Evaluating the Quality of Science Communication on r/science

DAN HICKEY, School of Information, University of California, Berkeley, United States
JULIAN STRIETZEL, Technical University of Munich, Germany
VETLE WIEDSWANG JAHR, University of Oslo, Norway
JARED MANTELL, University of California, Berkeley, United States

Clear and accurate science communication is critical for helping the public become better informed about scientific topics. Importantly, scholars have recently identified several issues with the quality of scientific discourse on social media, since the results of scientific research are often distorted or sensationalized by major media outlets. However, little is known about how the qualities of science communication such as accuracy or clarity are rewarded in online environments, how those qualities have changed over time, and how those qualities differ across diverse media sources. Such insights would help inform how recommendation algorithms and social media interfaces could be designed to optimize for more effective science communication. To address this knowledge gap, we systematically evaluate posts shared on r/science from 2016 to 2022, measuring the level of jargon, sensationalism, and factual consistency in each post, and exploring how these metrics correspond with the type of source used in the post and the engagement metrics of the posts. We find that posts that link to news sources are more sensational and contain less jargon than posts that link to academic papers directly. Furthermore, we find that moderate levels of sensationalism and factual consistency are associated with posts receiving high numbers of upvotes, while jargon is negatively associated with upvotes. Finally, we observe that in 2020, the quality of science communication on the subreddit improved, associated with shifts in topical content during the COVID-19 pandemic. Ultimately, our results suggest that solely optimizing for engagement in social media platforms will not reward the most effective science communication, but our work offers critical insights into how platforms can be reimagined to support the platforming of high-quality science communication.

Our code and data is available on GitHub.

Additional Key Words and Phrases: reddit, science communication, jargon, sensationalism, factual consistency, r/science

1 INTRODUCTION

The dissemination of scientific research provides several benefits to society, including increased public enthusiasm for science [6], and the affordance for the public to make informed decisions about their lives [23]. The shifting media landscape in the 21st century means that science communication is primarily taking place on online platforms now, which come with unique affordances, such as the ability for users to discuss science in comments sections or platforms' personalized forms of content recommendation that respond to signals of user engagement [5]. Previous work addressing the implications of the changing media environment for science communication has often focused on the subreddit r/science as a case study, showing how the subreddit's affordances affect users' perceptions of scientific content [26], how users view the purpose of the subreddit [12], and how the complexity of the language on the subreddit differs from other subreddits [2]. The focus on the complexity of the language in the subreddit is particularly important when determining the quality of science communication and scientific discourse online, as scholars have emphasized

Authors' addresses: Dan Hickey, School of Information, University of California, Berkeley, Berkeley, United States, dan_hickey@berkeley.edu; Julian Strietzel, Technical University of Munich, Munich, Germany, julian.strietzel@tum.de; Vetle Wiedswang Jahr, University of Oslo, Oslo, Norway, vetlewj@uio.no; Jared Mantell, University of California, Berkeley, Berkeley, United States, jaredm@berkeley.edu.

that effective science communication should be understandable to a large audience (i.e., free of words that only experts know the meaning of) [27].

The quality of science communication can also be illustrated in other ways – for example, in the process of providing summaries for laypeople, communicators of science often make the mistake of sensationalizing science, which can give readers the wrong impression of the degree of the impact of a given piece of research [24], or misrepresenting results from the primary literature, which can give readers the wrong idea of what a given study found [7]. While similar qualities have been used to extensively evaluate machine-generated science communication [10], the quality of science communication on social media nonetheless plays a key role in whether or not the public will correctly interpret the findings of scientific literature, and no study has systematically evaluated science communication quality in online communities and how quality is improved or diminished due to the unique affordances of social media platforms.

Research Questions. To address the gap in research concerning the quality of online science communication, we measure three different aspects of posts in `r/science`, each capturing a different element of the quality of the posts: the level of *jargon* in the posts (indicating how clear or easy the posts are for non-specialist audiences to understand), the *sensationalism* of the posts (indicating how exaggerated, shocking, or dramatic the posts are), and how *factually consistent* the posts are with the original studies they reference. We use these metrics to answer three research questions concerning science communication quality on `r/science`:

RQ1: How does science communication quality differ by the type of source used in the original Reddit post?

While `r/science` has some strict rules regarding what type of content can be posted to the platform, the most notable of which being that all content must refer to peer-reviewed research published within six months of the Reddit post, there is nonetheless flexibility in what kinds of sources of scientific reporting can be posted to the community. Namely, users can post news articles that reference a research paper or they can post direct links to research papers. Some users may also deviate from those norms and post links to other sources such as social media websites. In all cases, users can generate an original title describing the work. Given prior work showing that the results of research become more distorted as they pass through more channels of communication [11], there is reason to believe that the quality of science communication may be reduced when shared from secondary sources such as news articles.

RQ2: Is high-quality science communication favored by users?

By default, Reddit ranks the posts in a given community using the net "karma" (number of upvotes minus downvotes) and the recency of the posts, favoring recent posts that have been upvoted highly [19]. Beyond Reddit, personalized recommendation algorithms of other social media platforms often "optimize for engagement," recommending content to users that the user is predicted to read, watch, like, or share [9]. Given that the number of upvotes a post receives has a large impact on whether future users will see the post, it is important to know whether there is an association between the quality of a post's science communication and its net karma. Answering this question will help us understand whether optimizing for engagement platforms posts that contain high-quality science communication.

RQ3: How has the quality of science communication changed over time?

R/science is not a static entity – it is an evolving community with changing norms, rules, and user bases. Similarly, the public engagement with science can change over time. For example, scholars argue the COVID-19 pandemic increased public engagement with science [21]. For this reason, it is important to understand the historical context of the community we are studying and how changing rules, norms, and societal shifts may have an impact on the quality of science communication in the community. For this reason, we examine how the quality of science communication has changed over time in `r/science`.

Ultimately, our results paint a complex picture of the quality of science communication in r/science. For instance, we find that posts referencing news articles are more sensational than posts that reference academic literature directly, but they also contain less jargon, hinting at a difficult trade-off between the clarity of a piece of science communication and its reliability (RQ1). There is a negative association between the level of jargon in a post and its total karma, and posts that are extremely low in factual accuracy are often penalized. However, low-sensationalism posts do not perform well, and the most factually accurate posts suffer a similar fate (RQ2). Analyzing the temporal trends of the subreddit shows us that the COVID-19 pandemic likely brought a higher caliber of science communication quality to the subreddit than was prevalent before (RQ3). Taken together, these results tell us that while the "status quo" of optimizing for engagement does not reward the most ideal science communication, there are many opportunities for designing platform features and recommenders that amplify higher-quality posts.

2 METHODS

2.1 Dataset

To understand online science communication, we must obtain a dataset with examples of social media posts describing the results of specific scientific research papers. Obtaining such a dataset from a broad range of online communities is difficult, as social media posts that link to scientific sources may not necessarily directly describe the research they are referencing. For this reason, we turn to the community of r/science[1], a subreddit dedicated to reporting results of scientific research [12]. At the time of writing this paper, r/science is the 11th most subscribed to subreddit on Reddit (33M subscribers), and it describes itself as a place to "find and submit new publications and popular science coverage of current research." Additionally, one of the current rules of r/science is that "submissions must directly link either to recently published peer-reviewed research or a professional media summary of that peer-reviewed research." Most submissions that follow this rule are accompanied by a short title (median length = 14 words) summarizing the research. Users are also able to discuss submissions to the subreddit in the comments section of each submission, and prior research has suggested that users often read the comments on a given submission before deciding to read the primary article linked in the submission [12], indicating that the quality of submission titles in the subreddit have serious implications for how readers of the community perceive the results of scientific research. In summary, analyzing the titles of submissions to r/science reveals itself to be a useful and viable case study, as the community is one of the largest and most influential arenas of online science communication while maintaining community norms that allow us to safely assume the vast majority of submissions will summarize the results of specific academic research papers.

To collect posts from r/science, we use the Pushshift dataset [3]. Pushshift is a system that stores posts from Reddit very shortly after they are created, allowing researchers to easily collect historical data from Reddit without having to wait for the slow rate limits of the official Reddit API. However, the historical nature of Pushshift presents a challenge for us – low-quality posts that are against the rules of the subreddit, and do not summarize academic research, may appear in our dataset. It is also possible that due to changing community norms or moderation practices, low-quality posts may have been prevalent earlier in the subreddit's history. To ensure the quality of our dataset, one author of this paper manually checked random samples of 100 submissions from different time periods of the subreddit, labeling whether or not each post described scientific research. While we observed that approximately 50% of the posts from a simple random sample of the subreddit described academic research, filtering out posts from before 2016 improved the percentage of valid posts to 80% with no additional filtering. For this reason, we chose to include submissions from

[1] https://www.reddit.com/r/science/

r/science made within the years 2016-2022 in our final dataset, amounting to a total of 198,000 submissions. Along with the titles and timestamps of each submission, we collected the total karma of the submission, the URL and domain of the primary source material, and the "flair" of the submission, which describes the category of science the submission falls under (e.g., biology, psychology, physics, etc.). In addition to constraining the time period of our dataset to improve its quality, we also build a domain classification pipeline to filter out low-quality domains, which we describe in further detail in Section 2.5.

2.2 Measuring Jargon

The use of jargon — discipline- or field-specific terms — can make scientific texts more complex and less accessible to a general audience. Furthermore, jargon is negatively correlated with interdisciplinary work [17]. For example, the title: '*Mural lymphatic endothelial cells regulate meningeal angiogenesis in the zebrafish*' [4] contains several specialized terms that are difficult to understand by outsiders. To quantify how this added complexity affects accessibility, we measure the amount of jargon in each post's title. We chose to focus on titles due to the rigid format of r/science, where each submission only has a title and a link, hence the title being the main customizable part where the Reddit users can control how to convey the information.

To quantify the amount of jargon in a post, we calculate the proportion of jargon words by dividing the number of such terms by the word count. Jargon words are defined using a 'jargon-dictionary' developed by Lucy et al [17]. We rely on Lucy et al's dictionary as it has been specifically made and validated against abstracts in scientific papers. This dictionary is organized into 264 categories, each containing words with an associated 'NPMI score'[2]. These word lists are derived from a corpus of abstracts in the S2ORC dataset [16].

These categories help capture how certain words are distinctive to particular fields, the word senses. For example, the common word 'tree' is considered jargon in computer science due to its specialized meaning in that field, differing from its meaning in day to day speech. To ensure that words are only considered jargon within the domains where they have a specialized meaning, we only count jargon words for the categories relevant to the post's field. Instead of relying on the entire dictionary, we only use the domain-specific categories that apply to the post's field.

As r/science uses around 25 relatively broad categories — after filtering away categories we don't wish to investigate e.g. 'Ask Me Anything' — while the jargon-dictionary has 264 more granular categories, we mapped the jargon-dictionary's categories to the subreddit's to be able to effectively analyze the Reddit posts without having to manually categorize each individual post. For example, 'Software Engineering' from the jargon dictionary is classified under the broader 'Computer Science' category in r/science. The methodology for this category alignment is described in the Appendix B.

For each top-level subreddit category, we merged all associated jargon-dictionary categories into a single list. Following Lucy et al [17], a word is classified as jargon if its NPMI score exceeds 0.1. Positive NPMI values indicate association, while values too close to 0 indicate independence, hence setting the threshold to 0.1 proves to be a good balance.

We calculate the jargon proportion for each post by using its subreddit category to determine the relevant set of jargon words.

Evaluation. Creating jargon lists has been extensively evaluated in previous work [17, 22]. Lucy et al [17] — the jargon dictionary that we opted to use — validated their pipeline by comparing their model's word associations with

[2]NPMI score is a measurement of how associated a word is to a category, on a scale of -1 (no association) to 1 (perfect association) [17]

subfield definitions from Wiktionary by checking if words that were strongly linked to a subfield also had high NPMI scores. Additionally, Rakedszon et al 2017, who use similar methods to quantify jargon, thoroughly evaluated their program for jargon identification by comparing against existing measures, based on established word frequency lists from academic texts and writing samples, transcripts from TED lectures (both science and non-science lectures) and academic abstracts and lay summaries.

Altogether, these studies validate that the methodology of creating and utilizing word lists to classify jargon is a useful approach.

2.3 Measuring Sensationalism

To analyze the degree of sensationalism in Reddit posts from the `r/science` subreddit, we employed the fine-tuned model developed by Wührl et al. for detecting sensationalism in science communication [28]. This model, based on a fine-tuned RoBERTa architecture [15], demonstrates strong performance in detecting fine-grained distortions when applied to annotated pairs of scientific findings and their reported versions in news articles and tweets.

Model Adaptation and Implementation. We utilized the fine-tuned sensationalism model as provided by Wührl et al. 2024, without modifications to its architecture or training data. The model had been fine-tuned on an annotated dataset of 1,600 instances of scientific findings paired with corresponding reports. For our study, the model was applied to standalone Reddit posts, rather than paired comparisons with scientific papers. This application was carried out with guidance from the original authors, who welcomed our adaptation of the model as a means to evaluate its generalization and applicability.

The sensationalism model was applied to all posts in the `r/science` subreddit dataset. In accordance with the methodology recommended by Wührl et al., the model was integrated into a pipeline optimized for large-scale data processing. This ensured consistency with their framework and leveraged their expertise for reliable application at scale.

Known Limitations. The model outputs a continuous score ranging from 0 to 1, quantifying the level of sensationalism in each Reddit post. A score of 0 indicates minimal sensationalism, while a score of 1 reflects high sensationalism. Analysis of the distribution of scores revealed a concentration around 0.5, suggesting that the majority of posts exhibit moderate levels of sensationalism (see Figure 15). This distribution raises the concern that the model may not fully capture the entire range of sensationalism levels, potentially compressing variations into a narrower band. This characteristic does not undermine the relative differences in scores across posts and therefore our analysis. It could nevertheless limit the interpretability of extreme values and estimated coefficients.

The subjective nature of sensationalism detection, especially for posts discussing controversial or widely debated scientific topics, presents a potential limitation. These subjective influences may introduce variability in the perception of sensationalism. While we did not explicitly analyze this issue in our study, we rely on the validation and discussion provided by Wührl et al. to address these challenges comprehensively.

Evaluation. To evaluate the performance of our sensationalism metric, three annotators, who are also authors of this paper, manually rated the sensationalism level of a random sample of 50 `r/science` titles. Each title was rated on an integer scale of 0-2, where 0 indicated no sensationalism and 2 indicated high sensationalism. The annotators

were given the Wiktionary definition of sensationalism.[3] The Krippendorff's alpha value of the annotations was 0.57, indicating moderate agreement among annotators.

By evaluating the calculated scores against the human annotations we get a Pearson correlation of 0.36, a Spearman correlation of 0.29 and a Kendall's Tau of 0.22, all statistically significant having a p-value < 0.05.

The calculated scores show less variance than the human annotations. Because the model outputs a continuous value between 0 and 1, it can capture more subtle nuances than the human annotators who only had three levels of sensationalism. This difference in granularity may partly explain the low correlations. Despite this, there appears to be some alignment at the higher end of calculated scores, while the variance remains high at the lower end.

These findings suggest that this metric can benefit from further research on improving the correlation with the human annotators. We suggest annotating a larger dataset, as well as a more granular scale.

2.4 Factual Consistency

We define factual consistency (FC) in our context as "the degree to which a summary can be inferred from an original." This incorporates the general aspects we want to reason about within this metric. The motivation behind this is that factually inconsistent headlines, which can not be inferred from a research paper should generally be discouraged within a science community as they might spread misinformation and reduce the general trust in the scientific method. Ideally, we want to use a metric that gives us a rating about factual consistency between the Reddit post and the original published paper to which the post is referring. In practice we run into multiple limitations though, as we 1) can not get all respective research papers, 2) calculating FC metrics respective to longer text seems more unreliable, and 3) using large language models (LLMs) is expensive. We therefore limit our analysis to focus on FC between the Reddit posts and the abstracts of a subset of the papers, which were readily available through basic web-scraping and regex extraction (see Appendix C).

LLMs have been shown to be reasonably good factual consistency evaluators. When compared to other NLP methods, like ROUGE, BERT or BART scores, they correlate best to human evaluators with Spearman correlation [29], Pearson correlation [20], and Kendall's Tau [13] of $\rho = 0.419$, $r = 0.517$, and $\tau = 0.389$ [18, 25]. Additionally, they are easy to employ, scalable, and interpretable by their instructions. We are running the following prompt (inspired by [25]) on gpt-4o-mini-2024-07-18 using openai batch api.

```
Score the following reddit post summarization given the referenced research abstract with
    ↪ respect to factual consistency or identical titles on a discrete scale from 1 to 5,
    ↪ where a score of 1 means 'inconsistencies or statements that can't be inferred' and
    ↪ score of one 5 means 'perfect factual consistency or identical titles'.
Note that factual consistency measures the 'degree to which the reddit summary can be inferred
    ↪ from the research abstract'.
Research Title: {...}
Abstract: {...}
Reddit Summary: {...}
Respond in the following format: {"score": *insert score here*}
```

[3] Wiktionary definition of sensationalism: "The use of sensational subject matter, style or methods, or the sensational subject matter itself; behavior, published materials, or broadcasts that are intentionally controversial, exaggerated, lurid, loud, or attention-grabbing. Especially applied to news media in a pejorative sense that they are reporting in a manner to gain audience or notoriety but at the expense of accuracy and professionalism."[1]

Evaluation. Apart from referencing the evaluation done by Wang et. al in [25] we conducted our own evaluation of the factual consistency metric. We therefore use human annotations towards generated summaries on the SummEval dataset [8] and comparing our results on factual consistency.

We are able to reproduce the results from the paper with a slight decrease in performance resulting in a Spearman correlation coefficient of $\rho = 0.39$, a Pearson correlation coefficient of $r = 0.48$, and a Kendall Tau correlation coefficient of $\tau = 0.36$. More importantly, though, we can show that 97.1% of the summaries classified as good (4 or higher) by the model were also ranked high by experts, and 81.3% the other way around. Our method deviates from the ground truth by 3 or more points in only 6% of the cases. In general, our model is more critical and nuanced about the consistency of a text, with more ratings in the mid-range and less perfect ratings. This might be considered problematic and could be caused by the way that the model is prompted compared to how human evaluators are ranking multiple measures in the same run. For our analysis and general trends this is not expected to be an issue.

2.5 Domain Classification

To facilitate the analysis of engagement factors, such as jargon, sensationalism, and factual consistency, across distinct domain types referenced in articles from `r/science`, we developed a systematic domain classification and clustering process. This approach categorizes domains into one of five predefined groups: *scam*, *social_media*, *news*, *scientific*, or *repo*, with an additional category, *unknown*, for ambiguous cases. Domains appearing fewer than two times in the dataset were excluded from the analysis to ensure statistical relevance and computational efficiency. This clustering supports targeted analyses of engagement factors across different reference channels, such as social media versus scientific repositories, and helps to generate meaningful insights during the analysis. To the best of our knowledge, this domain classification and clustering approach is novel in the context of analyzing scientific communication on online platforms.

Prompting Strategy for Domain Classification. Domain classification was performed using `GPT-4o-mini-2024-07-18`, with the following structured prompt:

```
Please classify the following domains as either:
"scam": scam or irrelevant domain (e.g., bit.ly, goo.gl, etc.)
"social_media": generic social media domain
   (e.g., youtube, twitter, facebook, pinterest, instagram)
"news": relevant news site (e.g., nytimes, wsj, cnn, phys.org)
"scientific": relevant science site (e.g., sciencedaily, phys.org, nature)
   and university sites ending in .edu that are not repositories fall under this category
"repo": science repository (direct link to paper such as doi, arxiv, pubmed)
"unknown": if unsure, please classify as "unknown."

The output should be a dictionary with the domain as the key and the rating as the value,
    ↪ formatted like this:
{
   "bit.ly": "scam",
   "youtube.com": "social_media",
   "nytimes.com": "news",
```

```
    "sciencedaily.com": "scientific",
    "doi.org": "repo"
}
Please adhere strictly to the labels provided above. If unsure, classify as "unknown".
```

To optimize computational efficiency, requests were batched to handle up to 64 domains in a single prompt. This strategy minimized overhead while ensuring consistent processing across the dataset.

Validation and Refinement. To reduce inconsistencies in the model's output, classifications were repeated across three independent iterations for each domain. The model classification, defined as the most frequent label among the three outputs, was selected as the final label. In cases where all three outputs differed, the domain was labeled as *indecisive*. Domains classified as *indecisive* or *unknown* were subjected to manual review, with special attention given to the most frequently occurring and ambiguous cases. For instance, `self.science`, initially classified as *unknown*, was reclassified as *social_media* after verifying that it referenced Reddit.

Manual validation was performed against a human-labeled subset of domains to ensure reliability. Domains appearing fewer than two times in the dataset were excluded from further analysis, allowing us to focus on statistically significant data. This process ensured that the classification framework was both precise and scalable. When manually labeling the 100 most frequent domains in our sample, we find our automated categorization of domains to be 97% accurate.

One significant challenge was addressing inconsistencies in the language model's outputs. The voting mechanism effectively mitigated this issue by balancing computational efficiency with classification accuracy. Gradual distinctions between categories, such as *scientific* and *repo* domains (e.g., `nature.com` classified as *scientific*), were preserved in the dataset to maintain granularity for subsequent analysis.

For further analysis, especially in the context of factual consistency, we focused primarily on the ten most prominent domains within each category. This targeted approach provides a clearer understanding of the specific characteristics and trends associated with key domains, enhancing the interpretability and relevance of our findings.

3 RESULTS

Within our analysis we are running classic data analytics tools with visualization and regression analysis against our data to answer our research questions.[4]

Looking at Figure 1 score distribution of posts closely follows a power law, a phenomenon frequently observed in social networks and graph-based applications. This indicates that a small number of posts accumulate a disproportionately large number of scores, while the majority receive significantly fewer. Furthermore, the data reveals that the popularity of posts, as measured by their scores, is strongly influenced by the category they belong to, with notable variations in average scores across different categories. Therefore, to address the skewness in the data, we will use log-transformed score variables in subsequent regression analyses.

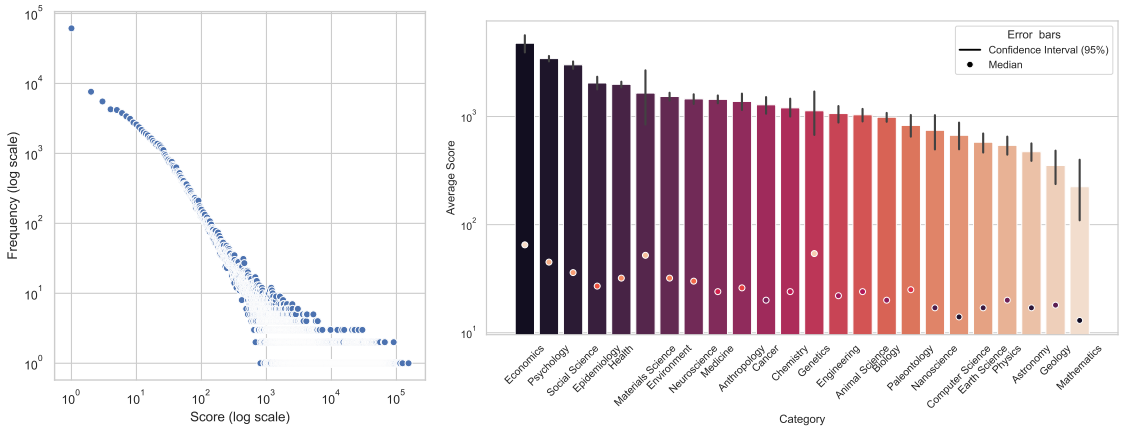

Fig. 1. Left: The distribution of scores follows a power law, with a small number of posts achieving disproportionately high scores. Right: The average popularity (score) of posts varies significantly across fields, indicating field-dependent engagement patterns.

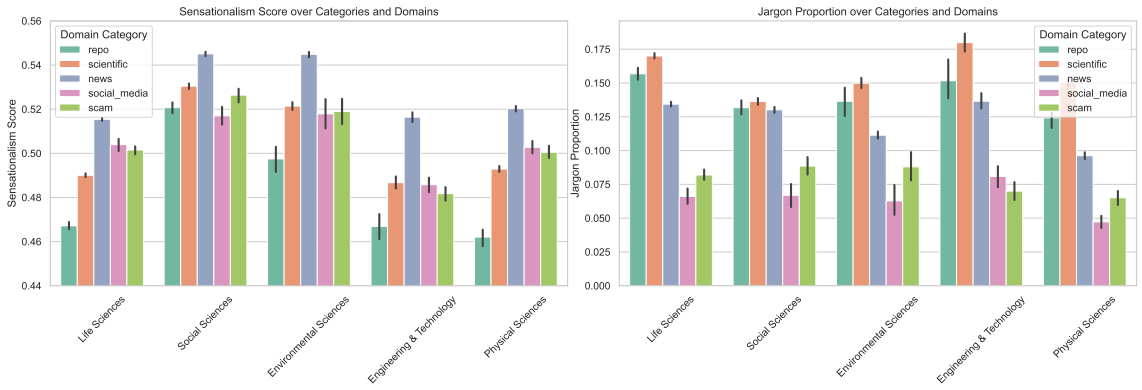

Fig. 2. Sensationalism and Jargon Proportions Across Categories and Domains

3.1 RQ1: Science Communication Quality by Domain

Our analysis (see Figure 2) reveals notable trends in sensationalism and jargon across various scientific categories and domain types. Sensationalism is lowest in repositories (repo), followed by scientific articles, with news showing slightly higher levels. Scam and social media domains fall in between scientific articles and news, suggesting moderate sensationalistic tendencies. Across scientific categories, sensationalism scores are relatively uniform, though Social Sciences and Environmental Sciences exhibit the highest levels of sensationalism, indicating a possible influence of topic-specific narratives.

[4]We are grouping scientific fields into broader top-level categories based on their link flair as follows:

- **Life Sciences:** Biology, Medicine, Health, Neuroscience, Cancer, Epidemiology, Genetics, Animal Science
- **Physical Sciences:** Chemistry, Physics, Earth Science, Geology, Astronomy, Nanoscience
- **Social Sciences:** Psychology, Social Science, Anthropology, Economics
- **Engineering & Technology:** Computer Science, Engineering, Materials Science, Mathematics
- **Environmental Sciences:** Environment, Paleontology

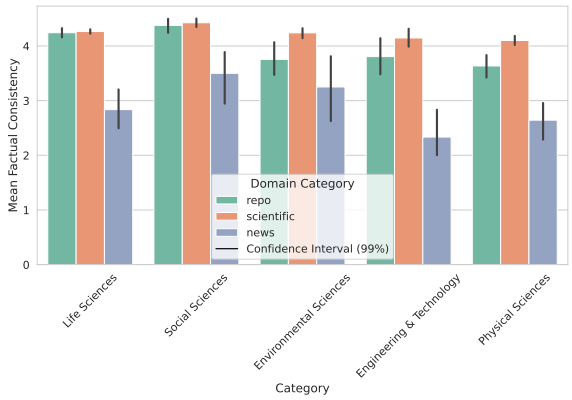

Fig. 3. Factual consistency scores across categories and domains

In terms of jargon, links to scientific articles are the most jargon-heavy, followed by repositories, and then news. This finding is somewhat surprising, as scientific, positioned between scientific articles and news in terms of content type, might be expected to exhibit intermediate levels of jargon but instead display more extreme trends. Scam and social media links show significantly lower jargon proportions, reflecting content that is less aligned with the field-specific language of the categories they are posted in. This discrepancy suggests that these domains often feature lower-quality content that diverges from the expected scientific discourse, which may influence their alignment with field-specific jargon dictionaries.

It is notable, that in both measurements, the general pattern of differences between the posted domains seems constant in all fields. This also holds, when splitting the data to more specific science fields and holds constant over years and seasonal changes.

Figure 3 displays the average factual consistency values by category and domain. While there do not appear to be substantial differences among fields with respect to mean factual consistency scores, it is clear that posts that link to news articles tend to be less factually consistent on average than posts that directly link to research papers or scientific news outlets.

Engagement. Looking at Figure 4, we observe that across all scientific fields, there is a consistent trend where news articles outperform direct links to original research hosted in repositories or scientific journals in terms of engagement. This performance disparity can be attributed to the characteristics of the news domain, which features less jargon and employs higher levels of sensationalism on `r/science`, as shown in the above sections. News articles are generally crafted to be more engaging and accessible, targeting a broader audience beyond domain experts, in contrast to research papers that are formatted for detailed and technical consumption by specialists. This alignment with audience preferences explains why news articles are more effective at fostering popular science communication in this setting.

Additionally, links to social media, including other subreddits or external platforms, perform exceptionally poorly in terms of engagement. This indicates that reposts, recycled content, or indirect sources fail to meet the expectations of the Reddit community, which values fresh and high-quality content. This result highlights the community's preference

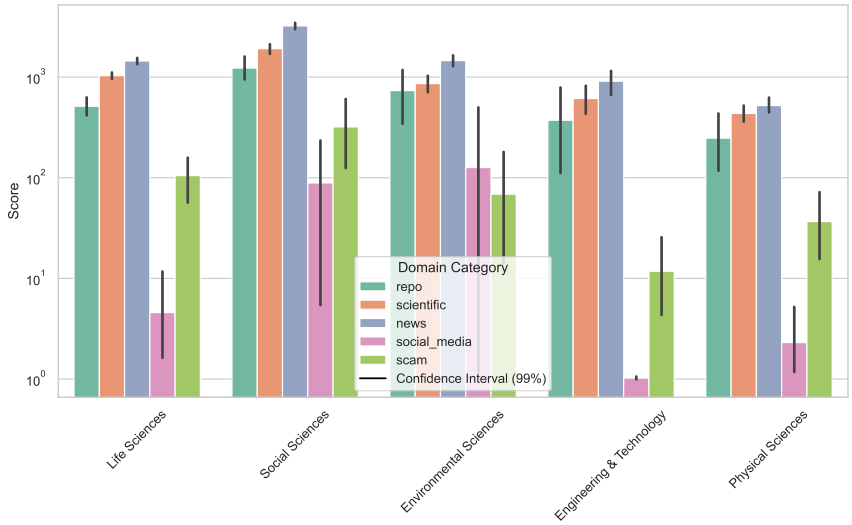

Fig. 4. Score (log-scale) across Science Categories and Domains

for originality and relevance, suggesting that social media reposts do not provide sufficient value to attract significant upvotes.

Lastly, while posts leading to domains classified as scams by our labeling system perform significantly worse than legitimate posts, they still manage to accumulate a notable amount of upvotes. This is particularly concerning given that the score variable accounts for downvotes as well. The ability of these scammy posts to garner substantial engagement despite community-driven downvotes suggests the involvement of bots or malicious actors actively upvoting such content to manipulate visibility and credibility. This underscores the challenges in moderating and removing low-quality or deceptive content from open online communities like Reddit. Addressing this issue requires robust detection mechanisms and community awareness to mitigate the influence of coordinated manipulation efforts.

3.2 RQ2: Relationships Between Science Communication Quality and Engagement

Our second research question is about the correlation between our metrics and the engagement measured in up- and down-votes on the subreddit. We will further dive into this in the following section.

Metric Correlation. Our three indicators of science communication quality are significantly associated with each other. The relationship between factual consistency and jargon is positive (Spearman's $r = 0.26$, $p < 0.001$), while the relationship between factual consistency and sensationalism is negative ($r = -0.43$, $p < 0.001$). The relationship between jargon and sensationalism is also negative ($r = -0.43$, $p < 0.001$). This indicates that when science communication quality improves with respect to the level of jargon in the post, on average, it tends to degrade with respect to the level of factual consistency and sensationalism in the post.

Jargon. The relationship between jargon proportion and post score reveals a nuanced pattern, as highlighted in figure 5 - Further details in the tables 3 and 1 in the Appendix. Overall, the regression analysis indicates a significant negative correlation between jargon proportion and score (-0.026^{***}. This suggests that as the proportion of jargon

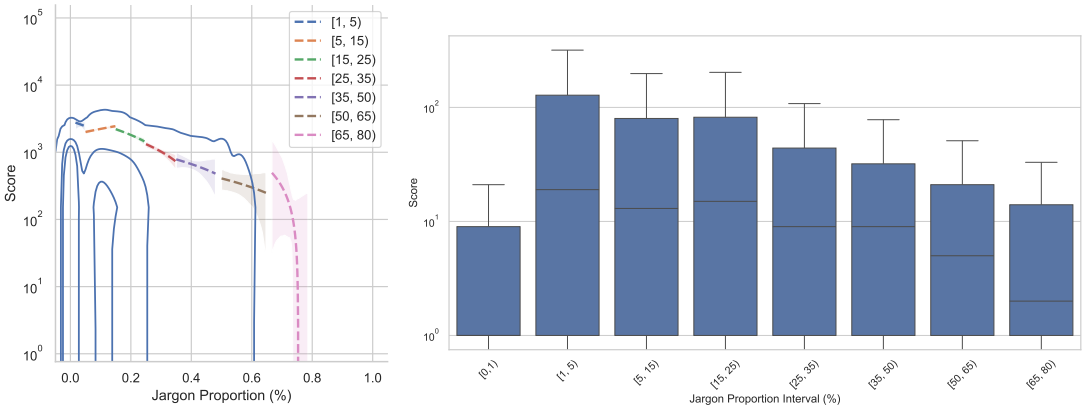

Fig. 5. Left: Score regression plotted on jargon intervals. Right: Boxplot of Scores for Each Interval of Jargon Proportion.

in a post increases, its overall engagement, as measured by the logarithm of the score, tends to decrease by 2.6% per percentage-point of jargon. This finding aligns with the notion that highly technical language may alienate a broader audience, resulting in lower engagement.

When examining specific jargon proportion intervals, we observe a more detailed pattern. Posts with a moderate level of jargon (5% – 15%) even exhibit a slight positive relationship with score (0.003 for log(score) and 35.550^{**} for raw score), indicating that a balanced use of jargon may resonate well with the audience. In contrast, higher levels of jargon (15% – 35%) show a significant negative impact on both log(score) (-0.045^{***} and -0.079^{***}) and raw score, emphasizing that overly technical content struggles to engage readers. Interestingly, the effect diminishes for extremely high jargon levels (> 50%), possibly due to fewer posts in this category or a niche audience actively engaging with such content.

These findings suggest that while some technical language can enhance engagement by signaling expertise, excessive jargon reduces accessibility and limits broader appeal. The analysis was filtered for posts with scores greater than 1 and corrected for potential confounders such as year, month, and domain or top category, ensuring robustness in the observed relationships. This highlights the importance of balancing technical accuracy with audience accessibility in science communication.

Sensationalism. The relationship between sensationalism and post score reveals a significant positive correlation, as detailed in Table 6 in the Appendix and illustrated in Figure 6. The overall linear regression indicates that a 1% increase in sensationalism is associated with a significant increase in engagement, with a coefficient of 0.103^{***}, reinforcing the idea that more sensationalist content tends to attract higher scores.

When analyzing sensationalism across different proportion intervals (Table 6), we observe a nuanced pattern. Moderate levels of sensationalism (35% – 60%) consistently exhibit a significant positive impact on engagement, with coefficients ranging between 0.093^{***} and 0.104^{***}. This indicates that content framed with a moderate degree of sensationalism resonates well with the audience, likely by drawing attention while still maintaining credibility.

However, the effect turns negative at higher levels of sensationalism (60% – 65%), with a coefficient of -0.105^{***}. This suggests that excessive sensationalism may detract from a post's credibility, resulting in lower scores. Beyond

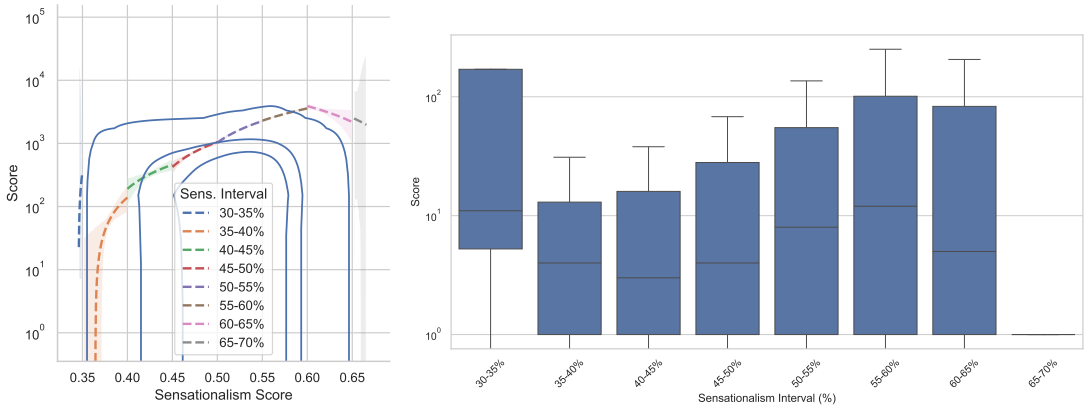

Fig. 6. Left: Score regression plotted on Sensationalism intervals. Right: Boxplot of Scores for Each Interval of Sensationalism Proportion.

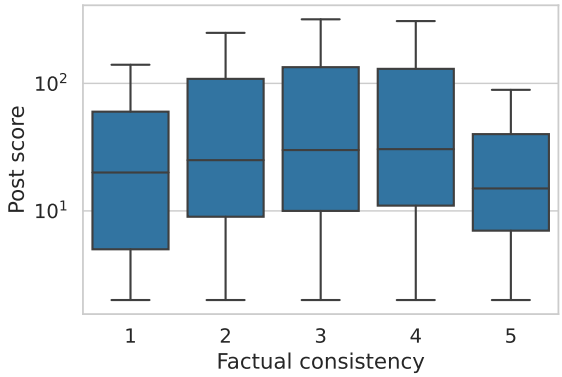

Fig. 7. Boxplot of engagement scores for each factual consistency score

this interval, the effect becomes more variable, possibly due to fewer observations or the niche appeal of extremely sensationalist content.

These results, corrected for year, month, and domain category, demonstrate a clear non-linear relationship: moderate sensationalism enhances engagement, while excessive sensationalism undermines it. Figure 6 visually supports this finding, showing a steady increase in scores with sensationalism until a tipping point is reached. This analysis highlights the importance of balancing attention-grabbing elements with credibility in science communication, as overly sensationalist content can alienate the audience. The findings were filtered to include only posts with scores greater than 1 to ensure that low-quality or irrelevant posts did not skew the results.

Factual Consistency. Overall, we find a negative relationship between factual consistency and engagement, shown in Table 7, where the regression coefficient for all posts is negative. Observing the trend in more detail, Figure 7 shows that the posts with the highest factual consistency scores receive the lowest engagement, but extremely low factual consistency values are also not highly upvoted, and differences are not substantial among posts with factual consistency

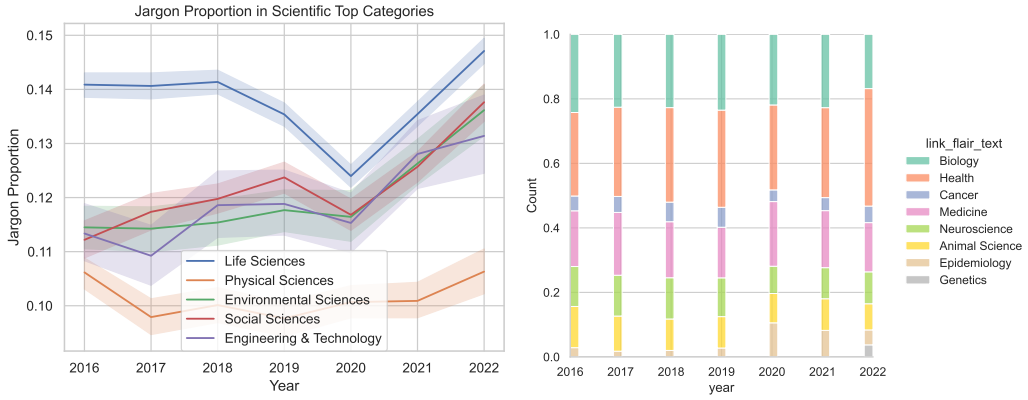

Fig. 8. Left: Jargon development over years, showing a dip in 2020, especially in the life science top-category. Right: Proportion of fields within Life Sciences

scores of 2, 3, or 4. Additionally, we observe different effects for posts with sources in different domains – for news, there is no significant correlation between engagement and factual consistency, but there is a negative relationship for scientific domains and repositories (Table 7).

3.3 RQ3: Temporal Trends in Science Communication

In the temporal analysis (RQ3), we are looking into how different metrics develop over different years and within a year. We link those to external events like the pandemic in 2020. Effects can be shown in the jargon metric. The score metric shows some temporal development, too, even when correcting for development over the growing community.

Jargon. The temporal analysis reveals a nuanced evolution of jargon proportions in `r/science` posts over time. While the overall use of jargon remains relatively stable across months within individual years, significant variations are observed when analyzed across different years. Notably, there is a marked drop in jargon usage during 2020 (−.8% points), coinciding with the COVID-19 pandemic. This dip can be attributed to a shift in the dominant topics being discussed, especially within the life sciences. There, epidemiology was trending (×4.5 compared to 2019) and increasing its proportion from 2.7% to 10.6%, which may inherently use less technical language compared to other scientific domains. But even within the epidemiology field, we see a drop of 45% in jargon. We interpret this as a more widespread interest in the field, which resulted in more non-niche people following and reporting on this topic. Additionally, experts on this topic were especially focused on making information about COVID-19 accessible to a broader public and might have communicated with less jargon. A last explanation could be that the topics discussed in epidemiology around the pandemic have established themselves in everyday speech usage and are therefore not considered jargon anymore, which is reflected in our metric.

Following this period, jargon levels begin to recover, with a notable increase by 2022 within science fields over all categories. This pattern suggests that external events and shifts in topical focus on the subreddit can influence the complexity of the language used in posts. The consistent variation in jargon across the years, despite relatively steady monthly trends, underscores the interplay between broader societal events and scientific communication norms within the subreddit.

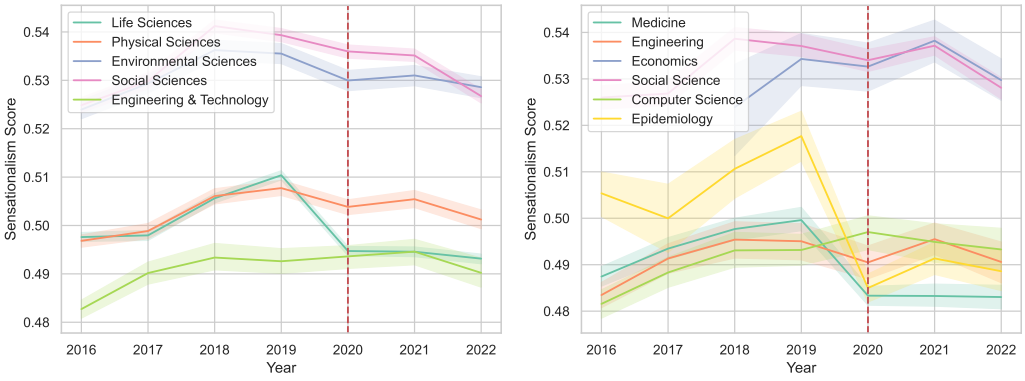

Fig. 9. Left: Jargon development over years grouped by relevant science fields. Right: Grouped by top science categories. Shaded regions represent 95% confidence intervals.

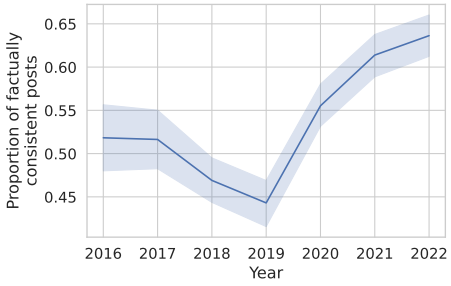

Fig. 10. Proportion of posts over time with a factual consistency score of 5. Shaded regions represent 95% confidence intervals.

Sensationalism. Similar developments can be observed in the Sensationalism metric (see figure 9), particularly during the pandemic period. This metric shows a notable dip in sensationalism within fields closely connected to the pandemic, such as medicine and epidemiology. We interpret this as a reflection of the immense public interest in these topics during that time, which may have reduced the perceived need for overly sensationalized posts to attract attention.

Additionally, the increased awareness of science communication challenges, especially regarding epidemiology, likely encouraged communicators to adopt a more measured tone. With a spotlight on the credibility and accuracy of scientific messaging during the pandemic, there may have been a conscious effort to avoid overly dramatic or misleading titles in favor of clear and responsible communication. This shift underscores how external societal factors can influence the style and framing of science communication on public platforms.

Factual Consistency. Figure 10 displays changes in the proportion of posts that receive the highest possible factual consistency score by year. Similar to how the the jargon and sensationalism metrics changed in 2020, the proportion of highly factually consistent posts increased dramatically in 2020. Due to the substantially smaller sample size of our dataset with factual consistency scores, the trends for individual fields are too noisy to obtain concrete conclusions.

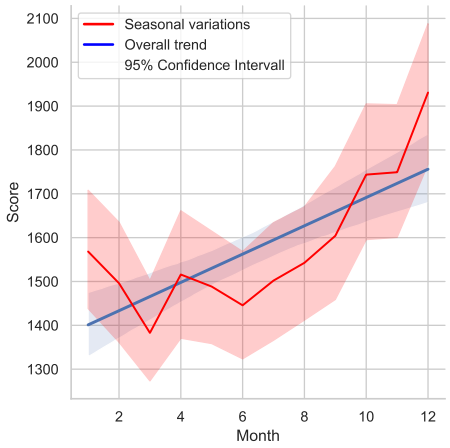

Fig. 11. Seasonal variations detrended by overall growth of `r/science`. Shaded regions represent 95% confidence intervals.

Given that a score of 5 as an indicator of "high factual consistency" is somewhat arbitrary, and we could have chosen other thresholds (e.g., a score of 4 or greater), we also repeated our analysis using different cutoffs for what determined high factual consistency and observed qualitatively similar trends.

Temporal development in Score. The temporal development in score over the months exhibits a clear increasing trend, as visualized in Figure 11. This upward trend likely reflects the inherent growth in community engagement and activity on the platform over time, which corresponds to the increasing number of users and posts contributing to higher aggregate scores.

To disentangle this baseline growth trend from changes in scores due to other factors, the detrended analysis reveals interesting insights. After removing the linear trend, the detrended scores show periodic fluctuations across the months, suggesting that while there is an overarching increase, specific months might see deviations in engagement or scoring patterns. These deviations are most likely driven by seasonal variations in activity, with increased involvement around the end of the year, when the predominantly English-speaking community in the Western Hemisphere experiences winter and holidays, contrasting with the typically slower engagement during warmer summer months. However, these fluctuations might also reflect shifts in the type of content being posted or its relevance to the community at specific times.

The confidence intervals, shown as shaded regions in the detrended plot, indicate that while the trend and fluctuations are statistically significant, there remains some uncertainty in the exact magnitude of the monthly variations. This analysis underscores the importance of accounting for temporal trends when interpreting score dynamics, ensuring that changes attributed to content or communication styles are not confounded by the overall growth in community size.

This proves more relevant for score than for metrics like jargon and sensationalism, as we did not observe a significant overarching growth trend in those metrics. Changes in jargon and sensationalism are more likely driven by shifts in topical focus, communication norms, and external societal factors rather than by platform-wide growth. Consequently, such metrics reflect qualitative changes in the nature of the content rather than quantitative increases in engagement, making detrending less critical for their interpretation.

4 DISCUSSION

Measuring jargon, sensationalism, and factual consistency offers significant insights into the nuances of science communication within online communities. Our results reveal nuanced trade-offs and trends in the type of language that drives engagement in these communities. Of particular value to the science communication community is our analysis of how these factors relate to engagement. However, it is important to recognize that these results reflect correlation, not causation, and are most applicable to community administrators and social media managers striving to enhance communication effectiveness. Nonetheless, individual users may also find these insights helpful for understanding the drivers of post engagement. We see for example that contributors shouldn't shy away from pointing out some novelty in their analysis, with a little sensationalism being beneficial. This does not only align with common sense but is also backed by the success of some more populistic news outlets and politicians. Furthermore, keeping it simple, reducing the jargon in broader public conversation will potentially outweigh the slight loss of detail in commentary.

Interestingly, all our correlations exhibit a U-shaped relationship. This illustrates the inherent challenges of science communication. Striking a balance is key: using specific terms without alienating readers with excessive jargon, sparking interest without resorting to exaggerated claims, and maintaining consistency without merely copying the titles of research papers. Our results need further exploration, ideally in collaboration with experts in psychology, sociology, and social media studies. While we can only give a broad interpretation, we hope they inspire additional research and experiments.

The moderate inter-annotator agreement among human evaluators underscores the subjective nature of some metrics, such as jargon and sensationalism, which in turn limits the precision of our models. Despite these constraints, we argue that the models still provide meaningful insights, capturing essential aspects of communication patterns within online science communities. Clear trends emerged in our quantitative and qualitative analyses, yet the inconsistencies in metric evaluation highlight the need for further refinement. Future studies could benefit from standardized annotation guidelines to improve reliability and robustness. We also suggest using multiple approaches to model the same metrics for improved reliability.

Implications for Moderation. Our findings suggest several ways to enhance moderation and optimize online science communities. By requiring evidence-based posts and statements, communities like `r/science` are uniquely positioned to enforce effective content moderation. Leveraging advancements in LLMs and implementing factual consistency checks toward referenced materials could significantly reduce the burden of human moderators. Additionally, the observed correlations in our study could inform predictive models to identify and flag potentially low-quality content, further streamlining moderation processes.

4.1 Future Work

While our analysis reveals important insights, it raises numerous questions and suggests promising directions for future research. We outline several areas which are of particular interest to us.

First, although we examined temporal trends in jargon and sensationalism, a more detailed analysis of how scientific narratives evolve on `r/science` could illuminate shifts in communication patterns over time. For instance, local events and global phenomena, such as the COVID-19 pandemic, significantly influence these dynamics and need further evaluation. Furthermore, exploring the evolution of specific topics within fields may reveal common patterns in how discussions trend and adapt over time. This could involve tracking specific scientific topics from their initial publication

through multiple posts and discussions, and examining how the community's understanding and communication styles adapt. Such analyses could guide strategies for introducing new scientific concepts to the public.

Second, our factual consistency metric could be expanded to address the broader science communication chain. Much of online discourse references research indirectly via scientific news articles, potentially amplifying inconsistencies relative to directly cited studies. It would be of interest to evaluate whether this intermediary step distorts intended findings compared to direct referencing. Additionally, analyzing Reddit post titles against full-text research papers, rather than abstracts, could deepen our understanding. Tackling this would necessitate solving challenges in large-scale document retrieval and semantic analysis, but the potential to pinpoint where and why misinterpretations occur is invaluable.

Impact of Moderation. Given dataset limitations and API access policies, our study could not analyze moderation's direct effects on Reddit posts. Nonetheless, moderation is a critical factor in shaping the quality of social media content, including on `r/science`. Evaluating the role of moderators through a systematic study of interventions, removed posts, and evolving rules—particularly during impactful periods like the pandemic could reveal strategies that effectively balance accessibility and communication quality. Working with methods from causal analysis could further increase the applicability of those findings.

These research directions aim to understand how scientific accuracy and public engagement can coexist. However, this pursuit assumes that engagement is the ultimate optimization goal. Is it? This fundamental question parallels broader discussions on what recommender systems should optimize for. What should we train our social media recommenders for? For a deeper dive into this topic we suggest "Six or Seven Things Social Media Can Do for Democracy" by Zuckerman [30].

5 CONCLUSION

We demonstrated how models from diverse domains can illuminate science communication patterns on social media, using `r/science`, the largest general science subreddit, as a case study. By defining and analyzing metrics for jargon, sensationalism, and factual consistency, we uncovered their distributions and correlations across scientific fields and their impact on engagement. The observed U-shaped effects suggest an optimal balance point for these metrics, offering guidance for science communicators, regulators, and moderators. Our findings also underscore the specificity of online science communication's target audience, with certain fields and communication styles being particularly successful.

Temporal analyses revealed the influence of global events, such as the COVID-19 pandemic, on science communication trends. By categorizing URLs linked in posts, we highlighted stark differences in our metrics across content types, reinforcing the necessity—and feasibility—of domain-specific content moderation.

Our work raises critical questions: How does science communication function in online communities? How does desired science communication conflict with engagement-driven social media optimization? And, fundamentally, what do we expect from science communication on social media? While our study only begins to address these vital questions, we hope it sparks further research and contributes to improving social media's role in fostering democratic discourse.

ACKNOWLEDGMENTS

We would like to thank Jonathan Stray for his insightful lecture and constructive criticism on our work. It was greatly appreciated! Lectures (or reading groups) like this make the academic experience and freedom in research at Berkeley worth the visit from Europe.

A DETAILS ON DATA ANALYSIS

In this appendix section we want to discuss the details about our data analysis. This includes how we chose our models, further insights that did not make it to the main article and justifications on ignoring outliers and filtering our data.

A.1 Jargon

As visible in figure 12 we have a strongly right-skewed distribution on the jargon, which is mostly caused by a lot of 0 values in our metric. This is not supposed to be, and when looking at this qualitatively, we noticed that many of the posts containing 0 jargon were low-quality spam posts. We therefore decided to filter out jargon with a value of 0 in any further analysis. Notably, this had a huge effect on our results. As the zero values on jargon mostly related to low quality or out of context post (spam), they were low in engagement, influencing the calculated relationship between jargon and engagement to be positive. However, removing posts with 0 jargon reveals a negative relationship between higher values of jargon and engagement.

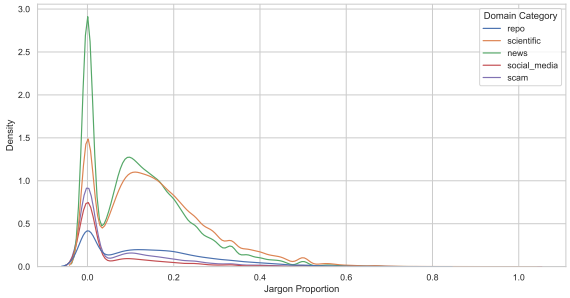

Fig. 12. Distribution over Jargon Proportion rating for different domain categories

Table 1. Log Score on Jargon Proportion Intervals

	Dependent variable:
	log(score)
Jargon Proportion (%)	-0.026^{***}
	(0.001)
Constant	3.474^{***}
	(0.047)
Observations	97,672
R^2	0.069
Adjusted R^2	0.069
Residual Std. Error	2.276 (df = 97649)
F Statistic	329.803^{***} (df = 22; 97649)
Note:	*p<0.1; **p<0.05; ***p<0.01
	Filtered for score > 1 and jargon proportion > 0.
	Corrected for year, month, and top category.

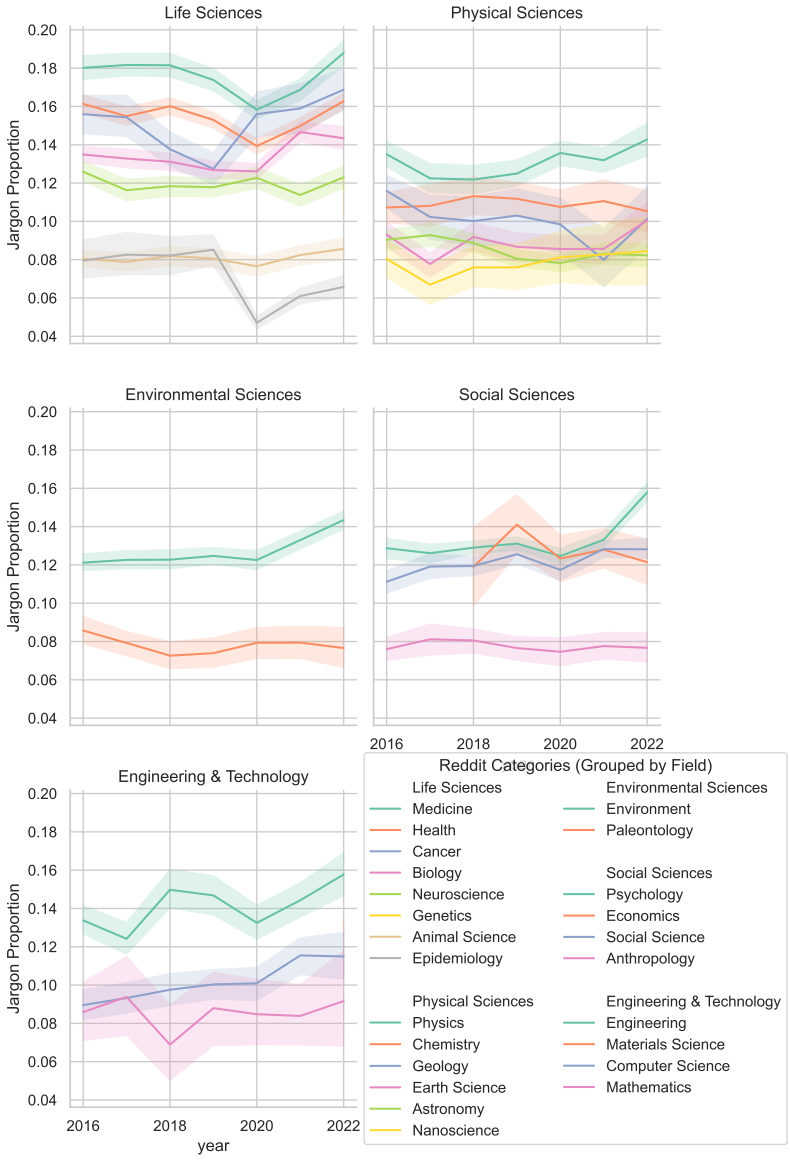

Fig. 13. Jargon proportion over years per science field grouped by top categories

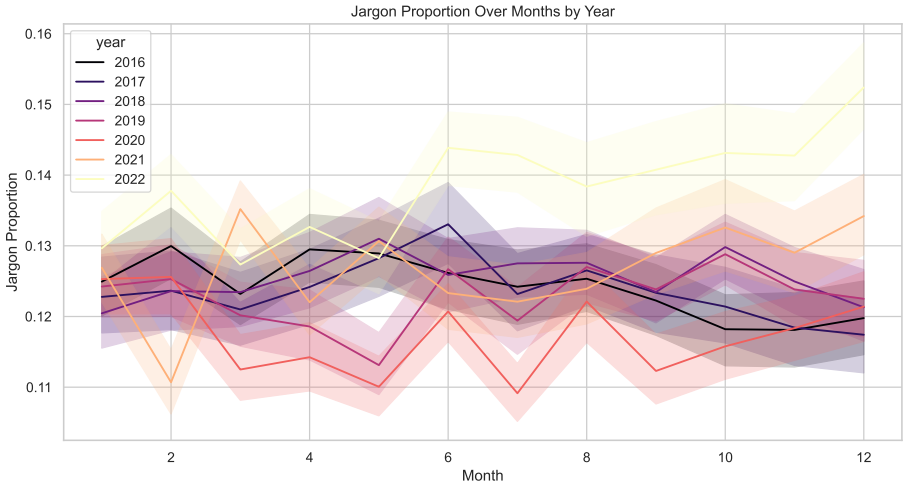

Fig. 14. Jargon Over Months by Year. This figure shows the variation in jargon proportion and sensationalism across different months for each year.

Table 2. Log(Score) on Jargon - Over different domain categories

	Dependent variable:			
	log(score)			
	Repo	Scientific	News	Combined
	(1)	(2)	(3)	(4)
Jargon Proportion (%)	−0.021***	−0.028***	−0.018***	−0.026***
	(0.002)	(0.001)	(0.001)	(0.001)
Constant	3.863***	4.321***	4.551***	4.431***
	(0.045)	(0.021)	(0.025)	(0.015)
Observations	8,225	41,138	38,987	88,350
R^2	0.015	0.021	0.005	0.015
Adjusted R^2	0.015	0.021	0.005	0.015
Residual Std. Error	2.045 (df = 8223)	2.202 (df = 41136)	2.507 (df = 38985)	2.341 (df = 88348)
F Statistic	123.176*** (df = 1; 8223)	879.049*** (df = 1; 41136)	203.674*** (df = 1; 38985)	1,312.398*** (df = 1; 88348)

Note: *p<0.1; **p<0.05; ***p<0.01
Base levels of jargon differ between the domains.
Filtered for score > 1 and repo, scientific, news domains.
Filtered for Jargon > 0.

Table 3. Log Score on Jargon Proportion Intervals

	Dependent variable:	
	log(score)	score
	(1)	(2)
Jargon x (0,5]	-0.018^{*}	-40.574
	(0.011)	(37.351)
Jargon x (5,15]	**0.003**	**35.550****
	(0.004)	**(14.044)**
Jargon x (15,25]	-0.045^{***}	-105.258^{***}
	(0.004)	(14.242)
Jargon x (25,35]	-0.079^{***}	-135.598^{***}
	(0.008)	(28.072)
Jargon x (35,50]	-0.018^{***}	-19.362
	(0.006)	(19.875)
Jargon x (50,65]	-0.002	-38.898
	(0.027)	(90.170)
Jargon x (65,80]	0.026	-70.473
	(0.053)	(179.551)
Constant		
Observations	97,672	97,672
R^2	0.073	0.018
Adjusted R^2	0.072	0.018
Residual Std. Error (df = 97634)	2.272	7,685.550
F Statistic (df = 37; 97634)	207.314^{***}	49.005^{***}

Note: *p<0.1; **p<0.05; ***p<0.01
Filtered for score > 1.
Repo, scientific, news, social media and scam domains.
Corrected for year, month, and domain category.

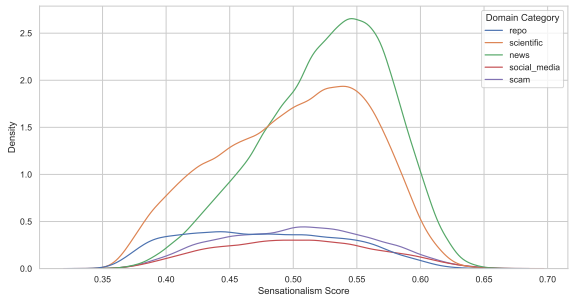

Fig. 15. Distribution over Sensationalism Score rating for different domain categories

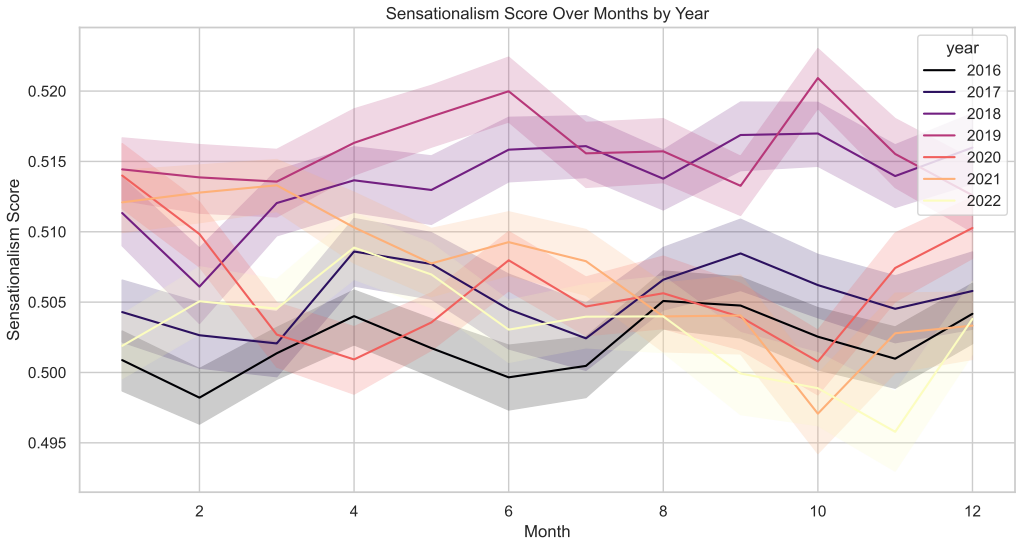

Fig. 16. Sensationalism Over Months by Year. This figure shows the variation in jargon proportion and sensationalism across different months for each year.

A.2 Sensationalism

We justify filtering for scores above 1 by recognizing that a single upvote can often be self-generated by the creator of the post, making posts with scores of 0 or 1 less reliable indicators of quality. Posts without meaningful engagement are more likely to be irrelevant or of insufficient quality, which makes them unsuitable for our analysis. By focusing on posts with scores greater than 1, we ensure that our dataset reflects content that has been positively received by at least one other community member.

The regression analysis (see table 4) supports this approach, demonstrating that excluding low-scoring posts does not fundamentally alter the results. Instead, it enhances the precision of the predictions by reducing noise from outliers and irrelevant data. This refinement is particularly evident in the increased explanatory power of the model, as shown by the higher R^2 values when applying the score > 1 filter. Thus, for the remainder of our analysis, we adopt this filter to focus on meaningful and impactful content without compromising the robustness of our findings.

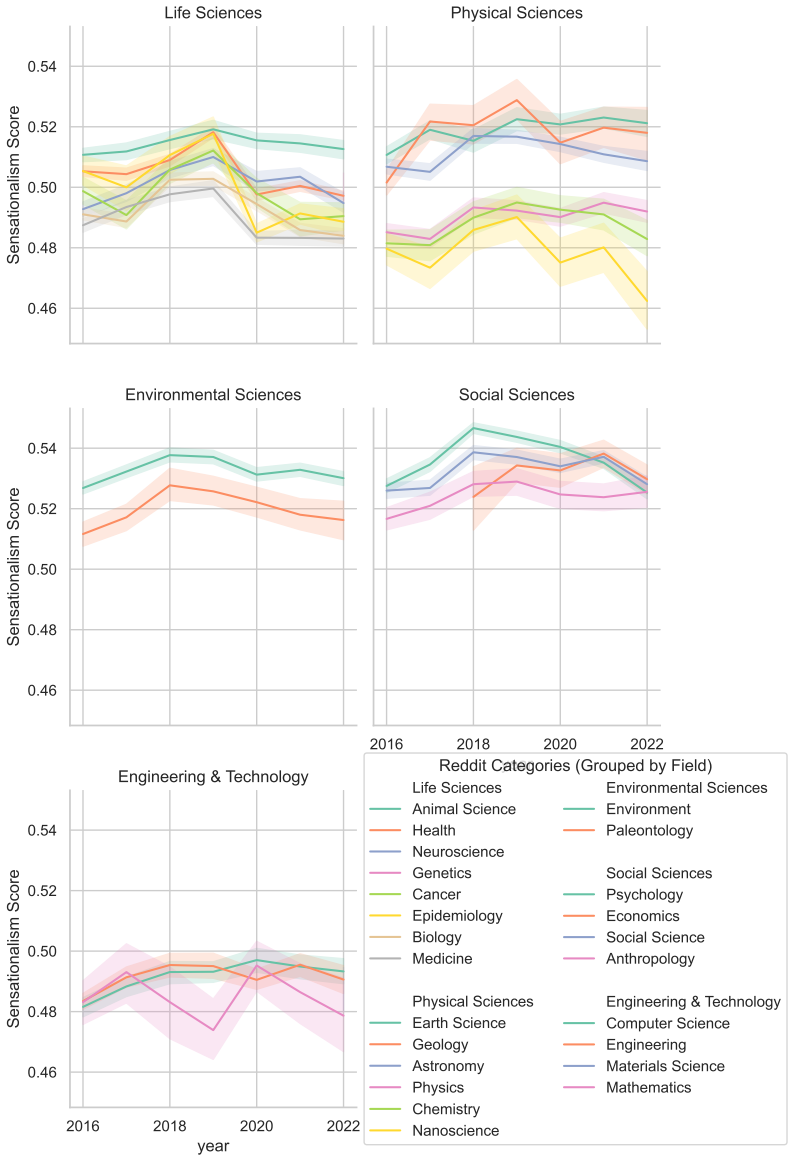

Fig. 17. Sensationalism score over years per science field grouped by top categories

Table 4. Log(Score) on Sensationalism - Justifying ignoring minimal score values

	Dependent variable:			
	log(score + 1)		log(score)	
	All	Score > 0	**Score > 1**	Score > 3
	(1)	(2)	**(3)**	(4)
Sens score (%)	0.076***	0.087***	**0.103*** **	0.104***
	(0.001)	(0.001)	**(0.001)**	(0.001)
Constant	−1.267***	−1.949***	**−1.539*** **	−1.251***
	(0.046)	(0.052)	**(0.057)**	(0.058)
Observations	197,622	184,025	122,655	109,513
R^2	0.035	0.039	0.066	0.073
Adjusted R^2	0.035	0.039	0.066	0.073
Res. Std. Error	2.318 (df = 197620)	2.520 (df = 184023)	2.222 (df = 122653)	2.113 (df = 109511)
F Statistic	7,094.507*** (df = 1; 197620)	7,416.523*** (df = 1; 184023)	8,689.553*** (df = 1; 122653)	8,594.380*** (df = 1; 109511)

Note: *p<0.1; **p<0.05; ***p<0.01
Excluding the posts that did not perform at all -> Effect is similar, but explanatory power higher.
Using score > 1 filter for the rest of the analysis.

Table 5. Log(Score) on Sensationlism - Over different domain categories

	Dependent variable:			
			log(score)	
	Repo	Scientific	News	Combined
	(1)	(2)	(3)	(4)
Sens score (%)	0.105***	0.112***	0.082***	0.105***
	(0.003)	(0.002)	(0.002)	(0.001)
Constant	−1.789***	−1.960***	−0.293**	−1.547***
	(0.141)	(0.081)	(0.115)	(0.060)
Observations	10,780	48,111	49,630	108,521
R^2	0.107	0.092	0.028	0.068
Adjusted R^2	0.107	0.092	0.028	0.068
Residual Std. Error	1.865 (df = 10778)	2.078 (df = 48109)	2.414 (df = 49628)	2.222 (df = 108519)
F Statistic	1,292.722*** (df = 1; 10778)	4,871.998*** (df = 1; 48109)	1,434.395*** (df = 1; 49628)	7,921.937*** (df = 1; 108519)

Note: *p<0.1; **p<0.05; ***p<0.01
Base levels of sensationalism differ between the domains.
Filtered for score > 1 and repo, scientific, news domains.

Table 6. Log(Score) on Sensationalism within Proportion Intervals

	Dependent variable:
	log(score)
Sensationalism x (30,35]	7.794
	(7.376)
Sensationalism x (35,40]	0.106***
	(0.031)
Sensationalism x (40,45]	0.094***
	(0.012)
Sensationalism x (45,50]	0.102***
	(0.009)
Sensationalism x (50,55]	0.093***
	(0.008)
Sensationalism x (55,60]	0.104***
	(0.009)
Sensationalism x (60,65]	**−0.105*** **
	(0.036)
Sensationalism x (65,70]	−1.754
	(1.503)
Constant	−268.312
	(256.966)
Observations	122,655
R^2	0.122
Adjusted R^2	0.122
Residual Std. Error	2.154 (df = 122615)
F Statistic	438.636*** (df = 39; 122615)
Note:	*p<0.1; **p<0.05; ***p<0.01
	Filtered for score > 1.
	Corrected for year, month, and domain category.

Table 7. Log(score) on factual consistency – over different domain categories.

	Dependent variable:			
	log(score)			
	Repo	Scientific	News	Combined
	(1)	(2)	(3)	(4)
fc_score	−0.219***	−0.261***	0.100	−0.257***
	(0.038)	(0.021)	(0.098)	(0.017)
Constant	4.131***	4.442***	3.818***	4.411***
	(0.167)	(0.091)	(0.296)	(0.075)
Observations	2,409	7,870	427	10,706
R^2	0.014	0.020	0.002	0.020
Adjusted R^2	0.013	0.019	0.0001	0.020
Residual Std. Error	1.959 (df = 2407)	1.946 (df = 7868)	2.298 (df = 425)	1.968 (df = 10704)
F Statistic	33.698*** (df = 1; 2407)	157.076*** (df = 1; 7868)	1.042 (df = 1; 425)	219.176*** (df = 1; 10704)

Note: *p<0.1; **p<0.05; ***p<0.01

Base levels of factual consistency differ between the domains.

Filtered for score > 1 and repo, scientific, news domains.

B LABELING OF CATEGORIES FROM JARGON-DICTIONARY TO SUBREDDIT-CATEGORIES

The categories in the jargon-dictionary from Lucy et al [17] included a lot of categories that were unfamiliar to us. As we did not know enough about each of the categories to be able to accurately label them into the correct `r/science`-categories, we used `ChatGPT o1-preview` with the following prompt to label them for us:

```
For each of the following categories that we call science-categories (Animal Science,
    ↪ Anthropology,Astronomy,Biology,Cancer,Chemistry,Computer Science,Earth Science,
    ↪ Economics,Engineering,Environment,Epidemiology,Genetics,Geology,Health,Materials
    ↪ Science,Mathematics,Medicine,Nanoscience,Neuroscience,Paleontology,Physics,Psychology,
    ↪ Social Science) I want you to find all subfields and categories from the list below
    ↪ that are related to each science-category. A subfield/category from the list below can
    ↪ be related to more than one science-category. The format should be a json dictionary
    ↪ where the science-category is the key, and the belonging subfields/categories are the
    ↪ values. e.g: { "Computer Science": [ "Software Engineering"] }.

These are the categories I want you to search through:
<List of the categories from Lucy et al., removed here for readability>
```

We performed an informal sanity check of some categories that we were familiar with to validate the correctness. The output of this prompt was a json file containing all the subreddit-categories at the top-level with all the associated categories from jargon-dictionary as lists of those subreddit-categories.

C SCRAPING

To obtain abstracts of research papers posted to `r/science`, we focused on attributes of the URLs that could be used as input into the Semantic Scholar API [14], which accepts DOIs and Pubmed IDs. We parse Pubmed IDs directly from the URLs of papers hosted on the ncbi.nlm.nih.gov website. Similarly, we search for all URLs in our dataset which directly and obviously include the doi or pubmed directly in the URL — 5757 abstracts almost exclusively from repo domains. Finally, we infer the DOI of all papers hosted by nature.com, which contains the unique identifier portion of the DOI in each paper link on the website (7086 abstracts). To compare the factual consistency of scientific websites and repositories with news articles, we also created a custom scraper to extract DOIs from articles from ScienceAlert.com, one of the most frequently posted news websites on `r/science` (570 abstracts). The scraper checked the HTML content of the page for the most frequently occurring DOI, which links to the original research paper in almost all cases due to the consistent format of news articles from this website. Altogether, we were able to collect 13,147 abstracts.

Identifier Extraction. The primary extraction layer focused on identifying scientific paper identifiers through pattern matching, supporting Digital Object Identifiers (DOIs), PubMed IDs (PMIDs), and PubMed Central IDs (PMCIDs). We optimized extraction patterns for repository domains like PubMed and arXiv, where standardized identifiers are consistently formatted. For Nature articles, we developed specialized extraction rules to handle their domain-specific URL structures.

Content Retrieval. For direct scientific sources, we employed BeautifulSoup for HTML parsing of abstract content, with customized rules for Nature's article structure. EurekAlert articles required separate parsing logic due to their

news-oriented format. When direct extraction failed, we leveraged the Semantic Scholar API as a fallback mechanism, particularly effective for repository content with valid DOIs.

Results and Performance. From our initial dataset of 56,507 submissions, we successfully extracted identifiers from 8,299 URLs (14.7% success rate). Through subsequent API queries, we obtained complete metadata for 5,236 papers. Nature articles showed a 68% successful extraction rate, repository DOIs achieved 72%, while EurekAlert articles managed 54%. This dataset forms the foundation for analyzing relationships between original scientific content and its public representation.

Evaluation. Domain-specific success rates varied by content type, with repository domains showing the highest reliability due to standardized structures. Nature articles provided consistent results through specialized parsing rules. News articles, particularly from EurekAlert, presented greater challenges due to varying formats but offered valuable comparative data for analyzing science communication patterns.

REFERENCES

[1] 2024. sensationalism. https://en.wiktionary.org/w/index.php?title=sensationalism&oldid=82788392 Page Version ID: 82788392.

[2] Tal August, Dallas Card, Gary Hsieh, Noah A Smith, and Katharina Reinecke. 2020. Explain like I am a Scientist: The Linguistic Barriers of Entry to r/science. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–12.

[3] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The Pushshift Reddit Dataset. *Proceedings of the International AAAI Conference on Web and Social Media* 14 (May 2020), 830–839. https://doi.org/10.1609/icwsm.v14i1.7347

[4] Neil I. Bower, Katarzyna Koltowska, Cathy Pichol-Thievend, Isaac Virshup, Scott Paterson, Anne K. Lagendijk, Weili Wang, Benjamin W. Lindsey, Stephen J. Bent, Sungmin Baek, Maria Rondon-Galeano, Daniel G. Hurley, Naoki Mochizuki, Cas Simons, Mathias Francois, Christine A. Wells, Jan Kaslin, and Benjamin M. Hogan. 2017. Mural lymphatic endothelial cells regulate meningeal angiogenesis in the zebrafish. *Nature Neuroscience* 20, 6 (June 2017), 774–783. https://doi.org/10.1038/nn.4558 Publisher: Nature Publishing Group.

[5] Dominique Brossard. 2013. New media landscapes and the science information consumer. *Proceedings of the National Academy of Sciences* 110, supplement_3 (2013), 14096–14101.

[6] Greg Clark, Josh Russell, Peter Enyeart, Brant Gracia, Aimee Wessel, Inga Jarmoskaite, Damon Polioudakis, Yoel Stuart, Tony Gonzalez, Al MacKrell, et al. 2016. Science educational outreach programs that benefit students and scientists. *PLoS Biology* 14, 2 (2016), e1002368.

[7] 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 (March 2022), A06. https://doi.org/10.22323/2.21010206 Publisher: SISSA Medialab srl.

[8] Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating Summarization Evaluation. *Transactions of the Association for Computational Linguistics* 9 (April 2021), 391–409. https://doi.org/10.1162/tacl_a_00373

[9] Jennifer Golbeck. 2020. Optimizing for engagement can be harmful. There are alternatives. *IEEE Intelligent Systems* 35, 4 (2020), 117–118.

[10] Yue Guo, Tal August, Gondy Leroy, Trevor Cohen, and Lucy Lu Wang. 2024. APPLS: Evaluating Evaluation Metrics for Plain Language Summarization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 9194–9211. https://doi.org/10.18653/v1/2024.emnlp-main.519

[11] Manoel Horta Ribeiro, Kristina Gligoric, and Robert West. 2019. Message distortion in information cascades. In *The World Wide Web Conference*. 681–692.

[12] Ridley Jones, Lucas Colusso, Katharina Reinecke, and Gary Hsieh. 2019. r/science: Challenges and Opportunities for Online Science Communication. https://www.semanticscholar.org/paper/r-science%3A-Challenges-and-Opportunities-for-Online-Jones-Colusso/c7ddc9213dddd3863feb66accde3a5047230970d

[13] M. G. Kendall. 1938. A New Measure of Rank Correlation. *Biometrika* 30, 1/2 (1938), 81–93. https://doi.org/10.2307/2332226 Publisher: [Oxford University Press, Biometrika Trust].

[14] Rodney Michael Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, Miles Crawford, Doug Downey, Jason Dunkelberger, Oren Etzioni, Rob Evans, Sergey Feldman, Joseph Gorney, David W. Graham, F.Q. Hu, Regan Huff, Daniel King, Sebastian Kohlmeier, Bailey Kuehl, Michael Langan, Daniel Lin, Haokun Liu, Kyle Lo, Jaron Lochner, Kelsey MacMillan, Tyler C. Murray, Christopher Newell, Smita R Rao, Shaurya Rohatgi, Paul Sayre, Zejiang Shen, Amanpreet Singh, Luca Soldaini, Shivashankar Subramanian, A. Tanaka, Alex D Wade, Linda M. Wagner, Lucy Lu Wang, Christopher Wilhelm, Caroline Wu, Jiangjiang Yang, Angele Zamarron, Madeleine van Zuylen, and Daniel S. Weld. 2023. The Semantic Scholar Open Data Platform. *ArXiv* abs/2301.10140 (2023). https://api.semanticscholar.org/CorpusID:256194545

[15] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv e-prints* (2019), arXiv–1907.

[16] Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The Semantic Scholar Open Research Corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 4969–4983. https://doi.org/10.18653/v1/2020.acl-main.447

[17] Li Lucy, Jesse Dodge, David Bamman, and Katherine A. Keith. 2023. Words as Gatekeepers: Measuring Discipline-specific Terms and Meanings in Scholarly Publications. https://doi.org/10.48550/arXiv.2212.09676 arXiv:2212.09676 [cs].

[18] Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2024. Factual consistency evaluation of summarization in the Era of large language models. *Expert Systems with Applications* 254 (Nov. 2024), 124456. https://doi.org/10.1016/j.eswa.2024.124456

[19] Alex Moehring. 2023. Personalized Rankings and User Engagement: An Empirical Evaluation of the Reddit News Feed. *Open Science Foundation Preprint* (2023).

[20] MM Mukaka. 2012. A guide to appropriate use of Correlation coefficient in medical research. *Malawi Medical Journal : The Journal of Medical Association of Malawi* 24, 3 (Sept. 2012), 69–71. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3576830/

[21] Simon Pollett and Caitlin Rivers. 2020. Social media and the new world of scientific communication during the COVID-19 pandemic. *Clinical Infectious Diseases* 71, 16 (2020), 2184–2186.

[22] Tzipora Rakedzon, Elad Segev, Noam Chapnik, Roy Yosef, and Ayelet Baram-Tsabari. 2017. Automatic jargon identifier for scientists engaging with the public and science communication educators. *PLOS ONE* 12, 8 (Aug. 2017), e0181742. https://doi.org/10.1371/journal.pone.0181742

[23] Debbie Treise and Michael F Weigold. 2002. Advancing science communication: A survey of science communicators. *Science Communication* 23, 3 (2002), 310–322.

[24] C Emmanuel Uzuegbunam and S Udeze. 2013. Sensationalism in the media: the right to sell or the right to tell. *Journal of Communication and Media Research* 5, 1 (2013), 69–78.

[25] Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is ChatGPT a Good NLG Evaluator? A Preliminary Study. https://arxiv.org/abs/2303.04048v3

[26] Spencer Williams and Gary Hsieh. 2021. The effects of user comments on science news engagement. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–29.

[27] Shannon D. Willoughby, Keith Johnson, and Leila Sterman. 2020. Quantifying scientific jargon. *Public Understanding of Science* 29, 6 (Aug. 2020), 634–643. https://doi.org/10.1177/0963662520937436

[28] Amelie Wuehrl, Dustin Wright, Roman Klinger, and Isabelle Augenstein. 2024. Understanding Fine-grained Distortions in Reports of Scientific Findings. In *Findings of the Association for Computational Linguistics ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand and virtual meeting, 6175–6191. https://doi.org/10.18653/v1/2024.findings-acl.369

[29] Jerrold H. Zar. 2005. Spearman Rank Correlation. In *Encyclopedia of Biostatistics*. John Wiley & Sons, Ltd. https://doi.org/10.1002/0470011815.b2a15150 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/0470011815.b2a15150.

[30] Ethan Zuckerman. 2018. Six or Seven Things Social Media Can Do For Democracy. https://medium.com/trust-media-and-democracy/six-or-seven-things-social-media-can-do-for-democracy-66cee083b91a