ARL-TR-9343 • JAN 2022

DEVCOM ARMY RESEARCH LABORATORY

You Can't Quarantine the Truth: Lessons Learned in Logical Fallacy Annotation of an Infodemic

by Claire Bonial, Taylor A Hudson, Austin Blodgett, Stephanie M Lukin, Jeffrey Micher, Douglas Summers-Stay, Peter Sutor, and Clare R Voss

Approved for public release; distribution is unlimited.

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

ARL-TR-9343 • JAN 2022

You Can't Quarantine the Truth: Lessons Learned in Logical Fallacy Annotation of an Infodemic

by Claire Bonial, Stephanie M Lukin, Jeffrey Micher, Douglas Summers-Stay, and Clare R Voss
Computational and Information Sciences Directorate, DEVCOM Army Research Laboratory

Taylor A Hudson
Oak Ridge Associated Universities

Austin Blodgett
Institute for Human & Machine Cognition

Peter Sutor
University of Maryland

Approved for public release; distribution is unlimited.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED (From - To)
January 2022	Technical Report	April 2021–June 2021

4. TITLE AND SUBTITLE
You Can't Quarantine the Truth: Lessons Learned in Logical Fallacy Annotation of an Infodemic

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Claire Bonial, Taylor A Hudson, Austin Blodgett, Stephanie M Lukin, Jeffrey Micher, Douglas Summers-Stay, Peter Sutor, and Clare R Voss

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
DEVCOM Army Research Laboratory
ATTN: FCDD-RLC-IT
2800 Powder Mill Rd, Adelphi, MD 20783

8. PERFORMING ORGANIZATION REPORT NUMBER
ARL-TR-9343

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
Contact author email: <claire.n.bonial.civ@army.mil>

14. ABSTRACT
Given the current COVID-19 infodemic that crosses multiple genres of text, we posit that flagging "potentially problematic information" (PPI) retrieved by a semantic search system will be critical to combating mis- or disinformation. This report describes the construction of a COVID-19 corpus and a two-level annotation of logical fallacies in these documents, supplemented with inter-annotator agreement results over two development phases. We also report a preliminary assessment of the corpus for training and testing a machine learning algorithm (Pattern-Exploiting Training) for fallacy detection and recognition. The agreement results and system performance underscore the challenging nature of this annotation task. We propose targeted improvements for fallacy annotation and conclude that a practical implementation may be to report a document's overall fallacy rate as a measure of its credibility.

15. SUBJECT TERMS
report, misinformation, logical fallacies, natural language processing

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Claire Bonial
Unclassified	Unclassified	Unclassified	UU	30	19b. TELEPHONE NUMBER (Include area code) 301-394-1431

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18

Contents

List of Figures

List of Tables

1. Introduction

With the outbreak of the COVID-19 pandemic, a parallel "infodemic" has emerged, defined by the World Health Organization as "too much information including false or misleading information in digital and physical environments during a disease outbreak".[1] Thus, we seek to identify misleading or "potentially problematic information" (PPI) in our efforts to develop a semantic search framework for COVID research, in which users can pose unconstrained natural language queries and receive a range of answers with explanations of their relevance. PPI can take many forms and has become a quickly growing area of natural language processing (NLP) research; here, we focus on an approach for annotating and automatically identifying logical fallacies. This area of research is particularly challenging because logical fallacies can be subtly encoded in the structure of a document across multiple sentences, making it difficult for human annotators or systems to recognize fallacies, yet also further motivating the need for assistance in finding such hidden, powerful sources of misinformation.

We base our fallacy annotation schema on that of Habernal et al.'s *Argotario*.[2,3] This allows us to leverage the authors' existing data set into a diverse training corpus for automatic identification of fallacies. Furthermore, we are able to explore differences in the realization of fallacies in the domain of COVID-19 documents, which is thought to be particularly prone to misinformation. While authors of the original corpus leverage a gamification approach to crowdsourcing the fallacy judgments, we instead follow a more traditional annotation approach. Two annotators and authors of this report, each with formal linguistic training, annotate the sentences in a set of documents on six COVID-19 topics that are known to be particularly rife with misinformation. Our report title includes a sentence from the corpus. Using our resulting corpus of 26 documents, we use the Pattern-Exploiting Training (PET) procedure[4] to train and evaluate automatic identification of the fallacies.

Our contributions here include the extension and evaluation of the *Argotario* annotation schema of five fallacies within the domain of COVID-related texts (Section 3), a novel data set of COVID texts annotated with fallacies (Section 4), and analyses of our annotated corpus (Section 5 and Section 6). We further conduct a preliminary evaluation of the PET approach for automatically identifying fallacies (Section 7).

2. Background

The exploration of fallacies of focus in this report fits into a broader research project on the development of an information search system, distinct from typical question-answering systems in that users are able to present a full, unconstrained natural language question (as opposed to restricting their search to keywords). The goal is not to return a single answer in a one-off interaction, but rather to encourage an ongoing interaction between user and system to forage for the range of relevant answers, where these may differ with respect to focus, genre, as well as truth value and mis- or disinformation status. The ability of a system to detect and identify potential fallacies becomes paramount in this envisioned interaction, as seen in Fig. 1, where a system may answer a user's question—"Do I need to sanitize my mask?"—with both the answer as identified in a document (i.e., "Here's an article claiming the importance of proper mask care".) as well as a warning alerting the user to potential fallacies present, supplementing the sentences in the retrieved document. This exchange portrays our longer term vision of how question-answering, information foraging, and mis- or disinformation detection can be unified under one framework.

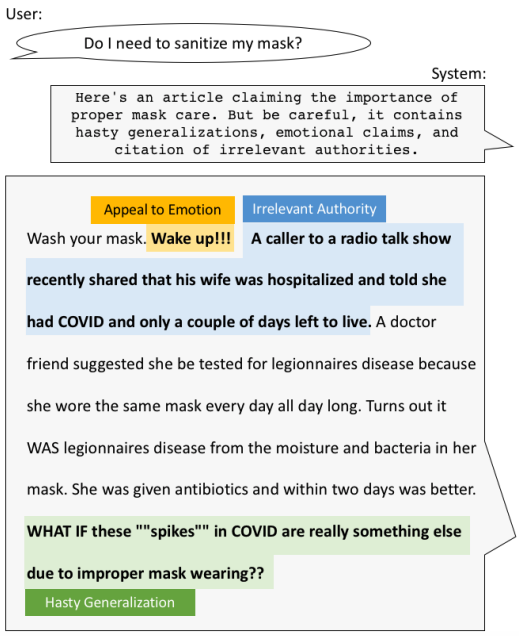

Fig. 1. An envisioned exchange between a user and a system where the user's question is answered through dialogue, document retrieval, and document annotation (e.g., fallacy annotation)

To support fallacy detection, we begin by annotating logical fallacies in our domain

of interest: scientific papers, general news, and talk radio, as well as health care sites and social media posts relating to COVID-19. We base our annotation schema on that of Habernal et al.,[2] which focuses on five logical fallacies: *Ad Hominem, Appeal to Emotion, Red Herring, Hasty Generalization,* and *Irrelevant Authority*. The authors crowdsource the *Argotario* corpus of 1,344 snippets of English text with fallacy annotations. They use a gamification approach to data collection, in which the first player writes a claim (ranging from a short sentence to a couple of sentences) and then indicates the "intended fallacy"—which of the five fallacies are present in their claim, or if none of those five fallacies are present. A second player is then presented with the first player's claim and attempts to guess the first player's intended fallacy. The majority label given by second players is termed the "voted fallacy". The data set includes information for each claim regarding the intended and voted fallacy, with an indication of how many second players voted. Instances that have five or more votes for a particular fallacy are added to a gold standard subset.

3. Fallacy Annotation Schema

While effective for collecting data in a fun and educational manner, the annotation procedure of the *Argotario* corpus, outside of the gold standard subset of the corpus, has the potential for annotation errors. Authors of the claims may not clearly recognize or label the intended fallacy in their own writing. Thus, in contrast to their crowdsourcing approach, we elected to extend fallacy annotation using the same five fallacies and their definitions, but relied on two linguistics-trained annotators to identify fallacies in documents related to COVID-19.

This difference in procedure allows us to annotate fallacies in existing scientific journal papers, news reports, as well as social media posts about COVID-19. However, this change in procedure poses a new challenge; as we are now annotating fallacies "in the wild", we encounter a mixture of well-hidden fallacies that serve a particular author's agenda, as well as fallacies that may be entirely unintentional as a result of faulty reasoning. In both cases, the fallacy may only be clear in the broader context of the document, given the primary thesis or claim being made, the evidence presented (often in multiple sentences across the document) to support that claim, and how these claims and pieces of evidence relate to the social and cultural context, including implicit assumptions and knowledge that would be clear to readers of a similar socio-cultural background.

On a practical level, we also had to determine the unit of annotation for this task uniquely for our procedure. In the *Argotario* setup, the unit of annotation is the claim authored by the first player, who may choose to express the claim in a short sentence or several sentences. When translating this task to existing documents, some of the challenges mentioned previously make determining the appropriate unit of annotation very difficult—a logical fallacy can be detectable within a single word (especially in the case of *Appeal to Emotion*), a clause or sentence, or, somewhat more commonly, as part of a set of sentences reflecting steps in reasoning where the fallacy is one step. Pilot experimentation demonstrated that leaving the unit open to interpretation resulted in vast disagreement in what should anchor the fallacy. For our purposes, we opted to separate each document into sentences and have both annotators annotate each and every sentence. The same sentence could be listed twice with different fallacies, where there were multiple fallacies exhibited in different parts of that sentence. Selecting this unit of annotation focuses the task and evaluation on the fallacy judgment, as opposed to the precise linguistic anchor of that fallacy, and also sets up our data nicely to serve as training prompts for PET.

Specifically, for each sentence of our document collection, annotators marked one of the following five choices, with labels and definitions adopted verbatim from the schema of Habernal et al.[2]; here, we have added examples and topics from our own corpus annotations:

1. Ad Hominem: The opponent attacks a person instead of arguing against the claims that the person has put forward. Example: *"It's just too convenient for vax-pimping scientists to claim that their precious vaccines don't work because not enough people are getting them"*. (Topic: General vaccine safety and efficacy)
2. Appeal to Emotion: This fallacy tries to arouse non-rational sentiments within the intended audience in order to persuade. Example: *"It is time for families to wake up to uncloaking of the new world order in its glory"*. (Topic: COVID-19 vaccine safety and efficacy)
3. Red Herring: This argument distracts attention away from the thesis that is supposed to be discussed. Example: *"Being a real scientist would be easy if it weren't for this 'needing evidence' stuff, just like being a professional golfer would be simple if it weren't for this 'having to put the ball in the hole thing".'* (Topic: SARS-CoV-2 virus origin)

4. Hasty Generalization: The argument uses a sample that is too small, or follows falsely from a sub-part to a composite or the other way around. Example: (Preceding sentences: *"They're not reporting the number of deaths per million. In other words, they're not reporting the survivability rate"*.) Annotation target: *"The answer here is don't mandate closures"*. (Topic: Herd immunity)
5. Irrelevant Authority: While the use of authorities in argumentative discourse is not fallacious inherently, appealing to authority can be fallacious if the authority is irrelevant to the discussed subject. Example: *"'Inside Edition' also lauded Biden, Mitt Romney, and Tom Cruise for double masking recently"*. (Topic: Mask safety and efficacy)

In our setup, as in the *Argotario* annotation procedure, a selection indicating "none" was a final annotation option. Note that this selection does not mean that no fallacy is present, but rather that none of the five fallacies of focus are present.

Our final annotated corpus has each sentence annotation unit accompanied by an "Level 1", two-way annotation of whether or not there is one of the five fallacies present. If there is such a fallacy present, then there is a "Level 2", five-way annotation indicating which one is present.

4. Approach

We describe the corpus constructed for developing and refining the fallacy annotation scheme and detail the annotation procedure as applied to that corpus.

4.1 COVID-19 Corpus

The annotation documents in our corpus are related to the topic of COVID-19, largely from US sources. Each article has one of six focus **topics** known to be particularly rife with misinformation: mask safety, long haulers, herd immunity, general vaccination safety and efficacy, COVID vaccination safety and efficacy, and the origin of the SARS-CoV-2 virus. For each of these topics, there were two opposing stances that were searched for and used in annotations. For example, on the topic of herd immunity, two articles were chosen—one from Fox News and one from *The New York Times*. The Fox News article indicates that many states in the United States were already at the level needed for herd immunity to take place, so it was unnecessary for people to get vaccinated. The New York Times article describes how far off the United States was from reaching herd immunity and without vaccines it would be impossible to reach.

We also paired the articles with different stances on a topic according to their **genre**. For example, two scientific articles are compared on the topic of the virus origin, where one argues for a manmade origin, while the other article argues for a natural origin. The resulting collection therefore allows for exploration of fallacies in documents demonstrating different perspectives on the same issues, and across different genres, while avoiding comparison between what we would expect to be very different genres with respect to fallacies, such as comparing social media posts to scientific articles.

The resulting corpus contains 26 documents that were manually selected based upon their main topic, the source genre, and the stance taken on the main topic.* The number of articles corresponding to a particular topic and genre are summarized in Table 1. Not all of our documents are full original texts; for example, the scientific journal sources only include the abstracts. The final corpus consists of 827 sentences, each of which is an annotated unit.

Table 1. Coverage of corpus topics and genres

Topic	Genre (no. of docs)
COVID vax safety	Online medical forum (2) Tabloid (1) Science magazine (1) Social media (1)
Herd immunity	General news (3) Talk radio (1)
Long haulers	Online medical forum (4)
Mask safety	General news (2) Social media (3)
General vax safety, efficacy	General news (2) Health care sites (2)
SARS-CoV-2 origin	General news (2) Scientific article (2)

4.2 Annotation Procedure

In the first pass of annotation, each individual document, with information on its genre and topic, was presented to two annotators (authors of this report and native English speakers with linguistics training living in the United States) in a separate spreadsheet, in which each sentence of the document was placed sequentially in its own row, and annotations were supplied in the adjacent column to each sentence instance within its row. The annotation process took place across the period of about one month, beginning with a training period of about a week, during which

*The corpus will be made available via data-sharing agreement pending publication.

annotators read and discussed the *Argotario* schema, annotated one of the documents, and then discussed these annotations.

For adjudication, after completing annotation of 23 of 26 documents independently ("first round" of annotation), the annotators met again to discuss annotations and established an agreed-upon gold standard subset of annotations for 9 documents containing 226 annotations. This gold standard subset was used for training the PET model described in Section 7. After this, two more documents (44 annotations) were annotated on the topics of mask and vaccine safety, from social media and general news outlets. This small subset was used for a final inter-annotator agreement (IAA) measurement on the "second round" of annotations of our two annotators, and its gold standard annotations were used as a test set for the PET model.

5. Results: Gold Standard Corpus Analysis

Our objective in building and annotating the full corpus and the gold standard subset was to find actual examples of the five types of logical fallacies over which to develop and refine the annotation scheme, not to make any generalized claims about the distribution of these types in the corpus.

5.1 Fallacies Across Genres

The gold standard corpus included genres of general news, social media, health care sites, online medical forums, and scientific articles (five of the full corpus' seven genres), where the majority of sentences came from general news and health care sites. The rate of fallacies varied greatly across these genres, as shown in Table 2 with their percentage in the gold standard corpus, made up of 226 annotations in total. The vast majority of sentences (77%) do not contain any of the five fallacy types. This is not unexpected, given that our approach to corpus construction, unlike that of *Argotario*, did not involve actively eliciting fallacies, but rather to search and sample actual documents from the relevant topic space for the purpose of identifying logical fallacies via annotations, and this process yielded documents that did not contain fallacious claims.

The general news documents, including articles from Fox News, Reuters News, *Time*, and *Forbes*, had the highest fallacy rate of 49%. Our small sample of social media had a similar rate of 43%. Health care sites (e.g., Children's Hospital of Philadelphia website) had the lowest rate of 2.2% and the online medical forums (e.g., BMJ.com

Table 2. Percent of annotations with fallacies by genre in gold standard corpus. Ratio specifies the number of sentences

Genre	%	Ratio
General news	49	41/83
Social media	43	3/7
Health care sites	2	2/89
Medical forum	6	2/31
Scientific articles	31	5/16
Total	100	53/226

forum for health care professionals) was also quite low at 6%. The fallacy rate of the scientific articles was 31%, which may seem surprisingly high on first glance, but one of the two scientific articles was a much-contested paper on the SARS-CoV-2 virus origin that was not peer-reviewed prior to publication on an open science site, where it is flagged as not following the norms of scientific rigor. Our single peer-reviewed scientific article contains 0 fallacies, while the non-peer-reviewed article has five fallacies for a fallacy rate of 45%, so that taken together, they yield the high average fallacy rate.*

5.2 Fallacy Types

The distribution of fallacy types in the gold standard corpus is shown in Table 3. There we see that the most frequent fallacy type annotated was *Hasty Generalization*, which occurred 19 times, 8.4% of the annotated sentences. The second most frequent fallacy annotated is *Appeal to Emotion*, which occurred 15 times, 6.6% of annotations. The next most frequent fallacy annotated is *Red Herring*, which occurred 13 times, 5.8% of the annotations. Both *Ad Hominem* and *Irrelevant Authority* were relatively infrequent in our gold standard corpus, both occurring three times each, and thereby each contributing to 1.3% of the corpus annotations.

6. Analyses

We provide analyses of the reliability of annotations, discuss annotation challenges, and propose modifications to the schema to improve reliability.

*We note again that documents were selected on topics thought to be rife with misinformation; our samples are small and we would not expect these genres to contain these levels of fallacies when treating other topics.

Table 3. Percentage of annotations by fallacy type in gold standard corpus (N refers to the number of sentences)

Annotation	N	% Gold
Hasty generalization	19	8.4
Appeal to emotion	15	6.6
Red herring	13	5.8
Ad hominem	3	1.3
Irrelevant authority	3	1.3
None of the above	173	76.5
Total	226	100.0

6.1 Reliability and Inter-Annotator Agreement

We compute IAA using Krippendorff's α.[5,6] Our choice of metric was motivated by the fact that the annotation categories are not equally distinct from one another and form two levels of hierarchical tagsets[7,8]: first whether or not one of the fallacies is present, and if one is present, which of the five fallacies. To tease out the reliability of each level of annotation, we compute IAA across the Level 1 "two-way" judgment level and then the subsequent Level 2 "five-way" judgment level. For Level 1 agreement, we simply compute IAA as to whether both annotators annotated that one of the five fallacies was or was not present. For Level 2 IAA, in only the subset of instances where both annotators agreed that a fallacy was present, we compute IAA as to whether annotators selected the same specific fallacy. IAA is summarized in Table 4.

We computed both levels of IAA for the full corpus, the first-round annotation portion of corpus, and the second-round/test portion of the corpus. This enables us to look for change over time after annotator discussion. We also computed Level 1 IAA for all documents by genre. Since the number of sentences when broken down by genre was small and so the number with fallacies even fewer, we did not compute Level 2 IAA within genre.

Overall, the IAA is quite low, demonstrating the challenging nature of this task even for annotators trained in linguistics who have exhibited reliable coding skills on other annotation tasks. Although there is no absolute value for high agreement, values below 0.67 are thought to be inconclusive.[5] The first-round corpus has the highest IAA of 0.54 for Level 2, while the second-round/test corpus has the highest IAA of 0.51 for Level 1. Thus, there is no evidence that the adjudication discussion establishing the gold standard corpus that occurred after the first round of annotation

Table 4. IAA (Krippendorff's α) across the full corpus, the first-round corpus, the second-round test corpus, and by genre

Data	Level 1 IAA	Level 2 IAA
Full corpus	0.47	0.51
First-round	0.46	0.54
Second-round/test	0.51	0.31
General news	0.23	-
Health care sites	0.48	-
Online med. forum	0.26	-
Talk radio	0.48	-
Science magazine	0.31	-
Scientific article	0.33	-
Tabloid	0.13	-

and before the second round of annotation of the test corpus led to any improvement in IAA, nor that there is any general improvement over time as annotators gain experience. Although our expectation was that Level 1 IAA (whether or not there is some fallacy present) would be higher than Level 2 IAA (which fallacy is present), this was not always the case. We posit that this may reflect a limitation of the choice to annotate at the sentence level, thereby causing some disagreement in where, precisely, a fallacy was present, especially in cases where the fallacious argument can only be identified as part of a broader context of the surrounding sentences setting up that argument, such as *Hasty Generalization*. Although further exploration is required, we hypothesize that the annotation unit was a main source of disagreements since otherwise the two annotators found roughly the same number of fallacies in each document.

It does not appear to be the case that the genre, and whether or not the genre tends to include fallacies, has any influence on whether or not annotators can reliably identify these fallacies. General news articles have the highest fallacy rates in our gold standard corpus; while this genre does have one of the lower Level 1 IAAs, it is quite similar to the IAA of online medical forums, which had a very low fallacy rate. Similarly, while health care sites have the lowest fallacy rate and have comparatively high Level 1 IAA, the talk radio genre has a high fallacy rate across both annotators and has an equally high Level 1 IAA as health care sites.

6.2 Annotation Challenges

Table 5 shows the confusion matrix between both annotators on the 120 sentences where it was agreed upon that a fallacy was present in the full corpus. The annotators agreed on the Level 2 fallacy label of 80 sentences, 67% of the corpus. Half of these agreements are *Appeal to Emotion* followed by *Red Herring* and *Hasty Generalization* with 24% and 20% agreement of Level 2 agreements, respectively.

Table 5. Confusion Matrix between both annotators on Hasty Generalization (HAST), Appeal to Emotion (EMOT), Red Herring (RED), Ad Hominem (AD), and Irrelevant Authority (IRR)

	HAST	**EMOT**	**RED**	**AD**	**IRR**
HAST	16	14	6	0	0
EMOT		40	13	3	1
RED			19	1	2
AD				3	0
IRR					2

Taking a closer look, the most common disagreement cases are *Hasty Generalization* $\times$ *Appeal to Emotion* and *Red Herring* $\times$ *Appeal to Emotion*. Annotators noted that the definition of *Hasty Generalization* seemed applicable to many claims that were not supported with adequate evidence or any evidence at all. *Appeal to Emotion* was noted to be realized not only in rhetorical structure, but also in single, emotion-evoking words. As a result, it could co-occur with other fallacies. In such cases, annotators could list the same instance twice with two annotations; however, within the gold standard corpus, the majority label was assigned, and the same sentence was never annotated twice with the same labels by both annotators, so there are no doubly annotated sentences in the gold standard. *Red Herring* annotations could be particularly difficult, given that determining a *Red Herring* or distracting argument relies upon an awareness of the author's stance and main thesis arguing that stance. Our documents were carefully selected for their focus on certain topics, and these topics are provided to the annotator in a clear labeling of each document. Thus, the annotator need only determine the stance on that topic. Nonetheless, it is likely that cultural, implicit knowledge of the annotators plays a role in making assumptions about the stance/thesis of a document and, in contrast, any distractor claims.

Table 6 gives examples of these disagreements from documents relating to the topic of COVID vaccine safety. There is a highly emotional, dramatic nature to each of these examples that certainly makes *Appeal to Emotion* a plausible annotation, but

there are also extreme conclusions being drawn that imply *Hasty Generalization*, as well as possible *Red Herring* arguments relating to authoritarianism as opposed to any direct evidence of vaccine safety.

Table 6. Disagreement examples relating to the topic of COVID vaccine safety

Label HASTY GEN. vs. **Label EMOTION**
I suspect these draconian organized crackdowns on health freedom will become a permanent reality. Government will use the Corona scare as a pretext to fast-track vaccine mandates into law everywhere.
Label RED HERRING vs. **Label EMOTION**
It is never wise to defer you personal decisions to an external authority, but especially now. This dilemma has already redefined the landscape, giving rise to new authoritarianism.

6.3 Domain Extension: Beyond COVID

In the course of leveraging the PET model described in Section 7 during the Joint Inter Agency Task Force 2021 SOCOM Data Challenge, we also annotated one document from their DOMEX data set, which is a data set comprising largely of images, where some images contained text that could under optical character recognition (OCR) in order to be converted into computer-readable text for further processing. Most of the documents were also in other languages besides English. We selected one document that was OCR'd and manually translated, resulting in the following short excerpt:

1. After boycotting Qatar..
2. Israel Today newspaper states that Saudi Arabia is an ally
3. Written by: Khaled Omar, last update on Friday June 09 2017 -14:02PM
4. Israel Today newspaper announced that Saudi Arabia is a close ally
5. This followed an announcement that Saudi Arabia, the UAE, Egypt and Bahrain will boycott Qatar, citing its support for terrorism.

Both annotators had found that the fallacy *Red Herring* was used here, but they did not choose the same placement for this fallacy. The first annotator said the *Red Herring* fallacies were found in the second and fourth line, while the second

annotator said the *Red Herring* fallacies were found in the first and fifth line. This reflects an underlying disagreement as to what the author's primary thesis or claim was, and, in contrast, what might be a "distractor" argument that pulls attention away from the argument supporting the main claim. Essentially, the first annotator thought that the primary claim related to the announcement of Saudi Arabia as an ally, thus the lines discussing the not-clearly-related boycott are the *Red Herring* fallacies; whereas the second annotator thought that the primary claim related to the boycott, so that the lines discussing the not-clearly-related announcement of Saudi Arabia as an ally were marked as *Red Herring* fallacies. Of course these two events (the boycott and the announcement of an ally relationship) are certainly related, but their relationship and interplay with respect to fallacies draws upon a great deal of social and cultural knowledge that the two annotators (as Americans with no other context for the article) did not have.

Thus, this example showed the importance of determining an agreed-upon main thesis or stance in advance so that the fallacies can be annotated with respect to the agreed upon primary thesis. We discuss other next steps for improving reliability, including a proposal for reconsidering the annotation unit, in the next section.

6.4 Next Steps: Improving Reliability

In the next phase of our research, we will make changes to our approach, distinguishing four sources of variation in the annotation process.[8] One of these arises from the similarity of annotation category definitions, where, in our case, *Hasty Generalization*, *Appeal to Emotion*, and *Red Herring* need to be more clearly distinguished. For this, we propose subdividing *Hasty Generalization* into subtypes defined by *Argotario* criteria.* Another source of variation comes for the differences in difficulty of the individual items. We will explore altering the span of annotation beyond single sentences, for example, to assist in identifying the premise sentences and then the *Hasty Generalization* made with respect to these premises. The diversity in the underlying data is yet another source of variation. We will add a separate category for purely unsupported claims with no evidence. We will also explore a subdivision of *Appeal to Emotion* into the annotation of single words that heighten the emotional or dramatic tone in a distracting way, making use of lexical connotation resources,[9] and full sentences that provide an argument that appeals specifically to fear instead

*Distinguishing 1) uses a sample that is too small, 2) follows falsely from a sub-part to a composite, and 3) follows falsely from a composite to a sub-part.

of providing an evidence-based argument. Finally, to address differences among annotators, as occurs in interpreting texts to annotate *Red Herrings*, we will explore annotating in advance and including the stance and thesis of the author, instead of just the topic area, as we saw that annotators disagreed on what text content could be the main thesis versus distracting arguments.

7. PET for Automatic Fallacy Annotation

We leverage our annotated data in the PET procedure described in Schick and Schütze.[4] PET is a method for training NLP models to solve a linguistic task that their original models were not specifically designed to solve. PET uses a pre-trained language model and is trained using a data set composed of Cloze-style (fill-in-the-blank) questions.

We use PET to classify fallacies in passages by constructing a pattern and verbalizers. The PET classifier for the ReCoRD data set,[10] in which named entities in a passage are possible answers to a question, was used as a template. As multiple fallacies could occur within a single passage, ReCoRD conveniently goes through each named entity and ascertains whether that entity is a possible answer with a Yes/No verbalizer. In our experiments, the named entities are instead possible fallacies, preceding the passage before the actual fallacy is presented, as if giving an example of a fallacy, or a statement without a fallacy.

We report our evaluation of the ability of the PET model to automatically annotate the five logical fallacies or provide the appropriate "none of the above" annotation. In Table 7, we summarize IAA, again using Krippendorff's α, for the PET model trained on the gold standard corpus (consisting of nine documents) and tested on the two held out documents in the test corpus. Here, we focus on IAA between PET and the gold standard label, and we also report for comparison the human IAA (repeated from Table 4), and the IAA between each of the human annotators and the gold standard label for the gold standard corpus and the test corpus. The IAA for PET versus the gold standard label on the test set is remarkably low, dipping into negative values for Level 2 IAA, demonstrating systematic (beyond chance) disagreements.

This underscores the possibility that if the annotation schema cannot be applied reliably by human annotators, then we cannot expect a system to be able to learn these distinctions reliably using training data annotated with that schema. PET is, as its name implies, exploiting patterns in text, but fallacies may not have easily

Table 7. Pairwise IAAs for Annotator 1 (A1) vs. Annotator 2 (A2), A1 vs. Gold, A2 vs. Gold, PET vs. Gold for both agreement levels on gold standard and test subsets

Data / Agreement Level	A1 vs. A2	A1 vs. Gold	A2 vs. Gold	PET vs. Gold
Gold standard corpus / level 1	0.46	0.71	0.82	-
Gold standard corpus / level 2	0.54	0.45	0.56	-
Test / level 1	0.51	0.77	0.73	0.22
Test / level 2	0.31	0.72	0.38	-0.07

exploitable patterns. Recognizing some kinds of logical fallacies may require the ability to follow the logical structure of an argument and see where it breaks down. In our future work, we will first explore a weighted approach. PET ranks the likelihood of each possible fallacy, and we noted that the correct answer was sometimes the second guess, with a relatively close difference. Second, we will explore altering the unit of annotation and perhaps training individual models for each fallacy, as each fallacy can differ greatly in the number of sentences involved in the realization of that fallacy.

8. Comparison to Related Work

There has been an explosion of activity in NLP on detecting misinformation and related tasks, including fake news detection and automatic fact-checking, resulting in various workshops and shared tasks (e.g., FEVER workshops). The broad goal of detecting misinformation automatically has motivated a wide variety of methods and tasks; thus, we briefly describe only the most relevant slice of this expanding research area, which treats annotating and detecting logical fallacies in particular.

In addition to the *Argotario* corpus, Da San Martino et al. annotate a corpus for various propaganda techniques, including the annotation of 12 fallacies, which only slightly overlap with the fallacies of interest in this report (*Red Herring* and *Appeal to Fear and Prejudice*).[11] As in our approach, the authors annotate journal articles as opposed to eliciting or seeking out particular fallacies. However, as a result, their corpus similarly suffers from an imbalance of fallacies that can be problematic for training data, further reinforcing the need for additional annotated fallacy data.

In Sahai et al., potential fallacies are collected automatically from Reddit by search-

ing for mentions of fallacies in comments, and then these are filtered through crowdsourced judgments.[12] The only fallacy of focus here included in their schema is *Hasty Generalization*, for which the authors report the lowest IAA, measured via Cohen's κ, of 0.38. The highest IAA reported is 0.64 for *Appeal to Authority*. We note that our own overall IAA for Level 2 agreement across the full corpus is comparable to this range, 0.51, when using Cohen's κ. This underscores the challenge of this annotation task. The authors explore several models for automatic prediction of the fallacies, including Bidirectional Encoder Representations from Transformers (BERT) and Multi-Granularity Network (MGN), with resulting F1 scores between 13% and 42% on the task most comparable to ours of labeling a comment with a particular fallacy. Unsurprisingly, given the correspondingly low IAA, the lowest F1 score is for *Hasty Generalization*.

9. Conclusions and Future Work

Our extension and evaluation of the *Argotario* logical fallacy schema has demonstrated the challenge of consistently recognizing these fallacies in documents relating to the COVID-19 pandemic, which has been accompanied by an "infodemic". We propose specific ways in which we the annotation schema can be more consistently applied and identify paths for future work in leveraging PET for automatic fallacy annotation. Although it is clear from this research and related work that agreeing upon these fallacies is difficult, we do see consistency in which documents had high fallacy rates and which documents had low fallacy rates—that is, although **where** a particular fallacy was realized was not agreed upon, annotators largely agreed on about the same number of fallacies overall for a given document. Thus, we posit that even with noisy annotations, plausibly the overall fallacy rate can be used to give a user in a search task a reliable metric of the credibility of a document.

As we work to incorporate fallacy detection in our search framework, we are leveraging the corpus annotated here to explore whether our search system is more likely to return a fallacious answer. We use Abstract Meaning Representation (AMR)[13] for our semantic search approach, in which a natural language user question is parsed into an AMR and then compared to the AMRs present in a parsed corpus of documents relevant to the question. We are currently exploring our hypothesis that scientific documents and other credible documents presented to the general public include thoughtful qualifiers and cautious hedging, while misinformative documents will tend to be stated plainly without any qualifiers or additional detail, and that this

may result in misinformative documents being retrieved and ranked more highly by our semantic search tools, which are designed to prioritize relevance. Preliminary results on 10 user questions with manually identified fallacious and non-fallacious answers demonstrate that both answer types are equally likely to be retrieved by our AMR-based search system. We will continue to explore this hypothesis to refine how and at what point in an interaction our search system should use fallacy information.

10. References

1. World Health Organization. Infodemic. Accessed November 2021 [online]. https://www.who.int/health-topics/infodemic; 2021.

2. Habernal I, Hannemann R, Pollak C, Klamm C, Pauli P, Gurevych I. Argotario: computational argumentation meets serious games. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations; Copenhagen, Denmark: Association for Computational Linguistics; 2017. p. 7–12.

3. Habernal I, Pauli P, Gurevych I. Adapting serious game for fallacious argumentation to German: pitfalls, insights, and best practices. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018); Miyazaki, Japan: European Language Resources Association (ELRA); 2018.

4. Schick T, Schütze H. Exploiting cloze-questions for few-shot text classification and natural language inference. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume; p. 255–269.

5. Krippendorff K. Content analysis: an introduction to its methodology. Sage Publications; 1980.

6. Passonneau RJ. Computing reliability for coreference annotation. In: Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04); Lisbon, Portugal: European Language Resources Association (ELRA); 2004.

7. Artstein R, Poesio M. Survey article: Inter-coder agreement for computational linguistics. Computational Linguistics. 2008;34(4):555–596.

8. Artstein R. Inter-annotator Agreement. In: Handbook of Linguistic Annotation; Dordrecht, Netherlands: Springer Netherlands; 2017; p. 297–313.

9. Allaway E, McKeown K. A unified feature representation for lexical connotations. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume; Online: Association for Computational Linguistics; 2021. p. 2145–2163.

10. Zhang S, Liu X, Liu J, Gao J, Duh K, Durme BV. ReCoRD: bridging the gap between human and machine commonsense reading comprehension. 2018.

11. Da San Martino G, Yu S, Barrón-Cedeño A, Petrov R, Nakov P. Fine-grained analysis of propaganda in news article. 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); Hong Kong, China: Association for Computational Linguistics; 2019. p. 5636–5646.

12. Sahai S, Balalau O, Horincar R. Breaking down the invisible wall of informal fallacies in online discussions. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers); Online: Association for Computational Linguistics; 2021. p. 644–657.

13. Banarescu L, Bonial C, Cai S, Georgescu M, Griffitt K, Hermjakob U, Knight K, Koehn P, Palmer M, Schneider N. Abstract meaning representation for sembanking. In: Proceedings of the 7th linguistic annotation workshop and interoperability with discourse; p. 178–186.

Appendix. Annotation Procedure

To annotate these fallacies, perform the following steps:

1. Open up one of the tab-delimited .txt annotation files in the current annotation directory using Excel.
2. Read through the text, pasted into cell A1 of the spreadsheet.
3. Copy the ENTIRE SENTENCE containing a fallacy into cell A2.
4. Copy and paste the appropriate fallacy type from the listing starting in cell E12 into cell B2 (copy and paste prevents typos and differences in format/capitalization).
5. Repeat (3) as needed into subsequent rows (e.g., A3, A4, A5) with appropriate fallacies in the adjacent column (e.g., B3, B4, B5) (see Fig. A-1 for proper annotation formatting).
6. List single sentences with more than one fallacy present twice, in separate rows, with a single fallacy indicated in the corresponding B-column cell.
7. When all fallacies in the text are listed with their sentence text, delete the fallacy listing starting in cell E12.
8. Save the file with your initials appended to the end of the file name in the format of "Tab delimited Text (.txt)".
9. Upload completed annotation files to appropriate shared folder location.

	A	B	C
1	Text text text…		
2	<Sentence 1>	<Fallacy 1>	
3	<Sentence 2>	<Fallacy 2>	
4	…	…	
5			
6			

Fig. A-1. Example spreadsheet demonstrating precise annotation procedure

List of Symbols, Abbreviations, and Acronyms

AD	Ad Hominem
AMR	Abstract Meaning Representation
BERT	Bidirectional Encoder Representations from Transformers
EMOT	Appeal to Emotion
HAST	Hasty Generalization
IAA	inter-annotator agreement
IRR	Irrelevant Authority
MGN	Multi-Granularity Network
NLP	natural language processing
OCR	optical character recognition
PET	Pattern-Exploiting Training
PPI	potentially problematic information
RED	Red Herring

1 (PDF)	DEFENSE TECHNICAL INFORMATION CTR DTIC OCA
1 (PDF)	DEVCOM ARL FCDD RLD DCI TECH LIB
1 (PDF)	DEVCOM ARL FCDD-RLC-IT C BONIAL