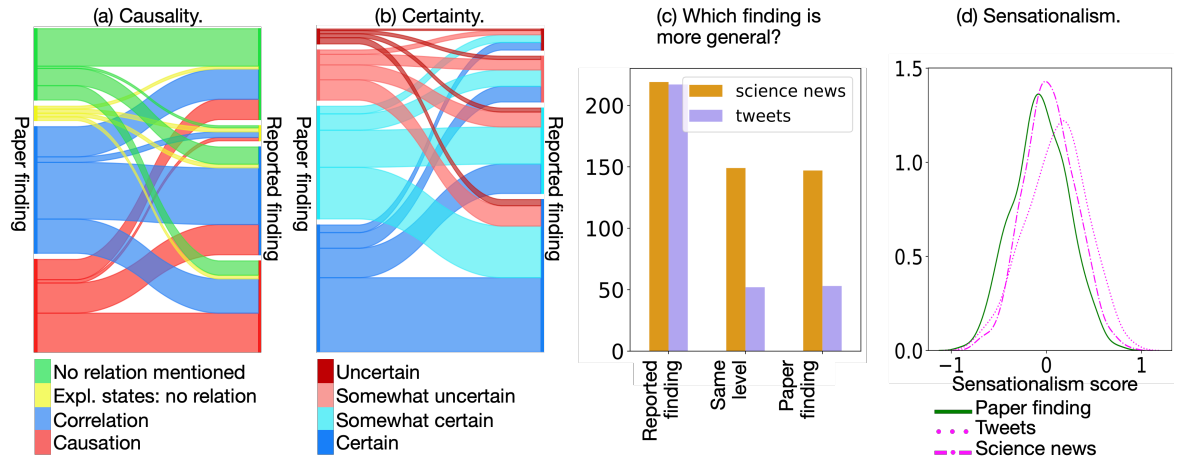


Figure 2: Sankey diagrams visualize changes in causality and certainty going from the paper finding to the reported finding. The bar plot shows the distribution of generalization labels. The density plot visualizes the difference in sensationalism scores across reporting source.

type	news	tweets	bio	cs	med	psy
caus↑	14.2	12.4	11.9	18.5	14.1	10.9
cert↑	33.0	32.9	31.9	43.5	29.1	30.4
gen↑	42.5	67.4	53.5	42.9	50.2	58.8
gen↓	28.5	16.5	23.2	30.4	25.6	18.7
sens↑	39.6	50.6	43.8	39.9	40.1	49.8

Table 2: Percentage of finding pairs affected by critical distortions across communication outlets (science news, tweets) and scientific disciplines (biology, computer science, medicine, psychology).

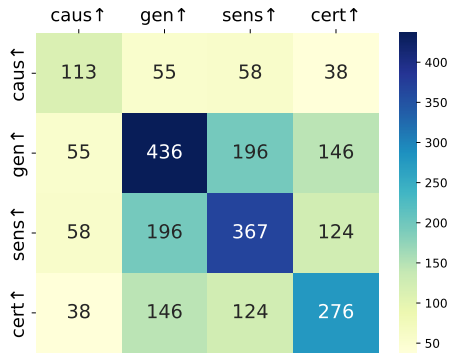


Figure 3: Co-occurrence matrix of critical distortions. Diagonals represent the number of paired findings affected by a particular distortion. All other counts represent distortion co-occurrences.

in tweets are distorted substantially more frequently than reported findings in science news. Similarly, findings reported in tweets more frequently sensationalize paper findings compared to reported findings from sciences news.

Changes across scientific disciplines. We investigate if the changes we observe are different wrt. the scientific discipline (biology, computer science, medicine, psychology) that the finding originates from. We report the percentage of finding pairs affected by critical distortions across the disciplines in Table 2. Overall, we observe a similar level of distortions across the disciplines biology, medicine and psychology. Findings from computer science show slightly different distortions: while they show increases in causal claim strength and increases in certainty more frequently compared to the other disciplines, the findings are generalized less often and they are less affected by increases in sensationalism compared to other disciplines.

Co-occurrence of changes. We analyze the co-occurrence of critical distortion labels to understand potential connections between them. For every paired finding, each critical change is a binary variable that is True when the pair is affected by the change, and False if not. We plot the co-occurrence of these variables in Fig. 3. The distortions which co-occur most frequently are generalizations and increased sensationalism (196 instances). This is intuitive as findings may be sensationalist, because they convey broad or generalized claims. Similarly, increased certainty frequently co-occurs with generalization (146 instances) as well as increases sensationalism (124). Findings that convey strong claims with heightened certainty may be perceived as sensational and vice versa.

4 Experiments

In RQ2, we investigate how reliably we can detect information changes automatically. For all model-

ing experiments, we collapse the causality labels *Expl. states: no relation* and *No relation mentioned* into the *Unclear relation* instances, and the certainty labels *Somewhat uncertain* into the *Uncertain* instances. We experiment with two modeling approaches which we describe in the following.

4.1 Setup

4.1.1 Task-specific Models

To establish baselines for automatically predicting fine-grained distortions of scientific findings, we fine-tune task-specific models for each distortion type. We model *causality*, *certainty* and *generalization* as classification tasks and predicting sensationalism scores as a regression. All models obtain as input a finding and learn to predict the distortion label. For *generalization* the input is a paired finding. For each task, we train a classifier/regressor on top of a transformer base model.

Experimental setting. We experiment with two base models (RoBERTa-base[10] and SciBERT[11]) to understand if domain-specific pretraining data is beneficial for our tasks. For details on model training refer to Appendix A.2.1. We train all models using a 80/20 train-test split of the dataset.

Evaluation. We evaluate the model performance using the task-specific evaluation metrics, i.e., macro F_1 and Pearson's r (see Sec. 3.3.1).

4.1.2 Few-shot Prompting

With the recent paradigm shift to in-context learning using instruction-tuned large language models (LLMs), we investigate the extent to which an LLM predicts, i.e., generates the correct change type label when prompted with the same instructions as human annotators.

Evaluation. To calculate performance metrics, we need to extract the label from the LLM's output. To this end, we assume the first mention of any label from the task-specific label space that is closest to the answer cue in the question to be the prediction. We evaluate[12] the model performance using the task-specific evaluation metrics (see Sec. 3.3.1).

Experimental setting. We experiment with three open, instruction-tuned LLMs for the few-shot prompting, varying in model size and architecture: LLaMa-2 13B[13] (Touvron et al., 2023), Mistral 7B[14] (Jiang et al., 2023) and Mixtral 8x7B[15] (Jiang et al., 2024). As prompts, we use the same task description and examples that we used to instruct the human annotators. We provide our prompt template in Fig. 5.

4.2 Results: How reliably can we detect distortions automatically? (RQ2)

Table 3 shows the results. Across all tasks, we observe that fine-tuned models exceed the performance of the few-shot prompting setup. For *causality*, the fine-tuned SciBERT achieves a macro F_1 of $0.54_{\pm.04}$ (avg. across 5 seeds, subscript denotes standard dev.), while the best results from prompting (LLaMa) achieves an F_1 of 0.46. Similarly, for *certainty*, SciBERT achieves an avg. F_1 of $0.53_{\pm.02}$, while all LLMs struggle to make meaningful predictions. For *generalization*, the avg. performance of the fine-tuned SciBERT is at $0.47_{\pm.04}$ F_1. For *sensationalism*, fine-tuning RoBERTa obtains the best results ($r = .61_{\pm.02}$). Notably, it is the only task in which the few-shot approach obtains comparable results. *Sensationalism* is also the only task for which the general domain model (RoBERTa) outperforms SciBERT. We presume that this is because detecting sensationalism is not strictly tied to scientific language, while the other tasks benefit from more specialized knowledge.

Overall, our results show that predicting fine-grained information changes is a very challenging task. Task-specific models produce more reliable results, while few-shot prompting performs poorly, even with large state of the art models. LLMs do not appear to be able to leverage the same instructions as humans, indicating that additional prompt engineering or fine-tuning may be required to obtain stronger results.

We introspect the confusion matrices and residual plot for the test predictions to identify potential error sources. Overall, we find no prevalent error patterns. For causality, we observe a slight tendency for instances expressing *Causation* being incorrectly classified as *Correlation* (18 out of 93

[10] https://huggingface.co/roberta-base

[11] https://huggingface.co/allenai/scibert_scivocab_uncased

[12] For *sensationalism*, we report correlations for the full dataset as opposed to only for the test portion. We mimic the best-worst-scaling setup from annotation and obtain the sensationalism score from the identical set of quad-tuples to allow for direct result comparison.

[13] https://huggingface.co/meta-llama/Llama-2-13b-chat-hf

[14] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.1

[15] https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1

model	causality				certainty				generalization				sensational.
	cau	corr	uncl	mF_1	c	s_c	unc	mF_1	same	r_f	p_f	mF_1	r
LLaMa-2	.52	.42	.43	.46	.04	.49	.37	.3	.37	.04	.26	.22	.24
Mistral7B	.47	.43	.25	.38	.5	.05	.34	.3	.04	.62	.32	.33	.10
MiXtral8x7B	.49	.38	.23	.4	.42	.02	.23	.22	.32	.47	.10	.3	.57
RoBERTa	.56	.62	.59	$.57_{\pm.01}$	.67	.48	.51	$.59_{\pm.04}$	.32	.69	.42	$.40_{\pm.06}$	$.61_{\pm.02}$
SciBERT	.58	.57	.60	$.54_{\pm.04}$	.70	.50	.50	$.53_{\pm.02}$	.32	.72	.49	$.47_{\pm.04}$	$.57_{\pm.03}$

Table 3: Performance for predicting distortion labels of a finding. We report per class F_1 scores, macro F_1 and correlation coefficients (r, where applicable). For few-shot prompting, we use LlaMa2-chat-hf-13B, Mistral7B and Mixtral8x7B. For fine-tuning, we use RoBERTa-base and SciBERT. For fine-tuning experiments, we report per class results of the best performing model; mF_1 scores denote avg. across 5 runs, including standard deviation.

instances). This happens slightly more frequently than such instances being confused with an *Unclear relation* (14 out of 93). For the certainty task, we see a slight tendency of the model to overestimate the level of certainty, i.e., to incorrectly classify *Certain* instances as *Somewhat certain* (40 out of 111). This is slightly more prevalent than the classifier confusing *Somewhat certain* instances with *Uncertain* instances (34 out of 111). We provide the plots in Appendix A.2.3.

5 Understanding Distorted Science Communication at Large

To gauge critical information changes more broadly, we use a large scale set of paper findings, news findings, and tweets and analyze the prevalence of distortions in science communication at large.

Data. The initial dataset is collected by pairing scientific papers from the S2ORC dataset (Lo et al., 2020) with news articles and tweets using Altmetric[16], an aggregator of mentions of scientific papers online. We automatically identify the result descriptions and select for each news and tweet finding the paper finding with the highest information matching score (Wright et al., 2022) above 4. Sec. A.3.1 describes the filtering process in detail. This gives us a set of 35,150 findings paired between papers and news and 72,032 findings paired between papers and tweets.

Using the best performing models from Sec.4, we estimate critical distortions wrt. *causality*, *certainty* and *sensationalism*.[17] Our goal is to understand which type of reports –news vs. tweets– are more susceptible to mis-reporting.

[16] https://www.altmetric.com/

[17] We exclude *generalization* from this analysis because of the varied classification performance across target classes.

Figure 4: Density plot visualizing distribution of sensationalism scores across 1,655,570 paper findings (paper F), 422,626 science news, and 356,275 tweet findings. Differences in the degree of sensationalism across different findings sources are statistically significant (see Fig. 12 in Appendix A.3).

Results. Overall, we find that science communication on Twitter is more frequently affected by mis-reporting. Tweets show pronounced critical changes in causality (Fig. 10b, Appendix A.3), while the vast majority of findings reported in science news accurately report the causal relation from the original finding (Fig. 10a, Appendix A.3). Science communicators on Twitter frequently overstate findings' certainty. Fig. 11 in Appendix A.3 shows changes in certainty levels. Notably, the vast majority of findings reported in tweets exhibit an increased level of certainty. This effect is less pronounced in the findings reported in science news. Most notably, reported scientific findings are presented with heightened levels in sensationalism, both in tweets and science news. The density plot visualizing sensationalism scores in Fig. 4 shows a large shift of towards increased sensationalism for reported findings compared to their counterparts in the original papers.

Analysis. To validate the robustness of these re-

sults, we annotate 108 instances of the unlabeled data for the properties of *causality* and *certainty*. Sec. A.3.3 provides details on the annotation. Evaluating the predicted distortions against the human labels, we find that the results in this setting are robust (macroF_1 of .65 and .59 causal distortions and certainty) and on par with the classifiers' performance on the original test set. Notably, given the relatively high precision for predicting the "certain" class (0.75) and the volume of reported findings which are predicted as such, these results provide further evidence that reports frequently overstate findings' certainty (Fig. 11). For *sensationalism* we validate the results by analyzing if the difference in distributions that we observe in Fig. 4 is in fact significant. We see a large statistically significant effect (Fig. 12), indicating that different sources are potentially perceived as more or less sensational.

6 Conclusion

Given both the societal impact and growing volume of scientific information online, it is crucial to understand how this information is presented to the public. In this work, we lay a foundation for performing large scale analysis and automatic detection of several types of distortions in the reporting of scientific findings. We contribute the first dataset of multiple fine-grained distortions in science communication, allowing us to study how scientific findings are changed when they are reported to lay audiences. We show both that findings frequently undergo subtle distortions when reported, and that detecting these distortions automatically poses a challenging problem. We find that fine-tuning custom models consistently outperforms LLM prompting, presumably because the models are not able to effectively leverage the annotation instructions and examples we provide as prompts. Using our best baseline models, we study the prevalence of distortions in science communication at large, observing that scientific findings are potentially frequently subject to distortions in terms of causality, level of certainty, and how sensationally they are presented.

Acknowledgments

This research has been conducted as part of the FIBISS project which is funded by the German Research Council (DFG, project KL 2869/5-1). Further, this work has been supported by a scholarship from the German Academic Exchange Service (DAAD, Forschungsstipendien für Doktorandinnen und Doktoranden, 2022 (57647563)). In addition, the work has been supported by a Post-doc fellowship grant awarded by the Danish Data Science Academy (2023-1425).

Limitations

While we extend existing work regarding the dimensions of distortions we study, this set may be further extended in the future. We focus on types of distortions which we consider to be applicable for a large set of disciplines, but there may be discipline-specific properties of misreporting that are out of scope of our current set of labels. Further, we focus on reports and scientific findings in English. Future work should extend this to other languages to understand the impact of the source language on the change types and their verbalization.

Our annotation study shows that the properties we are investigating are highly complex and to a certain extent subjective as indicated by the mixed agreement scores. We argue that the concepts themselves are well defined, however, we hypothesize that the mixed inter-coder agreement is a result of the fact that concepts such as sensationalism and generalization are challenging to identify on the textual level, especially for the scientific domain. We account for this for example by choosing a best-worst-scaling setup for labeling sensationalism, however, this does not (and should not) fully remove the subjective nature of the task. Sensationalism for example can be encoded in a variety of linguistic cues, but it is also connected to certain topics: someone might perceive a finding about Covid vaccines to be more sensationalist as opposed to a different treatment, because it is more frequently the topic of recent discussion. This may lead to variance in the labels without one of them necessarily being incorrect. Further analysis and modeling of the factors that constitute the perception of each of these properties should be the focus of future work.

With respect to our few-shot prompting experiments, we do not explore elaborate prompt engineering as the current work constitutes a baseline to explore in-context learning for our task. In the future, we aim to experiment with different prompts, and models and methods to extract labels from the generated output of the model as this is potentially error-prone. Further, we currently model each distortion type separately, however, to a certain degree the properties we investigate are related to each

other which a joint model may be able to leverage. We see this as an opportunity for future work.

For the large-scale analysis in Sec. 5, the results have to be contextualized with the performance of the models we use to automatically label the dataset. Our analyses validate that those findings are robust and serve as a starting point for large-scale analyses, showing that distortions are prevalent in science communication at large. However, future work is needed to determine their full extent with further developed models.

Ethical Considerations

Inaccurate science reporting is a form of misinformation, our work can therefore contribute to detecting and counter-acting false information online. This being said, while creating the resource to better understand this very task, annotators may be exposed to false information. We educate annotators about this possibility before they start the task. They can stop working on the task at any time.

References

Tal August, Lauren Kim, Katharina Reinecke, and Noah A. Smith. 2020. Writing strategies for science communication: Data and computational analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5327–5344, Online. Association for Computational Linguistics.

Christopher Bagdon, Prathamesh Karmalker, Harsha Gurulingappa, and Roman Klinger. 2024. You are an expert annotator: Automatic best-worst-scaling annotations for emotion intensity modeling. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Mexico City, Mexico. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3613–3618. Association for Computational Linguistics.

Laura Biester, Vanita Sharma, Ashkan Kazemi, Naihao Deng, Steven Wilson, and Rada Mihalcea. 2022. Analyzing the effects of annotator gender across NLP tasks. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 10–19, Marseille, France. European Language Resources Association.

Katarina Boland, Pavlos Fafalios, Andon Tchechmedjiev, Stefan Dietze, and Konstantin Todorov. 2022. Beyond facts–a survey and conceptualisation of claims in online discourse analysis. *Semantic Web – Interoperability, Usability, Applicability*, 13(5):793–827.

Luke Bratton, Rachel C Adams, Aimée Challenger, Jacky Boivin, Lewis Bott, Christopher D Chambers, and Petroc Sumner. 2019. The association between exaggeration in health-related science news and academic press releases: a replication study. *Wellcome open research*, 4.

Kathi Canese and Sarah Weis. 2013. PubMed: the Bibliographic Database. *The NCBI handbook*, 2(1).

Georgia Dempster, Georgina Sutherland, and Louise Keogh. 2022. Scientific Research in News Media: A Case Study of Misrepresentation, Sensationalism and Harmful Recommendations. *Journal of Science Communication*, 21(1):A06.

Baruch Fischhoff. 2012. Communicating uncertainty fulfilling the duty to inform. *Issues in Science and Technology*, 28(4):63–70.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

P Sol Hart and Lauren Feldman. 2016. The impact of climate change–related imagery and text on public opinion and behavior change. *Science Communication*, 38(4):415–441.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia. Association for Computational Linguistics.

George Hripcsak and Adam S. Rothschild. 2005. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association*, 12(3):296–298. _eprint: https://academic.oup.com/jamia/article-pdf/12/3/296/2429751/12-3-296.pdf.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian,

Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

Svetlana Kiritchenko and Saif Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 465–470, Vancouver, Canada. Association for Computational Linguistics.

Lauren M Kuehne and Julian D Olden. 2015. Lay summaries needed to enhance science communication. *Proceedings of the National Academy of Sciences*, 112(12):3585–3586.

Ozan Kuru, Dominik Stecula, Hang Lu, Yotam Ophir, Man-pui Sally Chan, Ken Winneg, Kathleen Hall Jamieson, and Dolores Albarracín. 2021. The effects of scientific messages and narratives about vaccination. *PLoS One*, 16(3):e0248328.

Ilia Kuznetsov, Jan Buchmann, Max Eichler, and Iryna Gurevych. 2022. Revise and Resubmit: An Intertextual Model of Text-based Collaboration in Peer Review. *Comput. Linguistics*, 48(4):949–986.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

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. *CoRR*, abs/1907.11692.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel S. Weld. 2020. S2ORC: the semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4969–4983. Association for Computational Linguistics.

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

Isabelle Mohr, Amelie Wührl, and Roman Klinger. 2022. Covert: A corpus of fact-checked biomedical COVID-19 tweets. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, pages 244–257. European Language Resources Association.

Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Apostolos Dedeloudis, Jackson Sargent, and David Jurgens. 2022. Potato: The portable text annotation tool. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Jiaxin Pei and David Jurgens. 2021a. Measuring sentence-level and aspect-level (un)certainty in science communications. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9959–10011, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jiaxin Pei and David Jurgens. 2021b. Measuring sentence-level and aspect-level (un)certainty in science communications. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9959–10011, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

David F Ransohoff and Richard M Ransohoff. 2001. Sensationalism in the Media: When Scientists and Journalists May be Complicit Collaborators. *Effective clinical practice*, 4(4).

Dennis Reidsma and Jean Carletta. 2008. Squibs: Reliability measurement without limits. *Computational Linguistics*, 34(3):319–326.

Victoria L. Rubin. 2007. Stating with certainty or stating with doubt: Intercoder reliability results for manual annotation of epistemically modalized statements. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pages 141–144, Rochester, New York. Association for Computational Linguistics.

Joselita T Salita. 2015. Writing for lay audiences: A challenge for scientists. *Medical Writing*, 24(4):183–189.

Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Jezek. 2023. Why don't you do it right? analysing annotators' disagreement in subjective tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2428–2441, Dubrovnik, Croatia. Association for Computational Linguistics.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Petroc Sumner, Solveiga Vivian-Griffiths, Jacky Boivin, Andy Williams, Christos A Venetis, Aimée Davies, Jack Ogden, Leanne Whelan, Bethan Hughes, Bethan

Dalton, et al. 2014. The association between exaggeration in health related science news and academic press releases: retrospective observational study. *Bmj*, 349.

Phillip J Tichenor, Clarice N Olien, Annette Harrison, and George Donohue. 1970. Mass Communication Systems and Communication Accuracy in Science News Reporting. *Journalism Quarterly*, 47(4):673–683.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Jan Trienes, Sebastian Joseph, Jörg Schlötterer, Christin Seifert, Kyle Lo, Wei Xu, Byron C. Wallace, and Junyi Jessy Li. 2024. Infolossqa: Characterizing and recovering information loss in text simplification.

Juraj Vladika and Florian Matthes. 2023. Scientific fact-checking: A survey of resources and approaches. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6215–6230, Toronto, Canada. Association for Computational Linguistics.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 7534–7550. Association for Computational Linguistics.

Dustin Wright and Isabelle Augenstein. 2021a. CiteWorth: Cite-Worthiness Detection for Improved Scientific Document Understanding. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 1796–1807. Association for Computational Linguistics.

Dustin Wright and Isabelle Augenstein. 2021b. Semi-supervised exaggeration detection of health science press releases. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10824–10836, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dustin Wright, Jiaxin Pei, David Jurgens, and Isabelle Augenstein. 2022. Modeling information change in science communication with semantically matched paraphrases. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1783–1807, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Bei Yu, Yingya Li, and Jun Wang. 2019. Detecting causal language use in science findings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4664–4674, Hong Kong, China. Association for Computational Linguistics.

Bei Yu, Jun Wang, Lu Guo, and Yingya Li. 2020. Measuring correlation-to-causation exaggeration in press releases. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4860–4872, Barcelona, Spain (Online). International Committee on Computational Linguistics.

A Appendix

A.1 Annotation

A.1.1 Setting

Participant filtering. To qualify for our study, participants have to have an undergrad degree in one of the following subjects: Computer Science: Computer Science, Computing (IT), Mathematics, Science; Biology & Medicine: Biochemistry (Molecular and Cellular), Biological Sciences, Biology, Biomedical Sciences, Chemistry, Health and Medicine, Medicine; Psychology: Psychology. Using Prolific's 'Balanced sample' option, we rely on the platform to distribute the study evenly to male and female participants. Participants are required to be fluent in English.

Payment. All studies are designed to take approx. 13 minutes. Annotators are payed £1.95 per study. This amounts to £9 per hour which PROLIFIC recommends as a fair compensation.

Number of annotations. For the change types *causality*, *certainty*, and *generalization*, we collect 3 sets of annotations for every instance. For *sensationalism*, we generate $1.5N$ quad-tuples, N being the total number of findings to be labeled, and collect 2 sets of annotations for the resulting quad-tuples.

A.1.2 Tasks

Table 4 provides an overview of the annotation tasks.

Causality Annotators are provided the following task description and examples: "Identify what type of causal relationship is described in a finding. Causality describes a cause-effect relationship between two things, variables, agents etc. A (directly) causes outcome B. Correlation describes relationships where two actions relate to each other, but one is not necessarily the effect or outcome of the other. Cues for causality are: *cause, direct connection, result in, lead to, trigger, produce, increase, decrease*. Cues for correlation: *associated with, association, connection, correlated with, linked to*. When in doubt or not clearly stated as a causal relation, it's usually a correlation. Let's go over some examples: FINDING 1: Low vitamin D levels cause tiredness. FINDING 2: Exposure to traffic noise at the office increases stress levels. → Both examples describe a causal relationship: The cause A (low vitamin D, traffic noise) causes outcome B (tiredness, increased stress level). Compared to that, this one describes a correlation: FINDING 3: Low Vitamin D levels are associated with tiredness. FINDING 4: Stress levels are higher in offices exposed to traffic noise. → Both examples describe a correlation. In both sentences the variables are related or associated to each other, but there it is unclear if one is the direct cause of the other. Sometimes no relation is stated: FINDING 5: We find evidence of biases across the majority of languages. → This finding presents a summary in which no causal relation or correlation is stated."

The labels are defined as follows; they follow Sumner et al. (2014) and annotators see them as label descriptions in the annotation environment:

- No mention of a relation: No mention of a relation of any kind. E.g., if the finding is a summary such as *We find evidence of biases across all languages*
- Correlation: The paper finding describes a correlation between two elements. E.g., *Vitamin D levels are associated with extensive tiredness*
- Causation: The paper finding describes a causal relation between two elements. E.g., *Low vitamin D levels cause extensive tiredness* or *Tiredness might be caused by lack of vitamin D.*
- Explicitly states: no relation: The finding explicitly states that there is no relation between the two elements. E.g., *We find no evidence that vitamin D levels are associated with tiredness.*

Certainty Annotators are provided the following task description and examples: "Rate the level of certainty that the author uses to describe a finding. Certainty means having complete confidence in something without any doubts. Uncertainty is the opposite: it refers to a state of doubt, lack of confidence, or absence of complete knowledge about something. Both can be expressed with respect to various aspects. Look out for: (un)certainty towards specific numbers/quantities: *They found that approximately 50% of participants...*, the extent to which a given finding applies: *The effect was mainly observed for teenagers.*), the probability that something applies, occurs or is associated: *possibly associated with*, hedging words: *seem, tend, appear to be, may, potentially, suggest, perhaps*. Let's go over some examples: FINDING 1: Now there is clear evidence that sunscreen prevents skin damage. → The description of the finding is very certain. No hedges or other indicators of uncertainty. Compared to that, this one is slightly less certain: FINDING 2: New study shows that sunscreen can prevent skin damage. → The description of the finding is pretty certain. The use of *can* indicates that the finding is limited in some way, but overall the finding is presented to be mostly certain. Let's look at some examples that express uncertainty: FINDING 3: New study suggests that sunscreen could prevent skin damage. → The finding is described to be pretty uncertain. The use of *could* and *suggests* are indicators that the findings are preliminary or very limited with regards to how impactful they may be. Let's go all the way to an uncertain finding: FINDING 4: Study presents potential indicators that sunscreen might have positive effects in preventing skin damage. → The finding is described to be very uncertain. It is stated that the results are indicators instead of a definite explanation. The use of the word *might* emphasizes the uncertainty of the finding."

Annotators are provided the following label descriptions in the annotation environment:

- Uncertain: E.g., 'Sunscreen might prevent skin cancer.' or 'Overconsumption of sugar may have negative effects on health.', 'Further research is necessary to understand...'

	Causality	Certainty	Generalization	Sensationalism
setup	classification	rating scale	comp. classification	best-worst-scaling
task	In the finding, what type of causal relationship is described?	How do you rate the level of certainty used to describe the finding?	Which finding is more general?	Which of the findings is the least/most sensational?
label space	Causation, Correlation, Expl. states: no relation, No relation mentioned	Certain, Somewhat certain, Somewhat uncertain, Uncertain	Reported Finding, Paper Finding, Same level of generality	Most sensational, Least sensational
aggregation	MACE	MACE	MACE	count-based BWS-score

Table 4: Overview of annotation tasks, annotation setup, label and aggregation strategies. MACE: Hovy et al. (2013), BWS-score (Kiritchenko and Mohammad, 2017): real-valued score obtained from the best-worst scaling.

- Somewhat uncertain: E.g., 'Sunscreen could prevent skin cancer.' or 'Overconsumption of sugar can cause diabetes.', 'The functionality possibly depends on...'

- Somewhat certain: E.g., 'Sunscreen can prevent skin cancer.' or 'The analysis suggests that papers with short titles receive more citations'

- Certain: E.g., 'Sunscreen prevents skin cancer.' or 'Papers with shorter titles get more citations', '...meaning that this treatment should be used...'

Generalization Annotators are provided the following task description and examples: "Identify if Finding A generalizes the results from Finding B. Generalizations claim something is always true, even if something is only valid in certain instances or occasionally. Let's go over an example: Read Finding A. FINDING A: Parent conversations with children about their weight connected to disordered eating Compare that to Finding B: FINDING B: Disordered eating was more prevalent in children whose fathers engaged in weight conversations. → Finding B specifies that it is conversation with fathers which were investigated. Finding A generalizes the statement from fathers to all parents. In the study, you are tasked to decide which of the findings is more general: Here, the correct solution is Finding A. Let's look at another example: FINDING A: Magnesium potentially has many health benefits. FINDING B: Increasing dietary magnesium intake is associated with a reduced risk of stroke and heart failure. → Here, Finding B specifies they researched dietary magnesium and the medical conditions that are affected. Finding A is more general, the correct solution is therefore Finding A. Both findings can be on the same level of generality: FINDING A: Dietary magnesium potentially has health benefits. FINDING B: We show that dietary magnesium had a positive influence on the participants' overall health. The correct solution is They are at the same level of generality."

Annotators are provided the following label descriptions in the annotation environment:

- Finding A: E.g., Finding A discusses a general finding about diabetes, while the report discusses the finding for a specific demographic only. Or generalizing from a specified set of medical conditions ('reduced risk of stroke, heart failure, diabetes') to a general statement ('has health benefits').

- Finding B: E.g., Finding B generalizes a finding about diabetes type 2 in seniors to all people with diabetes. Or generalizing from a specified set of medical conditions ('reduced risk of stroke, heart failure, diabetes') to a general statement ('has health benefits').

- They are at the same level of generality: Finding B accurately reports the Finding A with regards to generality.

Sensationalism Annotators are provided the following task description and examples: "Rate how sensational the language of a finding is. Sensational text intends to spark the interest of a reader, make them curious or elicit an emotional reaction. Cues for sensationalism could be: dramatic, urgent, exaggerated language: *life-changing*, *unparalleled performance*, *revolutionary*, *transformative*, *miracle treatment* or use informal or colloquial language: *amps up the efficiency*, *They ran some solid experiments to back this up*. Let's go over an example: FINDING A: Looks like vitamin D intake influences stress levels: New study by @username. FINDING B: Urban Green Spaces and Mental Health: A Positive Correlation Revealed. FINDING C: Exciting new research suggests that upping

your daily step count could be a simple solution to alleviate insomnia. FINDING D: We observe improved plant growth through positive human energy in our controlled study setup. Read all findings carefully. Some of the text contain sensational or entertaining language like *revealed*, *exciting*, or *simple solution*. Based on that you determine which finding uses the most sensational language and which one the least: Here, Finding D is the LEAST sensational. FINDING C is the MOST sensational."

Annotators are provided the following label descriptions in the annotation environment:

- A: Finding A is the least sensational.
- B: Finding B is the least sensational.
- C: Finding C is the least sensational.
- D: Finding D is the least sensational.

and

- A: Finding A is the most sensational.
- B: Finding B is the most sensational.
- C: Finding C is the most sensational.
- D: Finding D is the most sensational.

A.2 Experiments

A.2.1 Experimental setting

Fine-tuning task-specific models. We train all models using a Nvidia Titan-RTX GPU. We train for 5 epochs with a learning rate of 2e-5, a batch size of 8, 200 warmup steps, a weight decay of 0.01. We use the Adamw optimizer.

Few-shot prompting. For each input prompt, we generate the output sequence by using top k sampling with k=10. We set the maximum number of generated tokens to 200. We use a NVIDIA Tesla V100 GPU to generate the responses with the LLaMa model. Generating the responses for the full dataset takes approx. 7 hours. For the Mistral7B model, generating all responses takes approx. 5 hours on an Nvidia GeForce RTX A6000 GPU. For Mixtral8x7B, generation takes approx. 20 hours split across 5 Nvidia GeForce RTX A6000 GPUs. We wrap each prompt with [INST][/INST] tags to mimic the chat format from pre-training and include a system prompt that instructs the model to 'act' like a reliable annotator.

You are a reliable annotator in an annotation study. You studied {SCIENTIFIC DISCIPLINE}
{TASK DESCRIPTION}
{TASK EXAMPLES}
Now consider the following finding:
{FINDING, FINDING PAIR, QUAD-TUPLE} {QUESTION}
What is the correct solution? Choose one option. Do not repeat the findings.

Figure 5: Prompt template. We provide the instantiated prompts along with the LLM-specific system prompts and markup in the supplementary material.

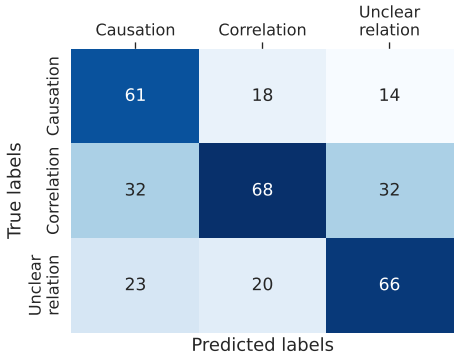


Figure 6: Confusion matrix of SciBERT test set predictions for causality.

A.2.2 Prompts

Fig. 5 shows our prompt template. For the BWS-prompts, we follow Bagdon et al. (2024) who experiment with automatically generating training data using best-worst-scaling. We provide the task-specific prompts in the supplementary material.

A.2.3 Error Analysis

Figures 6, 7, 8 and 9 visualize the confusion matrices and residual values for the test set predictions of the best classification/regression models, i.e., SciBERT for causality, certainty and generalization and RoBERTa for sensationalism.

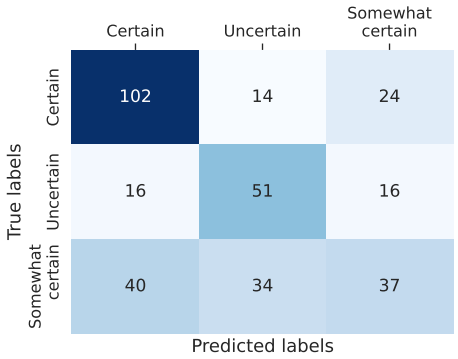


Figure 7: Confusion matrix of SciBERT test set predictions for certainty.

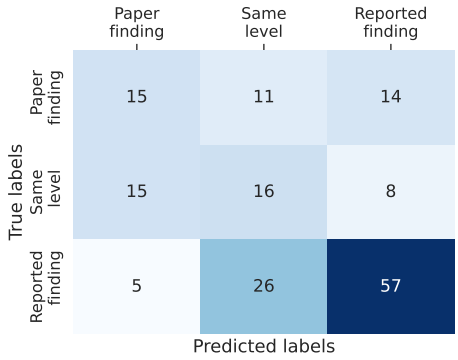


Figure 8: Confusion matrix of SciBERT test set predictions for generalization.

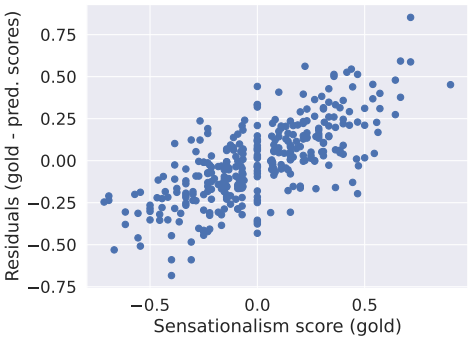


Figure 9: Residual plot for RoBERTa test set estimates for sensationalism score.

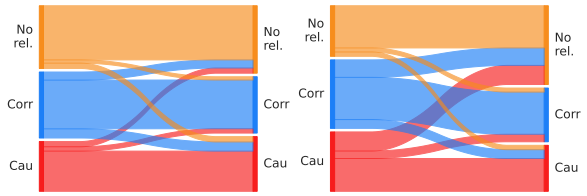
A.3 Additional Analyses

A.3.1 Filtering Process

The initial dataset is collected by pairing scientific papers from the S2ORC dataset (Lo et al., 2020) with news articles and tweets using Altmetric, an aggregator of mentions of scientific papers online.[18] The scientific papers and news articles are initially parsed to predict which sentences correspond to either a result or conclusion using a RoBERTa model (Liu et al., 2019)[19] trained on 200K paper abstracts from PubMed that are self-labeled with paper section categories (Canese and Weis, 2013). We then pass all pairs of conclusion and result sentences from papers with conclusion and result sentences from news articles and tweets through the model from **?** to measure the similarity of the pair of findings, and select for each news and tweet finding the paper finding with the highest IMS above 4.

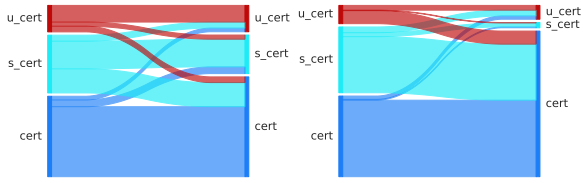
[18] https://www.altmetric.com/

[19] roberta-base



(a) Science news (b) Tweets

Figure 10: Changes in causality across paired findings from science news (35,150) and tweets (72,032).



(a) Science news (b) Tweets

Figure 11: Changes in certainty across paired findings from science news (35,150) and tweets (72,032).

A.3.2 Distortions in Large Scale Data

Visualizations of changes in causality (Fig. 10) and certainty (Fig. 11) across paired findings from science news (35,150) and tweets (72,032) as Sankey diagrams.

A.3.3 Validating the Results from the Large-scale Analysis

For annotating the subset of the unlabeled data, we sample 108 instances across all four scientific disciplines and collect three sets of labels for each instance. One set is annotated by one of the authors, the other two sets are obtained from independent annotators who we provide with the same annotation instructions that we used to instruct the crowdworkers.

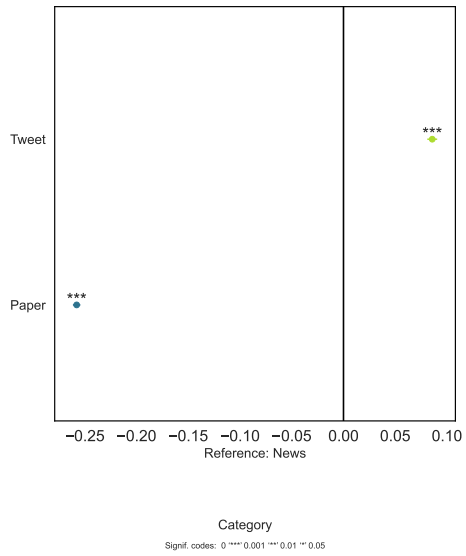


Figure 12: Plot of regression coefficients accompanying Fig. 4 for predicting sensationalism score based on findings source (Paper, News, or Tweet). We see a large statistically significant effect, indicating that different sources are potentially perceived as more or less sensational.