Towards a Perspectivist Turn in Argument Quality Assessment

Julia Romberg[1], **Maximilian Maurer**[1,3], **Henning Wachsmuth**[2] and **Gabriella Lapesa**[1,3]
[1]GESIS - Leibniz Institute for the Social Sciences
[2]Leibniz University Hannover
[3]Heinrich-Heine University Düsseldorf
[1]first.last@gesis.org, [2]h.wachsmuth@ai.uni-hannover.de

Abstract

The assessment of argument quality depends on well-established logical, rhetorical, and dialectical properties that are unavoidably subjective: multiple valid assessments may exist, there is no unequivocal ground truth. This aligns with recent paths in machine learning, which embrace the co-existence of different perspectives. However, this potential remains largely unexplored in NLP research on argument quality. One crucial reason seems to be the yet unexplored availability of suitable datasets. We fill this gap by conducting a systematic review of argument quality datasets. We assign them to a multi-layered categorization targeting two aspects: (a) What has been annotated: we collect the quality dimensions covered in datasets and consolidate them in an overarching taxonomy, increasing dataset comparability and interoperability. (b) Who annotated: we survey what information is given about annotators, enabling perspectivist research and grounding our recommendations for future actions. To this end, we discuss datasets suitable for developing perspectivist models (i.e., those containing individual, non-aggregated annotations), and we showcase the importance of a controlled selection of annotators in a pilot study.

1 Introduction

The question of "what makes an argument good" is at the core of computational argumentation (CA), the area of natural language processing (NLP) dealing with the mining, assessment, and generation of arguments (Stede and Schneider, 2019; Lauscher et al., 2022). While rooted in well-established theories, argument quality (AQ) still exhibits a high degree of subjectivity in perception. This degree may vary across quality aspects; for example, evaluating an argument's benefit in agreement-seeking discussions is considered less subjective than assessing its effectiveness (Wachsmuth et al., 2017b).

The CA community is aware of the variance in annotators' perception (Stab and Gurevych, 2014; Teruel et al., 2018; Hautli-Janisz et al., 2022). Not least, this is documented by the generally moderate inter-annotator agreement in AQ annotations — a widely accepted condition that authors commonly attribute to the subjective nature of the task (e.g., Wachsmuth et al. 2017b; Gretz et al. 2020; Ng et al. 2020; Ziegenbein et al. 2023).[1] Wachsmuth and Werner (2020) explicitly question whether an aggregated ground truth is suitable to model AQ.

Meanwhile, the NLP community has started to undergo a fundamental change in the way it deals with subjective tasks. While aggregated ground truth and an according model alignment were long standard, more recent work calls for this course to be reconsidered (Basile, 2020; Plank, 2022; Cabitza et al., 2023; Frenda et al., 2024): Rather than eradicating any existence of annotator disagreement, the *perspectivist turn* embraces the co-existence of perspectives (Uma et al., 2021a; Davani et al., 2022; Leonardelli et al., 2023). This transformation implies the acceptance of variations in data annotation (through non-aggregated datasets) as well as the consideration of heterogeneity in modeling and evaluation (Uma et al., 2021b; Basile et al., 2021; Plank, 2022).

We postulate that the perspectivist turn in NLP lends itself as a natural solution to face the issue of subjectivity in modeling AQ. Not only does it have a better shot at promoting diversity and fairness in AQ assessment, such as allowing for valid but minority voices (Noble, 2012; Prabhakaran et al., 2021). It is also likely to be more robust in modeling perceptions of AQ across (changing) societies (e.g., today's minority groups may become tomorrow's majority) and target audiences.

Yet, the perspectivist turn so far had only minimal impact on AQ. Presumably, one reason for the limited modeling of perspectives in AQ is the lack of datasets designed for this purpose. Preference

[1]Certainly, not all disagreement is due to subjectivity. We refer the reader to the Limitations section for a discussion.

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 7458–7485

has been given to aggregated annotations, whereas individual labeling decisions were often not communicated (e.g., Persing and Ng 2013; Park and Cardie 2018; Toledo et al. 2019; Goffredo et al. 2022). While a solution may be new datasets, the annotation of argumentation phenomena is highly complex and costly. We therefore deem it essential to first *gain an overview of the options that existing datasets already offer for developing perspectivist models*. This is the goal of the paper at hand.

We provide a systematic literature review of 103 AQ datasets and their properties.[2] Crucially, to support the perspectivist turn, our collection includes meta-information about annotators and the availability of non-aggregated annotations. While only 24 datasets come with the latter, 14 of them seem relevant to the perspectivist turn. In a pilot study, we conduct a statistical analysis of the disagreement patterns in four of them. We conclude by highlighting the opportunities of available datasets and discuss challenges, for example a lack of transparency and socio-demographic diversity.

Contributions (1) We release an extensive database with 32 types of meta-information about 103 AQ datasets. (2) We review the multitude of annotated AQ categories (*what is annotated*) and consolidate them in an overarching taxonomy to foster comparability and interoperability. (3) We perform a comprehensive meta-analysis of annotators (*who annotates*) across the datasets, uncovering a lack of transparency and socio-demographic diversity, promoting bias in AQ datasets and models. (4) We deep-dive into the 24 datasets with non-aggregated labels both qualitatively and quantitatively, and discuss their potential for a perspectivist turn in AQ.

2 Related Work

Surveys of Computational Argumentation In the last 20 years, the field of CA has witnessed a constant development driven by the potential for real world applications, but also by the increasingly interdisciplinary shape that the field has assumed. The number of surveys on CA is a clear sign of this progress: ranging from foundational work that set up or updated the conceptual coordinates for the field (Peldszus and Stede, 2013; Stede and Schneider, 2019; Lawrence and Reed, 2020; Lauscher et al., 2022), to surveys with a data-driven focus (Cabrio and Villata, 2018; Schaefer and Stede, 2021), to specific advances in NLP, such as generation (Wang et al., 2023).

Specifically for AQ, Wachsmuth et al. (2017b) introduced a first holistic systematization of the field according to AQ dimensions. Wachsmuth et al. (2024) update the survey, taking into account the challenges and potentials for the employment of large language models (LLMs) in AQ assessment. While focused on applications of CA for social good as a whole, the survey by Vecchi et al. (2021) puts a strong interdisciplinary focus on AQ and its interface with deliberation quality.

No survey so far has targeted a systematic categorization of datasets.[3] This is the gap we fill:[4] we survey datasets, focusing on the consolidation of covered AQ categories into an overarching taxonomy, and who the annotators are.

Dimensions of Argument Quality Wachsmuth et al. (2017b)'s taxonomy is the most commonly adopted one for AQ assessment. Rooted in argumentation theory, it emphasizes three aspects:

Logical cogency and its subcategories promote a valid reasoning process at the level of individual arguments. An argument is considered cogent if its premises are rationally worthy of being believed to be true (*local acceptability*), its premises contribute to the acceptance or rejection of its conclusion (*local relevance*), and if they provide enough support to make the conclusion rational (*local sufficiency*).

Rhetorical effectiveness and its subcategories mirror the persuasive power of an author's argument towards a target audience. Characteristics are a clear style (*clarity*), maintaining a tone appropriate to the issue (*appropriateness*), presenting components of the argument in a proper order (*arrangement*), establishing the author's credibility (*credibility*), and evoking emotions that make the audience more receptive (*emotional appeal*).

Dialectical reasonableness and its subcategories evaluate the contribution to resolving differences

[2]The resulting database can be accessed publicly here: https://github.com/juliaromberg/perspectivist-turn-aq

[3]Dataset repositories for CA do exist, i.e., *ARGLU* and the Webis database, but are by no means comprehensive and lack our conceptual categorization and focus on annotators.

[4]While this paper was under review, a survey on AQ was published by Ivanova et al. (2024), highlighting the timeliness of this topic. The authors examined the state of AQ research in general, whereas we focus on its future transition into a task where human label variation plays a significant role. Due to a different search strategy, our survey covers a broader range of datasets (103 compared to 32), and offers a more in-depth analysis (with 32 manually annotated meta-categories, compared to 10). Crucially, we adopt a timely interdisciplinary taxonomy that integrates argument and deliberation quality, which Ivanova et al. (2024) also hint at for future research.

of opinions on a discussion level. Argumentation is deemed reasonable if the consideration and presentation of the arguments put forward for the issue are acceptable to the target audience (*global acceptability*), contribute to the issue's resolutions (*global relevance*), and adequately rebut the contestable counterarguments (*global sufficiency*).

Vecchi et al. (2021) proposed to include *deliberative norms* as a further aspect of AQ. This dimension incorporates democratic values into the dialectical view, adherence to which is particularly relevant to political arguments, but also applies to broader contexts like online communication. While the authors resorted specifically to the *Discourse Quality Index* (Steenbergen et al., 2003), communication science has come up with various instruments to empirically measure deliberation quality (e.g., Stromer-Galley 2007; Black et al. 2011; Graham and Witschge 2003).

The exact criteria of (good) deliberation and consequently the instruments for measuring it are matter of controversial discussion (Delli Carpini et al., 2004). Friess and Eilders (2015) identified seven dimensions that are prevalent across various frameworks: Deliberative discourse should be an exchange grounded in *rationality*. The exchange should take place through listening, understanding and actively responding to each other's opinions in a substantive way (*interactivity*). Furthermore, deliberation should foster *equality* by equipping all sides with the same opportunity to participate in the discussion and *civility* for a respectful interaction. Arguments should be oriented towards the *common good* of the community, and *constructive* in finding a consensus decision for the issue of discussion. The last dimension relates to the use of *alternative forms of communication* (e.g., storytelling).

Perspectivism and Argument Quality AQ assessment is a prime example of a subjective task: beyond logical well-formedness (and even there), the question of good arguments is bound to be answered in conflicting ways by annotators with different features (e.g., socio-demographics, life experiences, personality, and values) (Lukin et al., 2017; Durmus and Cardie, 2019; El Baff et al., 2020). This makes AQ an ideal perspectivist topic.

Yet, *perspectivist AQ assessment is only at its beginning, also because of the need for suitable data.* As datasets will be reviewed in the remainder of the paper, we focus here on the few works that have specifically targeted the modeling of annotator perspectives in AQ, i.e., by integrating label variation in the machine learning workflow. The first explicit step was taken by Romberg (2022), who predicted the subjectivity of the annotation as an indicator for trustworthiness of majority vote models. What is more, Heinisch et al. (2023) compared approaches for modeling annotator-specific behavior.

3 Systematic Review of Datasets

Search Methodology We searched in all major publication organs that, according to our experience in the field, cover the topic of CA. We included the leading conferences in computational linguistics (the entire ACL anthology, including the Argument Mining Workshop), artificial intelligence (all from AAAI.org, IJCAI, and ICAIL), information retrieval (SIGIR and ECIR), and the specialized computational argumentation series COMMA.

To pre-filter a set of candidates, we used all pairs of search terms from *{argument, argumentation, argumentative, debate, deliberation, deliberative}* × *{quality, strength, persuasiveness, fallacies}*. The retrieval was carried out with the Google site search including all papers that had been published until August 20, 2024, resulting in 238 candidate papers of which we found 42 to be relevant. Additionally, we employed a snowballing method (Wohlin, 2014) to ensure that the field is covered as completely as possible: looking at studies that either cite one of the previously identified papers (forward snowballing, with Google Scholar) or are cited by those papers (backward snowballing) led to further 56 relevant papers. In total, we identify 98 relevant papers, distributed among research communities as follows: NLP (73), artificial intelligence (6), information retrieval (7), CA (1), and further venues from computer science (4) and the social sciences (7). Appendix A describes the process in detail.

Categorization Taxonomy We assess the relevance of datasets by drawing from the taxonomies of Wachsmuth et al. (2017b) and Friess and Eilders (2015). A paper is considered relevant if it introduces a new dataset (or extends an existing one) that at least loosely matches one or more of the following categories (whose theoretical background has been introduced in Section 2): i) *logical cogency* with subcategories *local acceptability*, *local relevance*, and *local sufficiency*; ii) *rhetorical effectiveness* with subcategories *clarity*, *appropriateness*, *arrangement*, *credibility*, and *emotional appeal*; iii) *dialectical reasonableness* with sub-

categories *global acceptability*, *global relevance*, and *global sufficiency*; iv) *deliberative norms* with subcategories *rationality*, *interactivity*, *equality*, *civility*, *common good reference*, *constructiveness*, and *alternative forms of communication*; and *overall argument quality*. Appendix D provides the taxonomy and exact definitions in full.

Categorization Reliability Mapping of dataset dimensions according to the AQ taxonomy was conducted by the first author of the paper. To validate this process, 10 papers (~10%) were reassigned to two other authors. For the high-level categories, we reached Fleiss' κ values of 1.0 for logical cogency, 0.73 for rhetorical effectiveness, 0.71 for dialectical reasonableness, and 0.70 for deliberative norms, demonstrating robust inter-annotator reliability. The mean agreement across all 23 categories was lower, 0.52. It is worth pointing out, though, that one of the other authors reached 0.72 with the first author, which is why we deem our categorization to be reasonably reliable. A clear source of disagreement arose from the categorization of fallacies into the taxonomy. Reasoning errors (i.e., fallacies) can affect all dimensions of AQ, and we refined the annotation guidelines accordingly.

Collection of Meta-information In addition to the AQ categories, we gathered further information about the datasets. This includes general details such as genre, modality, language, and the availability of the dataset and annotation guidelines. While dataset availability was generally good (84 public or upon request), information on annotation guidelines was less available (32). We contacted authors of datasets without clear indications, encouraging public release in line with open science principles. As a result, 14 additional datasets now have publicly accessible guidelines, for a total of 11 unique guidelines made available.

We also collected a variety of characteristics that are of interest when looking through the perspectivist glasses. Most notable for the paper at hand are meta-information about annotators and the availability of non-aggregated annotations. Appendix B lists all information contained in the database.

4 Datasets: Annotations & Annotators

The 98 identified papers introduce 103 AQ datasets in total. A complete list including the mapped AQ categories is in Appendix E. In what follows, we provide an overview of the quantitative properties

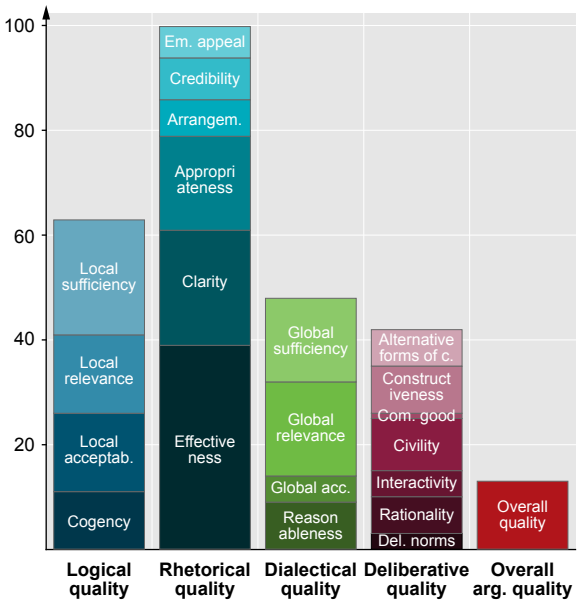

Figure 1: Frequency and distribution of AQ categories (major and sub-categories) as assigned to datasets, grouped by the four major categories and overall AQ.

arising from the comparison of their *annotation*, and we then focus on *annotator* meta-information.

4.1 Annotations: What Argument Quality?

Figure 1 shows the distribution of datasets among the categories of AQ. Particular interest can be observed for the rhetorical effectiveness of arguments, likely driven by its practical relevance and the availability of pre-annotated resources, such as the Reddit forum ChangeMyView and other online debate platforms where users rate the persuasiveness of each other's arguments. Increased attention has also been paid to the logical validity of arguments. However, the two dialectically-driven dimensions of reasonableness and deliberative norms received less attention, a finding that coincides with the CA community's constant call for a greater focus on dialogical argumentation (Ruiz-Dolz et al., 2024).

Looking into the individual sub-categories of AQ, we find that almost all of them are covered. The only exception is *equality*. However, we acknowledge that measuring whether participants have equal opportunities in a deliberation is challenging, as it extends beyond simply assessing active participation to the potential for participation depending on the socioeconomic capital that the participants hold (Friess and Eilders, 2015).

The upper part of Table 1 provides an overview of selected dataset properties. In terms of genre, the datasets cover a great variety, with social media,

Meta-Categories	Specifications (Counts, Sorted Descending)
Genre	social media (20), online debate portal (19), persuasive essays (12), crowd-sourced (10), public participation (8), news articles (7), political debate (6), web (5), collaborative online discussions (5), news comments (4), reviews (3), educational debate (2), fact-checking portals (2), QA forums (2), e-mail communication (2), Wikipedia (2), online educational material (1), classroom discussions (1), business model pitches (1), LLM-sourced (1)
Modality	text (100), multimodal (3)
Language	en (89), de (13), fr (4), jp (2), es (1), it (1), nl (1), pt-br (1), zh (1)
Manual annotation	manual (82), automatic (18), manual+automatic (3)
Selection method of annotators	students/available (22), consistency with experts (18), expertise (16), language competence (16), reliability checks (11), performance in prior tasks (8), educational level (5), balanced sample wrt. to some property (5), consistency with fellow annotators (3)
In-house annotators or crowd-workers	in-house (39: experts 13, novice 11, mixed 5, n.a. 10), crowd (27; task expertise unknown), in-house+crowd (5: in-house experts 4, in-house mixed 1), n.a. (14)
Annotator attributes (across in-house and crowd-sourcing datasets)	no indication at all (45); education (25), age (18), native language (14), gender (11), profession (9), professional background (8), stance (5), country of origin (4), country of residence (3), occupation (3), political view (3), nationality (2), personality traits (2), annotation time (1), civic engagement (1), competence (1), employment status (1), ethnicity (1), income (1), race (1), religion (1), role (1), spirituality (1)

Table 1: Counts of specifications for different meta-categories on datasets (top) and annotations (bottom).

online debate portals, and persuasive essays being the most prominent. Seven datasets draw from multiple genres (Xu et al., 2014; Napoles et al., 2017; Lauscher et al., 2020; Ziegenbein et al., 2023; Falk and Lapesa, 2023; Helwe et al., 2024; Li et al., 2024). While most of the datasets focus solely on text, three are multimodal (Liu et al., 2022, 2023; Mancini et al., 2024). With respect to languages, we observe a very imbalanced situation with English accounting for over 85% of the datasets. Four of the datasets contain multiple languages (Gerber et al., 2018; Toledo-Ronen et al., 2020; Falk and Lapesa, 2022; Reveilhac, 2023).

4.2 Annotators: Whose Perspectives?

We now take a closer look at the individuals that provide the AQ assessments, in order to understand whose perspectives current datasets cover. Table 1 shows the statistics on manual annotation, annotator selection and attributes. The majority of datasets were created through coordinated manual data annotation; fewer than 20% of datasets were generated automatically by parsing existing internet resources, with the ground truth labels derived from a natural sample of platform users.

In manual annotation, authors indicated a variety of reasons for the selection of annotators, among them predominantly consistency with experts, expertise of the annotators themselves, and language competence. Students were also a common choice (in one quarter of the datasets), which might be an indicator for selection upon availability. Only five datasets had annotators selected with the aim of weighting socio-demographic characteristics according to certain standards, such as the representation of a country's population (Lukin et al., 2017; Brenneis et al., 2021), balancing political ideologies or gender (El Baff et al., 2018; Falk et al., 2024), and annotators from diverse debating circuits (Joshi et al., 2023). Also noteworthy is that in three datasets, annotators were excluded if inconsistent with fellow annotators' label decisions.

Looking more closely at the socio-demographic background of AQ annotators, we find that authors only occasionally provide information (in the papers or datasets). In case of *in-house annotators*, we find a higher education in all cases indicated and often a background in NLP and related fields. A key differentiator is expertise in AQ, which separates in-house annotators into two groups: *experts* and *novice* annotators (usually students). Additionally, we find seven explicit mentions of gender (two datasets include both binary genders without specifying proportions, two use balanced samples, two have significantly more male annotators, and one includes two female and one male annotator). Age was reported twice, with ranges of 18–53 and 18–22 years. In case of *annotators recruited on crowd-sourcing platforms*, socio-demographic information is reported sparsely, only for 8 datasets. In these cases, it is either used to draw a more representative sample or serves to narrow down the selection of annotators to the language of data.

On a more general note, characteristics that may invoke some bias in assessment such as political views (and related stances) were rarely collected, and there is similarly little information on cultural diversity among annotators. Individual characteristics that go beyond socio-demographic features are hardly at issue, except from Lukin et al. (2017) and El Baff et al. (2018) who collect personality traits, recognizing the potential impact on AQ perception.

5 Towards Perspectivist Argument Quality Assessment

Developing perspectivist models requires the existence of multiple assessment perspectives. Among the 103 datasets we found, only 24 come with non-

Dataset	Size	Per-item	Total	Category	Annotators' Attributes Provided in Dataset
Introduced as non-aggregated to facilitate perspectivist machine learning or to promote diversity in annotations					
CrowDEA Ideas (Baba et al., 2020)	16,000	20	257	crowd	-
Argument Concreteness (Romberg et al., 2022)	1,127	5	5	novice	-
TYPIC (Naito et al., 2022)	197	1–2	4	in-house	-
Argument Validity Novelty (Heinisch et al., 2023)	1,474	3	5	expert	-
MAFALDA (Helwe et al., 2024)	268	1–4	4	expert	-
UMOD (Falk et al., 2024)	1,000	9	90	crowd	race, gender, age, annotation time, role, competence, stance
Built to explore how argument perception differs between groups and individuals					
Persuasion & Personality (Lukin et al., 2017)	100	20	637	crowd	personality traits, age, gender, political view, education, civic engagement, religion, spirituality, employment status, income, stance
Webis-Editorial-Quality-18 (El Baff et al., 2018)	1,000	6	24	crowd	political view, personality traits
Personalization					
n.a. (Hunter and Polberg, 2017)	30	50	50	crowd	-
SIGIR-19 (Potthast et al., 2019)	494	1	40	in-house	age, gender, stance
argumentation-attitude (Brenneis et al., 2021)	946	1–147	674	crowd	stance
Aggregated ground truth datasets that were released together with the individual labeling decisions					
Dagstuhl-ArgQuality (Wachsmuth et al., 2017b)	320	3	3	expert	-
n.a. (Wachsmuth et al., 2017a)	320	10	102	crowd	-
n.a. (Mirzakhmedova et al., 2024)	320	$\leq$ 10	108	novice	-
GAQCorpus (Lauscher et al., 2020)	5,285	1–13	27	exp, crowd	-
EuropolisAQ (Falk and Lapesa, 2022)	513	1–2	2	expert	-
ArgQ! Silva et al. (2021)	352	4	4	expert	-
UKPConvArg1 (Habernal and Gurevych, 2016b)	16,000	5	3,900	crowd	stance
UKPConvArg2 (Habernal and Gurevych, 2016a)	70,000	5	776	crowd	-
Essay Argument Organization (Persing et al., 2010)	1,003	1-6	6	novice	-
Appropriateness Corpus (Ziegenbein et al., 2023)	2,191	3	3	crowd	-
UKP-InsufficientArgs (Stab and Gurevych, 2017)	433	3	3	expert	-
Webis-ArgRank-17 (Wachsmuth et al., 2017c)	110	7	7	expert	-
StoryARG (Falk and Lapesa, 2023)	2,451	1–4	4	in-house	-

Table 2: Overview of AQ datasets that come with non-aggregated annotations. In each case, we provide annotator counts *per-item* and *total*, *categorize* them as in-house (experts, novice, or in-house; if expertise is unspecified) or crowd workers, and specify the *annotators' attributes* contained directly in the datasets at the individual level.

aggregated annotations. We detail these datasets, before we exemplify the potential impact of annotator groups on AQ assessment.

5.1 Non-Aggregated Datasets

Table 2 lists the datasets with properties relevant to perspectivist model development. We identify four conceptual blocks: Six datasets were exclusively introduced as non-aggregated for perspectivist approaches or to promote annotation diversity. Two were developed to study how argument perception varies based on group-level or individual characteristics. Three stem from personalization in argument retrieval, and 13 are aggregated datasets released together with the individual labeling decisions. An extensive description of all 24 datasets is provided in Appendix C. Here, we focus on those that we deem most useful for the perspectivist turn.

Populations For developing well-generalizable models, it is integral that the datasets represent a specific population, whose composition of perspectives can be learned. The annotations of four datasets were collected in a controlled setup in this regard: The *Persuasion & Personality* corpus (Lukin et al., 2017) was created to study differences in the perception of argument effectiveness. Stance changes elicited by social media arguments were recorded from a representative sample of the US population. The *argumentation-attitude* dataset (Brenneis et al., 2021) covers personalized views of strong arguments from a political opinion platform, rated by a representative sample of the German online population, in terms of age, gender, and education. *Webis-Editorial-Quality-18* (El Baff et al., 2018) captures differing perceptions of effectiveness in US news editorials on a balanced sample of liberals and conservatives, *UMOD* (Falk et al., 2024) annotates characteristics of user-driven online moderation (including comment constructiveness), using a gender-balanced population.

Crowd-platform annotations sourced from a sufficiently large group of workers may also be assumed to approximate the broader population from the respective platform to a certain extent: *UKPConvArg1* and *UKPConvArg2* (Habernal and Gurevych, 2016a,b) capture argument convincingness and the AQ reasons behind, with 16k and 70k items, respectively, and over 4k crowd annotators from the US. *CrowDEA Ideas* (Baba et al., 2020) contains preference labels of 257 crowd workers for 16k solutions to an issue. These three datasets also stand out in their size; they are the only datasets with tens of thousands of annotated items.

Lastly, we highlight two datasets that bring together different groups of annotators for the same data. The *Dagstuhl-ArgQuality* dataset (Wachsmuth et al., 2017b) of online debate arguments was rated by experts, novice student annotators (with no prior experience in CA) (Mirzakhmedova et al., 2024),[5] and crowd workers (Wachsmuth et al., 2017a) across the 15 dimensions of their taxonomy. The *GAQCorpus* (Lauscher et al., 2020), which includes diverse arguments annotated for cogency, effectiveness, reasonableness, and overall AQ, was annotated by a mix of 3 expert annotators and 24 crowd workers. Some items were annotated only by the crowd, others by the experts, and part of the dataset was jointly annotated by both groups. The different groups can be regarded as different types of populations and thus represent an interesting testing ground for group-specific analysis.

Individuals To model the perspectives of individuals more accurately, meta-information about them is needed. Such attributes can also help to build perspectivist models at the group level. The Persuasion & Personality corpus, the argument-attitude dataset, WEBIS-Editorial-Quality-18, UMOD, and UKPConvArg1 provide several relevant attributes, including socio-demographics, stances on certain topics, and personality traits. In addition, the *SIGIR-19* dataset (Potthast et al., 2019), which codes logical, rhetorical, and dialectical AQ, includes gender, age, and stance for each annotator.

Besides, two datasets from our overview have already been used successfully in modeling human label variation, the non-aggregated version of the *Argument Validity and Novelty* dataset (Heinisch et al., 2023) and the *Argument Concreteness* corpus (Romberg et al., 2022). Both lack background information on the annotators, but this limitation was initially secondary to the goal of developing perspectivist models for personalization (Heinisch et al., 2023).

5.2 Potential Impact of Annotator Groups

In Section 4, we found that in-house annotators form a relatively homogeneous group concerning education and work background, with expertise being a key differentiator between experts and novices. In contrast, annotators recruited via crowdsourcing platforms can be assumed to represent a much more diverse sample in terms of socio-demographic attributes and lived experiences. We thus study two research questions on the impact of annotator groups on AQ annotation and prediction:

RQ1. How comparable are annotations across annotator groups, how stable within them?

RQ2. How does this impact the performance bounds of models trained on group-specific annotations when transferred across groups?

Data We use the two mentioned non-aggregated datasets with a mix of annotator groups: Dagstuhl-ArgQuality (Wachsmuth et al., 2017b) and its extensions (Mirzakhmedova et al., 2024; Wachsmuth et al., 2017a) (summarized as *Dagstuhl*; annotated by experts, novice student annotators and crowd-platform workers), and the GAQCorpus (Lauscher et al., 2020) (*GAQ*; annotated by experts and crowd-platform workers). For Dagstuhl, we resort to the 304 arguments deemed argumentative in the original corpus. For GAQ, we use the 538 arguments annotated by both experts and crowd annotators. We focus on cogency, effectiveness, reasonableness, and overall AQ as categories.

Experimental Setup For RQ1, we compute inter-annotator agreement (IAA) within and across annotator groups, using Krippendorff's α. For a full picture, we include all annotators per group.

To answer RQ2, we assume a situation in which a *perfect model*, trained on the annotations of one group, is evaluated on another group, effectively performing a population transfer. We opted for simulation rather than real training of a model in order to minimize confounding factors, such as model deficits due to limited training data. This way, we can clearly illustrate the discrepancy that arises when population characteristics, and the differences in perspectives they encode, are ignored. We investigate the actual upper performance bounds in two scenarios, a traditional *aggregated approach* with a single regression output per argument, and a *perspectivist approach* in which we assume to obtain a learned regression label distribution per argument as the system output.

To compare label distributions (i.e., the perfectly predicted one and that of a target population), we calculate the Wasserstein distance (WS) between the label distributions per item. We report the mean across the whole dataset, respectively. For aggregated regression outputs (i.e., mean ratings), we calculate the mean average error (MAE). We report results per dataset and quality dimension.

[5]The official release includes only a subset of annotators, but the authors kindly provided the full set upon request.

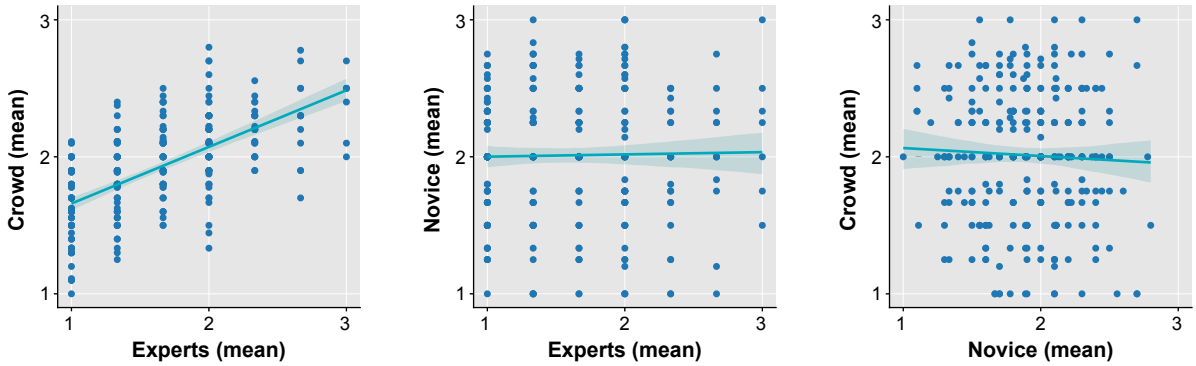

Figure 2: Instance-based aggregation of label decisions for overall AQ, assessed on a scale from 1 (low) to 3 (high), between two annotator groups on Dagstuhl, with a fitted linear regression model highlighting their relationship.

	Group	Cogency	Effectiveness	Reasonableness	Overall
Dagstuhl	E	**.372**	**.314**	**.437**	**.443**
	N	.230	.208	.197	.233
	C	.099	.107	.111	.140
	E, N	.114	.098	.134	.126
	E, C	.129	.121	.143	.180
	N, C	.060	.072	.071	.083
	E, N, C	.083	.085	.098	.115
GAQ	E	**.175**	**.272**	**.258**	**.254**
	C	.156	.148	.154	.173
	E, C	.142	.142	.150	.165

Table 3: Krippendorff's α for different groups of annotators; experts (E), novice (N), and crowd workers (C).

		Cogency		Effectiven.		Reasonab.		Overall	
	Transfer	MAE	WS	MAE	WS	MAE	WS	MAE	WS
Dagstuhl	E, N	.697	.625	.714	.663	.605	.538	.686	.582
	E, C	.499	.558	.530	.581	.463	.534	.430	.485
	N, C	.507	.482	.470	.450	.454	.471	.480	.460
GAQ	E, C	.697	.800	.751	.866	.629	.740	.659	.762

Table 4: Group transfer evaluation for the aggregated approach (MAE) and the perspectivist approach (mean WS); experts (E), novice (N), and crowd workers (C).

Results Regarding RQ1, Table 3 shows the IAA for each annotator group and their different combinations for both corpora. For in-group, across corpora and quality dimensions, we find that expert annotators have the highest agreement, though still comparably low. For Dagstuhl, crowd annotators have the lowest IAA. Across groups, IAAs drop compared to the involved group with the highest agreement for all combinations of groups and both corpora. This indicates varied annotations within groups and high disagreement across groups.

Regarding RQ2, Table 4 shows the results for both the aggregated and the perspectivist approach. For the former, we find that even if a model perfectly learns one group's aggregated annotation behavior, the minimum MAE achievable, in the worst case, is 0.714 for Dagstuhl and 0.751 for GAQ (both effectiveness). Figure 2 exemplifies what this means on Dagstuhl for predictions of overall AQ: while transfer from experts to crowd annotators and vice versa retains the same general trend, the other combinations behave effectively at random. This is in line with IAA per group combination.

The perspectivist evaluation shows that the WS and MAE scores' orderings partly align. However, the N,C groups always exhibit the lowest WS in Dagstuhl (compared to the aggregated approach). In contrast, according to Krippendorff's α, the N,C group combination consistently performs worst. This indicates systematic patterns of disagreement present to a different extent between groups.

In sum, we find considerable annotation variation within and across groups (RQ1), which causes limited transferability between groups (RQ2).

6 Discussion

Imbalance of AQ Datasets We identified a substantial number of datasets, covering nearly the entire taxonomy of quality categories. The two dialectically-driven dimensions (reasonableness and deliberative norms) are less represented, likely due to the reduced focus on dialogical argumentation in CA and the more recent attention to deliberation. Unfortunately but not unexpectedly, we found a very uneven representation of language. Together with the sparsity of languages comes a lack of cultural diversity. As arguments are perceived differently across cultures (Han and Shavitt, 1994; Shen, 2023), this gap should be closed.

Potential Bias in Annotator Representation The analysis of meta-information about annotators revealed that current datasets provide little and mostly fragmentary documentation about whose

perception is captured. This lack of transparency raises the question of whether existing AQ datasets facilitate models that are biased to certain populations. We argue that certain characteristics of the annotators should be openly communicated, such as demographics (while adhering to privacy regulations); providing such data statements has been long called for by Bender and Friedman (2018), but we recognize rare application in the field of AQ.

A prime example is gender, rarely specified in the reviewed datasets. More than half of the manually annotated datasets stem from the highly homogeneous group of students and experienced researchers with a background in computational linguistics or computer science (i.e., the in-house annotators). Given that these fields remain heavily male-dominated (Schluter, 2018), such a selection strategy may unintentionally amplify gender bias. Equal considerations apply to other demographic attributes such as education and age. We also critically point to the practice of excluding annotators for inconsistency with their peers, as it wrongly assumes that diversity is per se an error and reinforces the formation of overly homogeneous groups.

Documenting such properties more transparently goes hand in hand with a raised awareness of the impact that annotator selection can have on the whole process. This is equally important for both the perspectivist turn and the established approach of aggregated ground truth (e.g., simple majority vote can exclude underrepresented groups entirely).

Potentials and Challenges for the Perspectivist Turn We identified a handful of non-aggregated AQ datasets to be suitable for enabling perspectivist model development, whereas many other datasets are insufficient in terms of a controlled selection of annotators, the number of annotators (in total and per item), and the dataset size.

As our experiments emphasize, dataset annotation decides who is represented by the model's output. We thus deem the collection of labeling decisions from annotators representative of a specific population crucial to building reasonable models for a desired target population. Likewise, the availability of individual-level annotator attributes is of high relevance as it facilitates modeling annotators more accurately, but also the development of perspectivist models at the group level.

It is furthermore vital to reflect on the number of annotators needed to build robust models. While 2–3 annotators may seem inadequate, there is no clear threshold for an appropriate group size. This also depends on the modeling goal; whether to develop perspectivist models in the literal sense, or to apply them for other purposes, such as personalization. Additionally, the number of dataset items and their coverage per annotator is essential. If items are sparsely labeled, the dataset may not provide enough information for individual annotators, an issue that has been raised by Davani et al. (2022).

In connection to the previous point of discussion, it is also crucial to keep in mind the risk of capturing spurious correlations between annotators' backgrounds and patterns in AQ assessment. This is especially important when working with datasets that were not explicitly created for perspectivist usage (e.g., in SIGIR-19). The question of how many data points per group are needed for the robust generalization of findings is an empirical one, usually constrained by budget limits. In such cases, the consideration of linguistic patterns alongside the task phenomenon may help support results.

7 Conclusion

A critical first step in developing perspectivist models for AQ assessment is suitable data. We identify several datasets as a learning ground for selected AQ categories in a non-aggregated way, while noting current shortcomings. Future datasets should (1) cover a diverse set of perspectives with respect to a reference population, (2) collect annotator-specific attributes, and (3) maintain an adequate size of total and individually labeled items. The listed desiderata indicate that the perspectivist turn in AQ assessment requires resources to be invested in annotation quality and quantity. In this context, recent research on the potential of active learning methods for subjective NLP tasks may be relevant (Wang and Plank, 2023; van der Meer et al., 2024).

Filling the resource gap will unlock the potential of the perspectivist turn in AQ assessment advocated in this paper. With suitable resources, future work can start leveraging the machine learning toolbox developed thus far in the perspectivist community (e.g., Davani et al., 2022; Casola et al., 2023).

Our extensive meta-information database will facilitate AQ research in general. One example is instruction-following LLMs, in which access to annotation guidelines is crucial (Wachsmuth et al., 2024). We make a significant contribution to this by expanding the number of publicly available guidelines, provided as a central listing in our database.

Limitations

Scope of the Dataset Search and Number of Annotators While we conducted a comprehensive and systematic search for datasets, we acknowledge that further datasets may be viewed as relevant that we did not cover. For example, the concepts of offensive or toxic language overlap with uncivil communication (Pachinger et al., 2023); or fallacies in propaganda detection. Moreover, while we believe that the validation on the 10% sub-sample demonstrates the reliability of AQ annotations, carried out by a single reviewer on the remaining datasets, we cannot exclude the impact of subjectivity and potential errors. For both cases (scope of the datasets collected and potential categorization errors), however, we believe that the fact that the collection will be publicly available in form of a website will allow authors to reach out to us for updates.

Assumed Fit of the Selected Taxonomies to the Whole Dataset Collection To categorize existing datasets, we have selected and applied two specific taxonomies (Wachsmuth et al., 2017b; Friess and Eilders, 2015), one for argument quality and one for deliberation quality. While our motivations for this choice are strong, as discussed in Section 2, we cannot in principle exclude that a different categorization would have had a better fit to the papers we collected. We believe, however, that our categorization approach, strictly based on consulting descriptions provided in papers and annotation guidelines (and contacting the authors directly when guidelines were not available) alleviates this limitation.

Disagreements: Between Valid Human Label Variation and Annotation Errors Disagreement among annotators can arise from various factors, among them subjectivity, but also annotation errors and ambiguity in the items to be labeled. Working with non-aggregated datasets thus always comes with the question of annotation reliability and how to distinguish potential annotation errors from valid label variations. First approaches are being developed to eventually complement perspectivist machine learning workflows (Weber-Genzel et al., 2024). Disentangling the different types of disagreements can improve data quality, which would not only benefit the perspectivist turn, but also the well-established approach of aggregated ground-truth. In both cases, the release of non-aggregated annotations is crucial.

While our pilot experiments in Section 5.2 exemplify the general consequences of pronounced differences in annotations across and within annotator groups, in this paper we do not specifically tackle the question of how much of these differences can be attributed to annotation error and how much to legitimately varied, subjective labeling decisions. However, the low to moderate IAA, even among the Dagstuhl experts — who themselves developed the underlying taxonomy — clearly indicates a significant degree of subjectivity in AQ perception.

Ethics Statement

This paper shares the inherent ethical concerns raised by argument mining in general, and by the assessment of AQ in particular. The first concern is dual use of NLP tools developed to assess / generate persuasive arguments, which could be then employed to manipulate the public opinion. The second concern is bias that such tools may contain, leading to assess the higher quality of arguments based on spurious cues in the data.

Additionally, while perspectivism advocates for the inclusion of as many perspectives as possible, it inevitably calls for a data-greedy approach, that is, the more annotators, the better. This may come with a human cost: it is typically achieved by crowd-sourcing, which is known to raise concerns about fair pay and treatment of the annotators. Also, the need of such expensive data collections may give an unfair advantage to highly funded researchers. Finally, the finer-grained the information about annotators is, the higher the privacy risks they are exposed to. Ideally, large-scale dataset surveys that aim at making resources aligned and comparable can enable data-greedy modeling without the need to annotate anew.

Generally, annotator attributes must be collected in compliance with privacy regulations and with the consent of participants. Some of the datasets we reviewed were parsed from existing internet resources, with ground truth labels derived from platform users. We would like to emphasize that using any information for profiling users, especially when it may contain personally identifiable content, risks privacy violations and may raise ethical concerns.

Acknowledgments

We would like to thank Ana Maria Lisboa dos Santos Cotovio for her support in the collection of the

dataset and for setting up the website.

References

Khalid Al Khatib, Michael Völske, Shahbaz Syed, Nikolay Kolyada, and Benno Stein. 2020. Exploiting personal characteristics of debaters for predicting persuasiveness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7067–7072, Online. Association for Computational Linguistics.

Khalid Al-Khatib, Henning Wachsmuth, Kevin Lang, Jakob Herpel, Matthias Hagen, and Benno Stein. 2018. Modeling deliberative argumentation strategies on Wikipedia. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2545–2555, Melbourne, Australia. Association for Computational Linguistics.

Tariq Alhindi, Tuhin Chakrabarty, Elena Musi, and Smaranda Muresan. 2022. Multitask instruction-based prompting for fallacy recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8172–8187, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

David Atkinson, Kumar Bhargav Srinivasan, and Chenhao Tan. 2019. What gets echoed? understanding the "pointers" in explanations of persuasive arguments. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2911–2921, Hong Kong, China. Association for Computational Linguistics.

Yukino Baba, Jiyi Li, and Hisashi Kashima. 2020. Crowdea: Multi-view idea prioritization with crowds. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 8(1):23–32.

Valerio Basile. 2020. It's the end of the gold standard as we know it. on the impact of pre-aggregation on the evaluation of highly subjective tasks. In *Proceedings of the AIxIA 2020 Discussion Papers Workshop*, pages 31 – 40. CEUR-WS.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online. Association for Computational Linguistics.

Beata Beigman Klebanov, Binod Gyawali, and Yi Song. 2017. Detecting good arguments in a non-topic-specific way: An oxymoron? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 244–249, Vancouver, Canada. Association for Computational Linguistics.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Laura W. Black, Howard T. Welser, Dan Cosley, and Jocelyn M. DeGroot. 2011. Self-governance through group discussion in wikipedia: Measuring deliberation in online groups. *Small Group Research*, 42(5):595–634.

Alexander Bondarenko, Maik Fröbe, Johannes Kiesel, Ferdinand Schlatt, Valentin Barriere, Brian Ravenet, Léo Hemamou, Simon Luck, Jan Heinrich Reimer, Benno Stein, Martin Potthast, and Matthias Hagen. 2023. Overview of touché 2023: Argument and causal retrieval. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 507–530, Cham. Springer Nature Switzerland.

Alexander Bondarenko, Maik Fröbe, Johannes Kiesel, Shahbaz Syed, Timon Gurcke, Meriem Beloucif, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, and Matthias Hagen. 2022. Overview of touché 2022: Argument retrieval. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 311–336, Cham. Springer International Publishing.

Alexander Bondarenko, Lukas Gienapp, Maik Fröbe, Meriem Beloucif, Yamen Ajjour, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, and Matthias Hagen. 2021. Overview of touché 2021: Argument retrieval. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 450–467, Cham. Springer International Publishing.

Liora Braunstain, Oren Kurland, David Carmel, Idan Szpektor, and Anna Shtok. 2016. Supporting human answers for advice-seeking questions in cqa sites. In *Advances in Information Retrieval*, pages 129–141, Cham. Springer International Publishing.

Markus Brenneis, Maike Behrendt, and Stefan Harmeling. 2021. How will I argue? a dataset for evaluating recommender systems for argumentations. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 360–367, Singapore and Online. Association for Computational Linguistics.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a Perspectivist Turn in Ground Truthing for Predictive Computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868.

Elena Cabrio and Serena Villata. 2012. Combining textual entailment and argumentation theory for supporting online debates interactions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 208–212, Jeju Island, Korea. Association for Computational Linguistics.

Elena Cabrio and Serena Villata. 2018. Five years of argument mining: a data-driven analysis. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 5427–5433. International Joint Conferences on Artificial Intelligence Organization.

Winston Carlile, Nishant Gurrapadi, Zixuan Ke, and Vincent Ng. 2018. Give me more feedback: Annotating argument persuasiveness and related attributes in student essays. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 621–631, Melbourne, Australia. Association for Computational Linguistics.

Silvia Casola, Soda Marem Lo, Valerio Basile, Simona Frenda, Alessandra Cignarella, Viviana Patti, and Cristina Bosco. 2023. Confidence-based ensembling of perspective-aware models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3496–3507, Singapore. Association for Computational Linguistics.

Jonathan P. Chang and Cristian Danescu-Niculescu-Mizil. 2019. Trouble on the horizon: Forecasting the derailment of online conversations as they develop. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4743–4754, Hong Kong, China. Association for Computational Linguistics.

Zaiqian Chen, Daniel Verdi do Amarante, Jenna Donaldson, Yohan Jo, and Joonsuk Park. 2022. Argument mining for review helpfulness prediction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8914–8922, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kevin Coe, Kate Kenski, and Stephen A. Rains. 2014. Online and Uncivil? Patterns and Determinants of Incivility in Newspaper Website Comments. *Journal of Communication*, 64(4):658–679.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Christine De Kock, Tom Stafford, and Andreas Vlachos. 2022. How to disagree well: Investigating the dispute tactics used on Wikipedia. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3824–3837, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Christine De Kock and Andreas Vlachos. 2021. I beg to differ: A study of constructive disagreement in online conversations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2017–2027, Online. Association for Computational Linguistics.

Michael X. Delli Carpini, Fay Lomax Cook, and Lawrence R. Jacobs. 2004. Public deliberation, discursive participation, and citizen engagement. a review of empirical literature. *Annual Review of Political Science*, 7(3):315–344.

Lorik Dumani and Ralf Schenkel. 2020. Quality-aware ranking of arguments. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 335–344, New York, NY, USA. Association for Computing Machinery.

Esin Durmus and Claire Cardie. 2019. A corpus for modeling user and language effects in argumentation on online debating. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 602–607, Florence, Italy. Association for Computational Linguistics.

Esin Durmus, Faisal Ladhak, and Claire Cardie. 2019. The role of pragmatic and discourse context in determining argument impact. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5668–5678, Hong Kong, China. Association for Computational Linguistics.

Roxanne El Baff, Henning Wachsmuth, Khalid Al-Khatib, and Benno Stein. 2018. Challenge or empower: Revisiting argumentation quality in a news editorial corpus. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 454–464, Brussels, Belgium. Association for Computational Linguistics.

Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2020. Analyzing the Persuasive Effect of Style in News Editorial Argumentation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3154–3160, Online. Association for Computational Linguistics.

Katharina Esau. 2022. *Kommunikationsformen und Deliberationsdynamik*, volume 21 of *Politische Kommunikation und demokratische Öffentlichkeit*. Nomos, Baden-Baden.

Neele Falk and Gabriella Lapesa. 2022. Scaling up discourse quality annotation for political science. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3301–3318, Marseille, France. European Language Resources Association.

Neele Falk and Gabriella Lapesa. 2023. StoryARG: a corpus of narratives and personal experiences in argumentative texts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2350–2372,

Toronto, Canada. Association for Computational Linguistics.

Neele Falk and Gabriella Lapesa. 2024. Stories and personal experiences in the COVID-19 discourse. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15320–15340, Torino, Italia. ELRA and ICCL.

Neele Falk, Eva Vecchi, Iman Jundi, and Gabriella Lapesa. 2024. Moderation in the wild: Investigating user-driven moderation in online discussions. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 992–1013, St. Julian's, Malta. Association for Computational Linguistics.

Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide Bernardi. 2024. Perspectivist approaches to natural language processing: a survey. *Language Resources and Evaluation*, 56:1574–0218.

Dennis Friess and Christiane Eilders. 2015. A systematic review of online deliberation research. *Policy & Internet*, 7(3):319–339.

Marlène Gerber, André Bächtiger, Susumu Shikano, Simon Reber, and Samuel Rohr. 2018. Deliberative abilities and influence in a transnational deliberative poll (europolis). *British Journal of Political Science*, 48(4):1093–1118.

Frauke Gerlach and Christiane Eilders, editors. 2022. *#meinfernsehen 2021*. Nomos, Baden-Baden.

Lukas Gienapp, Benno Stein, Matthias Hagen, and Martin Potthast. 2020. Efficient pairwise annotation of argument quality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5772–5781, Online. Association for Computational Linguistics.

Martin Gleize, Eyal Shnarch, Leshem Choshen, Lena Dankin, Guy Moshkowich, Ranit Aharonov, and Noam Slonim. 2019. Are you convinced? choosing the more convincing evidence with a Siamese network. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 967–976, Florence, Italy. Association for Computational Linguistics.

Max Glockner, Yufang Hou, Preslav Nakov, and Iryna Gurevych. 2024. Missci: Reconstructing fallacies in misrepresented science. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4372–4405, Bangkok, Thailand. Association for Computational Linguistics.

Pierpaolo Goffredo, Mariana Chaves, Serena Villata, and Elena Cabrio. 2023. Argument-based detection and classification of fallacies in political debates. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11101–11112, Singapore. Association for Computational Linguistics.

Pierpaolo Goffredo, Shohreh Haddadan, Vorakit Vorakitphan, Elena Cabrio, and Serena Villata. 2022. Fallacious argument classification in political debates. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4143–4149. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Todd Graham and Tamara Witschge. 2003. In search of online deliberation: Towards a new method for examining the quality of online discussions. *Communications*, 28(2):173–204.

Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2020. A large-scale dataset for argument quality ranking: Construction and analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7805–7813.

Ivan Habernal and Iryna Gurevych. 2016a. What makes a convincing argument? empirical analysis and detecting attributes of convincingness in web argumentation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1214–1223, Austin, Texas. Association for Computational Linguistics.

Ivan Habernal and Iryna Gurevych. 2016b. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599, Berlin, Germany. Association for Computational Linguistics.

Ivan Habernal, Raffael Hannemann, Christian Pollak, Christopher Klamm, Patrick Pauli, and Iryna Gurevych. 2017. Argotario: Computational argumentation meets serious games. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 7–12, Copenhagen, Denmark. Association for Computational Linguistics.

Ivan Habernal, Patrick Pauli, and Iryna Gurevych. 2018a. Adapting serious game for fallacious argumentation to German: Pitfalls, insights, and best practices. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018b. Before name-calling: Dynamics and triggers of ad hominem fallacies in web argumentation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages

386–396, New Orleans, Louisiana. Association for Computational Linguistics.

Sang-pil Han and Sharon Shavitt. 1994. Persuasion and culture: Advertising appeals in individualistic and collectivistic societies. *Journal of Experimental Social Psychology*, 30(4):326–350.

Annette Hautli-Janisz, Ella Schad, and Chris Reed. 2022. Disagreement space in argument analysis. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 1–9, Marseille, France. European Language Resources Association.

Dominique Heinbach, Lena Wilms, and Marc Ziegele. 2022. Effects of empowerment moderation in online discussions: A field experiment with four news outlets. In *72nd Annual Conference of the International Communication Association (ICA)*.

Philipp Heinisch, Anette Frank, Juri Opitz, Moritz Plenz, and Philipp Cimiano. 2022. Overview of the 2022 validity and novelty prediction shared task. In *Proceedings of the 9th Workshop on Argument Mining*, pages 84–94, Online and in Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Philipp Heinisch, Matthias Orlikowski, Julia Romberg, and Philipp Cimiano. 2023. Architectural sweet spots for modeling human label variation by the example of argument quality: It's best to relate perspectives! In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11138–11154, Singapore. Association for Computational Linguistics.

Chadi Helwe, Tom Calamai, Pierre-Henri Paris, Chloé Clavel, and Fabian Suchanek. 2024. MAFALDA: A benchmark and comprehensive study of fallacy detection and classification. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4810–4845, Mexico City, Mexico. Association for Computational Linguistics.

Ruixin Hong, Hongming Zhang, Xinyu Pang, Dong Yu, and Changshui Zhang. 2024. A closer look at the self-verification abilities of large language models in logical reasoning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 900–925, Mexico City, Mexico. Association for Computational Linguistics.

Anthony Hunter and Sylwia Polberg. 2017. Empirical methods for modelling persuadees in dialogical argumentation. In *2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 382–389.

Rositsa V Ivanova, Thomas Huber, and Christina Niklaus. 2024. Let's discuss! quality dimensions and annotated datasets for computational argument quality assessment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20749–20779, Miami, Florida, USA. Association for Computational Linguistics.

Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schoelkopf. 2022. Logical fallacy detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7180–7198, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yohan Jo, Seojin Bang, Emaad Manzoor, Eduard Hovy, and Chris Reed. 2020. Detecting attackable sentences in arguments. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1–23, Online. Association for Computational Linguistics.

Omkar Joshi, Priya Pitre, and Yashodhara Haribhakta. 2023. ArgAnalysis35K : A large-scale dataset for argument quality analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13916–13931, Toronto, Canada. Association for Computational Linguistics.

Georgi Karadzhov, Tom Stafford, and Andreas Vlachos. 2023. Delidata: A dataset for deliberation in multiparty problem solving. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW2).

Zixuan Ke, Hrishikesh Inamdar, Hui Lin, and Vincent Ng. 2019. Give me more feedback II: Annotating thesis strength and related attributes in student essays. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3994–4004, Florence, Italy. Association for Computational Linguistics.

Anastassia Kornilova, Vladimir Eidelman, and Daniel Douglass. 2022. An item response theory framework for persuasion. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 77–86, Seattle, United States. Association for Computational Linguistics.

Anne Lauscher, Lily Ng, Courtney Napoles, and Joel Tetreault. 2020. Rhetoric, logic, and dialectic: Advancing theory-based argument quality assessment in natural language processing. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4563–4574, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Anne Lauscher, Henning Wachsmuth, Iryna Gurevych, and Goran Glavaš. 2022. Scientia Potentia Est—On the Role of Knowledge in Computational Argumentation. *Transactions of the Association for Computational Linguistics*, 10:1392–1422.

John Lawrence and Chris Reed. 2020. Argument Mining: A Survey. *Computational Linguistics*, 45(4):765–818.

Elisa Leonardelli, Gavin Abercrombie, Dina Almanea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Verena Rieser, Alexandra Uma, and Massimo Poesio. 2023. SemEval-2023 task 11: Learning with disagreements (LeWiDi). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2304–2318. Association for Computational Linguistics.

Hao Li, Yuping Wu, Viktor Schlegel, Riza Batista-Navarro, Tharindu Madusanka, Iqra Zahid, Jiayan Zeng, Xiaochi Wang, Xinran He, Yizhi Li, and Goran Nenadic. 2024. Which side are you on? a multi-task dataset for end-to-end argument summarisation and evaluation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 133–150, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Shengjie Li and Vincent Ng. 2024. ICLE++: Modeling fine-grained traits for holistic essay scoring. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8465–8486, Mexico City, Mexico. Association for Computational Linguistics.

Zhexiong Liu, Mohamed Elaraby, Yang Zhong, and Diane Litman. 2023. Overview of ImageArg-2023: The first shared task in multimodal argument mining. In *Proceedings of the 10th Workshop on Argument Mining*, pages 120–132, Singapore. Association for Computational Linguistics.

Zhexiong Liu, Meiqi Guo, Yue Dai, and Diane Litman. 2022. ImageArg: A multi-modal tweet dataset for image persuasiveness mining. In *Proceedings of the 9th Workshop on Argument Mining*, pages 1–18, Online and in Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Stephanie Lukin, Pranav Anand, Marilyn Walker, and Steve Whittaker. 2017. Argument strength is in the eye of the beholder: Audience effects in persuasion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 742–753, Valencia, Spain. Association for Computational Linguistics.

Eleonora Mancini, Federico Ruggeri, and Paolo Torroni. 2024. Multimodal fallacy classification in political debates. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 170–178, St. Julian's, Malta. Association for Computational Linguistics.

Santiago Marro, Elena Cabrio, and Serena Villata. 2022. Graph embeddings for argumentation quality assessment. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4154–4164, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nailia Mirzakhmedova, Marcel Gohsen, Chia Hao Chang, and Benno Stein. 2024. Are large language models reliable argument quality annotators? In *Robust Argumentation Machines*, pages 129–146, Cham. Springer Nature Switzerland.

Elena Musi, Myrto Aloumpi, Elinor Carmi, Simeon Yates, and Kay O'Halloran. 2022. Developing fake news immunity: Fallacies as misinformation triggers during the pandemic. *Online Journal of Communication and Media Technologies*, 12(3):e202217.

Shoichi Naito, Shintaro Sawada, Chihiro Nakagawa, Naoya Inoue, Kenshi Yamaguchi, Iori Shimizu, Farjana Sultana Mim, Keshav Singh, and Kentaro Inui. 2022. TYPIC: A corpus of template-based diagnostic comments on argumentation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5916–5928, Marseille, France. European Language Resources Association.

Courtney Napoles, Joel Tetreault, Aasish Pappu, Enrica Rosato, and Brian Provenzale. 2017. Finding good conversations online: The Yahoo News annotated comments corpus. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 13–23, Valencia, Spain. Association for Computational Linguistics.

Lily Ng, Anne Lauscher, Joel Tetreault, and Courtney Napoles. 2020. Creating a domain-diverse corpus for theory-based argument quality assessment. In *Proceedings of the 7th Workshop on Argument Mining*, pages 117–126, Online. Association for Computational Linguistics.

Vlad Niculae and Cristian Danescu-Niculescu-Mizil. 2016. Conversational markers of constructive discussions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 568–578, San Diego, California. Association for Computational Linguistics.

Kenia Nieto-Benitez, Noe Alejandro Castro-Sanchez, Hector Jimenez Salazar, Gemma Bel-Enguix, Dante Mújica Vargas, Juan Gabriel González Serna, and Nimrod González Franco. 2023. Elements for automatic identification of fallacies in mexican election campaign political speeches. *Programming and Computer Software*, 49:762–774.

Jennifer A. Noble. 2012. Minority voices of crowdsourcing: Why we should pay attention to every member of the crowd. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work Companion*, CSCW '12, page 179–182, New York, NY, USA. Association for Computing Machinery.

Pia Pachinger, Allan Hanbury, Julia Neidhardt, and Anna Planitzer. 2023. Toward disambiguating the definitions of abusive, offensive, toxic, and uncivil

comments. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 107–113, Dubrovnik, Croatia. Association for Computational Linguistics.

Joonsuk Park and Claire Cardie. 2018. A corpus of eRulemaking user comments for measuring evaluability of arguments. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Andreas Peldszus and Manfred Stede. 2013. From argument diagrams to argumentation mining in texts: A survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 7(1):1–31.

Isaac Persing, Alan Davis, and Vincent Ng. 2010. Modeling organization in student essays. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 229–239, Cambridge, MA. Association for Computational Linguistics.

Isaac Persing and Vincent Ng. 2013. Modeling thesis clarity in student essays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 260–269, Sofia, Bulgaria. Association for Computational Linguistics.

Isaac Persing and Vincent Ng. 2014. Modeling prompt adherence in student essays. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1534–1543, Baltimore, Maryland. Association for Computational Linguistics.

Isaac Persing and Vincent Ng. 2015. Modeling argument strength in student essays. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 543–552, Beijing, China. Association for Computational Linguistics.

Isaac Persing and Vincent Ng. 2017. Why can't you convince me? modeling weaknesses in unpersuasive arguments. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 4082–4088.

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

Peter Potash, Adam Ferguson, and Timothy J. Hazen. 2019. Ranking passages for argument convincingness. In *Proceedings of the 6th Workshop on Argument Mining*, pages 146–155, Florence, Italy. Association for Computational Linguistics.

Martin Potthast, Lukas Gienapp, Florian Euchner, Nick Heilenkötter, Nico Weidmann, Henning Wachsmuth, Benno Stein, and Matthias Hagen. 2019. Argument search: Assessing argument relevance. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 1117–1120, New York, NY, USA. Association for Computing Machinery.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On releasing annotator-level labels and information in datasets. In *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 133–138, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Maud Reveilhac. 2023. Comparing and mapping difference indices of debate quality on twitter. *Methodological Innovations*, 16(2):234–249.

Julian Risch, Anke Stoll, Lena Wilms, and Michael Wiegand. 2021. Overview of the GermEval 2021 shared task on the identification of toxic, engaging, and fact-claiming comments. In *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments*, pages 1–12, Duesseldorf, Germany. Association for Computational Linguistics.

Julia Romberg. 2022. Is your perspective also my perspective? enriching prediction with subjectivity. In *Proceedings of the 9th Workshop on Argument Mining*, pages 115–125, Online and in Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Julia Romberg, Laura Mark, and Tobias Escher. 2022. A corpus of German citizen contributions in mobility planning: Supporting evaluation through multidimensional classification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2874–2883, Marseille, France. European Language Resources Association.

Ramon Ruiz-Dolz, John Lawrence, Ella Schad, and Chris Reed. 2024. Overview of DialAM-2024: Argument mining in natural language dialogues. In *Proceedings of the 11th Workshop on Argument Mining (ArgMining 2024)*, pages 83–92, Bangkok, Thailand. Association for Computational Linguistics.

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

Sara Salamat, Negar Arabzadeh, Amin Bigdeli, Shirin Seyedsalehi, Morteza Zihayat, and Ebrahim Bagheri. 2023. Don't raise your voice, improve your argument: Learning to retrieve convincing arguments. In

Advances in Information Retrieval, pages 589–598, Cham. Springer Nature Switzerland.

Robin Schaefer and Manfred Stede. 2021. Argument mining on twitter: A survey. *it - Information Technology*, 63(1):45–58.

Nils-Jonathan Schaller, Andrea Horbach, Lars Ingver Höft, Yuning Ding, Jan Luca Bahr, Jennifer Meyer, and Thorben Jansen. 2024. DARIUS: A comprehensive learner corpus for argument mining in German-language essays. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4356–4367, Torino, Italia. ELRA and ICCL.

Natalie Schluter. 2018. The glass ceiling in NLP. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2793–2798, Brussels, Belgium. Association for Computational Linguistics.

Lin Shen. 2023. Culture and explicitness of persuasion: Linguistic evidence from a 51-year corpus-based cross-cultural comparison of the united nations general debate speeches across 55 countries (1970-2020). *Cross-Cultural Research*, 57(2-3):166–192.

Cássio Silva, Amanda Rassi, Jackson Souza, Renata Ramisch, Roger Antunes, and Helena Caseli. 2021. Quality of argumentation in political tweets: what is and how to measure it / qualidade da argumentação em tweets de política: o que e como avaliar. *REVISTA DE ESTUDOS DA LINGUAGEM*, 29(4):2537–2586.

Wei Song, Ziyao Song, Lizhen Liu, and Ruiji Fu. 2020. Hierarchical multi-task learning for organization evaluation of argumentative student essays. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3875–3881. International Joint Conferences on Artificial Intelligence Organization. Main track.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1501–1510, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2017. Recognizing insufficiently supported arguments in argumentative essays. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 980–990, Valencia, Spain. Association for Computational Linguistics.

Manfred Stede and Jodi Schneider. 2019. *Argumentation Mining*. Synthesis Lectures on Human Language Technologies. Springer International Publishing, Cham.

Marco R. Steenbergen, André Bächtiger, Markus Spörndli, and Jürg Steiner. 2003. Measuring political deliberation: A discourse quality index. *Comparative European Politics*, 1:21–48.

Jennifer Stromer-Galley. 2007. Measuring deliberation's content: A coding scheme. *Journal of Public Deliberation*, 3(1):1–35.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 613–624, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Milagro Teruel, Cristian Cardellino, Fernando Cardellino, Laura Alonso Alemany, and Serena Villata. 2018. Increasing argument annotation reproducibility by using inter-annotator agreement to improve guidelines. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Assaf Toledo, Shai Gretz, Edo Cohen-Karlik, Roni Friedman, Elad Venezian, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2019. Automatic argument quality assessment - new datasets and methods. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5625–5635, Hong Kong, China. Association for Computational Linguistics.

Orith Toledo-Ronen, Matan Orbach, Yonatan Bilu, Artem Spector, and Noam Slonim. 2020. Multilingual argument mining: Datasets and analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 303–317, Online. Association for Computational Linguistics.

Alexandra Uma, Tommaso Fornaciari, Anca Dumitrache, Tristan Miller, Jon Chamberlain, Barbara Plank, Edwin Simpson, and Massimo Poesio. 2021a. SemEval-2021 task 12: Learning with disagreements. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 338–347, Online. Association for Computational Linguistics.

Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021b. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Michiel van der Meer, Neele Falk, Pradeep K. Murukannaiah, and Enrico Liscio. 2024. Annotator-centric active learning for subjective NLP tasks. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18537–18555,

Miami, Florida, USA. Association for Computational Linguistics.

Eva Maria Vecchi, Neele Falk, Iman Jundi, and Gabriella Lapesa. 2021. Towards argument mining for social good: A survey. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1338–1352, Online. Association for Computational Linguistics.

Henning Wachsmuth, Gabriella Lapesa, Elena Cabrio, Anne Lauscher, Joonsuk Park, Eva Maria Vecchi, Serena Villata, and Timon Ziegenbein. 2024. Argument quality assessment in the age of instruction-following large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1519–1538, Torino, Italia. ELRA and ICCL.

Henning Wachsmuth, Nona Naderi, Ivan Habernal, Yufang Hou, Graeme Hirst, Iryna Gurevych, and Benno Stein. 2017a. Argumentation quality assessment: Theory vs. practice. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 250–255, Vancouver, Canada. Association for Computational Linguistics.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017b. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187, Valencia, Spain. Association for Computational Linguistics.

Henning Wachsmuth, Benno Stein, and Yamen Ajjour. 2017c. "PageRank" for argument relevance. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1117–1127, Valencia, Spain. Association for Computational Linguistics.

Henning Wachsmuth and Till Werner. 2020. Intrinsic quality assessment of arguments. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6739–6745, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Thiemo Wambsganss and Christina Niklaus. 2022. Modeling persuasive discourse to adaptively support students' argumentative writing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8748–8760, Dublin, Ireland. Association for Computational Linguistics.

Xiaoou Wang, Elena Cabrio, and Serena Villata. 2023. Argument and counter-argument generation: A critical survey. In *Natural Language Processing and Information Systems*, pages 500–510, Cham. Springer Nature Switzerland.

Xinpeng Wang and Barbara Plank. 2023. ACTOR: Active learning with annotator-specific classification heads to embrace human label variation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2046–2052, Singapore. Association for Computational Linguistics.

Leon Weber-Genzel, Siyao Peng, Marie-Catherine De Marneffe, and Barbara Plank. 2024. VariErr NLI: Separating annotation error from human label variation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2256–2269, Bangkok, Thailand. Association for Computational Linguistics.

Matti Wiegmann, Khalid Al Khatib, Vishal Khanna, and Benno Stein. 2022. Analyzing persuasion strategies of debaters on social media. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6897–6905, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Claes Wohlin. 2014. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*, EASE '14, New York, NY, USA. Association for Computing Machinery.

Xiaoxi Xu, Tom Murray, Beverly Park Woolf, and David A. Smith. 2014. Identifying social deliberative behavior from online communication — a cross-domain study. In *Proceedings of the Twenty-Seventh International Florida Artificial Intelligence Research Society Conference*, pages 237–242. Association for the Advancement of Artificial Intelligence.

Wonsuk Yang, Jung-Ho Kim, Seungwon Yoon, Chae-Hun Park, and Jong C. Park. 2019a. A corpus of sentence-level annotations of local acceptability with reasons. In *Proceedings of the 33rd Pacific Asia Conference on Language, Information and Computation*.

Wonsuk Yang, Seungwon Yoon, Ada Carpenter, and Jong Park. 2019b. Nonsense!: Quality control via two-step reason selection for annotating local acceptability and related attributes in news editorials. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2954–2963, Hong Kong, China. Association for Computational Linguistics.

Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. 2018. Conversations

gone awry: Detecting early signs of conversational failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1350–1361, Melbourne, Australia. Association for Computational Linguistics.

Justine Zhang, Ravi Kumar, Sujith Ravi, and Cristian Danescu-Niculescu-Mizil. 2016. Conversational flow in Oxford-style debates. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 136–141, San Diego, California. Association for Computational Linguistics.

Timon Ziegenbein, Shahbaz Syed, Felix Lange, Martin Potthast, and Henning Wachsmuth. 2023. Modeling appropriate language in argumentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4344–4363. Association for Computational Linguistics.

A Details of the Dataset Search

Search Process Using Google site search, we identified 238 candidate papers. Based on a review of the titles, abstracts, and a skim of the content, we excluded papers that neither introduced novel data nor addressed AQ. We filtered duplicates — papers that introduce the same dataset — and kept the first published paper in such cases.

Datasets come in various formats, such as machine learning corpora, user studies, or manual content analyses. Our focus is on annotated datasets initially created as resources for machine learning. We also collect extensions of these datasets, even if introduced primarily for annotation studies (i.e., extensions of the Dagstuhl-ArgQuality corpus). Additionally, we include datasets originally developed for other contexts, such as qualitative content analysis in the social sciences, if they have later been used to train machine learning models. We collect datasets that were purposefully annotated through coordinated efforts in the context of scientific work, as well as such that have been crawled from existing sources that provide some sort of annotation (e.g., online debate portals).

Relevance For a dataset (and corresponding paper) to be considered relevant, it must encode at least one category (either major or sub-category) according to the taxonomy introduced in Section 3. We do not require the entire dataset to focus exclusively on AQ assessment; it is sufficient if certain dimensions of the annotation code AQ.

Some works we reviewed consider topical relevance as part of AQ, either in terms of a premise's relevance to a conclusion (as local relevance) or an argument's relevance to the issue at hand (as global relevance). According to the definitions we follow, local relevance focuses on "the contribution towards the acceptance or rejection of the argument's conclusion", while global relevance addresses the "contribution towards the issue's resolution". However, it is debatable whether topical relevance consistently aligns with these goals. As briefly discussed by Potthast et al. (2019), although the notions of relevance in information retrieval and global relevance share similarities, the authors chose to keep the concepts distinct. In our review, we similarly excluded such cases.

Reddit CMV (ChangeMyView) Several works use the Reddit CMV subreddit as a source for pre-annotated data on persuasion. Due to the at times significant overlap between these datasets, in case of duplicate content we focus on the CMV crawl that was introduced first. We include additional datasets in our database only if they (a) offer distinct value by including more recent data from the subreddit, or (b) provide further annotations of AQ for the same time span.

Dataset Extensions Nine datasets extend other AQ datasets directly by adding further samples (4), by re-annotating the same samples (3), by adding a further modality (1) or by releasing the non-aggregated version of a previously aggregated dataset release (1).

B Collection of Properties in the Database

We publish the results of our systematic literature review in the form of an online database to inform future research in AQ. This database solely contains meta-information on the identified datasets. Each dataset is represented as a row, while the columns contain a comprehensive set of characteristics describing the datasets. First, we collect basic information about each dataset, namely the **dataset name**, whether the dataset is an **extension of another dataset** (*extension of samples*, *extension of annotators*, *multimodal extension*, and *non-aggregated version*), the **availability of the dataset** (*online*, *upon request*, *not publicly available*, *no indication*), — if available online — the **link to the resource** and **status of the link** (*accessible* or *not accessible*; as of August 2024), the **license**, whether the data contains **manual annotation**[6], and — in case of manual annotation — the **availability of annotation guidelines** (*online* (with reference to location), *upon request*, *not available in full*, *not publicly available*, and *no indication*). Corresponding information on the paper that introduced the dataset is provided through **paper title**, the **paper authors**, the **year** of publication, and the **paper URL**, together with the targeted **research community** (based on the publication venue).

Going more into detail on the single datasets, we indicate the **size** of the dataset in terms of **units of annotation**, the **genre**, the **modality** (either *text* or *multimodal*), and the **language** represented. In line with our goal of identifying AQ datasets, we provide a **textual description of AQ**

[6]We differentiate between datasets produced through coordinated manual annotation studies and those extracted from existing resources, like debate forums, where labels for certain AQ phenomena are inherently present.

categories contained as described in the words of the respective paper and indicate via **AQ taxonomy codes** (see Table 5) which categories of the taxonomy the dataset covers. For manual annotation datasets, we track the **aggregation method** that was used to form an aggregated ground truth (in case of aggregated ground truth datasets), the inter-annotator-agreement through **IAA score** and **IAA measure**, and whether there is a **non-aggregated version** available (*no*, *yes (annotator-specific)*, *yes (distribution-based)*, and *no indication*).

For exploring whose perspectives (in terms of perceiving AQ) are represented in current datasets, we gathered details on the **selection method of annotators** from the authors' descriptions, whether they are **in-house or crowd** workers, whether they are **expert or novice** annotators, and any available **annotator attributes** provided in the associated paper or dataset. Additionally, we account for the **number of annotators per item** and the **number of annotators in total**.

We also examined the authors of the arguments, referred to as "argument producers". For these producers, we give the number of distinct individuals represented in the dataset (**number of producers**) and reviewed the available socio-demographic information provided in the corresponding papers or datasets (**producer attributes**).

C Details of the Non-Aggregated Datasets

CrowDEA Ideas dataset (Baba et al., 2020) This Japanese-language corpus contains preference labels for solution proposals to everyday life questions. A total of 16k argument pairs were annotated by 20 different workers, drawn from a pool of 257 crowd workers (of which no further information is provided).

CIMT PartEval Argument Concreteness Corpus (Romberg et al., 2022) As a tool to support public institutions in Germany in evaluating citizen contributions, this dataset provides annotations on how concrete an argument is introduced. Five student annotators fully labeled a total of 1127 argument units. Acknowledging the subjectivity of the task, the dataset was explictely published in a non-aggregated way.

TYPIC (Naito et al., 2022) To offer feedback on flaws in Japanese students' arguments, the authors took an approach that first provides diagnostic comments describing weaknesses in the arguments. These comments are then mapped to AQ criteria: local acceptability, local sufficiency, local relevance, global relevance, and global sufficiency. Four experts generated 1,082 diagnostic comments for 197 Japanese-language arguments, with each argument receiving two labels. The categorization was conducted by one to two annotators.

Argument Validity and Novelty Prediction Shared Task (Heinisch et al., 2023) The non-aggregated version of the dataset from the Argument Validity and Novelty Prediction Shared Task (Heinisch et al., 2022), co-located with the ArgMining workshop 2022, was published in later work for use in multi-annotator models. 1474 premise-conclusion pairs from English-language online debate portals come with three annotations per item, drawing from a pool of five student experts.

MAFALDA (Helwe et al., 2024) The authors developed a hierarchical taxonomy of fallacies, resulting in the MAFALDA corpus, which contains 268 argumentative spans drawn from English-language news articles, social media, and political debates. Four of the authors annotated the spans, with each item receiving up to four labels.

UMOD dataset (User Moderation in Online Discussions) (Falk et al., 2024) The study focuses on annotating characteristics of user-driven moderation in online discussions, among them the constructiveness of such comments. The dataset, sourced from English Reddit's Change My View (CMV), contains 1,000 comment-reply pairs. Each pair was annotated by nine crowd workers, with a total of 90 workers participating. To provide a more nuanced understanding of the annotators, socio-demographic information including race, sex, age, and role) are collected too.

The Persuasion and Personality Corpus (Lukin et al., 2017) An explicitly perspectivist corpus was introduced to investigate differences in the perception of argument effectiveness. To this end, for 637 crowd workers — representative of the U.S population — stance changes elicited by presented social media arguments were recorded. At the same time, Big Five personality traits were collected, alongside further socio-demographic information (age, gender, political view, education, civic engagement, religion, spirituality, employment status, and income). The resulting dataset contains 100 items, each annotated by 20 workers.

Webis-Editorial-Quality-18 corpus (El Baff et al., 2018) This corpus was created for assessing AQ of news editorials. Resembling Lukin et al. (2017), different perceptions were considered using the Big Five personality traits and the political leaning. 1000 news editorials were annotated by three liberals and three conservatives each, 24 crowd workers from the U.S. in total.

n.a. (Hunter and Polberg, 2017) To study the personalization of argumentation, 50 crowd workers (no further details about the workers were provided) annotated a set of 30 English-language arguments for believability, convincingness, and appeal.

SIGIR-19 (Potthast et al., 2019) A resource that codes the logical, rhetorical, and dialectical quality of arguments in the context of information retrieval. 40 student volunteers annotated 494 online debate portal arguments in English, with one annotator per item. While gender and age are indicated annotator-specific, from the paper we learn about the political leaning (80% vote for left-wing, green parties) which may impact the perspectives of annotators.

argumentation attitude dataset (Brenneis et al., 2021) The German-language dataset was collected from a deliberation platform for political opinion-forming. A total of 946 arguments were rated in four waves based on individual conviction and argument strength. The 674 crowd workers constitute a representative selection of the German online population in terms of age, gender and education. Each argument received ratings from one to 147 individuals. A unique feature of this dataset is that it captures not only the perception of argument effectiveness but also the writing style of the individual members of the crowd, as they contribute part of the arguments as well.

Dagstuhl-15512-ArgQuality (Wachsmuth et al., 2017b) Along with introducing their taxonomy of AQ, the authors provide a dataset: 320 English-language arguments from online debate portals were rated across 15 categories by three expert annotators. While there is no attribution to the annotator id, distributional information about gender, education, and employment is given, providing some information about the annotators' sociodemographics.

Dagstuhl-15512-ArgQuality: extension 1 (Wachsmuth et al., 2017a) The same dataset was annotated by 102 crowd workers (no further details about the workers were provided), with 10 workers assigned to each argument.

Dagstuhl-15512-ArgQuality: extension 2 (Mirzakhmedova et al., 2024) The dataset was also annotated by 108 novice in-house annotators, all undergraduate students without prior experience in computational argumentation, with up to 10 raters per argument. *Note: The official release includes a subset of annotators only, which are not uniquely identifiable.*

GAQCorpus (Lauscher et al., 2020) The authors introduced one of the largest corpora for assessing multiple key aspects of AQ. The corpus includes 5,285 arguments from diverse domains (debate forums, question-answering forums, and review forums), annotated for cogency, effectiveness, reasonableness, and overall quality. Annotation was conducted using a mix of three expert annotators and 24 crowd workers. Some items were annotated exclusively by the crowd (10 per item), others exclusively by the experts (up to three annotations per item), and a portion of the dataset was jointly annotated by both groups (up to 13 annotations per item). *Note: The corpus contains varying numbers of annotators for different parts of the data.*

EuropolisAQ (Falk and Lapesa, 2022) The dataset contains 513 transcribed speeches from a transnational deliberative poll. It builds on the Europolis corpus (Gerber et al., 2018), labeled with deliberative norms, and extends it with annotations of the other dimensions of AQ, namely cogency, effectiveness, reasonableness, and overall AQ. Each item was annotated by one or both of two expert annotators. To the best of our knowledge, this is the only corpus that provides comprehensive coverage of the four major aspects of AQ according to our taxonomy — while not all dimensions are available in non-aggregated format.

ArgQ! (Silva et al., 2021) The authors adapt the rhetorical effectiveness dimension of Wachsmuth et al. (2017b)'s taxonomy for use with Twitter data and expert annotate 352 argumentative tweets from the Brazilian political context accordingly. Each argument contains four labels from a total of four different annotators.

UKPConvArg1 (Habernal and Gurevych, 2016b) The dataset consists of 16k pairwise comparisons of arguments from online debate

portals with respect to their convincingness. Each pair was annotated by five crowd workers, with about 3,900 workers participating, all from the U.S. Additionally, the workers' stance towards the discussed topic was tracked, as it could likely influence their perspective during the assessment.

UKPConvArg2 (Habernal and Gurevych, 2016a) In addition to the overall assessment of effectiveness in UKPConvArg1, this dataset provides a categorization into finer attributes based on textual decision rationales formulated by the individual workers in the previous annotation. The attributes give reasons for why an annotator found an argument to be convincing, covering primarily logical and rhetorical categories of the taxonomy. Each of the 70k reason units comes with 5 annotations from a pool of 776 crowd workers.

Essay Argument Organization Dataset (Persing et al., 2010) Targeting the automatic evaluation of persuasive essays, this study addresses the evaluation dimension of organization, which assesses how well an essay is structured to logically develop an argument. 1003 persuasive essays, written by a diverse set of English learners from 15 native languages, were rated by one to six annotators. *Note, that in this datasets there is only a distribution of annotations given per item, no attribution to unique annotator.*

Appropriateness Corpus (Ziegenbein et al., 2023) The authors introduce a refined definition of the subcategory "appropriateness" by offering a more sophisticated interpretation. Using 2,191 arguments from existing AQ corpora, they had appropriateness annotated by three crowd workers for each argument. While there is no attribution to the annotator id, distributional information about gender and mother tongue is given, providing some information about the annotators' sociodemographics.

UKP-InsufficientArguments (Stab and Gurevych, 2017) The authors present a corpus of persuasive essays annotated for local sufficiency. For inter-annotator-agreement, 433 of the 1029 arguments were annotated by three expert annotators.

Webis-ArgRank-17 Dataset (Wachsmuth et al., 2017c) This dataset is on the global relevance of arguments. Mainly created automatically, it contains a manually annotated ground truth of 110 arguments annotated by seven experts from computational linguistics and information retrieval each.

StoryARG (Falk and Lapesa, 2023) The authors developed a corpus focused on narratives and personal experiences in argumentative texts. Each argument (2451 in total) was annotated by one to four annotators, with a total of four annotators participating in the project. Distributional information on gender, education, and country of origin is provided for the annotators.

D Definition of Argument Quality

Table 5 shows the original definitions of all AQ dimensions considered in this work.

E List of AQ Datasets

Table 6 lists the 103 datasets for AQ that we identified in our systematic literature search.

Category	Description
Logical cogency	An argument is cogent if it has acceptable premises that are relevant to its conclusion and that are sufficient to draw the conclusion.
Local acceptability	A premise of an argument is acceptable if it is rationally worthy of being believed to be true.
Local relevance	A premise of an argument is relevant if it contributes to the acceptance or rejection of the argument's conclusion.
Local sufficiency	An argument's premises are sufficient if, together, they give enough support to make it rational to draw its conclusion.
Rhetorical effectiveness	Argumentation is effective if it persuades the target audience of (or corroborates agreement with) the author's stance on the issue.
Clarity	Argumentation has a clear style if it uses correct and widely unambiguous language as well as if it avoids unnecessary complexity and deviation from the issue.
Credibility	Argumentation creates credibility if it conveys arguments and similar in a way that makes the author worthy of credence.
Appropriateness	Argumentation has an appropriate style if the used language supports the creation of credibility and emotions as well as if it is proportional to the issue.
Emotional appeal	Argumentation makes a successful emotional appeal if it creates emotions in a way that makes the target audience more open to the author's arguments.
Arrangement	Argumentation is arranged properly if it presents the issue, the arguments, and its conclusion in the right order.
Dialectical reasonableness	Argumentation is reasonable if it contributes to the issue's resolution in a sufficient way that is acceptable to the target audience.
Global acceptability	Argumentation is acceptable if the target audience accepts both the consideration of the stated arguments for the issue and the way they are stated.
Global relevance	Argumentation is relevant if it contributes to the issue's resolution, i.e., if it states arguments or other information that help to arrive at an ultimate conclusion.
Global sufficiency	Argumentation is sufficient if it adequately rebuts those counterarguments to it that can be anticipated.
Deliberative norms	Argumentation adheres to deliberative norms if it promotes a respectful and inclusive exchange of rational or alternative forms of argument, with the aim of reaching mutual understanding.
Rationality	Deliberation is rational if it is centered on arguments that are supported by solid evidence (either through facts that can be verified or through a shared understanding of moral or normative behavior), arguments and further information that are put forward in the discourse are relevant to the topic, and an informed ground for discussion is built (e.g., through providing an information base in the beginning of the discussion, or information requests by participants to make the discourse more informed). With respect to the dimensions of argumentation quality, the focus is on normatively well-reasoned arguments and not on how good these are perceived by the target audience.
Interactivity	Deliberation is interactive if the participants actively engage with each other by exchanging arguments in a way where they listen to the other participants, understand their perspective, and relate to it in a substantive way (e.g., by valuing, critiquing, or countering other's arguments, or question asking).
Equality	Deliberation is equal if all participants (irrespective of their background) have the same opportunity to participate by putting forward their own arguments and responding to other's claims. This dimension of deliberation quality tackles inclusiveness and accessibility.
Civility	Deliberation is civil if the participants show respect to the other participants by recognizing them as equal actors in the discourse and acknowledging the value of opposing claims. Respectful interaction is regarded as a prerequisite for participants to be convincable by other opinions and to reach a consensus decision in the sense of deliberation.
Common good reference	Deliberation is oriented towards the common good if arguments are justified by promoting the well-being of the community as a whole rather than serving the interests of narrow interest groups. What exactly is considered the common good can include different goals, such as achieving the best outcome for the greatest number of people or prioritizing the needs of the most disadvantaged members of society. The joint focus on a common good is regarded as a basis for participants with diverse interests to be able to convince each other.
Constructiveness	Deliberation is constructive if it contributes to finding a consensus decision for the issue of discussion through actions like proposing new solutions, searching for common ground, appeals for mobilisation, or questions addressed to the community.
Alternative forms of communication	In scenarios in which not all participants are able to adhere to the rigid concept of rational argumentation based on verifiable facts, other forms of communication can provide a valuable resource for good deliberation. These include storytelling, testimonies, narratives, emotional talk, casual talk, humor, or even gossip.
Overall quality	An overarching measure of the quality of arguments.

Table 5: Taxonomy of argument quality. The definitions of the first three dimensions are taken verbatim from Wachsmuth et al. (2017b). The definitions of the last dimension are based on Friess and Eilders (2015).

Dataset Name	Extension of Previous	Paper	Community	AQ
Essay Argument Organization Dataset		Persing et al. (2010)	NLP	Arrangement
n.a.		Cabrio and Villata (2012)	NLP	Global acceptability
Essay Thesis Clarity Dataset		Persing and Ng (2013)	NLP	Clarity
n.a.		Xu et al. (2014)	AI	Deliberative norms
Essay Prompt Adherence Dataset		Persing and Ng (2014)	NLP	Clarity
n.a.		Coe et al. (2014)	SocSci	Civility
Essay Argument Strength Dataset		Persing and Ng (2015)	NLP	Effectiveness
Intelligence Squared Debates Corpus		Zhang et al. (2016)	NLP	Effectiveness
n.a.		Niculae and Danescu-Niculescu-Mizil (2016)	NLP	Constructiveness
n.a.		Braunstain et al. (2016)	IR	Local Relevance
UKPConvArg1		Habernal and Gurevych (2016b)	NLP	Effectiveness
UKPConvArg2		Habernal and Gurevych (2016a)	NLP	Local relevance, local sufficiency, clarity, credibility, appropriateness, emotional appeal, overall quality
CMV		Tan et al. (2016)	Web	Effectiveness
Webis-ArgRank-17 Dataset		Wachsmuth et al. (2017c)	NLP	Global relevance
n.a.		Habernal et al. (2017)	NLP	Local sufficiency, clarity, appropriateness, global sufficiency, global relevance
n.a.	more samples	Habernal et al. (2018a)	NLP	Local sufficiency, clarity, appropriateness, global sufficiency, global relevance
The Persuasion and Personality Corpus		Lukin et al. (2017)	NLP	Effectiveness
Dagstuhl-15512-ArgQuality		Wachsmuth et al. (2017b)	NLP	Cogency, local acceptability, local relevance, local sufficiency, effectiveness, clarity, credibility, appropriateness, emotional appeal, arrangement, reasonableness, global acceptability, global relevance, global sufficiency, overall quality
n.a.	more annotators	Wachsmuth et al. (2017a)	NLP	Cogency, local acceptability, local relevance, local sufficiency, effectiveness, clarity, credibility, appropriateness, emotional appeal, arrangement, reasonableness, global acceptability, global relevance, global sufficiency, overall quality
n.a.	more annotators	Mirzakhmedova et al. (2024)	CA	Cogency, local acceptability, local relevance, local sufficiency, effectiveness, clarity, credibility, appropriateness, emotional appeal, arrangement, reasonableness, global acceptability, global relevance, global sufficiency, overall quality
n.a.		Beigman Klebanov et al. (2017)	NLP	Overall quality
n.a.		Hunter and Polberg (2017)	AI	Effectiveness
YNACC		Napoles et al. (2017)	NLP	Overall quality
UKP-InsufficientArguments		Stab and Gurevych (2017)	NLP	Local sufficiency

Continued on next page

Table 6 – continued from previous page

Dataset Name	Extension of Previous	Paper	Community	AQ
Debate Argument Persuasiveness Data		Persing and Ng (2017)	AI	Effectiveness
CDCP		Park and Cardie (2018)	NLP	Clarity
n.a.		Habernal et al. (2018b)	NLP	Appropriateness, global sufficiency
Webis-Editorial-Quality-18		El Baff et al. (2018)	NLP	Global acceptability
CGA-WIKI		Zhang et al. (2018)	NLP	Appropriateness, civility
n.a.	more samples	Chang and Danescu-Niculescu-Mizil (2019)	NLP	Appropriateness, civility
n.a.		Gerber et al. (2018)	SocSci	Rationality, interactivity, civility, common good reference, alternative forms of communication
Essay Argument Persuasiveness Dataset		Carlile et al. (2018)	NLP	Effectiveness
Webis-WikiDebate-18		Al-Khatib et al. (2018)	NLP	Rationality, constructiveness
DDO		Durmus and Cardie (2019)	NLP	Effectiveness
n.a.		Yang et al. (2019a)	NLP	Local acceptability
n.a.	more annotators	Yang et al. (2019b)	NLP	Local acceptability
IBM-EviConv		Gleize et al. (2019)	NLP	Effectiveness
SIGIR-19		Potthast et al. (2019)	IR	Cogency, effectiveness, reasonableness
IBM-ArgQ-Args		Toledo et al. (2019)	NLP	Overall quality
IBM-ArgQ-Pairs		Toledo et al. (2019)	NLP	Overall quality
Essay Thesis Strength Dataset		Ke et al. (2019)	NLP	Effectiveness
n.a.		Potash et al. (2019)	NLP	Effectiveness
n.a.		Durmus et al. (2019)	NLP	Global relevance
CGA-CMV		Chang and Danescu-Niculescu-Mizil (2019)	NLP	Appropriateness, civility
n.a.		Atkinson et al. (2019)	NLP	Effectiveness
IBM-ArgQ-Rank-30kArgs		Gretz et al. (2020)	AI	Overall quality
CrowDEA Ideas		Baba et al. (2020)	HCI	Overall quality
n.a.		Jo et al. (2020)	NLP	Global sufficiency
Webis-ArgQuality-20		Gienapp et al. (2020)	NLP	Cogency, effectiveness, reasonableness
Webis-CMV-20		Al Khatib et al. (2020)	NLP	Effectiveness
Chinese-Essay-Dataset-For-Organization-Evaluation		Song et al. (2020)	AI	Arrangement
XArgMining Dataset ArgsHG		Toledo-Ronen et al. (2020)	NLP	Overall quality
n.a.		Dumani and Schenkel (2020)	IR	Cogency, effectiveness, reasonableness
GAQCorpus		Lauscher et al. (2020)	NLP	Cogency, effectiveness, reasonableness, overall quality
n.a.		Sahai et al. (2021)	NLP	Local acceptability, local relevance, local sufficiency, clarity, global relevance
argumentation-attitude-dataset		Brenneis et al. (2021)	NLP	Effectiveness
WikiDisputes		De Kock and Vlachos (2021)	NLP	Constructiveness
GermEval 2021		Risch et al. (2021)	NLP	Rationality, interactivity, civility
Touché21-Argument-Retrieval-for-Controversial-Questions		Bondarenko et al. (2021)	IR	Effectiveness
Touché21-Argument-Retrieval-for-Comparative-Questions		Bondarenko et al. (2021)	IR	Effectiveness
ArgQ!		Silva et al. (2021)	NLP	Effectiveness, clarity, credibility, emotional appeal, arrangement
#meinfernsehen2021		Gerlach and Eilders (2022)	SocSci	Rationality, interactivity, civility, alternative forms of communication

Continued on next page

Table 6 – continued from previous page

Dataset Name	Extension of Previous	Paper	Community	AQ
CIMT PartEval Argument Concreteness Corpus		Romberg et al. (2022)	NLP	Clarity
Advocacy Campaign Corpus		Kornilova et al. (2022)	NLP	Effectiveness
Webis-Persuasive-Debaters-on-Reddit-CMV-2022		Wiegmann et al. (2022)	NLP	Effectiveness
AM2		Chen et al. (2022)	NLP	Effectiveness
n.a.		Musi et al. (2022)	SocSci	Local relevance, local sufficiency, clarity, credibility, global relevance
KODIE		Heinbach et al. (2022)	SocSci	Rationality, interactivity, civility, alternative forms of communication
ElecDeb60To16-fallacy		Goffredo et al. (2022)	AI	Local acceptance, local sufficiency, appropriateness, global sufficiency
ElecDeb60to20	more samples	Goffredo et al. (2023)	NLP	Local acceptance, local sufficiency, appropriateness, global sufficiency
MM-USED-fallacy	multimodality	Mancini et al. (2024)	NLP	Local acceptance, local sufficiency, appropriateness, global sufficiency
Persuasive Essays - Argument Quality Dataset		Marro et al. (2022)	NLP	Cogency, effectiveness, reasonableness
WikiTactics		De Kock et al. (2022)	NLP	Constructiveness
ImageArg		Liu et al. (2022)	NLP	Effectiveness
ImageArg-Shared-Task	more samples	Liu et al. (2023)	NLP	Effectiveness
n.a.		Esau (2022)	SocSci	Rationality, civility, constructiveness, alternative forms of communication
Logic		Jin et al. (2022)	NLP	Cogency, local acceptability, local relevance, local sufficiency, clarity, appropriateness, global relevance, global sufficiency
LogicClimate		Jin et al. (2022)	NLP	Cogency, local acceptability, local relevance, local sufficiency, clarity, appropriateness, global relevance, global sufficiency
ABMPC		Wambsganss and Niklaus (2022)	NLP	Effectiveness
CLIMATE		Alhindi et al. (2022)	NLP	Local relevance, local sufficiency, clarity, credibility, global sufficiency, global relevance
Argument Validity and Novelty Prediction Shared Task		Heinisch et al. (2022)	NLP	Local sufficiency, global relevance
n.a.	non-aggregated	Heinisch et al. (2023)	NLP	Local sufficiency, global relevance
Touché22-Argument-Retrieval-for-Controversial-Questions		Bondarenko et al. (2022)	IR	Effectiveness
Touché22-Argument-Retrieval-for-Comparative-Questions		Bondarenko et al. (2022)	IR	Effectiveness
EuropolisAQ		Falk and Lapesa (2022)	NLP	Cogency, effectiveness, reasonableness, overall quality
TYPIC		Naito et al. (2022)	NLP	local acceptability, local relevance, local sufficiency, clarity, global relevance, global sufficiency
ArgAnalysis35K		Joshi et al. (2023)	NLP	Overall quality
n.a.		Reveilhac (2023)	SocSci	Deliberative norms, rationality, interactivity, civility, constructiveness, alternative forms of communication

Continued on next page

Table 6 – continued from previous page

Dataset Name	Extension of Previous	Paper	Community	AQ
DeliData		Karadzhov et al. (2023)	HCI	Constructiveness
CoRe		Salamat et al. (2023)	IR	Effectiveness
Fallacies of Appeal to Emotions Corpus		Nieto-Benitez et al. (2023)	CS	Local sufficiency
Appropriateness Corpus		Ziegenbein et al. (2023)	NLP	Appropriateness
Touché23-Argument-Retrieval-for-Controversial-Questions		Bondarenko et al. (2023)	IR	Effectiveness
StoryARG		Falk and Lapesa (2023)	NLP	Alternative forms of communication
FALLACIES		Hong et al. (2024)	NLP	Local acceptability, local relevance, local sufficiency, clarity, appropriateness, credibility, emotional appeal, global relevance, global sufficiency
DARIUS		Schaller et al. (2024)	NLP	Local acceptability, local relevance, clarity
ICLE++		Li and Ng (2024)	NLP	effectiveness, clarity, arrangement
MAFALDA		Helwe et al. (2024)	NLP	Local acceptability, local relevance, local sufficiency, clarity, appropriateness, global sufficiency, global relevance
MISSCI		Glockner et al. (2024)	NLP	Local acceptability, local relevance, local sufficiency, clarity, global relevance
UMOD		Falk et al. (2024)	NLP	Constructiveness
COVID-19 Discourse Corpus		Falk and Lapesa (2024)	NLP	Alternative forms of communication
ArgSum Dataset		Li et al. (2024)	NLP	Effectiveness

Table 6: Overview of AQ datasets, including the name of the dataset (if provided, otherwise n.a.), whether and how it extends the previously listed dataset, the publication in which the dataset was introduced, the research community targeted (NLP, AI: artificial intelligence, CA: computational argumentation, CS: computer science, HCI: human computer interaction, IR: information retrieval, SocSci: Social Sciences, Web) and the assigned categories of AQ.