MAFALDA: A Benchmark and Comprehensive Study of Fallacy Detection and Classification

Chadi Helwe[1], **Tom Calamai**[1,2], **Pierre-Henri Paris**[1], **Chloé Clavel**[3], **and Fabian M. Suchanek**[1]
[1] Télécom Paris, Institut Polytechnique de Paris, France
[2] INRIA Saclay, Amundi, France
[3] INRIA Paris, France
firstName.lastName@telecom-paris.fr,firstName.lastName@inria.fr

Abstract

We introduce MAFALDA, a benchmark for fallacy classification that merges and unites previous fallacy datasets. It comes with a taxonomy that aligns, refines, and unifies existing classifications of fallacies. We further provide a manual annotation of a part of the dataset together with manual explanations for each annotation. We propose a new annotation scheme tailored for subjective NLP tasks, and a new evaluation method designed to handle subjectivity. We then evaluate several language models under a zero-shot learning setting and human performances on MAFALDA to assess their capability to detect and classify fallacies.

1 Introduction

A fallacy is an erroneous or invalid way of reasoning. Consider, e.g., the argument "*You must either support my presidential candidacy or be against America!*". This argument is a *false dilemma* fallacy: it wrongly assumes no other alternatives. **Fallacies can be found in various forms of communication**, including speeches, advertisements (Danciu et al., 2014), Twitter/X posts (Macagno, 2022), and political debates (Balalau and Horincar, 2021; Goffredo et al., 2022). They are also part of propaganda techniques employed to shape public opinion and promote specific agendas. Most notably, fallacies played a role in the 2016 Brexit referendum (Zappettini, 2019), and the debate about COVID-19 vaccinations (Elsayed et al., 2020), where fake news spread on news outlets and in social networks (Martino et al., 2019; Sahai et al., 2021; Balalau and Horincar, 2021). Detecting and identifying these fallacies is thus a task of broad importance.

The recent advances in deep learning and the availability of more data have given rise to approaches for detecting and classifying fallacies in text automatically (Martino et al., 2019; Al-Omari et al., 2019; Balalau and Horincar, 2021; Sahai et al., 2021; Abdullah et al., 2022). And yet, this work is fragmented: **most approaches focus on specific types of corpora** (e.g., only speeches) or **specific types of fallacies** (e.g., only *ad hominem* fallacies). Furthermore, **not all works use the same types of fallacies**, there is no consensus on a common terminology (Hansen, 2020), and fallacies come at different levels of granularity: an *appeal to emotion* can be, for instance, an *appeal to anger*, *fear*, *pride*, or *pity*. Most importantly, **annotating fallacies is an inherently subjective task**. While previous works acknowledge the subjectivity, none explicitly embraces it. On the contrary, the annotators typically aim for a unique annotation – by discussion or vote. Additionally, **existing works do not give human performances** on the benchmarks and evaluate only models.

This paper addresses these drawbacks by introducing the Multi-level Annotated Fallacy Dataset MAFALDA – a manually created fallacy classification benchmark. Our contributions are:

1. A taxonomy of fallacies that aligns, consolidates, and unifies existing public fallacy collections (Section 3).
2. A new annotation scheme – coined disjunctive annotation scheme – that accounts for the inherent subjectivity of fallacy annotation by permitting several correct annotations (Section 4).
3. A corpus that merges existing corpora, with 9,545 non-annotated texts and 200 manually annotated texts with 260 instances of fallacies, each with a manual justification (Section 5).
4. A study of the performance of state-of-the-art language models and humans on our benchmark (Section 6).

All our code and data are publicly available under a CC-BY-SA license[1] at https://github.com/ChadiHelwe/MAFALDA, allowing our study to be reproduced and built upon. We start our paper by

[1] as imposed by the dataset from Goffredo et al. (2022)

discussing related work in Section 2.

2 Related Work

2.1 Datasets

Numerous works have created datasets of fallacies. Habernal et al. (2018) created a dataset for *ad hominem* fallacies from the "Change My View" subreddit. Martino et al. (2019) created a news article dataset featuring 18 fallacies such as *red herring, appeal to authority, bandwagon, etc.*. Barlalau and Horincar (2021) trained models for propaganda technique identification using online forums. Sahai et al. (2021) compiled a Reddit-based corpus for fallacy detection with eight types of fallacies. Goffredo et al. (2022) introduced a dataset from American political debates with six different fallacy types. Along the same line, Jin et al. (2022) curated a claim dataset, containing 13 types of fallacies, based on online quizzes and the Climate Feedback website, employing a novel approach that mimics first-order logic. To address data annotation challenges, Habernal et al. (2017) created the Argotario game for fallacy detection in QA pairs. It created a corpus of 5 fallacy types. In the domain of (dis/mis)information, Musi et al. (2022) and Alhindi et al. (2022) annotated fallacies in COVID-19 and climate change articles with ten types of fallacies. Lastly, Payandeh et al. (2023) developed LOGICOM to evaluate large language models' (LLMs) robustness against logical fallacies in debate scenarios.

While all of these works advanced the understanding of fallacy detection, the studied fallacies are not the same across different works and are sometimes outright disjoint. The only work that creates a comprehensive taxonomy of fallacies is the (not yet peer-reviewed) work of Hong et al. (2023). However, this work enumerates 232 fallacies, which is clearly too many to be handled by a human. And indeed, their dataset is composed only of toy examples generated by GPT-4.

In this paper, we propose a benchmark that not only unifies public datasets on fallacy detection in a handy yet all-embracing taxonomy, but also comes with human annotations, human explanations, and evaluations for both language models and humans.

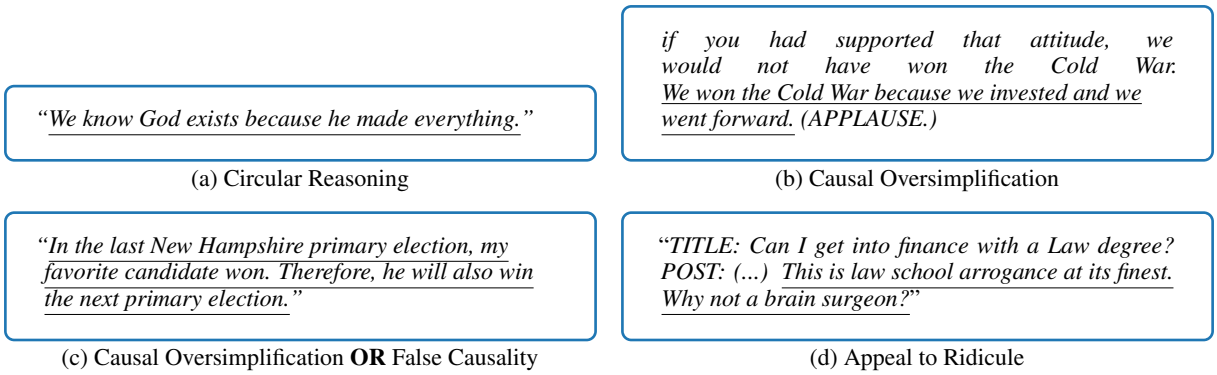
2.2 Subjectivity and Annotation Challenges

Human label variation is inherently part of annotating complex and subjective tasks (Plank, 2022). It is usually addressed with strategies such as simplifying the task, majority votes, or reconciliation of discrepancies. Goffredo et al. (2022) computed the Krippendorff's α on a subset of fallacies and reached inter-annotator agreements (IAAs) ranging from 0.46 to 0.60, which is a moderate agreement. On simpler tasks such as identifying only *ad hominem* using two groups of 6 workers, Habernal et al. (2018) reported a Cohen's κ of 0.79, which is a good agreement. However, they acknowledge the difficulty of annotated sub-categories such as *tu quoque* and *guilt by association* (they found a low IAA, but the value is not provided). When annotating spans of propaganda techniques, a complex task, Martino et al. (2019) found a γ IAA of 0.26, which is low. However, they could increase the IAA up to 0.60 when adding a reconciliation step. In Sahai et al. (2021), the annotator had to identify one fallacy at a time, and they reached a Cohen's κ of 0.515 (ranging from 0.38 to 0.64 based on the fallacy), which is a moderate agreement. They also computed the γ for the span selection per fallacy type and found values between 0.60 to 0.80, which is a good agreement. This was expected since it is a binary classification task. Sahai et al. (2021); Jin et al. (2022); Musi et al. (2022); Alhindi et al. (2022) used a reconciliation step too to tackle discrepancies in the annotations.

In summary, IAA in related work is usually only moderate. Disagreements are interpreted as noise, and are removed with various strategies. In this paper, we propose not just to acknowledge the subjectivity of fallacy annotation but actually to follow through with it. We contend that there are cases where **multiple, equally valid annotations can coexist for the same textual span**. Therefore, we propose a new subjective annotation scheme that allows for several alternative labels for the same span.

2.3 Taxonomies of Fallacies

Logical fallacies have been studied and classified since the time of Aristotle. There is a notable diversity in approaches and contents across various sources. The works of Aristotle (see Wikipedia (2023b)) and Whately (1826), despite their historical significance, present limitations in terms of the breadth of fallacies covered, listing only 13 fallacies each (our work, in contrast, finds more than 20). Downes (2003) offers a more extensive list with 36 fallacies. However, it still fails to mention common fallacies such as *appeal to nature*, *appeal to tradition*, and *guilt by association*. Curtis (2003)



(a) Circular Reasoning

(b) Causal Oversimplification

(c) Causal Oversimplification **OR** False Causality

(d) Appeal to Ridicule

Figure 1: Examples of Fallacies. The spans of the fallacies are underlined. Example 1a is from Jin et al. (2022), 1b from Goffredo et al. (2022), and 1d from Sahai et al. (2021). Detailed annotations are in Appendix C.

provides an exhaustive list of 87 fallacies. Yet, it provides only a rudimentary hierarchy (classifying, e.g., *no true Scotsman* as a sub-category of *equivocation*). Fallacies (2008) lists 48 fallacies – but lacks a hierarchical framework altogether. At the other end of the spectrum, Dowden (2010), Bennett (2012), Hong et al. (2023), and Wikipedia (2023a) offer extensive compilations of 231, 300+, 232, and 149 fallacies respectively. Yet, such a sheer volume of fallacies would be challenging in practical annotation tasks, as the annotator would have to scan (or memorize) hundreds of different fallacies.

Our work, in contrast, is driven by today's practical application scenarios. It aims to systematize and classify the fallacies used in current works on fallacy annotation, detection, and classification.

3 A Unified Taxonomy of Fallacies

3.1 Definitions

We start with the definition of an argument, following Copi et al. (2018); Britannica (2023):

Definition 3.1: Argument

An argument consists of an assertion called the *conclusion* and one or more assertions called *premises*, where the premises are intended to establish the truth of the conclusion. Premises or conclusions can be implicit in an argument.

Thus, an argument is typically of the form "***Premise***$_1$***:*** *All humans are mortal.* ***Premise***$_2$***:*** *Socrates is human.* ***Conclusion:*** *Therefore, Socrates is mortal.*". However, premises and conclusion can also appear in the opposite order and/or in the same sentence, as in "*Socrates is mortal because he is a human and all humans are mortal*". In many real-world arguments, the premise and the conclusion are spread apart (as in "*Of course, Socrates is mortal! How can you doubt this? After all, he's human, and all humans are mortal!*"). Sometimes, premises are left implicit (as in "*Socrates is mortal because he is human*"). Even the conclusion can be implicit (as in "*Socrates is human and all humans are mortal*"). In the context of a discussion, an argument can attack another argument (Dung, 1995), in which case the conclusion is implicitly negated ("*Socrates is immortal! – But he is human!*").

A valid argument is one where the truth of the premises guarantees the truth of the conclusion; otherwise, following Copi et al. (2018); Britannica (2023), the argument is a fallacy:

Definition 3.2: Fallacy

A fallacy is an argument where the premises do not entail the conclusion.

We refer the reader to Helwe et al. (2022) for a discussion of a formal definition of textual entailment. Appendix D.1 further distinguishes fallacies from other types of erroneous statements.

3.2 Taxonomy of Fallacies

In this paper, we propose a taxonomy that unifies and consolidates all types of fallacies used in current work on fallacy detection. We built our taxonomy manually, starting with a collection of fallacy types that are used in related work. Since the same fallacy can appear in different datasets under different names, we aligned equivalent fallacies manually. We used the definitions and guidelines in the source datasets to determine whether two fallacies are equivalent. We removed fallacies that were too broad (e.g., *appeal to emotion* could cover

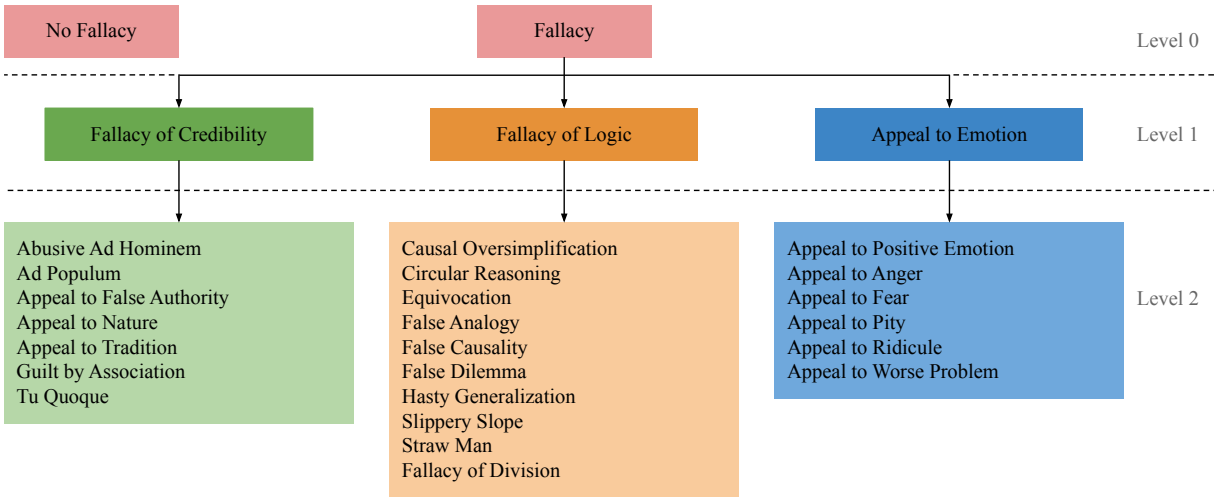


Figure 2: Tree structure of our taxonomy. Detailed definitions of the fallacies are in Appendix A.

many emotions), fallacies that appeared in only a single work (e.g., *confusion fallacy* appears only in Martino et al. (2019)), and we merged fallacies that were too similar in their definitions (like *begging the question* and *circular reasoning*). Some fallacies were not taken into account because they were not actually fallacies in our definition. These are, e.g., rhetorical techniques such as *flag waving* or *repetition*. Our list can obviously be extended in the future with new fallacies. Details on how our collection unifies existing works are in Appendix B.

We grouped our fallacies into broader categories to create a taxonomy on top of our collection. We chose the categories that Aristotle originally proposed (Wisse, 1989), because it has been shown to be applicable across various forms of communication – from political speeches to advertisements (Wisse, 1989). This yields the following taxonomy:

1. **Level 0** is a binary classification, categorizing text as either fallacious or non-fallacious.
2. **Level 1** groups fallacies into Aristotle's categories: 'Pathos' (appeals to emotion), 'Ethos' (fallacies of credibility), and 'Logos' (fallacies of logic, relevance, or evidence).
3. **Level 2** contains fine-grained fallacies within the broad categories of Level 1. For instance, under *fallacy of credibility*, we have specific fallacies such as *appeal to tradition*, *ad populum*, and *guilt by association*.

Our taxonomy is shown in Figure 2. For each fallacy, we provide **both a formal and an informal definition** in Appendix A (inspired by Bennett (2012)). For instance, the *appeal to ridicule* is informally defined as "an argument that portrays the opponent's position as absurd or ridiculous with the intention of discrediting it.". Formally, it is defined as "E_1 claims P. E_2 makes P look ridiculous, by misrepresenting P (P'). Therefore, $\neg P$.", where E_i are entities (e.g., people, organizations, etc.), and P is a proposition.

4 Tackling Subjectivity in Annotations

4.1 Subjectivity in Fallacy Annotation

Annotating fallacies is an inherently subjective endeavor. To see this, consider Example (c) in Figure 1. The argument goes that the candidate has to win again because he won last time. This can be seen as a *false causality* fallacy: a cause-effect relationship is incorrectly inferred between two events that have nothing to do with each other. However, it can also be seen as a *causal oversimplification* fallacy. This is because we can contend that having won the last election gives the candidate an edge over other candidates in terms of visibility, and thus makes it more likely that he wins this year's election as well. The argument is thus fallacious mainly because it fails to acknowledge other factors that play a role in re-election.

This simple example already shows subjectivity in fallacy annotations, where several annotations can be defended. It would be counter-productive if the annotators converged on, say, *causal oversimplification*, so that every approach of fallacy annotation is penalized for proposing an (equally plausible) *false causality*. There are other cases of legitimately differing opinions: One annotator may see implicit assertions that another annotator does not see. In "*Are you for America? Vote*

for me!" one reader may see the implicit "*or you must be against America*" (which makes this a *false dilemma*), while another annotator may see no such implicature. Annotators may also have different thresholds for fear (*appeal to fear*) or insults (*ad hominem*). Finally, different annotators have different background knowledge: A sentence such as "*Use disinfectants or you will get Covid-19!*" may be read as a plausible warning by one annotator but as an *appeal to fear* fallacy by an annotator who knows that Covid-19 does not spread via contaminated surfaces. We will now present a disjunctive annotation scheme that accounts for this inherent subjectivity.

4.2 Disjunctive Annotation Scheme

Before presenting our annotation scheme, we need to establish some common ground:

Definition 4.1: Text

A text is a sequence of sentences $st_1, \ldots, st_n$.

A *span* on a text is a contiguous sequence of sentences. The set S of all spans of a text $st_1, \ldots, st_n$ is thus $S = \{st_i \ldots st_j \mid 0 < i \leq j \leq n\}$.

Definition 4.2: Span

The span of a fallacy in a text is the smallest contiguous sequence of sentences that comprises the conclusion and the premises of the fallacy. If the span comprises a pronoun that refers to a premise or to the conclusion, that premise or conclusion is not included in the span.

We work on the level of sentences, because previous work has shown that agreement on the token level is even harder to achieve (Jin et al., 2022). We allow the use of pronouns to decrease the size of the spans: When a sentence refers to another sentence by a pronoun, that other sentence does not have to be part of the span. For instance, in Example (d) of Figure 1, the premise of the fallacy is in the title of the post, and the conclusion is at the end of the text. Thus, a span that covers the entire fallacy would have to cover the entire post from title to end. However, the pronoun "This" refers to the title, and thus we can omit the title from the span. Nevertheless, a span can comprise several sentences.

A span can be annotated with a label (such as a fallacy type) by an annotator (or a group of them) or by a system. We now propose the key element of our disjunctive annotation scheme, in which subjectivity is not projected away, but explicitly embraced by allowing for several equally valid labels for the same span.

Definition 4.3: Gold Standard

Let $\mathcal{F}$ be the set of fallacy types and $\perp$ be a special label that means "no fallacy".
Given a text and its set of spans S, a gold standard G is a set of pairs of a span $s \in S$ and a set of labels from $\mathcal{F} \cup \{\perp\}$:

$$G \subseteq S \times (\mathcal{P}(\mathcal{F} \cup \{\perp\}) \setminus \{\emptyset, \{\perp\}\})$$

Here, $\mathcal{P}(\cdot)$ denotes the powerset.

The gold standard associates a given span with one or more fallacy labels. If more than one label is present, this means that any label is acceptable as an annotation. The gold standard can also associate a span with $\perp$, which means that the annotation of this span is optional. However, in this case, the gold standard has to associate the span also with at least one other label, as we are not interested in annotating non-fallacious sentences. The gold standard can also contain the same span twice, which means that the span has to be annotated with two labels. The alternative labels for a span can be generated through various methods during the annotation process, e.g., one annotator giving alternatives, a group of annotators proposing different labels due to lack of consensus, or multiple independent annotators combining their labels (see Example 4.1).

We define a prediction as the annotation of a text by a system or a user:

Definition 4.4: Prediction

Given a set of fallacy types $\mathcal{F}$, a text, and its set of spans S, a prediction P is a set of pairs of a span $s \in S$ and a label $l \in \mathcal{F}$:

$$P \subseteq S \times \mathcal{F}$$

The following example gives substance to these definitions:

Example 4.1

Let "$a\ b\ c\ d$" be a text where a, b, c, and d are sentences.
Suppose $S = \{a\ b, d\}$ (i.e., the sentences a and b are one fallacious span, and d is a span of one fallacious sentence), $a\ b$ has labels $\{l_1, l_2\}$, and d has label $\{l_3\}$. In that case, $G = \{(a\ b, \{l_1, l_2\}), (d, \{l_3\})\}$

An example of prediction P could be $P = \{(a, \{l_1\}), (a, \{l_2\}), (b, \{l_3\}), (c, \{l_4\}), (d, \{l_1\})\}$

4.3 Evaluation Metrics.

To compare two annotated spans, we adapt the precision and recall of Martino et al. (2019) to alternatives. Given two spans, p with its label l_p, and g with its set of labels l_g, respectively, and a normalizing constant h, these metrics compute a comparison score as follows:

$$C(p, l_p, g, l_g, h) = \frac{|p \cap g|}{h} \times \delta(l_p, l_g)$$

δ is a similarity function. We use $\delta(x, y) = [x \in y]$, where $[\cdot]$ is the Iverson bracket.

Let G be the gold standard, and let P be the prediction of a user or a system. The precision for P is computed by comparing each span in P against all spans in G, and taking the score of the best-matching one:

$$\textit{Precision}(P, G) = \frac{\sum_{(p,l_p) \in P} \max_{(g,l_g) \in G} C(p, l_p, g, l_g, |p|)}{|P|}$$

If there are no annotations in P (i.e., $|P| = 0$), we set precision to 1. This choice is inspired by the intuition that a loss in precision should result only from false predictions. If there are no such false predictions, then precision should not be harmed.

To calculate the recall, we exclude all the spans from the gold standard that contain a $\perp$. The rationale for this choice is that when a span has also been marked as "no fallacy", its annotation is considered optional. Therefore, we do not want to penalize models that do not provide an annotation for such a span. We define the set $G\ \ $, which is G restricted to the spans that do not map to $\perp$, i.e., $G\ \ = \{(s, L) \in G \mid \perp \notin L\}$. The recall is then computed as:

$$\textit{Recall}(P, G) = \frac{\sum_{(g,l_g) \in G} \max_{(p,l_p) \in P} C(p, l_p, g, l_g, |g|)}{|G\ \ |}$$

If $|G\ \ | = 0$, we set the recall to 1. The intuition is that a model should be penalized in recall only for the annotations it misses from the gold standard. If there are no such missed annotations, recall should not suffer. Appendix L shows how our metrics handle various edge cases.

The F1-score is computed as usual as the harmonic mean of precision and recall. It is easy to see that our definitions of precision and recall fall back to the standard definitions if G does not have alternatives and all spans comprise only a single sentence. In that case, the score C is 1 if the spans are identical and identically labeled. If there are no alternatives, our measures are also identical to the ones in Martino et al. (2019), with one difference: we use the $\max$ instead of a sum in the definitions to select the best matching span. In this way, two neighboring spans with the same label do not achieve full precision or recall if the gold standard requires one contiguous span with that label. Using $\max$ instead of a sum also avoids the case where Martino et al. (2019)'s metric yields precision or recall scores exceeding one. This occurs when the gold standard contains overlapping spans with identical labels, affecting precision, or when the prediction includes such overlaps, affecting recall. However, their metric is equivalent to ours as long as (1) there are no alternatives in the gold standard, (2) neither the gold standard nor the prediction contains overlapping spans with the same label and (3) each span from the gold standard overlaps with at most one span with the same label from the predictions and vice versa. Examples, proofs, and further details are in Appendix I.

5 MAFALDA Dataset

5.1 Source Datasets

We used four publicly available fallacy datasets to construct our benchmark:

We imported all 8,576 texts from Sahai et al. (2021). These are online discussions from Reddit, as in Example (d) of Figure 1. We reconstructed the texts by concatenating the post of interest, the previous post (if available), and the title. The title was considered as a citation and was thus not annotated. This dataset contains sentences that were labeled as negative examples.

We imported all 336 texts from Martino et al. (2019), which are from news outlets. We imported all 583 texts from Jin et al. (2022), which are either toy examples gathered from online quizzes (as in

Example (a)), or longer climate-related texts originating from news outlets. Finally, we imported all 250 texts from Goffredo et al. (2022), which are American political debates (Example (b)). We split these longer texts into shorter texts by concatenating the previous and following sentences of allegedly fallacious texts.

This gives us an English-language corpus of 9,745 texts, which is diverse in terms of linguistic terms and text length. We removed URLs, emails, and phone numbers globally.

5.2 Annotation

The existing annotation schemes on our corpus varied a lot among papers: for example, only Sahai et al. (2021) approached the annotation task as a binary classification, where annotators determine if a given text contains a specified type of fallacy. The annotations process also varied greatly w.r.t. how consensus was obtained, as explained in Section 2.2.

Therefore, we removed all annotations, and manually re-annotated, from scratch, 200 randomly selected texts from our merged corpus. Our sample mirrors the distribution of sources and the original labels in our corpus: it contains 124 texts from Sahai et al. (2021), 59 texts from Jin et al. (2022), and 17 political debate texts from (Goffredo et al., 2022). We did not use the texts from Martino et al. (2019) we initially planned to use because they were more than 5,000 characters long. Thus, annotating a single text would considerably bias the work towards Martino's. However, the texts are part of our cleaned and homogenized dataset, and our goal is to include the annotations of these texts as we enlarge our manual annotation.

LLMs were not involved in the annotation process. We did **not involve crowd workers** either, because 33%-46% of Mturk workers are estimated to use ChatGPT (Veselovsky et al., 2023). Hence, we annotated the texts ourselves. Our task was (*i*) identifying each argument in a text, (*ii*) determining whether it is fallacious, (*iii*) determining the span of the fallacy (as defined in Definition 4.2), and (*iv*) choosing the fallacy type(s). We discussed each fallacious span together, and either converged on one annotation or permitted several alternative annotations for the same span. **We provide an explanation for each annotation**, and **we provide a completed template for each annotation**, as defined in Section 3.2. For instance, Example (d) from Figure 1, which shows an *appeal to ridicule*: the post argues against the possibility of working in finance with a law degree by exaggerating the position and thus portraying it as ridicule. Breaking down the example with the formal definition yields:

- E_1= The original poster
- P= It might be possible to work in finance with a law degree
- E_2= The author of the post.
- $P^{\emptyset}$= Law school students are so intelligent that they can do any job, even surgeons.

Here, E_i are entities (persons, organizations) or groups of entities, P and $P^{\emptyset}$ are premises, properties, or possibilities.

The process took around 40 hours. This corresponds to an average of 12 minutes per example, ranging from less than a minute for toy examples to half an hour when disagreement raised a debate. The total number of person-hours was 130. Details about the annotators can be found in Appendix E.

5.3 Statistics

Our dataset comprises 9,745 texts, of which 200 texts have been annotated manually, with a total of 268 spans. Among these, 137 texts contained at least one span identified as fallacious, while the remaining texts did not contain any fallacious spans. The mean number of spans per text is 1.34.

Among the 200 texts, 71 were initially labeled as non-fallacious. However, our annotation found fallacies in some of these texts. This can be explained by the methodology from Sahai et al. (2021), where crowd workers check only one specific type of fallacy. If that fallacy is not present, the text is annotated as non-fallacious. Our annotation, however, spotted other fallacies in the text, and labeled them. In the end, we have 63 non-fallacious texts.

The dataset contains all the fallacies presented in Section 3.2. The three most frequent fallacies represent 1/4 of the dataset, while the least frequent fallacies appear less than three times. **71.5% of the texts were annotated with a similar fallacy as the original one** (at least one fallacy of the source annotation was in the new annotation, or we agreed on a non-fallacious text). The difference is mainly because our taxonomy introduced new fallacies, such as *appeal to ridicule*, and removed fallacies that we considered too vague or broad, such as *intentional fallacy*. In some cases, we used a different granularity than in the source: while the source might say *appeal to emotion*, we annotated, e.g., with *appeal to fear*. We also permit several

alternative annotations per span, which entails that the new annotations have, on average, more annotations per text than the source annotations (Original: 0.665, Ours: 1.34).

The dataset contains 203 spans, of which 65 (i.e., 28%) contain at least two different (alternative) labels (see Example 1c). We computed the co-occurrence matrix of the fallacies. Most fallacies do not co-occur too frequently (less than 30% of the time) with another particular fallacy, which indicates that **our definitions of fallacies are broadly orthogonal**. However, there are two fallacies with high co-occurrence frequency: *appeal to pity* has a 100% co-occurrence with *strawman* and *appeal to worse problem*. However, this is because there is only one occurrence of *appeal to pity* in our dataset. The second one is *guilt by association*, which is 38% of the time associated with *abusive ad hominem*. This is explained by the fact that *guilt by association* and *abusive ad hominem* are two types of the *ad hominem* fallacy. Complete statistics about our dataset are provided in Appendix F.

6 Experiments

We will now evaluate the ability of state-of-the-art LLM to detect the fallacies in our benchmark. Our benchmark is not intended for training or fine-tuning, and hence, we study a zero-shot setting with basic prompts. We are interested in the task of fallacy detection and classification of a given text, i.e., the input is a text, and the output is a list of annotated spans.

6.1 Settings

We study ChatGPT as well as 12 open-source models, covering different model sizes (Table 1). We use a bottom-up approach to evaluate our models starting at Level 2 granularity and extrapolate labels for Levels 1 and 0 based on these predictions, as our dataset includes three levels of granularity.

We employ a basic prompting approach that presents the model with our definition of a fallacy, the instruction to annotate the fallacies, the list of fallacies without their definitions, the corresponding text example, and the sentence to be labeled. The detailed prompt can be found in Appendix H. Our experiments are conducted at the sentence level; spans are formed by grouping consecutive sentences with the same label. A significant challenge with generative models is their inconsistent format output. Thus, we deem an output correct if it includes the name of the correct fallacy (or a part of it). Details about the models can be found in Appendix G.

6.2 Results

	MAFALDA		
Model	F1 Level 0 *	F1 Level 1 *	F1 Level 2
Baseline random	0.435	0.061	0.010
Falcon 7B	0.397	0.130	0.022
LLAMA2 Chat 7B	0.572	0.114	0.068
LLAMA2 7B	0.492	0.148	0.038
Mistral Instruct 7B	0.536	0.144	0.069
Mistral 7B	0.450	0.127	0.044
Vicuna 7B	0.494	0.134	0.051
WizardLM 7B	0.490	0.087	0.036
Zephyr 7B	0.524	0.192	0.098
LLaMA2 Chat 13B	0.549	0.160	0.096
LLaMA2 13B	0.458	0.129	0.039
Vicuna 13B	0.557	0.173	0.100
WizardLM 13B	0.520	0.177	0.093
GPT 3.5 175B	**0.627**	**0.201**	**0.138**
Avg. Human on Sample	**0.749**	**0.352**	**0.186**

* Labels were extrapolated from Level 2.

Table 1: Performance results of different models across different granularity levels in a zero-shot setting. Avg. human on sample concerns only the 20 subsamples of MAFALDA for the user study. Metrics are explained in Section 4.2.

Table 1 shows the results across different granularity levels in a zero-shot setting, as evaluated using our metric (see Section 4.2). We added *Baseline random*, a dummy model that predicts labels randomly following a uniform distribution.

At all levels of granularity, all models surpass the performance of the baseline model (except for Falcon on Level 0), indicating that they are successfully identifying certain patterns or features. GPT 3.5 outperforms all other models at all levels. At Level 1, Zephyr 7B achieves comparable results to GPT 3.5, possibly thanks to the quality of its training dataset and/or engineering tricks, challenging the assumption that larger models are always more effective. More surprisingly, LLaMA2 performs better in its 7B version than in its 13B version for Levels 0 and 1. This phenomenon is in line with findings from Wei et al. (2023).

We also investigate whether it makes a difference to prompt the models directly on Level 1 (as opposed to extrapolating Level 1 from Level 2). For Mistral Instruct and Zephyr, there is no significant difference: Mistral Instruct obtains an F1 score of 0.149, and Zephyr achieves an F1 score of 0.185.

Gold Standard	F1 Level 0 *	F1 Level 1 *	F1 Level 2
User 1	0.616	0.310	0.119
User 2	0.649	0.304	0.098
User 3	0.696	0.253	0.093
User 4	0.649	0.277	0.144
MAFALDA	**0.749**	**0.352**	**0.186**

* Labels were extrapolated from Level 2.

Table 2: Cross-comparison of user annotations and the gold standard. Each annotation of the user study has been alternatively used as a gold standard to demonstrate the superiority of our own gold standard.

We also measure human performance on our dataset (which constitutes, to our knowledge, the first such study in the fallacy classification literature). We aim to establish (*i*) whether humans outperform language models for the task at hand and (*ii*) whether humans agree more with our gold standard than among themselves. As human effort is more costly than running a language model (and even more so since we need engaged annotators who do not resort to ChatGPT or other LLMs), we asked four other annotators to annotate 20 randomly chosen examples on the same task as the systems. On these 20 samples, we compared the results of human annotators and LLMs. The low scores of human annotators reported in Table 2 show that the task is difficult. Still, **human participants outperform the language models** as shown in Table 1: Contrary to what previous work has demonstrated (Gilardi et al., 2023), GPT-3.5 does not perform better than humans.

Next, we study whether they agree more with our gold standard than among themselves. We treat each annotator's work as a gold standard and assess the precision, recall, and F1 scores of the other annotators. Our gold standard achieves an F1 score of 0.186 on average for humans (see Table 2), outperforming the best alternative, which scores 0.144. More details about the annotators and the results are in Appendix E and J, respectively.

We analyze the errors in the two models, GPT-3.5 and Falcon, focusing on their performance at Levels 1 and 2, alongside user study annotations. Our first goal is to determine whether the models, specifically the highest-performing GPT-3.5 and the lowest-performing Falcon, exhibit controlled or uncontrolled behavior in their output generation. While both models can produce nonsensical outputs, Falcon often predicts multiple irrelevant fallacies and significantly more unknown labels than GPT-3.5.

Our second goal is to determine which fallacy type is the most challenging. The analysis reveals that humans and models struggle with the fallacies of appeal to emotion. We hypothesize that emotions often appear in texts without necessarily constituting a fallacy, which complicates the distinction between emotional texts and fallacious texts that use appeals to emotion. More details about this analysis are in Appendix K.

7 Conclusion

We have presented MAFALDA, a unified dataset designed for fallacy detection and classification. This dataset integrates four pre-existing datasets into a cohesive whole, achieved through developing a new, comprehensive taxonomy. This taxonomy aligns publicly available taxonomies dedicated to fallacy detection. We manually annotated 200 texts from our dataset and provided an explanation in the form of a completed template for each of them. The disjunctive annotation scheme we proposed embraces the subjectivity of the task and allows for several alternative annotations for the same span. We have further demonstrated the capabilities of various large language models in zero-shot fallacy detection and classification at the span level. While Level 0 classification shows good results, Levels 1 and 2 are largely out of reach of LLMs in zero-shot settings. We hope that our benchmark will enable researchers to improve the results of this challenging task.

Future work includes expanding into few-shot settings and exploring advanced prompting techniques, such as chain-of-thought, using the template-based definitions of fallacy and the taxonomy we provided. Furthermore, we believe that using a top-down approach, i.e., from Level 0 to Level 2 of our taxonomy, may provide better results than the bottom-up approach we used in our experiments. Regarding our disjunctive annotation scheme, we are interested in exploring its use in other NLP domains. Lastly, we plan to enrich the dataset with more annotated examples for model fine-tuning.

Acknowledgement. This work was partially funded by the NoRDF project (ANR-20-CHIA-0012-01) and the SINNet project (ANR-23-CE23-0033-01) and Amundi Technology.

Limitations

Our work provides a dataset that may contain sensitive content such as racism, apologies for committing crimes, and misogyny. We believe that this is unavoidable in our fight against fake news and manipulative content.

One limitation of our benchmark is that we may have been biased while annotating it. Our collective bias is limited because we come from diverse cultural backgrounds, countries, religions, and political convictions, but it may still exist. We have also made efforts to mitigate bias during our annotation, with clear guidelines, or by achieving a consensus or at least providing strong arguments in the form of a completed template as described in the formal definitions of Appendix A.

Our dataset has the potential for misuse in training systems that could be exploited for manipulative purposes, such as crafting more convincing fallacious arguments or disinformation campaigns. Finally, models trained on this problem may wrongly label a text as fallacious. They must thus not be used to flag a text as fallacious without manual verification.

A further consideration is the size of our dataset. It is relatively small due to the time-intensive nature of the annotation process. It is thus not suited for fine-tuning, but rather intended for evaluating large language models in zero-shot and few-shot settings.

References

Malak Abdullah, Ola Altiti, and Rasha Obiedat. 2022. Detecting Propaganda Techniques in English News Articles using Pre-trained Transformers. In *International Conference on Information and Communication Systems*.

Hani Al-Omari, Malak Abdullah, Ola AlTiti, and Samira Shaikh. 2019. JUSTDeep at NLP4IF 2019 task 1: Propaganda detection using ensemble deep learning models. In *Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*.

Tariq Alhindi, Tuhin Chakrabarty, Elena Musi, and Smaranda Muresan. 2022. Multitask instruction-based prompting for fallacy recognition. In *Conference on Empirical Methods in Natural Language Processing*.

Oana Balalau and Roxana Horincar. 2021. From the Stage to the Audience: Propaganda on Reddit. In *Conference of the European Chapter of the Association for Computational Linguistics*.

Bo Bennett. 2012. *Logically Fallacious: The Ultimate Collection of over 300 Logical Fallacies (Academic Edition)*.

Britannica. 2023. *Fallacy*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Irving M Copi, Carl Cohen, and Victor Rodych. 2018. *Introduction to logic*. Routledge.

Gary N. Curtis. 2003. Logical Fallacies: The Fallacy Files.

Victor Danciu et al. 2014. Manipulative marketing: persuasion and manipulation of the consumer through advertising. *Theoretical and Applied Economics*, 21(2).

Pieter Delobelle, Murilo Cunha, Eric Massip Cano, Jeroen Peperkamp, and Bettina Berendt. 2019. Computational Ad Hominem Detection. In *Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*.

Bradley Dowden. 2010. Fallacies | Internet Encyclopedia of Philosophy.

Stephen Downes. 2003. Stephen's Guide to the Logical Fallacies.

Phan Minh Dung. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2).

Shadia Abdelhameed Elsayed, Osama Abu-Hammad, Albraa B Alolayan, Yasmin Salah Eldeen, and Najla Dar-Odeh. 2020. Fallacies and facts around covid-19: the multifaceted infection. *The Journal of craniofacial surgery*.

Logical Fallacies. 2008. Logical Fallacies - List of Logical Fallacies with Examples.

Harry J. Gensler. 2010. *The A to Z of Logic*. The A to Z Guide Series. Scarecrow Press.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks.

Pierpaolo Goffredo, Shohreh Haddadan, Vorakit Vorakitphan, Elena Cabrio, and Serena Villata. 2022. Fallacious Argument Classification in Political Debates. In *International Joint Conference on Artificial Intelligence*.

Ivan Habernal, Raffael Hannemann, Christian Pollak, Christopher Klamm, Patrick Pauli, and Iryna Gurevych. 2017. Argotario: Computational argumentation meets serious games. In *Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. Before Name-Calling: Dynamics and Triggers of Ad Hominem Fallacies in Web Argumentation. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Hans Hansen. 2020. Fallacies. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*.

Chadi Helwe, Simon Coumes, Chloé Clavel, and Fabian M. Suchanek. 2022. TINA: Textual Inference with Negation Augmentation. In *Conference on Empirical Methods in Natural Language Processing Findings*.

Ruixin Hong, Hongming Zhang, Xinyu Pang, Dong Yu, and Changshui Zhang. 2023. A closer look at the self-verification abilities of large language models in logical reasoning. *arXiv preprint arXiv:2311.07954*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schölkopf. 2022. Logical Fallacy Detection. *arXiv preprint arXiv2202.13758*.

Fabrizio Macagno. 2022. Argumentation profiles and the manipulation of common ground. the arguments of populist leaders on twitter. *Journal of Pragmatics*, 191.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-Grained Analysis of Propaganda in News Article. In *Conference on Empirical Methods in Natural Language Processing*.

Elena Musi, Myrto Aloumpi, Elinor Carmi, Simeon Yates, and O'Halloran Kay. 2022. Developing fake news immunity: Fallacies as misinformation triggers during the pandemic. In *Online Journal of Communication and Media Technologies*.

Amirreza Payandeh, Dan Pluth, Jordan Hosier, Xuesu Xiao, and Vijay K Gurbani. 2023. How susceptible are llms to logical fallacies? *arXiv preprint arXiv:2308.09853*.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Barbara Plank. 2022. The 'problem' of human label variation: On ground truth in data, modeling and evaluation. *arXiv preprint arXiv:2211.02570*.

Paul Reisert, Benjamin Heinzerling, Naoya Inoue, Shun Kiyono, and Kentaro Inui. 2019. Riposte! a large corpus of counter-arguments. *arXiv preprint arXiv:1910.03246*.

Saumya Sahai, Oana Balalau, and Roxana Horincar. 2021. Breaking Down the Invisible Wall of Informal Fallacies in Online Discussions. In *Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, et al. 2023. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*.

Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. 2023. Artificial artificial artificial intelligence: Crowd workers widely use large language models for text production tasks. *arXiv preprint arXiv:2306.07899*.

Jason Wei, Najoung Kim, Yi Tay, and Quoc V. Le. 2023. Inverse scaling can become u-shaped. In *Conference on Empirical Methods in Natural Language Processing*.

Richard Whately. 1826. *Elements of Logic*. International Debate Education Association.

Wikipedia. 2023a. List of fallacies.

Wikipedia. 2023b. Sophistical refutations.

J. Wisse. 1989. *Ethos and Pathos: From Aristotle to Cicero*. Hakkert.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

Franco Zappettini. 2019. The brexit referendum: How trade and immigration in the discourses of the official campaigns have legitimised a toxic (inter) national logic. *Critical Discourse Studies*, 16(4).

A Definitions of the fallacies

In the following, we provide, for each fallacy, its informal definition, its formal definition, and a toy example. We start by describing the variables/-placeholders used in the formal templates.

- $A =$ attack
- $E =$ entity (persons, organizations) or group of entities
- $P, P_i =$ premises, properties, or possibilities
- $C =$ conclusion

The following definitions are inspired by Bennett (2012) and have been adapted to be more generic.

A.1 Fallacies of Emotion

Appeal to Anger

Informal: This fallacy involves using anger or indignation as the main justification for an argument, rather than logical reasoning or evidence.

Formal: E claims P. E is outraged. Therefore, P. Or E_1 claims P. E_2 is outraged by P. Therefore, P (or $\neg P$ depending on the situation).

Example: The victim's family has been torn apart by this act of terror. Put yourselves in their terrible situation, you will see that he is guilty.

Annotation with Variables: E (the speaker) claims P (the accused is guilty) and expresses outrage. Therefore, P (guilt).

Appeal to Fear

Informal: This fallacy occurs when fear or threats are used as the main justification for an argument, rather than logical reasoning or evidence.

Formal: If $\neg P_1$, something terrible P_2 will happen. Therefore, P_1.

Example: If you don't support this politician, our country will be in ruins, so you must support them.

Annotation with Variables: If $\neg P_1$ (not supporting the politician), then P_2 (country in ruins) will happen. Therefore, P_1 (must support the politician).

Appeal to Pity

Informal: This fallacy involves using sympathy or compassion as the main justification for an argument, rather than logical reasoning or evidence.

Formal: P which is pitiful, therefore C, with only a superficial link between P and C

Example: He's really struggling, so he should get the job despite lacking qualifications.

Annotation with Variables: P (he's struggling) is presented as a pitiful situation, leading to C (he should get the job), despite a merely superficial link between P and C.

Appeal to Positive Emotion

Informal: This fallacy occurs when a positive emotion – like hope, optimism, happiness, or pleasure – is used as the main justification for an argument, rather than logical reasoning or evidence.

Formal: P is positive. Therefore, P.

Example: Smoking a cigarette will make you look cool, you should try it!

Annotation with Variables: P (smoking cigarettes looks cool) leads to P (try smoking).

Appeal to Ridicule

Informal: This fallacy occurs when an opponent's argument is portrayed as absurd or ridiculous with the intention of discrediting it.

Formal: E_1 claims P. E_2 makes P look ridiculous, by misrepresenting P (P'). Therefore, $\neg P$.

Example: There's a proposal to reduce carbon emissions by 50% in the next decade. What's next? Are we all going to stop breathing to reduce CO2?

Annotation with Variables: E_1 (unspecified entity) claims P (proposal to reduce the carbon emissions). E_2 (the speaker) makes P looks ridiculous by suggesting an extreme scenario $P^{\emptyset}$ (stop breathing). Therefore, $\neg P$ (reducing carbon emissions is unreasonable).

Appeal to Worse Problems

Informal: This fallacy involves dismissing an issue or problem by claiming that there are more important issues to deal with, instead of addressing the argument at hand. This fallacy is also known as the "relative privation" fallacy.

Formal: P_1 is presented. P_2 is presented as a best-case. Therefore, P_1 is not that good. OR P_1 is presented. P_2 is presented as a worst-case. Therefore, P_1 is very good.

Example: Why worry about littering when there are bigger problems like global warming?

Annotation with Variables: P_1 (littering) is compared to P_2 (global warming), which is a worse problem, leading to the conclusion that P_1 is not important.

A.2 Fallacies of Logic

Causal Oversimplification

Informal: This fallacy occurs when a complex issue is reduced to a single cause and effect, oversimplifying the actual relationships between events or factors.

Formal: P_1 caused C (although P_2, P_3, P_4, etc. also contributed to C.)

Example: There is an economic crisis in the country, the one to blame is the president.

Annotation with Variables: P_1 (the president) caused C (economic crisis), while ignoring other contributing factors (P_2 (worldwide economical context), P_3 (previous policies), etc.).

Circular Reasoning

Informal: This fallacy occurs when an argument assumes the very thing it is trying to prove, resulting in a circular and logically invalid argument.

Formal: C because of P. P because of C. OR C because C.

Example: The best smartphone is the iPhone because Apple creates the best products.

Annotation with Variables: C (iPhone is the best smartphone) because P (Apple creates the best products), which in turn is justified by the claim C.

Equivocation

Informal: This fallacy involves using ambiguous language or changing the meaning of a term within an argument, leading to confusion and false conclusions.

Formal: No logical form: P_1 uses a term T that has a meaning M_1. P_2 uses the term T with the meaning M_2 to mislead.

Example: The government admitted that many cases of credible UFOs (Unidentified flying objects) have been reported. Therefore, that means that Aliens have already visited Earth.

Annotation with Variables: P_1 (many cases of credible UFOs have been reported) uses the term UFO with the meaning M_1 (unidentified flying objects). P_2 (aliens have already visited Earth) uses UFO with a different meaning M_2 (implying that aliens = UFOs), misleading the conclusion.

Fallacy of Division

Informal: This fallacy involves assuming that if something is true for a whole, it must also be true of all or some of its parts.

Formal: E_1 is part of E, E has property P. Therefore, E_1 has property P.

Example: The team is great, so every player on the team must be great.

Annotation with Variables: E_1 (every player) is part of E (the team). E has the property P (great), then E_1 also has P.

False Analogy

Informal: This fallacy involves making an analogy between two elements based on superficial resemblance.

Formal: E_1 is like E_2. E_2 has property P. Therefore, E_1 has property P. (but E_1 really is not too much like E_2)

Example: We should not invest in Space Exploration. It's like saying that a person in debt should pay for fancy vacations.

Annotation with Variables: E_1 (a country in debt plans to explore space) is linked to E_2 (a family in debt plans fancy vacations). E_2 has property P (expensive and not advisable), implying E_1 should also have P.

False Causality

Informal: This fallacy involves incorrectly assuming that one event causes another, usually based on temporal order or correlation rather than a proven causal relationship.

Formal: P is associated with C (when the link is mostly temporal and not logical). Therefore, P causes C.

Example: After the rooster crows, the sun rises; therefore, the rooster causes the sunrise.

Annotation with Variables: P (rooster crows) is associated with C (sunrise), but the link is temporal, not causal, leading to the false conclusion that P causes C.

False Dilemma

Informal: This fallacy occurs when only two options are presented in an argument, even though more options may exist.

Formal: Either P_1 or P_2, while there are other possibilities. OR Either P_1, P_2, or P_3, while there are other possibilities.

Example: You're either with us, or against us.

Annotation with Variables: Presents a choice between P_1 (with us) and P_2 (against us), excluding other possibilities.

Hasty Generalization

Informal: This fallacy occurs when a conclusion is drawn based on insufficient or unrepresentative evidence.

Formal: Sample E_1 is taken from population E. (Sample E_1 is a very small part of population E.) Conclusion C is drawn from sample E_1.

Example: I met two aggressive dogs, so all dogs must be aggressive.

Annotation with Variables: A small sample E_1 (two aggressive dogs) is taken from a larger population E (all dogs). Therefore C (all dogs are aggressive).

Slippery Slope

Informal: This fallacy occurs when it is claimed that a small step will inevitably lead to a chain of events, resulting in a significant negative outcome.

Formal: P_1 implies P_2, then P_2 implies P_3,... then C which is negative. Therefore, $\neg P_1$.

Example: If we allow kids to play video games, they will see fights, guns, and violence, and then they'll become violent adults.

Annotation with Variables: P_1 (allowing kids to play video games) implies P_2 (seeing fights, guns, and violence), which in turns implies P_3 (to like violence, etc.) leading to C (kids becomes violent adults). Therefore, $\neg P_1$.

Strawman Fallacy

Informal: This fallacy involves misrepresenting an opponent's argument, making it easier to attack and discredit.

Formal: E_1 claims P. E_2 restates E_1's claim (in a distorted way P^0). E_2 attacks (A) P^0. Therefore, $\neg P$.

Example: He says we need better internet security, but I think his panic about hackers is overblown.

Annotation with Variables: E_1 (an unspecified person (He)) claims P (need for better internet security), E_2 (the speaker) distorts the claim as P^0 (panic about hackers). Therefore $\neg P$.

A.3 Fallacies of Credibility

Abusive Ad Hominem

Informal: This fallacy involves attacking a person's character or motives instead of addressing the substance of their argument.

Formal: E claims P. E's character is attacked (A). Therefore, $\neg P$.

Example: "John says the earth is round, but he's a convicted criminal, so he must be wrong."

Annotation with Variables: E (John) claims P (the earth is round). John's character is attacked (A) (being a criminal). Therefore, $\neg P$ (the earth is not round).

Ad Populum

Informal: This fallacy involves claiming that an idea or action is valid because it is popular or widely accepted.

Formal: A lot of people believe/do P. Therefore, P. OR Only a few people believe/do P. Therefore, $\neg P$.

Example: Millions of people believe in astrology, so it must be true.

Annotation with Variables: Many people believe in P (astrology). Therefore, P (astrology is true).

Appeal to Authority

Informal: This fallacy occurs when an argument relies on the opinion or endorsement of an authority figure who may not have relevant expertise or whose expertise is questionable. When applicable, a scientific consensus is not an appeal to authority.

Formal: E claims P (when E is seen as an authority on the facts relevant to P). Therefore, P.

Example: A famous actor says this health supplement works, so it must be effective.

Annotation with Variables: E (famous actor) claims P (the health supplement works). Therefore, P (it must be effective).

Appeal to Nature

Informal: This fallacy occurs when something is assumed to be good or desirable simply because it is natural, while its unnatural counterpart is assumed to be bad or undesirable.

Formal: P_1 is natural. P_2 is not natural. Therefore, P_1 is better than P_2. OR P_1 is natural, therefore P_1 is good.

Example: Herbs are natural, so they are better than synthetic medicines.

Annotation with Variables: P_1 (herbs are natural) and P_2 (synthetic medicines are not natural), leading to P_1 is better than P_2.

Appeal to Tradition

Informal: This fallacy involves arguing that something should continue to be done a certain way because it has always been done that way, rather than evaluating its merits.

Formal: We have been doing P for generations. Therefore, we should keep doing P. OR Our ancestors thought P. Therefore, P.

Example: We've always had a meat dish at Thanksgiving, so we should not change it.

Annotation with Variables: P (always had a meat dish at Thanksgiving) should continue. Therefore, continue P.

Guilt by Association

Informal: This fallacy involves discrediting an idea or person based on their association with another person, group, or idea that is viewed negatively.

Formal: E_1 claims P. Also E_2 claims P, and E_2's character is attacked (A). Therefore, $\neg P$. OR E_1 claims P. E_2's character is attacked (A) and is similar to E_1. Therefore $\neg P$.

Example: Alice believes in climate change, just like the discredited scientist Bob, so her belief must be false.

Annotation with Variables: E_1 (Alice) claims P (belief in climate change). E_2 (Bob) also claims P. However E_2's character (A) is attacked (being discredited). Therefore $\neg P$.

Tu Quoque

> **Informal:** This fallacy occurs when someone's argument is dismissed because they are accused of acting inconsistently with their claim, rather than addressing the argument itself.
>
> **Formal:** E claims P, but E is acting as if $\neg P$. Therefore $\neg P$.
>
> **Example:** Laura advocates for healthy eating but was seen eating a burger, so her advice on diet is invalid.
>
> **Annotation with Variables:** E (Laura) claims P (advocates for healthy eating), but E is acting as if $\neg P$ (eating a burger, which is unhealthy eating). Therefore $\neg P$ (advice on diet is invalid).

Our categorization of fallacies into logic, emotion, and credibility is based on the primary aspect of the fallacy that leads to an invalid or weak argument. In practice, some fallacies could be argued to fit into more than one category.

B Comparison of Fallacy Types

Previous works have studied a large number of different fallacy types. The earliest works focused on *ad hominem*, while later works included dozens of other types. To build our taxonomy, we tried to unify most fallacy types in the literature. Table 3 shows each type of fallacy studied by each paper that proposed a dataset. Most fallacies from our taxonomy are part of at least two already existing datasets. Based on our definition (Section 3), rhetorical techniques that are not based on an actual argument are not considered fallacies. Thus, we did not include techniques such as *repetition* or *slogans*. During the initial annotation phase, we observed that the *red herring* fallacy was too vague, so we replaced it with more precise sub-categories, such as *appeal to worse problem*. This explains why *appeal to worse problem* , which is present in only one other dataset, is part of our taxonomy. Similarly, during the annotation, we found multiple examples of *fallacy of division*, which is related to *hasty generalization* but does not fit its description. Hence, we added *fallacy of division* in the taxonomy.

C Additional Examples

Example 1b is an *appeal to positive emotion*. Using our formal definition, it has the following key component:

- P= pride in the military's strength and state of preparedness

Example 1c illustrates a case of alternative labels, where the text can be a *causal oversimplification* or a *false causality*. In the context of *causal oversimplification*, the scenario can be deconstructed according to the formal definition as follows:

- P_1= Winner of the last primary election
- C= He/she will win the general election
- $P_1, P_2, \ldots$: Factors including political dynamics, the opponent in the race, etc.

As a *false causality*, the structure is:

- P= Winner of the last primary election
- C= He/she will win the general election

D Annotation Guidelines for Identifying Fallacious Arguments

The task of annotating a text with fallacies can be decomposed into several steps: First, determine if the text contains an argument and what the premises and conclusion are. Then, the span must be delimited. Finally, an adequate label must be chosen. For the construction of our gold standard, annotators used Doccano[2], and followed these guidelines:

1. **Consensus Requirement:** Before finalizing annotations for any given text, annotators should try to reach a consensus. This collaborative approach ensures consistency and accuracy in the identification of fallacious arguments. In instances where consensus is unattainable, the differing viewpoints regarding potential fallacies should be noted as alternative interpretations, as detailed in Section 4.2.

2. **Resource Utilization:** Annotators are encouraged to consult various resources, including Google Search, Wikipedia, and books on argumentation. However, using Large Language

[2] https://github.com/doccano/doccano

Our Taxonomy				Alhindi et al. (2022)								
Level 1	Level 2	Habernal et al. (2018)	Delobelle et al. (2019)	Martino et al. (2019) (Balalau and Horincar, 2021)	Jin et al. (2022)	Musi et al. (2022)	Habernal et al. (2017)	Goffredo et al. (2022)		Reisert et al. (2019)	Sahai et al. (2021)	Hong et al. (2023)
Fallacy of Credibility	Abusive Ad Hominem	abusive	abusive attack	whataboutism	Ad Hominem			General				Probability Proposition Quantification Syllogism Ambiguity Inconsistency Irrelevance Insufficiency Inappropriate Presumption
	Tu Quoque	tu quoque						Tu quoque				
		circumstantial bias					ad hominem	ad hominem	Bias ad hominem			
	Guilt by Association	guilt by association		reductio ad hitlerum								
			name calling	name calling					Name-calling			
	Ad Populum			doubt bandwagon	Ad Populum				Popular opinion		Appeal to Majority	
	Appeal to Nature										Appeal to Nature	
	Appeal to Tradition										Appeal to Tradition	
	Appeal to False Authority			appeal to false authority	Fallacy of Credibility	Appeal to Inappropriate Authority	irrelevant authority	Appeal to authority	False authority without evidence		Appeal to Authority	
Fallacy of Logic	Causal Oversimplification			causal oversimplification		Evading the Burden of Proof Cherry Picking of Evidence						
	Hasty Generalization				Faulty Generalization	Hasty Generalization	Hasty Generalization			Hasty Generalization	Hasty Generalization	
	False Causality				False Causality	False Cause Post Hoc (Correlation presented as Causation)		False cause		Questionable Cause		
	Circular Reasoning				Circular Claim					Begging the Question		
	False Dilemma			black-and-white fallacy	False Dilemma						Black-or-White	
	Slippery Slope								Slippery Slope		Slippery Slope	
	False Analogy					False Analogy						
	Straw Man			straw man	Fallacy of Extension	Strawman						
	Fallacy of Division				Deductive Fallacy							
	Equivocation			red herring obfusc. int. vagueness confusion	Fallacy of Relevance Equivocation	Red Herring Vagueness	Red Herring			Red Herring		
		Technique based on use of vocabulary		thought-terminating cliches exaggeration/minimization repetition slogans loaded language				Slogan	Loaded Language			
Fallacy of Emotion	Appeal to Positive Emotion			flag-waving					Flag waving			
	Appeal to Anger Appeal to Fear			appeal to fear/prejudice	Appeal to Emotion		Appeal to Emotion	Appeal to Emotion	Appeal to fear Appeal to pity			
	Appeal to Pity											
	Appeal to Ridicule											
	Appeal to Worse Problem										Appeal to Worse Problems	
				red herring	Intentional Fallacy of Relevance	Red Herring				Red Herring		

Table 3: List of fallacies per paper and in our taxonomy.

Models, such as ChatGPT, is prohibited to prevent potential bias or contamination in the annotations.

3. **Reference Material:** For definitions and clarifications:
    - Refer to the definitions of an argument and a fallacy as outlined in Section 3 and Appendix D.1.
    - Consult the Appendix A for detailed formal and informal definitions of individual fallacies.
    - Follow the definition of spans detailed in Section 4.2
4. **Annotation Protocol:**
    - Upon reaching a consensus, annotators must document their rationale, aligning their reasoning with the formal definitions provided.
    - Annotators are encouraged to add useful comments. This includes identifying text segments that may require special post-processing or additional review.

Our annotation guidelines are also available on the Web page of our project, https://github.com/ChadiHelwe/MAFALDA/.

D.1 Edge Cases

We provide additional information to help the annotators with edge cases.

In our definition (as in Gensler (2010)), a fallacy is always an argument in the sense of Definition 3.1, i.e., **a fallacy is always of the form "*A, B, C, ...* therefore *X*" or of the form "*X* because *A, B, C, ...*", or it can be rephrased into these forms**. Hence, false assertions are not, *per se*, fallacies. For example, "*Paris is the capital of England*" is a false claim. However, it is not a fallacy because it is not an argument. The same goes for generalizations: "*All Americans love Trump*" is false, but not a fallacy. An insult (such as "*You are too stupid*"), likewise, is not a fallacy[3]. Slogans (such as "*America first!*"), likewise, are not fallacies in our definition, even if other works classify them as propaganda (Martino et al., 2019). An *appeal to emotion* (such as "*Think of the poor children!*"), likewise, is not a fallacy by itself. It becomes a fallacy only when used as the premise of a fallacious argument, as in: "*Think of the poor children, and [therefore] vote for me!*". But not every argument that appeals to emotion is automatically fallacious. For instance, the argument "*During a Covid-19 pandemic, you should wear a mask in public transport because otherwise you could get infected*" appeals to fear. However, it is still a valid argument because the premise does entail the conclusion. Even if the premises of an argument are factually false, the argument is not necessarily fallacious. For example, "*All Americans love Trump, and therefore Biden loves Trump*" is an argument that rests on a false premise – but it is not fallacious because the premise indeed entails the conclusion in the sense of (Helwe et al., 2022): if the premise were true, the conclusion would be true as well. **The fallaciousness of an argument is thus largely independent of the truth values of its components.**

Finally, **the description of fallacious reasoning is not automatically fallacious.** For example, "*You should wear a tin foil hat because it protects you against mind control*" is a fallacy (because the tin foil hat does not protect against mind control of any known form). However, the following is a factual assertion, not a fallacy: "*Some people wear tin foil hats because they are afraid of mind control*".

[3] It becomes a fallacy when it is used as the premise of an argument, as in "*You are stupid, therefore what you say can't be true.*"

E Annotators

E.1 Annotators

We conducted two annotation phases, one for the annotation of the gold labels, which resulted in the MAFALDA dataset, and one for the user study. Here is a description of the background of the annotators.

E.1.1 Gold Standard Annotators

The gold standard was produced by the authors of the paper, who have the following characteristics:

- Nationality: Lebanese, Gender: Male, Native language: Arabic, Education: Master's degree, Occupation: Ph.D. student in computer science.
- Nationality: French, Gender: Male, Native language: French, Education: Master's degree, Occupation: Ph.D. student in computer science.
- Nationality: French, Gender: Male, Native language: French, Education: Ph.D. degree,

Occupation: Post-doctoral researcher in computer science.

- Nationality: French, Gender: Female, Native language: French, Education: Ph.D. degree, Occupation: Professor in computer science.
- Nationality: German, Gender: Male, Native language: German, Education: Ph.D. degree, Occupation: Professor in computer science.

Compensation: The annotators are the paper's authors and did not receive compensation for the annotations.

Biases and Limitations: The annotators are all authors of the paper. They are working in the computer science field

E.1.2 User Study Annotators

The user study annotations were provided by the following 4 persons:

- Nationality: Lebanese, Gender: Male, Native Language: ArabicEducation: Master's degree in mechanical engineering, Occupation: Statistics Expert.
- Nationality: French, Gender: Male, Native Language: French, Education: Master's degree in big data and data science, Occupation: Ph.D. Student in computer science.
- Nationality: Morrocan, Gender: Female, Native Language: French, Education: Ph.D. degree, Occupation: Data scientist.
- Nationality: French, Gender: Male, Native Language: French, Education: Master's degree in machine learning, Occupation: Ph.D. Student in computer science.

Compensation: The annotators were volunteers and were not compensated for the annotations.

E.2 Insights from the User Study Annotators

The annotation process was very time-consuming, with some annotators taking up to four hours to complete their task for the 20 examples. One annotator humorously questioned their normality, stating, "*I don't know if I'm a normal human, but I found it difficult! :)*" while another jokingly expressed regret over accepting the task. These comments reflect the general sentiment about the task's complexity. The annotators often struggled with specific examples, such as *"Reasonable regulations don't lead to the fed keeping lists and someday coming after all gun owners to suppress the working class"*, which has been annotated differently by each user such as an *ad populum* and *false causality* fallacy while it is not a fallacy. This is often due to over-complicated sentences.

F Dataset

Source Dataset	Non-annotated	Annotated
Sahai et al. (2021)	640 (7,812)	71 (53)
Jin et al. (2022)	524	59
Martino et al. (2019)	336	0
Goffredo et al. (2022)	233	17
TOTAL	1733 (7,812)	137 (53)

Table 4: Distribution of text from the initial source and from the final re-annotated dataset. Numbers in parenthesis are for non-fallacious texts.

Number of Spans	0	1	2	3	4	5	6
w/ counting alternatives	63	70	32	18	8	6	3
w/o counting alternatives	63	95	23	15	3	1	0

Table 5: Number of text with N spans. The first line considers alternatives, i.e., a disjunction of two labels for a span will count as two annotations. Conversely, in the second line, an alternative will count as one annotation. This allows for comparing the usage of alternatives in our annotations.

Table 4 presents statistics about our dataset: the source of each text, Table 5 displays the number of annotations in the 200 texts, and Table 6 presents the frequency of each fallacy.

The left-hand side of Figure 3 shows the diversity of the vocabulary used by the source datasets. The right-hand side shows the average length of the texts. The large diversity our benchmark results from the merger of the four source datasets.

Figure 4 shows the co-occurrence frequency of each fallacy in the MAFALDA dataset.

G Description of the Models

The computational budget was around 144 GPU hours for all models except GPT 3.5 and around $2 for GPT 3.5 experiments. The GPU was an NVIDIA A 100. We used a temperature of 0.8.

GPT-3.5 (Brown et al., 2020), developed by OpenAI, is a transformer-based language model with 175 billion parameters, pre-trained on

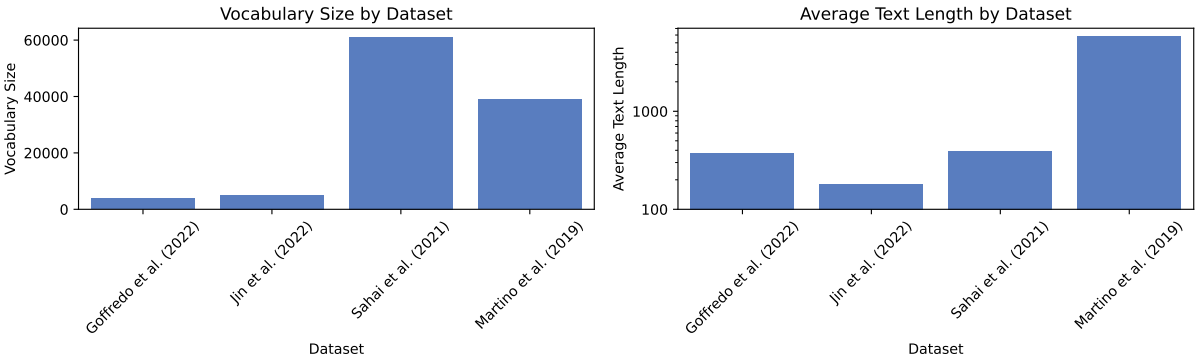


Figure 3: Statistics about our source datasets. The left graphic shows the vocabulary size, while the right graphic shows the average length of the texts.

Figure 4: Co-occurrence of labels (frequency)

	annotations	sources
non-fallacious	63	71
hasty generalization	28	33
causal oversimplification	23	0
Appeal to Ridicule	20	0
false dilemma	18	7
ad hominem	16	8
nothing	14	0
ad populum	14	13
straw man	13	0
false causality	13	8
false analogy	12	0
slippery slope	11	6
appeal to fear	11	0
appeal to nature	11	10
circular reasoning	11	10
appeal to (false) authority	9	10
appeal to worse problems	8	8
guilt by association	8	0
equivocation	7	1
appeal to tradition	6	6
appeal to anger	6	0
appeal to positive emotion	3	0
tu quoque	3	0
fallacy of division	2	0
appeal to pity	1	0
fallacy of relevance *	0	2
intentional *	0	1
appeal to emotion *	0	10

* Fallacies not included in MAFALDA.

Table 6: Number of spans for each fallacy: this table presents the distribution of fallacies in our dataset, comparing MAFALDA annotations with source annotations.

an extensive dataset encompassing a diverse range of texts. GPT-3.5 employs a technique known as Reinforcement Learning from Human Feedback (RLHF) for fine-tuning. In this process, human trainers review and provide feedback on the model's outputs, ensuring the model responses are accurate and aligned with human judgment and values.

Falcon (Penedo et al., 2023) is a large language model primarily pre-trained on the RefinedWeb – a curated dataset extracted from CommonCrawl and refined for quality through filtering and deduplication. The model has two versions: 40B and 7B parameters.

LLaMA-2 (Touvron et al., 2023), developed by Meta, is a transformer-based language model pre-trained on 2 trillion tokens from various public sources. This model has multiple versions, including LLaMA 2-chat, tailored for dialogue applications. LLaMA-2 Instruct, another variant, has been fine-tuned using human instructions, LLaMA-2 generated instructions, and datasets like BookSum and Multi-document Question Answering. LLaMA-2 models come in different sizes, with parameters ranging from 7B to 70B.

Vicuna (Chiang et al., 2023) is a model based on LLaMA, fine-tuned using a dataset comprising user conversations with ChatGPT. This model is available in two different sizes: 7B and 13B.

Mistral (Jiang et al., 2023) is a 7B-parameter transformer-based model. It uses two attention mechanisms to improve inference speed and memory requirements: grouped-query attention (GQA) and sliding window attention (SWA). Specific details regarding the training data and hyperparameters are not disclosed. An alternative model version is also provided, fine-tuned to follow instructions. This refined model was trained using publicly available instruction datasets from the Hugging Face repository.

WizardLM (Xu et al., 2023) is a model based on LLaMa. It has been fine-tuned with a dataset comprising instructions that vary in complexity. The dataset was generated through a method known as Evol-Instruct, which systematically evolves simple instructions into more advanced ones. WizardLM is available in two sizes: 7B and 13B.

Zephyr (Tunstall et al., 2023) is a model based on Mistral and was fine-tuned on a variant of the UltraChat dataset, a synthetic dataset of dialogues generated by ChatGPT. Zephyr was further trained using the UltraFeedback dataset, which encompasses 64,000 ranked prompts and responses evaluated by GPT-4 to enhance its alignment.

H Level 2 Prompt

Definitions:

- *An argument consists of an assertion called the conclusion and one or more assertions called premises, where the premises are intended to establish the truth of the conclusion.*

Premises or conclusions can be implicit in an argument.

- *A fallacious argument is an argument where the premises do not entail the conclusion.*

Text: "{complete_example_input}"
Based on the above text, determine whether the following sentence is part of a fallacious argument or not. If it is, indicate the type(s) of fallacy without providing explanations. The potential types of fallacy include:

- *appeal to positive emotion*
- *appeal to anger*
- *...*
- *guilt by association*
- *tu quoque*

Sentence: "{sentence_input}"
Output:

An example and the generated output using GPT-3.5:
Definitions:

- *An argument consists of an assertion called the conclusion and one or more assertions called premises, where the premises are intended to establish the truth of the conclusion. Premises or conclusions can be implicit in an argument.*
- *A fallacious argument is an argument where the premises do not entail the conclusion.*

Text: "I lost my phone in the living room, so it will always be in the living room when it is lost."
Based on the above text, determine whether the following sentence is part of a fallacious argument or not. If it is, indicate the type(s) of fallacy without providing explanations. The potential types of fallacy include:

- *appeal to positive emotion*
- *appeal to anger*
- *...*
- *guilt by association*
- *tu quoque*

Sentence: "I lost my phone in the living room, so it will always be in the living room when it is lost."
Output: This sentence is an example of the fallacy of hasty generalization.

I Metrics

Figure 5 and Figure 6 show an example for the calculation of our precision and recall metrics are computed. We now prove some properties of our metrics.

Proposition I.1. *Given a gold standard G, where each span comprises only a single sentence, and where each fallacy set contains only one element, and given a prediction P, where each span comprises only a single sentence, our precision coincides with the standard precision.*

Proof: By definition, we have, for any spans p, g, and for any sets l_p, l_g:

$$C(p, l_p, g, l_g, |p|) \tag{1}$$
$$= \frac{|p \cap g|}{|p|} \times \delta(l_p, l_g) \tag{2}$$
$$= \frac{|p \cap g|}{|p|} \times [l_p = l_g] \tag{3}$$

If p and g are singleton spans, this boils down to

$$= [p = g] \times [l_p = l_g] \tag{4}$$
$$= [p = g \wedge l_p = l_g] \tag{5}$$

Thus, we have, for any singleton span s and any label l:

$$[(s, \{l\}) \in G] \tag{6}$$
$$= [\exists (s^0, \{l^0\}) \in G : s^0 = s \wedge l^0 = l] \tag{7}$$
$$= [\exists (s^0, \{l^0\}) \in G : C(s, l, s^0, l^0, |s|) = 1] \tag{8}$$
$$= \max_{(s^0, fl^0g) 2 G} C(s, l, s^0, l^0, |s|) \tag{9}$$

This entails that the number of true positives (TP) is

$$|\{(s, l) \in P \mid (s, \{l\}) \in G\}| \tag{10}$$
$$= \sum_{(s,l) 2 P} [(s, \{l\}) \in G] \tag{11}$$
$$= \sum_{(s,l) 2 P} \max_{(s^0, fl^0g) 2 G} C(s, l, s^0, l^0, |s|) \tag{12}$$

The standard precision is the ratio of true positives (TP) out of the sum of true positives and false positives (FP):

$$\text{Standard Precision} \tag{13}$$
$$= \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{14}$$

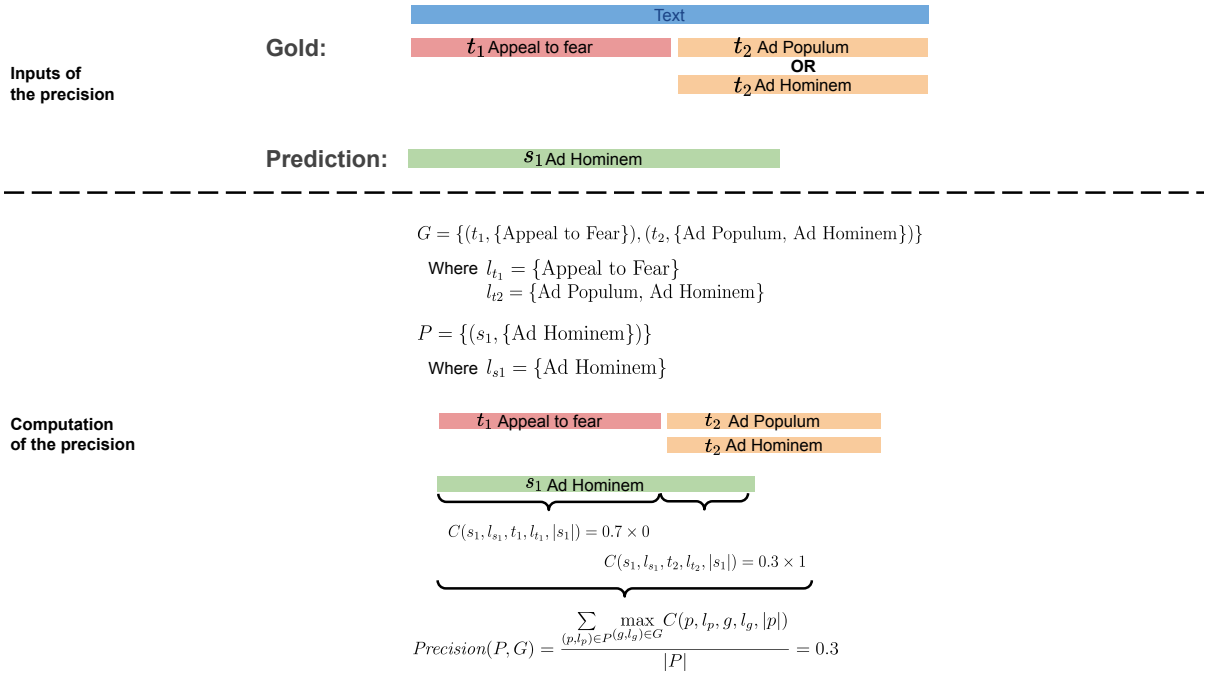


Figure 5: Example of Precision computation with alternatives.

With $|P| = \text{TP} + \text{FP}$ and Equation 12, this is equivalent to

$$= \frac{\sum_{(s,l)2P} \max_{(s^0,fl^0g)2G} C(s, l, s^0, l^0, |s|)}{|P|}$$

□

Proposition I.2. *Given a gold standard G, where each span comprises only a single sentence, and where each fallacy set contains only one element, and given a prediction P, where each span comprises only a single sentence, our recall coincides with the standard recall.*

Proof: As previously, for any p, g that are singleton spans, we have:

$$C(p, l_p, g, l_g, |g|) \quad (15)$$
$$= [p = g \wedge l_p = l_g] \quad (16)$$

Thus, we have, for any singleton span s and any label l:

$$[(s^0, l^0) \in P] \quad (17)$$
$$= [\exists (s, l) \in P : s = s^0 \wedge l = l^0] \quad (18)$$
$$= [\exists (s, l) \in P : C(s, l, s^0, l^0, |s^0|) = 1] \quad (19)$$
$$= \max_{(s,l)2P} C(s, l, s^0, l^0, |s^0|) \quad (20)$$

This entails that the number of true positives (TP) is

$$|\{(s^0, \{l^0\}) \in G \mid (s^0, l^0) \in P\}| \quad (21)$$
$$= \sum_{(s^0,fl^0g)2G} [(s^0, l^0) \in P] \quad (22)$$
$$= \sum_{(s^0,fl^0g)2G} \max_{(s,l)2P} C(s, l, s^0, l^0, |s^0|) \quad (23)$$

The standard recall is the ratio of true positives (TP) out of the sum of true positives and false negatives (FN):

$$\text{Standard Recall} \quad (24)$$
$$= \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (25)$$

With $|G| = \text{TP} + \text{FN}$ and Equation 23, this is equivalent to

$$= \frac{\sum_{(s,flg)2G} \max_{(s^0,l^0)2P} C(s, l, s^0, l^0, |s|)}{|G|}$$

□

Proposition I.3. *In cases where a system predicts multiple labels for a single span, and the corresponding gold standard also contains multiple alternative labels for that span, the system's recall does not increase with the number of correctly predicted labels. Our recall formula ensures that multiple predictions for the same span do not artificially inflate the recall metric.*

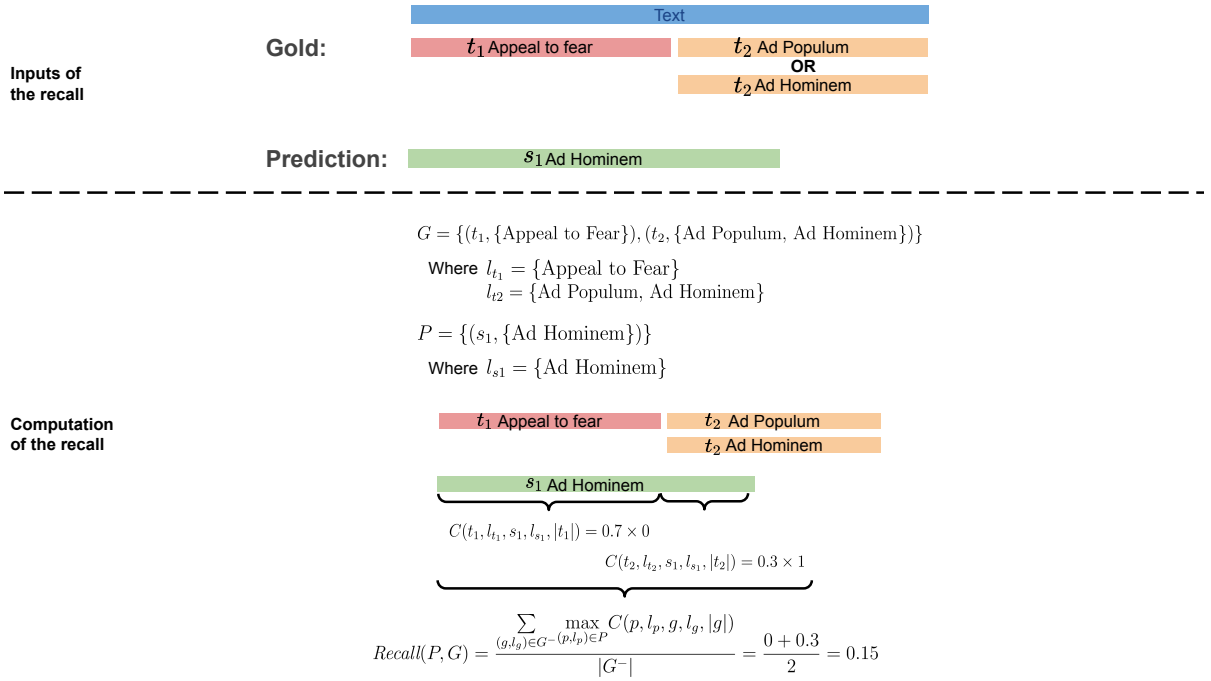


Figure 6: Example of Recall computation with alternatives.

Proof: By our definition, we have, for any prediction P and any gold standard G:

$$Recall(P,G) = \frac{\sum_{(g,l_g) 2 G} \max_{(p,l_p) 2 P} C(p, l_p, g, l_g, |g|)}{|G \quad |} \tag{26}$$

By definition, we have, for any spans p, g, and for any sets l_p, l_g:

$$C(p, l_p, g, l_g, |g|) \tag{27}$$
$$= \frac{|p \cap g|}{|g|} \times \delta(l_p, l_g) \tag{28}$$

If p and g are the same spans, this boils down to

$$= \delta(l_p, l_g) \tag{29}$$
$$= [l_p \cap l_g \neq \emptyset] \tag{30}$$

The max operation ensures that the score contribution for the span in the recall is based on the best match between predicted labels and gold standard labels, capped at 1 regardless of the number of labels correctly predicted. So even if multiple labels in l_p for p in P match with different alternative labels in l_g in G, the contribution to the recall for the span remains 1.

$$\max(C(p1, l_{p1}, g, l_g, |g|), \tag{31}$$
$$C(p2, l_{p2}, g, l_g, |g|), \tag{32}$$
$$, ..., \tag{33}$$
$$C(pn, l_{pn}, g, l_g, |g|)) \tag{34}$$
$$= 1 \tag{35}$$

□

I.1 Metric equivalence

The metric proposed in this paper is similar to the metric proposed in (Martino et al., 2019). This metric supposes that there is no overlap of spans with the same label. However, such spans are very frequent in a multi-level taxonomy, when evaluating Levels 0 and 1. Consider the following example:

Example I.1

You are a liar. Therefore you are wrong.

In this example, there is only one *abusive ad hominem*, which is a *Fallacy of Credibility* on Level 1. Now assume that the model outputs: (You are a liar, *abusive ad hominem*), (You are a liar, therefore you are wrong, *tu quoque*). Using the recall from (Martino et al., 2019), and $G = \{([0, 40], \text{CREDIBILITY})\}$, $P = \{([0, 10], \text{CREDIBILITY}), ([0, 40], \text{CREDIBILITY})\}$

we get the following recall:

$$\begin{aligned} Recall_m(P,G) &= \frac{1}{|G|}\sum_{p \in P, g \in G} C_m(p,g,|g|) \\ &= \frac{1}{1} * (\frac{10}{40} * 1 + \frac{40}{40} * 1) \\ &= 1.25 \end{aligned}$$

Instead of computing one score for each element of G as it would be expected for the recall, the metrics is computing all scores between all spans with the same label. We thus get a score larger than one.

Hence, we propose to sum only the best match for each element of G.

$$\begin{aligned} Recall(P,G) &= \frac{\sum_{(g,l_g) \in G} \max_{(p,l_p) \in P} C(p,l_p,g,l_g,|g|)}{|G|} \\ &= \frac{1}{1} * (max(\frac{10}{40}, \frac{40}{40})) \\ &= 1 \end{aligned}$$

Proposition I.4. *Given a gold standard G, where for each span there are no alternatives, and there is only one span from a prediction P that overlaps with one span from the gold standard, our recall metric $Recall(P,G)$ coincides with Martino et al. (2019)'s $Recall_m(P,G)$.*

Proof: Given a gold standard G with no alternatives so:

$$G = G \tag{36}$$

By definition, we have, for any spans p, g, and for any sets l_p, l_g:

$$\begin{aligned} & C(p,l_p,g,l_g,|g|) && (37)\\ = & \frac{|p \cap g|}{|g|} \times \delta(l_p,l_g) && (38)\\ = & \frac{|p \cap g|}{|g|} \times [l_p = l_g] && (39) \end{aligned}$$

In case of Martino et al. (2019)'s comparison score C_m, we have:

$$\begin{aligned} & C_m(p,g,|g|) && (40)\\ = & \frac{|p \cap g|}{|g|} \times \delta(l(p),l(g)) && (41) \end{aligned}$$

In Martino et al. (2019)'s comparison score C_m, l represents a labeling function, so in this case:

$$\begin{aligned} & C_m(p,g,|g|) && (42)\\ = & \frac{|p \cap g|}{|g|} \times \delta(l(p),l(g)) && (43)\\ = & \frac{|p \cap g|}{|g|} \times [l_p = l_g] && (44)\\ = & C(p,l_p,g,l_g,|g|) && (45) \end{aligned}$$

For a given label, there is either no prediction span that overlaps or a unique prediction span p that overlaps with a gold standard span g, such that their overlap and label agreement is non-zero: If there is no span p in P that overlaps, then both recalls are equal to zero. The other case is:

$$\begin{aligned} & \forall (g,l) \in G, \exists! (p,l) \in P : \frac{|p \cap g|}{|g|} \times [l = l] && (46)\\ = & C(p,l,g,l,|g|) && (47)\\ = & C_m(p,g,|g|) && (48)\\ > & 0 && (49) \end{aligned}$$

This implies that for each gold standard annotation, the maximum score of C between each annotation in the gold standard and the annotations in the prediction is achieved exactly once:

$$\begin{aligned} & \forall (g,l_g) \in G, \max_{(p,l_p) \in P} C(p,l_p,g,l_g,|g|) && (50)\\ & = C_m(p,g,|g|) && (51) \end{aligned}$$

Additionally, for each gold standard annotation, the sum of scores C across all predictions equals the maximum score C, since only one prediction per annotation contributes a non-zero score, so:

$$\begin{aligned} & \forall (g,l_g) \in G, \sum_{(p,l_p) \in P} C(p,l_p,g,l_g,|g|) && (52)\\ & = C_m(p,g,|g|) && (53) \end{aligned}$$

Finally,

$$\begin{aligned} & \frac{\sum_{(p,l_p) \in P, (g,l_g) \in G} C(p,l_p,g,l_g,|g|)}{|G|} && (54)\\ = & \frac{1}{|G|}\sum_{p \in P, g \in G} C_m(p,g,|g|) && (55) \end{aligned}$$

$$\frac{\sum_{(g,l_g)2G} \max_{(p,l_p)2P} C(p, l_p, g, l_g, |g|)}{|G \quad |} \tag{56}$$

$$= \frac{1}{|G|} \sum_{p2P, g2G} C_m(p, g, |g|) \tag{57}$$

$$\tag{58}$$

We can conclude that:

$$Recall_m(P, G) = Recall(P, G) \tag{59}$$

□

A similar demonstration can be done for precision.

There is another difference between our metrics and the original one: If two disjoint spans from the prediction overlap with the gold standard, in (Martino et al., 2019)'s metrics, they both contribute to the score. In our metrics, only the best match contributes to the score. Our metric thus rewards models that output the correct span without splitting it into multiple spans (see Figure 7).

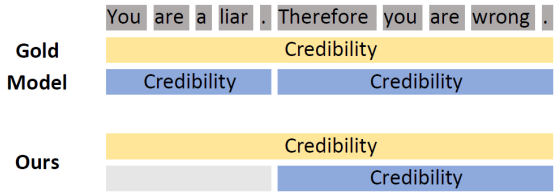


Figure 7: Illustration of the difference between our metric and the one from (Martino et al., 2019). In Martino's metric, both annotated spans count, and we get a recall of 1. Our metric counts only the largest overlapping span (light blue), and gives a recall of 0.6.

J Results

Table 7 displays the F1-scores for our experiments on both the complete dataset and the subset from the user study, across all three task levels, with the best scores for each level highlighted in bold. For more in-depth analysis, Table 8 provides detailed results, including Recall, Precision, and F1-score for Levels 0, 1, and 2 (referenced as Tables 8a, 8b, and 8c, respectively), along with corresponding data from the user study (Table 8d).

K Error Analysis

We conduct an error analysis on two models, GPT-3.5 and Falcon, which exhibit the best and worst performance on Level 2. Our analysis also includes the annotations of the users study. **Our first goal in this analysis is to compare whether the best model has better controlled behavior than the worst model when generating outputs**. The Falcon model identifies 625 fallacious spans, with an average of 4.8 fallacies per span, while the GPT-3.5 model detects only 199 fallacious spans, with an average of 1.07 fallacies per span. However, we have 203 fallacious spans in the gold standard. The distribution of fallacies for the fallacious span at Level 2 for each model is presented in Table 9. Based on our analysis, we have observed that the Falcon model tends to predict multiple fallacies that are irrelevant to a fallacious span. In contrast, the GPT-3.5 model displays a more controlled behavior, which explains why Falcon has a low precision score. It is also worth noting that GPT-3.5 never predicted a span as *tu quoque*. We observe that both models produce nonsensical outputs, such as SQL code like "*select name color order from tag where the name*," or incomplete classification of fallacies such as "*the sentence it's a mistake being considered as part of a fallacious argument.*". Falcon has 115 spans labeled as unknown, while GPT-3.5 has only 5. **Our second goal in this analysis is to analyze the exact matching performance of detecting fallacies and the type of fallacies that models and humans struggle with at Level 1**. Out of the 625 fallacious spans identified by Falcon, only 60 match the gold standard exactly, while out of 199 fallacious spans detected by GPT-3.5, only 55 match the gold standard exactly. Both models struggle mainly with fallacies categorized as fallacies of emotion, as shown in Figure 8.

For the annotators of the user study, we use a small sample of 20 examples with 24 spans. User 2 performs the best with 17 exactly matched spans, while User 4 performs worst with only 8 exactly matched spans. Based on the exact matched results, the analysis of Figure 10 reveals that all the annotators struggle mainly with the fallacies of appeal to emotion. This difficulty can be partly attributed to these fallacies being less prevalent in our sample compared to the other types of fallacies. Interestingly, Users 1 and 3 correctly predict more fallacies of logic. Conversely, Users 2 and 4 correctly predict more fallacies of credibility than the others. It is worth noting that none of the users used all 23 fallacies of the taxonomy during the annotations, as shown in Table 11.

In conclusion, models and humans tend to struggle

	MAFALDA			Sample of MAFALDA		
Model	F1 Level 0	F1 Level 1	F1 Level 2	F1 Level 0	F1 Level 1	F1 Level 2
Baseline random	0.435	0.061	0.010	0.211	0.013	0.004
Falcon 7B	0.397	0.130	0.022	0.274	0.099	0.019
LLaMA2 Chat 7B	0.572	0.114	0.068	0.356	0.065	0.030
LLaMA2 Chat 13B	0.549	0.160	0.096	0.364	0.103	0.043
LLaMA2 7B	0.492	0.148	0.038	0.347	0.145	0.037
LLaMA2 13B	0.458	0.129	0.039	0.309	0.109	0.003
Mistral Instruct 7B	0.536	0.144	0.069	0.404	0.089	0.004
Mistral 7B	0.450	0.127	0.044	0.393	0.102	0.017
Vicuna 7B	0.494	0.134	0.051	0.258	0.061	0.049
Vicuna 13B	0.557	0.173	0.100	0.293	0.121	0.032
WizardLM 7B	0.490	0.087	0.036	0.233	0.036	0.0
WizardLM 13B	0.520	0.177	0.093	0.246	0.123	0.021
Zephyr 7B	0.524	0.192	0.098	0.312	0.109	0.025
GPT 3.5 175B	**0.627**	**0.201**	**0.138**	0.338	0.095	0.034
Avg. Human	-	-	-	**0.749**	**0.352**	**0.186**

Table 7: Performance results of different models across different granularity levels in a zero-shot setting. The right part concerns only the user study with a subsample of 20 texts from MAFALDA. The best results for each level are highlighted in bold.

more with fallacies of appeal to emotion, which could be expected since not every expression of emotion is necessarily a fallacy. The difficulty of the task lies in distinguishing valid arguments accompanied by emotions from fallacious arguments. This is supported by Figures 8 and 10. Despite the underrepresentation of the fallacies of the appeal to emotion in our user study sample, our findings indicate that humans often fail to exactly identify the specific fallacious spans classified under appeal to emotion fallacies. Moreover, even when humans correctly identify such fallacious spans, they are frequently misclassified. In contrast, models tend more to find these fallacious spans although they, too, frequently misclassify them. The only instances where the models can correctly predict the fallacious spans and their labels are when they involve an *appeal to ridicule* or an *appeal to a worse problem*. These cases can be observed in Figures 9.

Model	Precision Level 0	Recall Level 0	F1 Level 0
Falcon 7B	0.427	0.655	0.397
LLAMA2 Chat 7B	0.506	0.837	0.572
LLAMA2 7B	0.456	0.758	0.492
Mistral Instruct 7B	0.570	0.651	0.536
Mistral 7B	0.444	0.691	0.450
Vicuna 7B	0.529	0.628	0.494
WizardLM 7B	0.565	0.567	0.490
Zephyr 7B	0.489	0.765	0.524
LLaMA2 Chat 13B	0.493	0.793	0.549
LLaMA2 13B	0.433	0.739	0.458
Vicuna 13B	0.591	0.670	0.557
WizardLM 13B	0.523	0.756	0.520
GPT 3.5 175B	0.701	0.669	0.627

(a) Performance results for Level 0 on MAFALDA

Model	Precision Level 1	Recall Level 1	F1 Level 1
Falcon 7B	0.134	0.164	0.130
LLAMA2 Chat 7B	0.134	0.136	0.114
LLAMA2 7B	0.158	0.185	0.148
Mistral Instruct 7B	0.176	0.152	0.144
Mistral 7B	0.136	0.159	0.127
Vicuna 7B	0.161	0.146	0.134
WizardLM 7B	0.121	0.093	0.087
Zephyr 7B	0.207	0.230	0.192
LLaMA2 Chat 13B	0.173	0.183	0.160
LLaMA2 13B	0.140	0.151	0.129
Vicuna 13B	0.200	0.191	0.173
WizardLM 13B	0.193	0.205	0.177
GPT 3.5 175B	0.233	0.203	0.201

(b) Performance results for Level 1 on MAFALDA

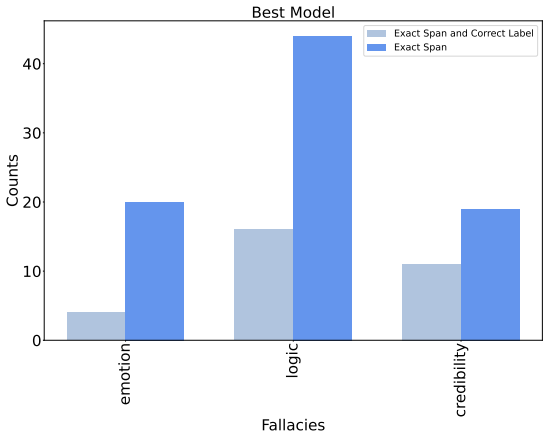
Model	Precision Level 2	Recall Level 2	F1 Level 2
Falcon 7B	0.016	0.078	0.022
LLAMA2 Chat 7B	0.070	0.095	0.068
LLAMA2 7B	0.038	0.073	0.038
Mistral Instruct 7B	0.086	0.076	0.069
Mistral 7B	0.046	0.072	0.044
Vicuna 7B	0.062	0.067	0.051
WizardLM 7B	0.056	0.041	0.036
Zephyr 7B	0.090	0.145	0.098
LLaMA2 Chat 13B	0.101	0.122	0.096
LLaMA2 13B	0.037	0.068	0.039
Vicuna 13B	0.115	0.118	0.100
WizardLM 13B	0.088	0.134	0.093
GPT 3.5 175B	0.162	0.138	0.138

(c) Performance results for Level 2 on MAFALDA

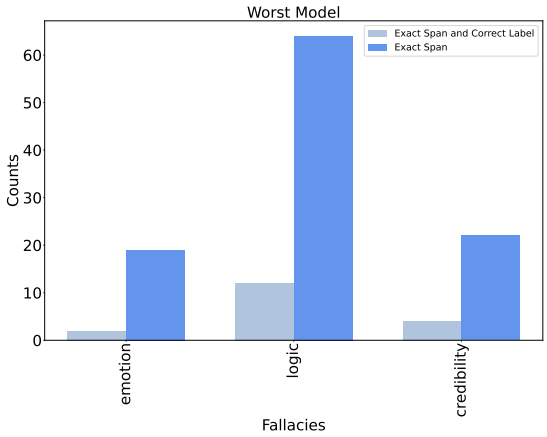
	Model	Precision	Recall	F1
Level 0	user1	0.732	0.847	0.760
	user2	0.785	0.892	0.821
	user4	0.728	0.809	0.728
	user5	0.704	0.767	0.694
	Average	0.737	0.829	0.749
Level 1	user1	0.326	0.342	0.322
	user2	0.399	0.402	0.397
	user4	0.311	0.364	0.319
	user5	0.375	0.394	0.371
	Average	0.353	0.376	0.352
Level 2	user1	0.192	0.248	0.204
	user2	0.162	0.172	0.164
	user4	0.186	0.239	0.194
	user5	0.170	0.211	0.180
	Average	0.177	0.217	0.186

(d) Performances results for the user Study

Table 8: Detailed results of the experiments, including Recall, Precision, and F1-score, for each level, for models and each user of the user study.



(a) Accuracy of fallacy labeling for spans that exactly match the gold standard at Level 1 for the **best model**



(b) Accuracy of fallacy labeling for spans that exactly match the gold standard at Level 1 for the **worst model**

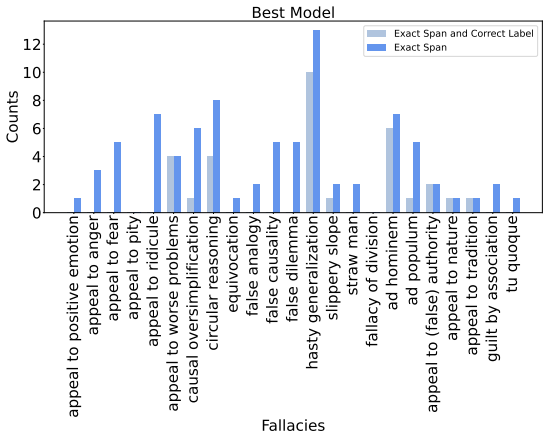
Figure 8: Accuracy of fallacy labeling for spans that **exactly match** the gold standard at Level 1 for the best and worst models. *Exact Span* corresponds to the number of spans correctly identified by the model, *Exact Span and Correct Label* corresponds to the number of correctly labeled spans out of the correctly identified spans.

Fallacy Type	Best Model	Worst Model	Gold Standard
Appeal to Positive Emotion	3	128	3
Appeal to Anger	6	119	6
Appeal to Fear	5	132	11
Appeal to Pity	1	198	1
Appeal to Ridicule	10	121	20
Appeal to Worse Problems	21	157	8
Causal Oversimplification	6	81	23
Circular Reasoning	8	132	11
Equivocation	1	106	7
False Analogy	6	127	12
False Causality	9	57	13
False Dilemma	6	169	18
Hasty Generalization	41	123	28
Slippery Slope	10	77	11
Straw Man	6	135	13
Fallacy of Division	2	102	2
Ad Hominem	32	135	16
Ad Populum	4	75	14
Appeal to (False) Authority	10	211	9
Appeal to Nature	7	143	11
Appeal to Tradition	4	156	6
Guilt by Association	4	91	8
Tu Quoque	0	111	3
Unknown	5	115	-

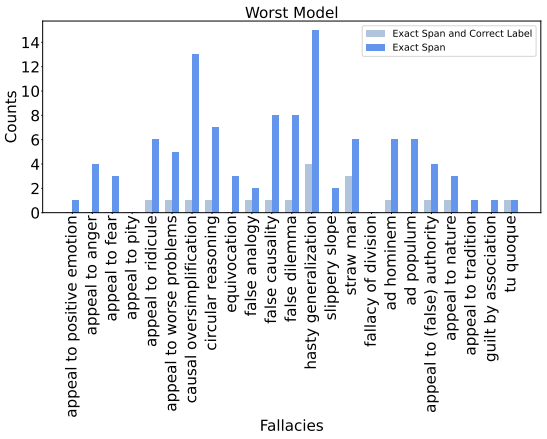
Table 9: Fallacy distribution at Level 2 of the Gold standard, Best model and Worst model

Fallacy Type	Best Model	Worst Model	Gold Standard
Emotion	46	855	49
Logic	95	1109	138
Credibility	61	922	67
Unknown	5	115	0

Table 10: Fallacy distribution at Level 1 of the Gold standard, Best model and Worst model



(a) Accuracy of fallacy labeling for spans that exactly match the gold standard at Level 2 for the **best model**

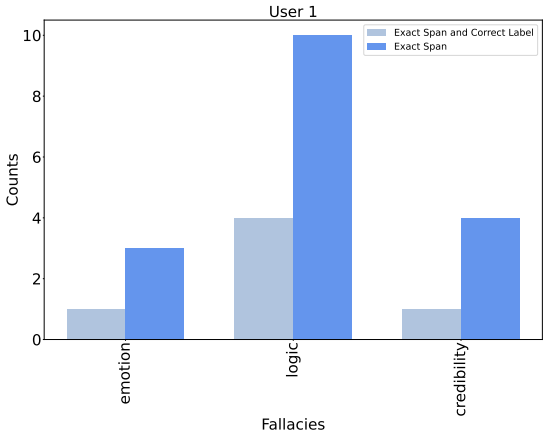


(b) Accuracy of fallacy labeling for spans that exactly match the gold standard at Level 2 for the **worst model**

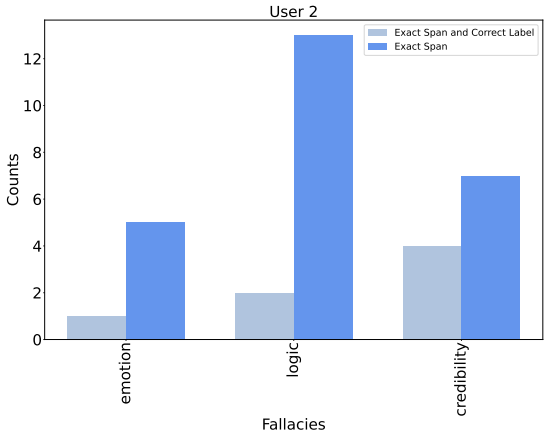
Figure 9: Accuracy of fallacy labeling for spans that **exactly match** the gold standard at Level 2 for the best and worst models. *Exact Span* corresponds to the number of spans correctly identified by the model, *Exact Span and Correct Label* corresponds to the number of correctly labeled spans out of the correctly identified spans.

Fallacy Type	User 1	User 2	User 3	User 4	Sample Gold Standard
Appeal to Positive Emotion	2	0	0	2	0
Appeal to Anger	1	0	0	0	0
Appeal to Fear	1	1	0	2	0
Appeal to Pity	0	0	0	0	0
Appeal to Ridicule	8	1	1	4	5
Appeal to Worse Problems	3	0	0	0	1
Causal Oversimplification	2	2	1	4	2
Circular Reasoning	2	0	2	0	1
Equivocation	1	0	0	5	1
False Analogy	1	1	1	0	0
False Causality	3	4	2	1	2
False Dilemma	1	1	0	1	2
Hasty Generalization	4	2	3	5	3
Slippery Slope	1	1	0	7	1
Straw Man	2	5	0	0	3
Fallacy of Division	3	0	0	0	0
Ad Hominem	4	1	3	2	4
Ad Populum	3	1	0	5	1
Appeal to (False) Authority	0	2	1	3	1
Appeal to Nature	0	1	0	0	0
Appeal to Tradition	1	1	1	2	2
Guilt by Association	1	3	1	1	1
Tu Quoque	0	1	2	0	0
Unknown	0	0	0	0	0

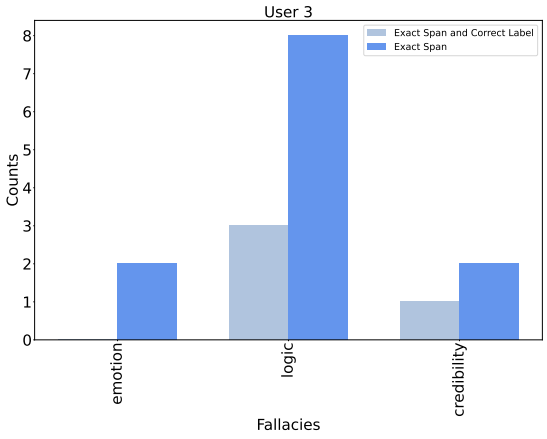
Table 11: Fallacies distribution at Level 2 of User 1, User 2, User 3, User 4, and the sample gold standard



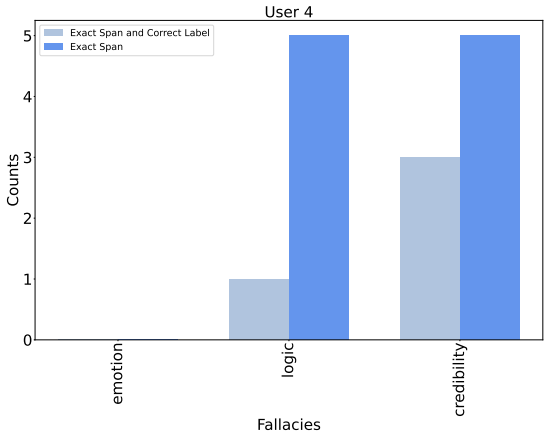
(a) Accuracy of fallacy labeling for spans that exactly match the gold standard at Level 1 for the **User 1**



(b) Accuracy of fallacy labeling for spans that exactly match the gold standard at Level 1 for the **User 2**

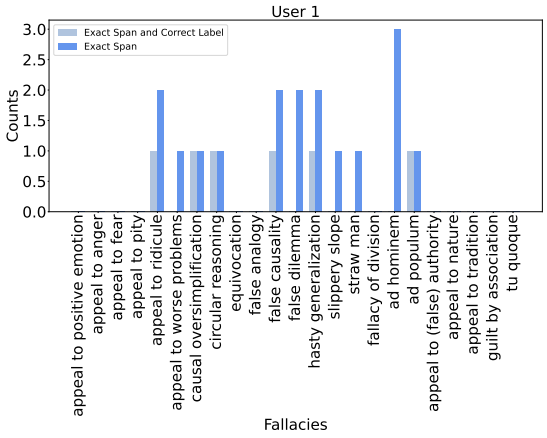


(c) Accuracy of fallacy labeling for spans that exactly match the gold standard at Level 1 for the **User 3**

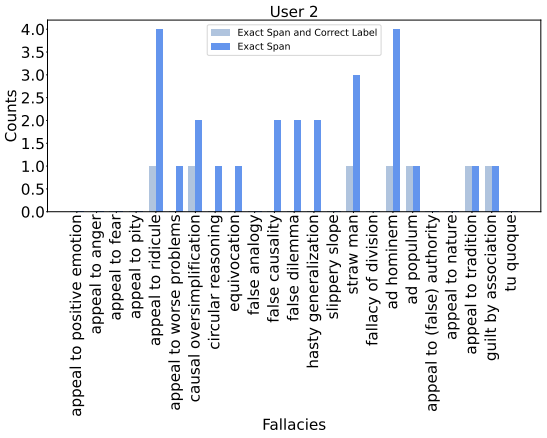


(d) Accuracy of fallacy labeling for spans that exactly match the gold standard at Level 1 for the **User 4**

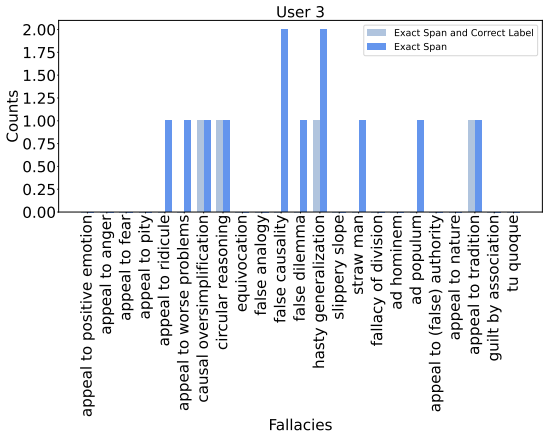
Figure 10: Accuracy of fallacy labeling for spans that **exactly match** the gold standard at Level 1 for the Users' annotations. *Exact Span* corresponds to the number of spans correctly identified by the user, *Exact Span and Correct Label* corresponds to the number of correctly labeled spans out of the correctly identified spans.



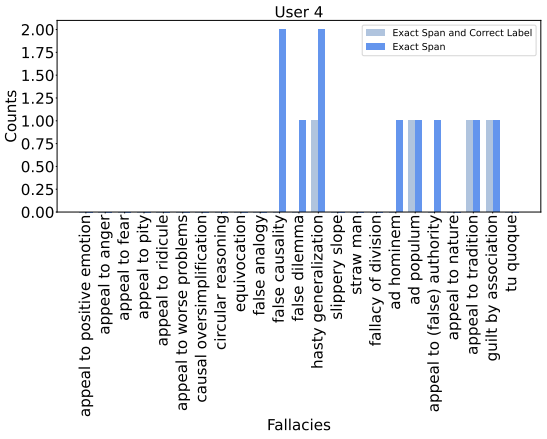
(a) Accuracy of fallacy labeling for spans that exactly match the gold standard at Level 2 for the **User 1**



(b) Accuracy of fallacy labeling for spans that exactly match the gold standard at Level 2 for the **User 2**



(c) Accuracy of fallacy labeling for spans that exactly match the gold standard at Level 2 for the **User 3**



(d) Accuracy of fallacy labeling for spans that exactly match the gold standard at Level 2 for the **User 4**

Figure 11: Accuracy of fallacy labeling for spans that **exactly match** the gold standard at Level 2 for the Users' annotations. *Exact Span* corresponds to the number of spans correctly identified by the user, *Exact Span and Correct Label* corresponds to the number of correctly labeled spans out of the correctly identified spans.

L Edge Cases of the Metrics

In this section, we show how our metrics handle edge cases.

Table 12: The model predicts at least one correct label

	Spans	Labels		
Gold	Lorem ipsum dolor sit amet. Ut enim ad minim veniam. Sed do eiusmod tempor incididunt.	$l1, \perp$ $l2$ $l3$		
Case	Prediction	Label	Recall	Precision
0.1	Ut enim ad minim veniam.	$l2$	0.5	1
0.2	Lorem ipsum dolor sit amet. Ut enim ad minim veniam.	$l1$ $l2$	0.5	1
0.3	Ut enim ad minim veniam. Sed do eiusmod tempor incididunt.	$l2$ $l4$	0.5	0.5
0.4	Lorem ipsum dolor sit amet. Ut enim ad minim veniam. Sed do eiusmod tempor incididunt.	$l1$ $l2$ $l4$	0.5	0.666

Table 13: The gold standard has only one span, which contains a "no fallacy" as an alternative

	Spans	Labels	Recall	Precision
Gold	Lorem ipsum dolor sit amet.	$l1, \perp$		
Case	Prediction	Label		
1.1	Lorem ipsum dolor sit amet.	$l1$	1	1
1.2	Lorem ipsum dolor sit amet.	$l3$	1	0
1.3	-	-	1	1
1.4	Ut enim ad minim veniam.	$l1$	1	0
1.5	Lorem ipsum dolor sit amet. Ut enim ad minim veniam.	$l3$	1	0

Table 14: The gold standard does not contain a "no fallacy"

	Spans	Labels	Recall	Precision
Gold	Lorem ipsum dolor sit amet.	$l1$		
Case	Prediction	Label		
2.1	Lorem ipsum dolor sit amet.	$l1$	1	1
2.2	Lorem ipsum dolor sit amet.	$l3$	0	0
2.3	Ut enim ad minim veniam.	$l1$	0	0
2.4	-	-	0	1

Table 15: The gold standard contains no fallacious span

	Spans	Labels	Recall	Precision
Gold	-	-		
Case	Prediction	Label		
3.1	Lorem ipsum dolor sit amet.	$l1$	1	0
3.2	-	-	1	1

Table 16: The gold standard contains a “no fallacy” and a required fallacy

	Spans	Labels	Recall	Precision
Gold	Lorem ipsum dolor sit amet. Ut enim ad minim veniam.	$l1, \perp$ $l2$		
Case	Prediction	Label		
4.1	Lorem ipsum dolor sit amet.	$l1$	0	1
4.2	Lorem ipsum dolor sit amet.	$l3$	0	0
4.3	Ut enim ad minim veniam.	$l2$	1	1
4.4	Ut enim ad minim veniam.	$l3$	0	0

Table 17: The gold standard spans across 2 sentences

	Spans	Labels	Recall	Precision
Gold	Lorem ipsum dolor sit amet. Ut enim ad minim veniam.	$l1, \perp$		
Case	Prediction	Label		
5.1	Lorem ipsum dolor sit amet. Ut enim ad minim veniam.	$l1$	1	1
5.2	-	-	1	1
5.3	Lorem ipsum dolor sit amet.	$l1$	1	1
5.4	Lorem ipsum dolor sit amet. Ut enim ad minim veniam.	$l3$	1	0
5.5	Ut enim ad minim veniam.	$l3$	1	0

Table 18: The gold standard spans across 2 sentences and there is overlap

	Spans	Labels	Recall	Precision
Gold	Lorem ipsum dolor sit amet. Ut enim ad minim veniam. Ut enim ad minim veniam.	$l1, \perp$ $l2$		
Case	Prediction	Label		
6.1	Lorem ipsum dolor sit amet. Ut enim ad minim veniam. Ut enim ad minim veniam.	$l1$ $l2$	1	1
6.2	Ut enim ad minim veniam.	$l2$	1	1
6.3	Lorem ipsum dolor sit amet. Ut enim ad minim veniam.	$l1$	0	1
6.4	-	-	0	1
6.5	Lorem ipsum dolor sit amet. Ut enim ad minim veniam.	$l1$ $l2$	1	1
6.6	Lorem ipsum dolor sit amet.	$l1$	0	1
6.7	Ut enim ad minim veniam.	$l3$	0	0

Table 19: Two gold standard spans overlap

	Spans	Labels	Recall	Precision
Gold	Lorem ipsum dolor sit amet. Ut enim ad minim veniam. Ut enim ad minim veniam.	$l1$ $l2$		
Case	Prediction	Label		
7.1	Lorem ipsum dolor sit amet. Ut enim ad minim veniam. Ut enim ad minim veniam.	$l1$ $l2$	1	1
7.2	Ut enim ad minim veniam.	$l2$	0.5	1
7.3	Lorem ipsum dolor sit amet. Ut enim ad minim veniam.	$l1$	0.5	1
7.4	-	-	0	1
7.5	Lorem ipsum dolor sit amet. Ut enim ad minim veniam.	$l1$ $l2$	0.75	1
7.6	Lorem ipsum dolor sit amet.	$l1$	0.25	1
7.7	Ut enim ad minim veniam.	$l3$	0	0

Table 20: Two labels have the same Level 0 or Level 1 fallacy category

	Spans	Labels	Recall	Precision
Gold	Lorem ipsum dolor sit amet. Ut enim ad minim veniam. Ut enim ad minim veniam.	fallacy (l1) fallacy (l2)		
Case	Prediction	Label		
8.1	Lorem ipsum dolor sit amet. Ut enim ad minim veniam. Ut enim ad minim veniam.	fallacy fallacy	1	1
8.2	Ut enim ad minim veniam.	fallacy	0.75	1
8.3	Lorem ipsum dolor sit amet. Ut enim ad minim veniam.	fallacy	1	1
8.4	Ut enim ad minim veniam. Ut enim ad minim veniam.	fallacy fallacy (duplicate)	0.75	0.5
8.5	Lorem ipsum dolor sit amet. Ut enim ad minim veniam.	fallacy fallacy	0.75	1
8.6	Lorem ipsum dolor sit amet.	fallacy (l1)	0.25	1
8.7	Lorem ipsum dolor sit amet.	fallacy (l2)	0.25	1

Table 21: Two labels have the same Level 0 or Level 1 fallacy category with an alternative “no fallacy”

	Spans	Labels	Recall	Precision
Gold	Ut enim ad minim veniam. Ut enim ad minim veniam.	fallacy (l1), $\perp$ fallacy (l2)		
Case	Prediction	Label		
9.1	Ut enim ad minim veniam. Ut enim ad minim veniam.	fallacy fallacy (duplicate)	1	1
9.2	Ut enim ad minim veniam.	fallacy (l1)	1	1
9.3	Ut enim ad minim veniam.	fallacy (l2)	1	1
9.4	Lorem ipsum dolor sit amet. Ut enim ad minim veniam.	fallacy	1	0.5
9.5	-	-	0	1
9.6	Lorem ipsum dolor sit amet.	fallacy	0	0

Table 22: The same obligatory fallacious span has different labels

	Spans	Labels	Recall	Precision
Gold	Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet.	$l1$ $l2$		
Case	Prediction	Label		
10.1	Lorem ipsum dolor sit amet.	$l1$	0.5	1
10.2	Lorem ipsum dolor sit amet.	$l3$	0	0
10.3	Ut enim ad minim veniam.	$l1$	0	0
10.4	-	-	0	1
10.5	Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet.	$l1$ $l2$	1	1

Table 23: The same fallacious span has two labels and a "no fallacy" alternative

	Spans	Labels	Recall	Precision
Gold	Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet.	$l1, \perp$ $l2$		
Case	Prediction	Label		
11.1	Lorem ipsum dolor sit amet.	$l1$	1	1
11.2	Lorem ipsum dolor sit amet.	$l3$	0	0
11.3	Ut enim ad minim veniam.	$l1$	0	0
11.4	-	-	0	1
11.5	Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet.	$l1$ $l2$	1	1