Summairnes

1. https://en.wikipedia.org/wiki/Wikipedia:Wikipedia_Signpost/2023-01-01/Recent_research

Graham's Hierarchy of Disagreement (сверху-вниз)

*"[English] Wikipedia recommends the hierarchy of disagreement formulated by Graham (2008) as a **guide for constructive dispute resolution**. Graham's hierarchy posits that there are **seven levels of disagreement**, ranging from namecalling (at the bottom) to refuting the central point. [...] Despite its popularity, this hierarchy has not been verified empirically."*

☐ Опровержение главного суждения [очевидная несостоятельность основного суждения]

☐ Ложность [поиск ошибок и обоснование, почему это ошибочно использованная цитата]

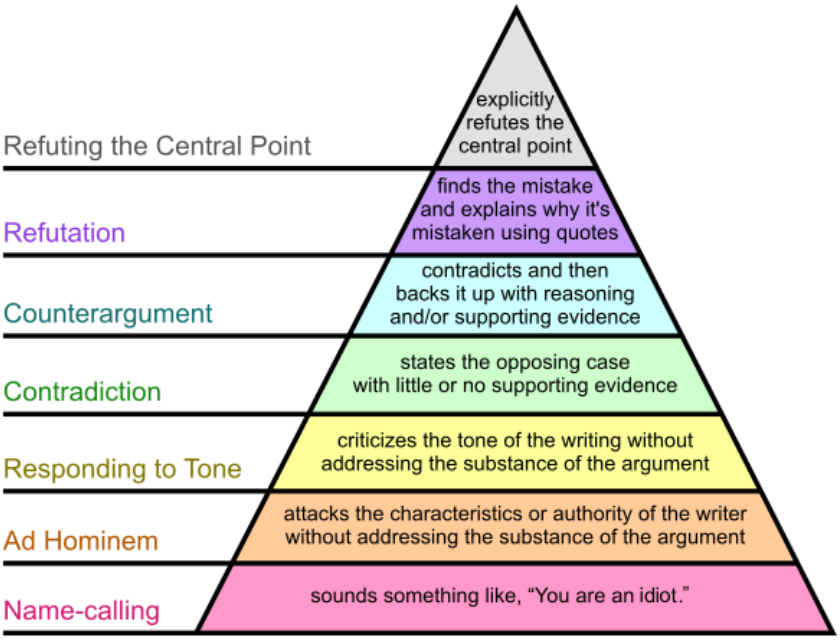
☐ Контраргументация [противоречие и затем снова приведение/поддержка аргумента]

☐ Противоречие [приведение противоположного случая с малой или же без подтверждающей аргументации]

☐ Суждение “книги по обложке” [критика формы высказывания и её подачи, но не сути]

☐ Переход на личности [осуждение самой личности автора высказывание, а не сути последнего]

☐ Оскорбления [”А ты дурак”]



2. https://github.com/christinedekock11/wikitactics

- *Dataset consists of **a list of dicts**, **each representing a disagreement.** Each disagreement dict has 4 keys:* `conv_id`, `utterances`, `split` *and* `escalation_label`.
- Датасет из предложений, выражающих несогласие. Каждый аргумент представления в виде словаря с 4 ключами, где `conv_id` - что-то типа индекса аргумента/блока аргументов по одной теме; `utterances` - непосредственно сами высказывания (которые тоже в свою очередь представляют словари из набора признаков); `split` - как я поняла окончание блока `conv_id`, обозначаемое "split": "train"; `escalation_label` - уровень конструктивности аргумента по **шкале De Kock и Vlachos.**

3. https://arxiv.org/pdf/1911.11408.pdf

▼ **A Large-scale Dataset for Argument Quality Ranking**

▼ Recently, IBM introduced **Project Debater**, the first AI system able to debate humans on complex topics. The system participated in a live debate against a world champion debater, and was able to mine arguments, use them for composing a speech supporting its side of the debate, and also rebut its human competitor. The underlying technology is intended to enhance decision-making.

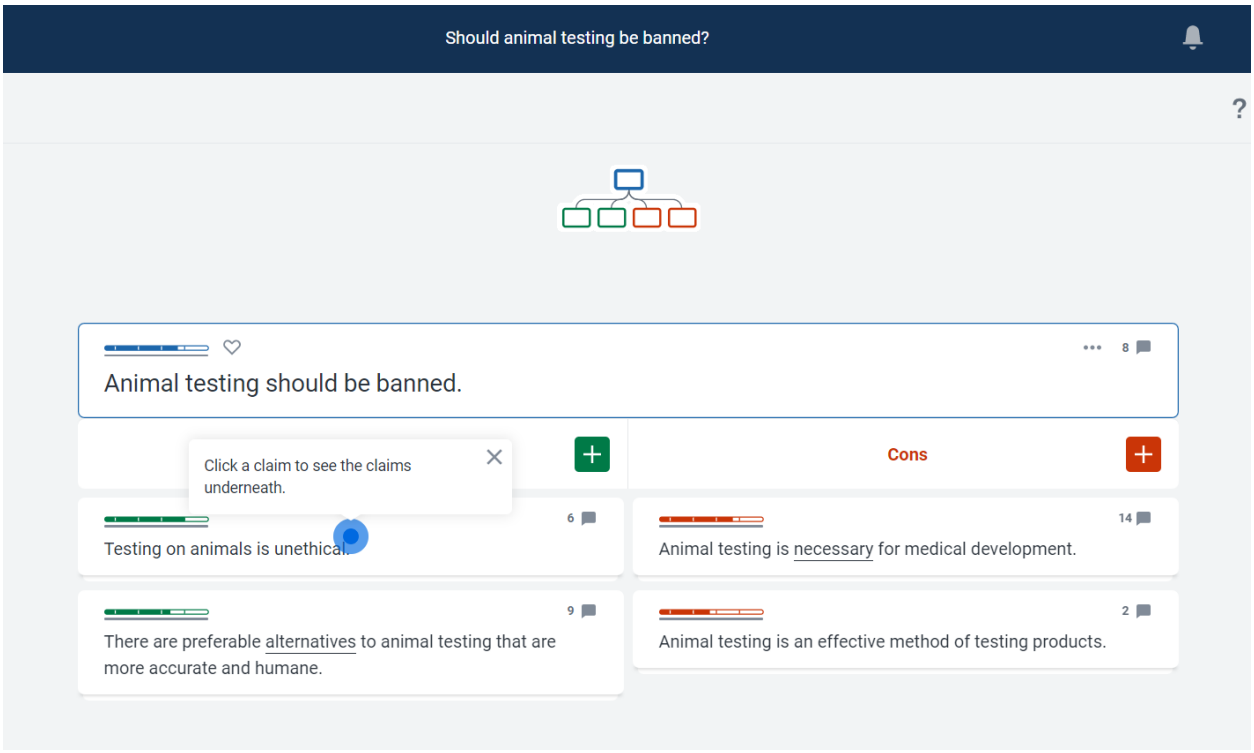
▼ "More recently, IBM also introduced **Speech by Crowd**, a service which supports the collection of free-text arguments from large audiences on debatable topics to

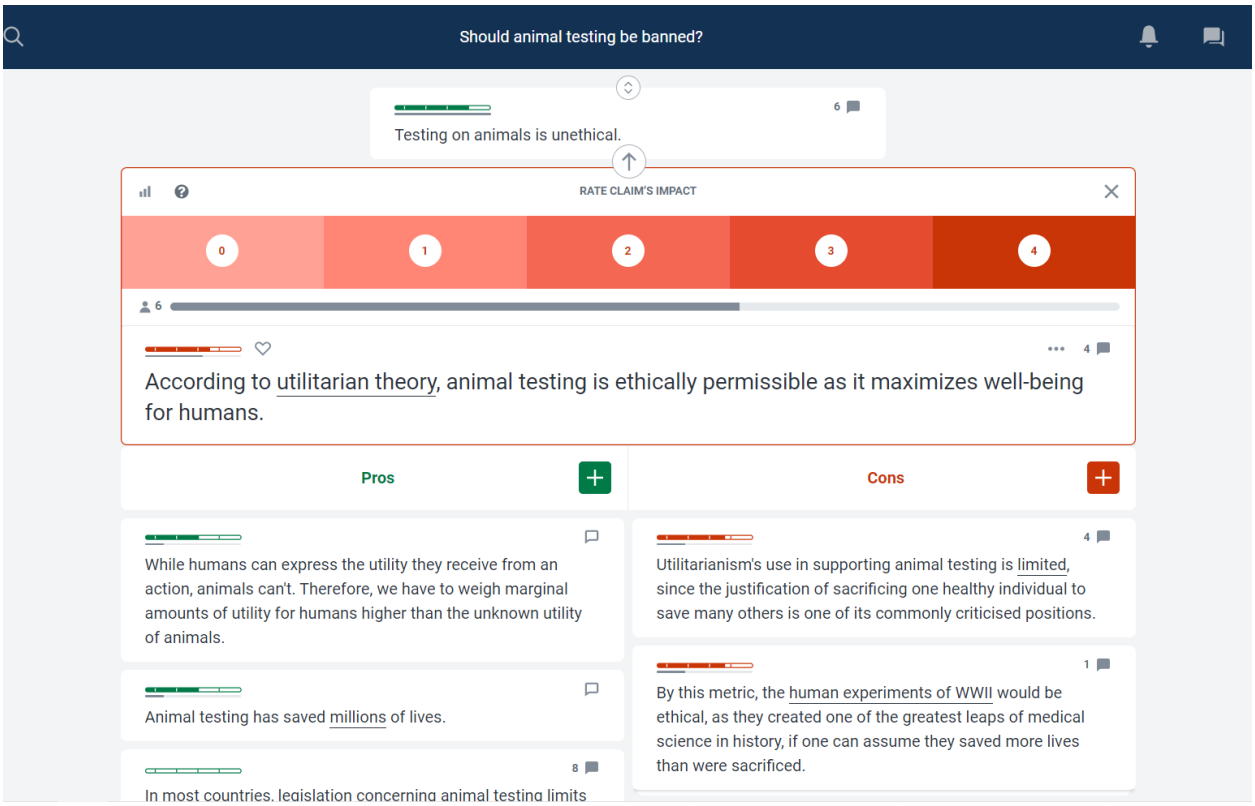
generate meaningful narratives. An important sub-task of this service is automatic assessment of argument quality, which is the focus of the present work. Detecting argument quality is a prominent task due to its importance in automated decision making...”

Argument	Topic	Label
the interest rates are too high and trap people in debt	Payday loans should be banned	1
racial profiling unfairly targets minorities and the poor	We should end racial profiling	1
we should subsidize student loans for reach excelent education	We should subsidize student loans	0.05
i think the same as you, they should ban	Payday loans should be banned	0.09

▼ Swanson, Ecker, and Walker (2015) approach argument quality as a point-wise ranking task, with the goal of selecting argument segments that clearly express an argument facet in a given dialogue. **Arguments are labeled by a real value in the range of [0, 1]**, where a score of 1 indicates that an argument can **be easily interpreted**. They then develop an **automatic regression method** using these labels. Their corpus, which we refer to henceforth as SwanRank, contains 5.3k labeled arguments.

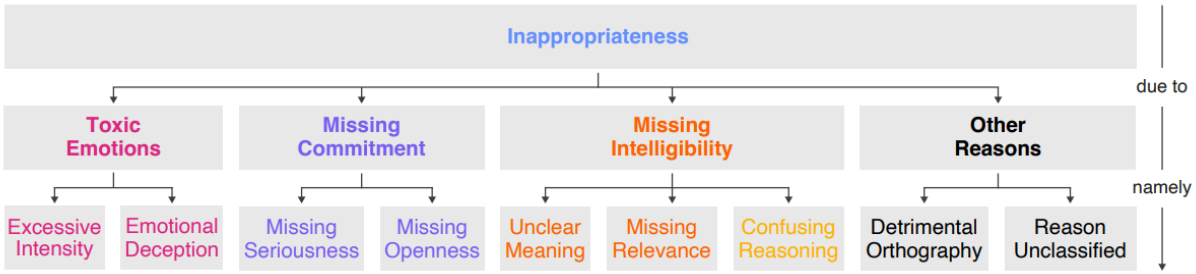
▼ Durmus, Ladhak, and Cardie (2019) present a new dataset comprised of over 47k claims in 471 topics from the website kialo.com, aimed at **evaluating the effect of pragmatic** and discourse context when determining argument quality. They propose models to **predict the impact value of each claim**, as determined by the users of the website. Their dataset is somewhat different from ours as it focuses on **argument impact**, rather than overall quality, and doing so in the context of an argumentative structure, instead of independently. In addition, their impact values are based on spontaneous input from users of the website, whereas our dataset was carefully annotated with clear guidelines. Still, it further highlights the importance of this field.





4. https://arxiv.org/pdf/2305.14935.pdf

- From these, we derive a new taxonomy of 14 dimensions that determine inappropriate language in online discussions. Building on three argument quality corpora, we then create a corpus of 2191 arguments annotated for the **14 dimensions**.
- These include the four subdimensions of rhetorical effectiveness (besides appropriateness), namely, **credibility** (.49), **emotional appeal** (.30), **clarity** (.45), and **arrangement** (.48), as well as **local acceptability** (.54) (sub-dimension of logical cogency) and **global acceptability** (.59)
(sub-dimension of dialectical reasonableness)
- While our goal is to model appropriate language in argumentation, we decided to define when an argument is not appropriate (as indicated above) to maintain freedom of speech as much as possible. Therefore, we define the **four core dimensions** and their sub-dimensions from Figure 3 in a “reverse” way, clarifying what is considered inappropriate:



- In particular, we collected all 2191 arguments on 1154 unique issues from existing corpora (Habernal and Gurevych, 2016b; Wachsmuth et al., 2017b; Ng et al., 2020).2 All corpora are used in research on **argument quality assessment** (Habernal and Gurevych, 2016a; Wachsmuth and Werner, 2020; Lauscher et al., 2020) and contain annotations that we identified as related to appropriateness… The corpus includes arguments of three genres, **1590 from debate portals, 500 from question answering forums, and 101 reviews.**
- In line with Table 1, we treat all annotations as **binary labels**. We performed five repetitions of 5-fold **cross-validation** (25 folds in total) and ensured a similar distribution of the labels in each fold. For each folding, we used 70% for training, 10% for selecting the best-performing approach in terms of the mean macro-F1 score, and 20% for testing.

5. https://arxiv.org/pdf/2301.09992.pdf

▼ Work in computational models for fallacy recognition is still in its infancy, with a limited set of relatively small datasets such as ARGOTARIO (Habernal et al., 2017), which consists of **question and answer dialog moves**; **name-calling** in social media debates (Habernal et al., 2018), fallacies as propaganda techniques in news (Da San Martino et al., 2019b); **logical fallacies** from educational websites (Jin et al., 2022), and fallacies used for **misinformation** in social media and news around **Covid-19** (Musi et al., 2022). Table 1, shows some examples of fallacies from these datasets.

▼ Fallacy recognition is a challenging task for **three main reasons**: **i)** the number of classification labels (fallacy types) and class imbalance in existing datasets is often very high; **ii)** existing datasets cover varying genres and are typically very small in size due to annotation challenges; and **iii)** models trained on individual data sets often show poor out of distribution generalization.

▼ Based on this success, we propose **a unified model based on multitask instruction-based prompting** using T5 (Raffel et al., 2020) to solve the above challenges **for fallacy recognition** (Section 3).

▼ We experiment with **five datasets** (4 existing and a new dataset) that cover **28 unique fallacy types** in multiple domains (e.g., covid-19, climate change, politics) and genres (e.g. news articles, QA turns in dialog, social media).

▼ Their scheme include five fallacy types: **Ad Hominem, Appeal to Emotion, Red Herring, Hasty Generalization, Irrelevant Authority.** We focus on 15 that are fallacies and frequent enough in the data: **Loaded Language, Name Calling or Labeling, Exaggeration or Minimization, Doubt, Appeal to Fear/Prejudice, Flag-Waving, Causal Oversimplification, Slogans, Appeal to Authority, Black-and-White Fallacy, Thought-Terminating Cliche, Whataboutism, Reductio ad Hitlerum, Red Herring, and Strawman.** The third dataset (LOGIC) is recently released by Jin et al. (2022) and contains 13 logical fallacies (**Faulty Generalization, False Causality, Circular Claim, Ad Populum, Ad Hominem, Deductive Fallacy, Appeal to Emotion, False Dilemma, Equivocation, Fallacy of Extension, Fallacy of Relevance, Fallacy of Credibility, Intentional Fallacy**) from educational websites on fallacy such as Quizziz and study.com. The final existing fallacy dataset (COVID-19) is about fact-checked content around Covid-19 (Musi et al., 2022). The authors identify 10 fallacies (**Evading the Burden of Proof, Cherry Picking, Strawman, Red Herring, False Authority, Hasty Generalization, Post Hoc, False Cause, False**

Analogy, Vagueness) through analysis of fact-checked social media posts and news by considering fallacies as indicators of misinformation.

▼ **Final Labels.** We **unify the labels of similar fallacies** (e.g., False Cause, False Causality, Causal
Oversimplification → Causal Oversimplification). We also **rephrase some fallacy types by removing words** such as “Appeal to” (e.g,. Appeal to Emotion → Emotional Language) that tend to throw off generative models causing over prediction of these types as observed in our initial experiments. Some fallacies have partial or full overlap with others across the four schemes. Therefore, we merge these types and use the label of the most frequent or the most representative label of the fallacy type (e.g., Fallacy of Relevance → Red Herring). We also **unify the definitions of fallacy types** in prompts across datasets. We end up with 28 unique fallacy types across five datasets ARGOTARIO: 5, PROPAGANDA: 15, LOGIC: 13, COVID-19 and CLIMATE: 9.

6. https://arxiv.org/pdf/2205.09803.pdf

- **In this work**, we close this gap by approaching argument quality estimation from multiple different angles: Grounded on rich results from thorough empirical evaluations, we assess the generalization capabilities of argument quality estimation across diverse domains, the interplay with related argument mining tasks, and **the impact of emotions on perceived argument strength.**
- … investigated which argument from the pair is **more convincing;** if they would **recommend** a
  friend to use the argument **in a speech supporting;** is labeled in the range of [0, 1], where 1 indicates an argument can be **easily interpreted**

7. https://arxiv.org/pdf/2212.07425.pdf

▼ **In this paper**, we formalize prior theoretical work on logical fallacies into a comprehensive three-stage evaluation framework of detection, coarse-grained, and fine-grained classification.

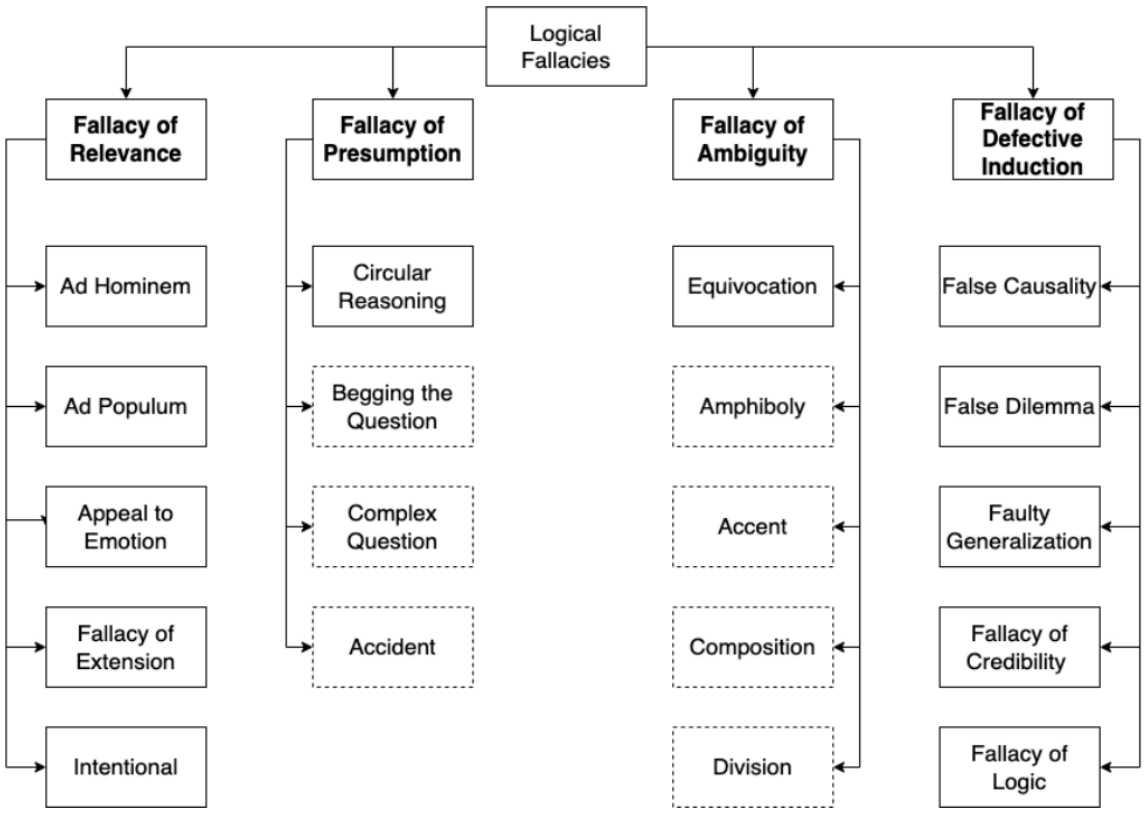
▼ There are two broad categories of fallacies: **formal**, involving the error in the logical structure of the argument, and **informal**, mostly concerned with the content of the argument or the latent error in their expression of logic [45].

▼ There are six classes of fallacies dependent on language: **Equivocation, Amphiboly, Combination of Words, Division of Words, Accent, and Form of Expression.**

▼ We design **a three-stage framework** (Figure 1) as an overarching testbed for prior research on logical fallacies. The first stage of the logical fallacy detection **aims to identify whether a logical statement contains a logical fallacy or not**. The detection is formalized as a binary classification task to identify the arguments that are logically fallacious in any sense. If a fallacy has been detected, the goal of the second stage is **to categorize the fallacy into one of a few broad classes** (e.g., Fallacy of Relevance). In the third stage, the aim is **to further classify a fallacy into a fine-grained class** (e.g., Ad Populum).

▼ Following [26], we consider the following four coarsegrained classes: **Fallacy of Relevance, Fallacy of Defective Induction, Fallacy of Presumption, and Fallacy of Ambiguity.**

Robust and Explainable Identification of Logical Fallacies in Natural Language Arguments



8. https://arxiv.org/pdf/2209.02062.pdf

- In this work, we take a first step in shedding light on the usage of ad hominem fallacies in the wild. First, we build a powerful **ad hominem detector** based on transformer architecture with high accuracy (F1 more than 83%, showing a significant improvement over prior work), even for datasets for which annotated instances constitute a very small fraction. We then used our detector on 265k arguments collected from the online debate forum – CreateDebate.
- **ad hominem - переход на личности**
- For our experiments, we used ad hominem argumentation in the ChangeMyView (CMV) dataset (Habernal et al. 2018) as benchmark. ChangeMyView is a popular subreddit in which a user (called OP, original poster) posts an opinion and other users write comments to change the perspective of OP about the posted opinion. OP can acknowledge convincing arguments by giving delta points.
- The CMV dataset contains 7242 comments from this subreddit (3622 instances with the label **‘ad hominem**’ and 3620 instances with the label **‘none’**).

9. https://arxiv.org/pdf/2202.13758.pdf

▼ **In this paper**, we propose the task of **logical fallacy detection**, and provide a new dataset (LOGIC) of logical fallacies generally found in text, together with an additional challenge set for detecting logical fallacies in climate change claims (LOGICCLIMATE).

▼ Our logical fallacy dataset consists of two parts: **a)** a set of common logical fallacies (LOGIC), and **b)** an additional challenge set of logically fallacious claims about climate change (LOGICCLIMATE).

▼ Data Collection The LOGIC dataset consists of common logical fallacy examples collected from
various online educational materials meant to teach or test the understanding of logical fallacies among students. We automatically crawled examples of logical fallacies from three **student quiz websites, Quizziz, study.com and ProProfs** (resulting in around 1.7K samples), and manually collected fallacy examples from some **additional websites** recommended by Google search (resulting in around 600 samples).

▼ For each news article, we ask two different annotators who are native English speakers to go through each sentence in the article, and label all logical fallacies if applicable. Since directly classifying the logical fallacies at the article level is too challenging, we let the annotators select the text span while labeling the logical fallacies, and we compose each sample using the sentence containing the selected text span as logical fallacies.